Probability and Statistics

Jordi Caballé

UAB - BSE

IDEA

- **Prof. Jordi Caballé**

  E-mail: jordi.caballe@uab.es

  Webpage: http://pareto.uab.es/jcaballe

- **Textbooks**

  - Ash, R.B. (1972), Real Analysis and Probability, Academic Press.
  - Bierens, H.J. (2004), Introduction to the Mathematical and Statistical Foundations of Econometrics, Cambridge University Press.
  - Billingsley, P. (1995), Probability and Measure, Wiley.
  - Casella, G.H. and Berger, R.L. (2002), Statistical Inference, Duxbury/Thomson Learning.
  - DeGroot, M.H. and Schervish, M.J. (2012), Probability and Statistics, Pearson.
  - Hogg, R.V., McKean, J. and Craig A.T. (2012), Introduction to Mathematical Statistics, Pearson.
  - Lindgren, B.V. (1993), Statistical Theory, Chapman and Hall/CRC.
  - Mood, A.M., Graybill, F.A. and Boes, D.C. (1974), Introduction to the Theory of Statistics, McGraw Hill.
  - Rice, J.A. (2007), Mathematical Statistics and Data Analysis, Cengage Learning.

- **Program**
  - ★ Descriptive Statistics
    - 1. Probability
  - ★ A Primer on Integration
  - ★ Measure Theory
    - 2. Random Variables and Distributions
    - 3. Expectation
    - 4. Special Distributions
    - 5. Functions of Random Variables
    - 6. Stochastic Processes and Limiting Distributions
    - 7. Sampling
    - 8. Estimation
    - 9. Hypothesis Testing

Graduate Program in Economics

IDEA

Course:	**Probability and Statistics**
Faculty:	**Jordi Caballé**
Term:	First Semester
E-mail:	jordi.caballe@uab.es
Web page:	http://pareto.uab.es/jcaballe/
Office Hours:	Upon request. Send an e-mail for an appointment (Office B3-146).

Description:

This course will cover the topics of probability theory, random variables, distributions, and stochastic processes. It will also cover the traditional topics of sampling, estimation, and hypothesis testing, which are needed for undertaking empirical research.

Slides and exercises can be downloaded from the Campus Virtual of the course https://cv.uab.cat/

Objective:

The objective of this course is to provide the mathematical foundations and the tools of Probability and Statistics that will be used in courses dealing with Economics of Uncertainty and Econometrics.

Outline:

1. Probability

Combinatorics. Events and measurable spaces. Probability. Conditional probability. Theorem of total probability. Bayes' theorem.

2. Random Variables and Distributions

Random variables. Probability distributions. Distribution function and quantile function of a random variable. Discrete random variables and probability functions. Absolutely continuous random variables and densities. Random vectors and marginal distributions. Independent random variables. Generalized conditional probability and distribution.

3. Expectation

Mathematical expectation. Moments. Chebyshev's inequality. The moment-generating function, the characteristic function, and the Laplace transform. Product moments and covariance. Mean and variance of linear combinations of random variables. Conditional expectation. The law of iterated expectations. Jensen's inequality.

4. Special Distributions

The discrete uniform, Bernoulli, binomial, Pascal, geometric and hypergeometric distributions. The multinomial and multivariate hypergeometric distribution. Integration by parts and by change of variable to polar coordinates. The uniform, gamma, exponential, chi-square and beta distributions. The normal and the multivariate normal distributions. Multivariate normality and linear models. The logistic and the generalized extreme value distributions.

5. Functions of Random Variables

The distribution of a function of a random variable. The probability function of a function of a random variable. Convolutions. The density of a function of a random variable. Characteristic function, moment-generating function, and Laplace transform of a function of a random variable. Mixture distributions.

6. Stochastic Processes and Limiting Distributions

Stochastic processes. Filtrations and martingales. Convergence in probability, in mean square, in distribution, and almost sure convergence. Convergence of distribution functions and of probability measures. Markov processes. The Poisson distribution as the limit of binomial distributions. The standard normal distribution as the limit of standardized binomial distributions. Laws of large numbers. The central limit theorem.

7. Sampling

Random samples and statistics. The distribution of the sample mean. The distribution of the variance of a random sample from a normal population. The *t* distribution. The *F* distribution. Order statistics and extreme value theory.

8. Estimation

Point estimation. Efficiency of estimators. The sample mean and sample variance as unbiased estimators. The Cramér-Rao lower bound for unbiased estimators. Asymptotic properties of estimators: consistent estimators. Sufficient estimators and the Rao-Blackwell theorem. The method of moments. Maximum likelihood estimation. Bayesian estimation. Interval estimation.

9. Hypothesis Testing

Statistical hypotheses and their tests. The power function of a test. Uniformly most powerful tests and the Neyman-Pearson lemma. The monotone likelihood ratio property and the Karlin-Rubin theorem. Likelihood ratio, Wald, and score tests. Acceptance intervals. The *p*-value. Contingency tables. Goodness of fit.

References:

Ash, R.B. (1972), Real Analysis and Probability, Academic Press.

Bierens, H.J. (2004), Introduction to the Mathematical and Statistical Foundations of Econometrics, Cambridge University Press.

Billingsley, P. (1995), Probability and Measure, Wiley.

Casella, G.H. and Berger, R.L. (2002), Statistical Inference, Duxbury/Thomson Learning.

DeGroot, M.H. and Schervish, M.J. (2012), Probability and Statistics, Pearson.

Hogg, R.V., McKean, J. and Craig A.T. (2012), Introduction to Mathematical Statistics, Pearson.

Lindgren, B.V. (1993), Statistical Theory, Chapman and Hall/CRC.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974), Introduction to the Theory of Statistics, McGraw Hill.

Rice, J.A. (2007), Mathematical Statistics and Data Analysis, Cengage Learning.

Grading:

Students must solve a series of problem sets. Problem sets will have a weight of 20% in the final grade. There will be a final exam, which will have a weight of 80% in the final grade.

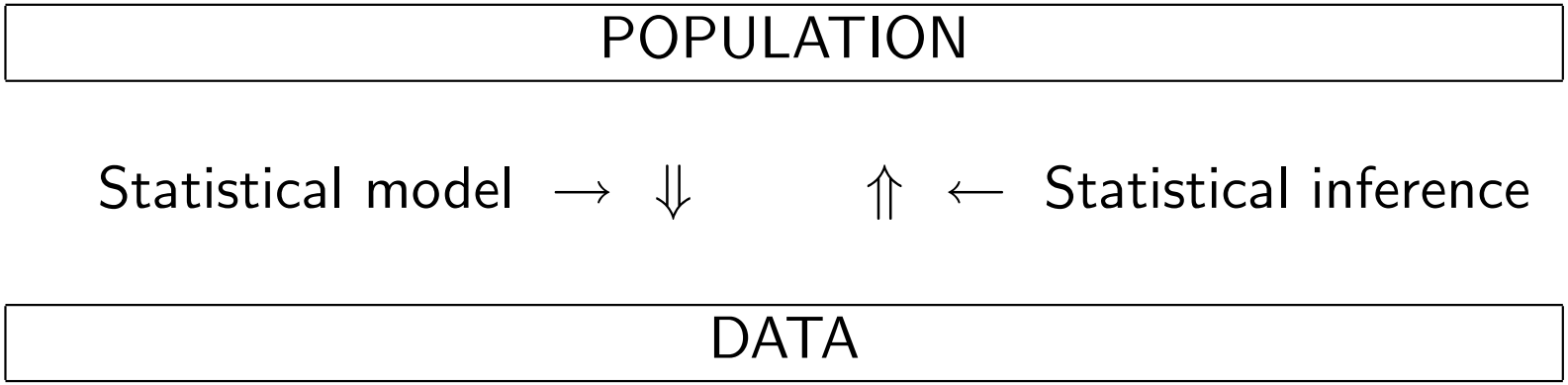
Note: Some of this year's exercises have appeared on this course previously. Thus, it is possible, even likely, that you might be able to obtain solutions to these exercises that I have handed out earlier. However, I strongly recommend you not to look at these solutions when solving the exercises. By handing in your answers, you declare that the solutions are your own and that they are not based on solutions from previous years. If I catch you cheating, I will give you 0 points from the exercises. Even more significantly, you will suffer a reputation loss within IDEA (and academia in general) by presenting someone else's work as your own. Check out the definition of "plagiarism" and how it is viewed in academic circles if you do not immediately grasp what the consequences of cheating will be.

Descriptive Statistics

1. Introduction to Statistics

- Statistics is a mathematical science pertaining to the collection, analysis, interpretation, and presentation of data. (Wikipedia!!).
- Data description (17th century: first censuses in national states).
- Probability theory (18th century: games of chance).
- During the 19th century Statistics is applied to experimental sciences (Biology, Physics, Chemistry).
- During the 20th century Statistics benefits from the development of computers.
- During the 21th century Statistics benefits from the development of both big data analysis and artificial intelligence.

- **Statistical method:**



- **Descriptive statistics** studies how to analyze and present data.
- **Probability theory** explains how data are generated from a population. This is achieved by means of a statistical (or probability) model.
- **Statistical inference** allows us to say something about the population from the available data. This is achieved by "inverting" the statistical model.
- This chapter is about descriptive statistics.

2. Types of variables

- A **population** is a set of people or objects. The elements of a population are called **individuals.**
- A **sample** is a subset of a population.
- Population and sample are relative concepts.
- A **variable** is a characteristic of a population which can take different values (examples: weight, age, color of eyes, income, etc.)

- Two types of variables:
  1. **Qualitative (or categorical) variables**, which are the ones that cannot be measured numerically.
  2. **Quantitative variables**, which are the ones that can be measured numerically, i.e., through numerical values.

- There are two types of quantitative variables:
  1. **Discrete** or **countable**, which are the ones that can take values from a countable set of numbers.
  2. **Continuous**, which are the ones that can take values from an uncountable set of numbers (for instance, the set of real numbers).

- There are two types of discrete (or countable) variables:
  1. **Finite**, which are the ones that take values from a finite set of numbers.
  2. **Infinite**, which are the ones that take values from a countable infinite set of numbers (for instance, the set of rational numbers).

- A variable is represented by a capital letter: X, Y, Z, ...
- The different K values taken by a variable are represented by small letters: $x_1, x_2, ..., x_K$.
- If N is the number of individuals or objects of the population for which we study a characteristic, then $K \leq N$.

3. Univariate frequency distributions: absolute, relative, and cumulative frequencies

- Let us assume that we only care about one single variable X of the population.
- Let us assume that the number of different values taken by this variable is small ($K \leq 10$).
- The absolute frequency $n(x)$ of the value x is the number of times that the value x appears in the data.
- Note that $\sum_x n(x) = N$.
- The distribution of absolute frequencies is the function $n(\cdot)$ defined for all the possible values of the variable X.

- The relative frequency $f(x)$ of the value x is the fraction (or percentage) of times that the value x appears in the data,

$$f(x) = \frac{n(x)}{N}.$$

- Note that

$$\sum_x f(x) = \sum_x \frac{n(x)}{N} = \frac{\sum_x n(x)}{N} = 1.$$

- The cumulative absolute frequency $N(x)$ is the number of times that the variable X takes values smaller or equal than x,

$$N(x) = \sum_{y \leq x} n(y).$$

- The cumulative relative frequency $F(x)$ is the fraction of times that the variable X takes values smaller or equal than x,

$$F(x) = \frac{N(x)}{N} = \frac{\sum_{y \leq x} n(y)}{N} = \sum_{y \leq x} \frac{n(y)}{N} = \sum_{y \leq x} f(y).$$

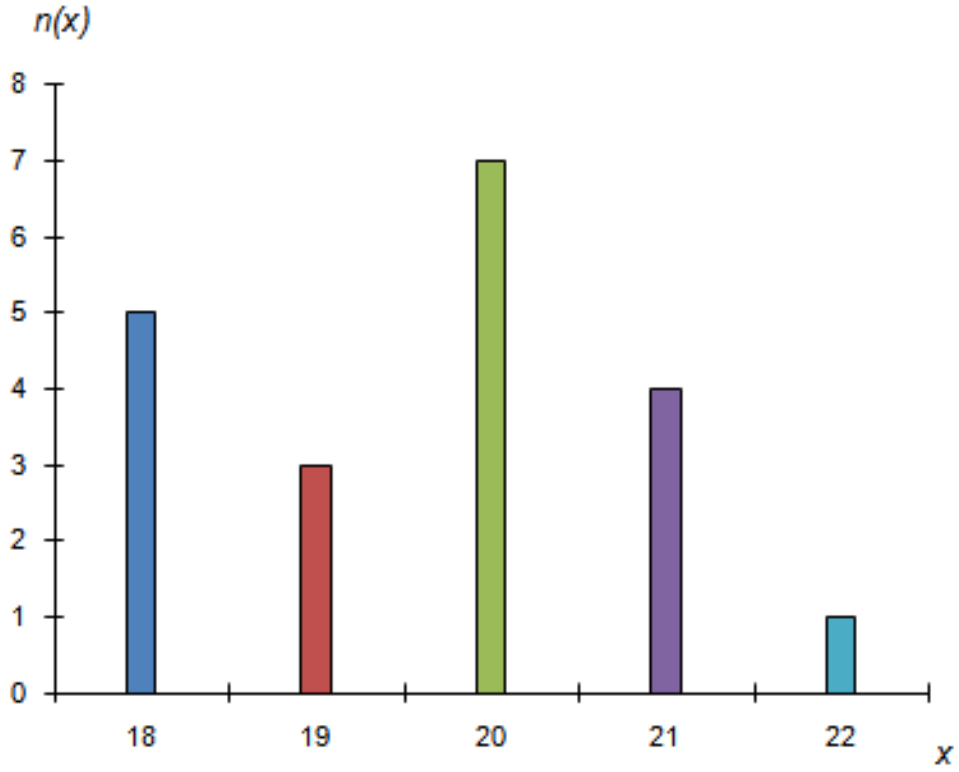
- Note that cumulative frequencies are only well-defined for quantitative variables.

The following table summarizes the distribution of absolute, relative, cumulative absolute and cumulative relative frequencies of the variable "age" for a population (or sample) of 20 students:

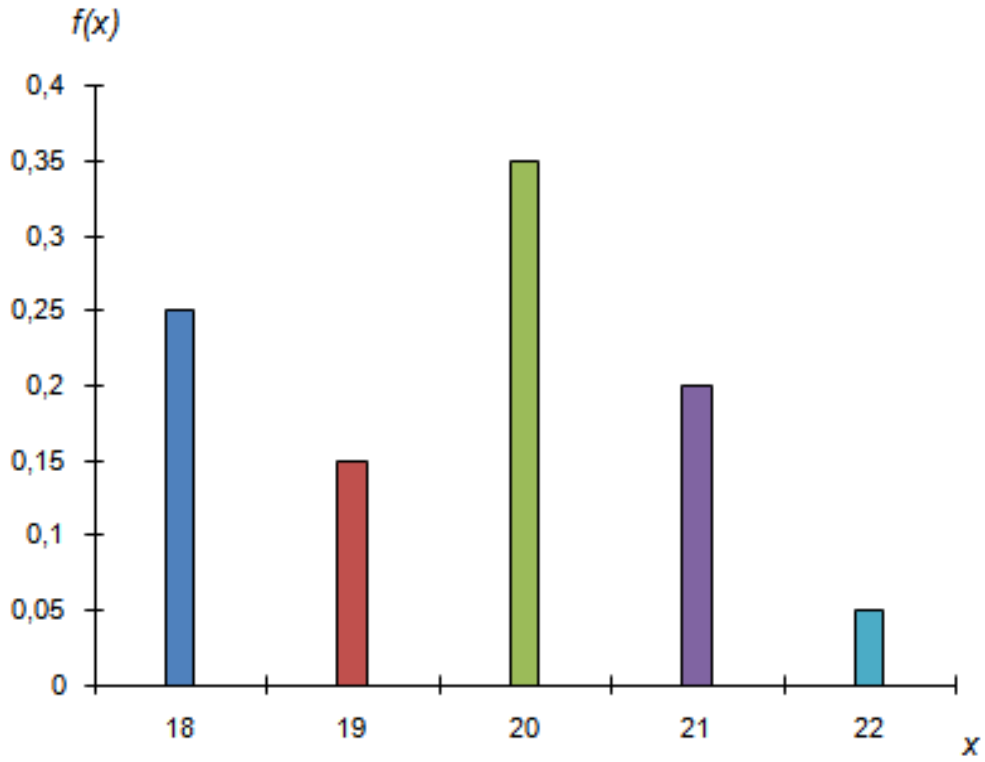
Age x	$n(x)$	$N(x)$	$f(x)$	$F(x)$
18	5	5	0.25	0.25
19	3	8	0.15	0.40
20	7	15	0.35	0.75
21	4	19	0.20	0.95
22	1	20	0.05	1
	$\sum_x n(x) = N = 20$		$\sum_x f(x) = 1$	

4. Graphic representation of frequency distributions: bar diagrams and histograms

Bar diagram of **absolute frequencies** of age**:**



Bar diagram of **relative frequencies** of age:



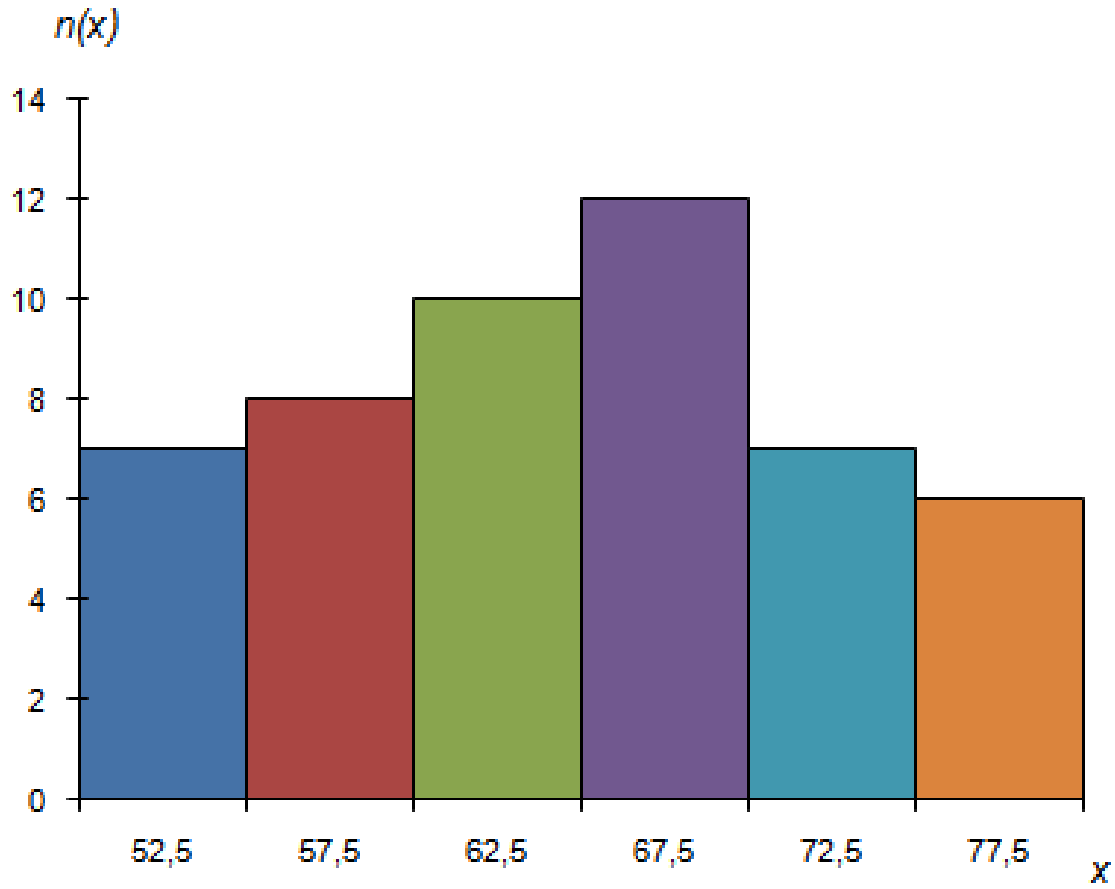
- If the number of values taken by a variable is large ($K > 10$), then we partition the set of values into classes, intervals, or bins. That is, we work with grouped data.
- Usually, we construct from 5 to 10 classes.

Example. Population (or sample) size: 50 students. Variable: weight.

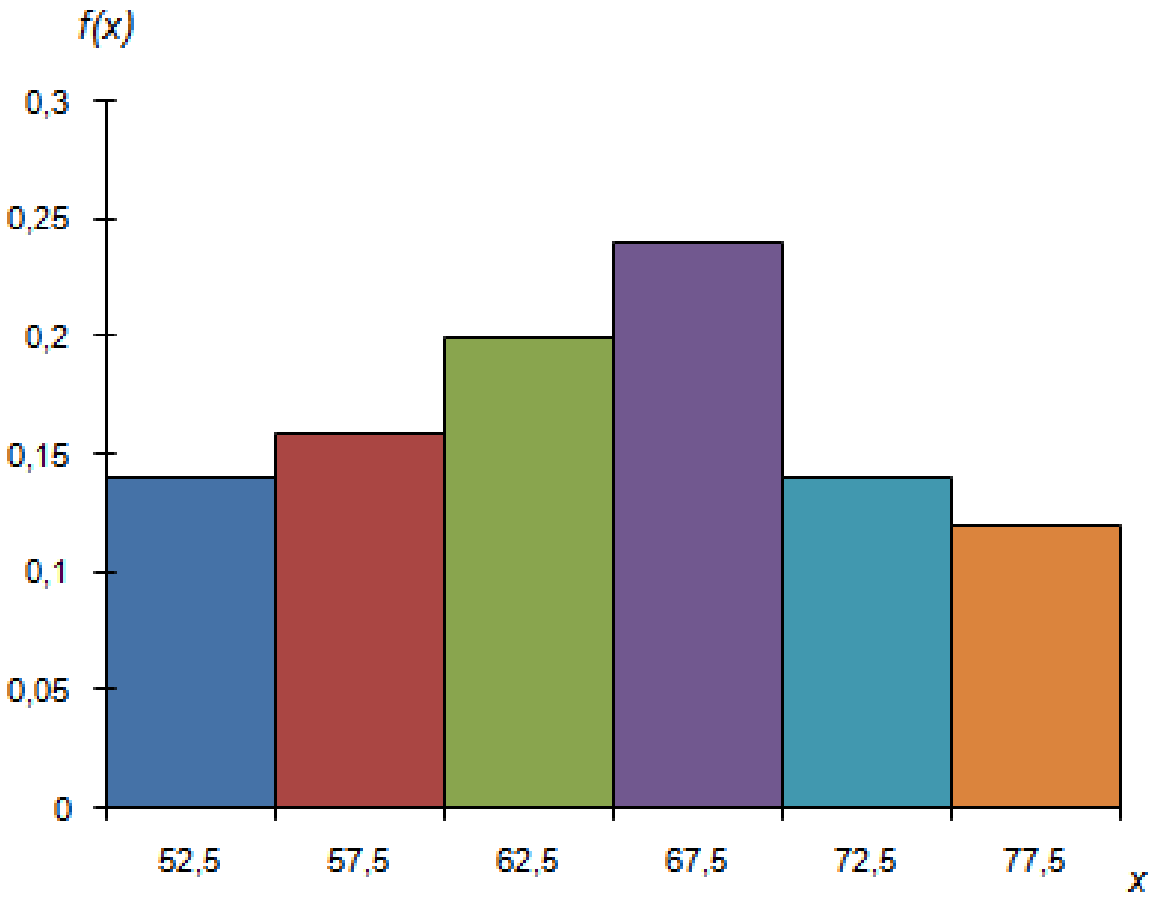
Interval	Midpoint x	Absolute freq. $n(x)$	Cumulat. absolute freq. $N(x)$	Relative freq. $f(x)$	Cumulat. relative freq. $F(x)$
$(50, 55]$	52.5	7	7	0.14	0.14
$(55, 60]$	55.5	8	15	0.16	0.30
$(60, 65]$	62.5	10	25	0.20	0.50
$(65, 70]$	67.5	12	37	0.24	0.74
$(70, 75]$	72.5	7	44	0.14	0.88
$(75, 80]$	77.5	6	50	0.12	1
		$N = 50$		$\sum_x f(x) = 1$	

- Note that the midpoint is the value that represents the interval. Therefore, we are committing an approximation error since we treat all the values of an interval as if they were equal to the value of the midpoint.

- When we work with grouped data we use histograms for the corresponding graphical representation.
- Histogram of absolute frequencies:



- Histogram of relative frequencies:



5. Measures of central tendency

- **Mean (or average value or arithmetic mean) of the variable X :**

$$\overline{X} = \frac{\sum_i x_i}{N} = \sum_x \frac{x \cdot n(x)}{N} = \sum_x x \cdot f(x),$$

where x_i is the value of the variable X for individual i.

- Note that in the sum $\sum_i$ we sum over individuals so that this sum has N terms.
- However, in the sum $\sum_x$ we sum over the different values taken by the variable X so that this sum has K terms.

- **Properties of the mean.**

1. $\sum_i \left(x_i - \overline{X}\right) = 0$ or $\sum_x \left(x - \overline{X}\right) n(x) = 0$ or $\sum_x \left(x - \overline{X}\right) f(x) = 0$.

   Proof.

$$\sum_i \left(x_i - \overline{X}\right) = \sum_i x_i - \sum_i \overline{X} = N\overline{X} - N\overline{X} = 0. \qquad Q.E.D.$$

2. $\overline{kX} = k\overline{X}$, where k is a constant (or scalar).

   Proof. Note that the values taken by the variable kX for the N individuals are $kx_1, kx_2, ..., kx_N$. Therefore,

$$\overline{kX} = \frac{\sum_i kx_i}{N} = k \cdot \left(\frac{\sum_i x_i}{N}\right) = k\overline{X}. \qquad Q.E.D.$$

- **Median:** We order all the values of the variable X taken by the N individuals from the smallest to the largest. We re-index the values accordingly so that

$$x_1 \leq x_2 \leq x_3 \leq ... \leq x_{N-1} \leq x_N.$$

Then, the median of X is

$$Median(X) = \begin{cases} x_{\frac{N}{2}+\frac{1}{2}} & \text{if } N \text{ is odd} \\ \dfrac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} & \text{if } N \text{ is even.} \end{cases}$$

- *Examples:* (a) $N = 7$ and the ordered values are $1, 1, 3, 6, 8, 8, 12$. The median is $x_4 = 6$.
- (b) $N = 8$ and the ordered values are $1, 1, 3, 6, 8, 8, 12, 3472$. The median is

$$\frac{x_4 + x_5}{2} = \frac{6 + 8}{2} = 7.$$

- The mean uses all the values of all individuals but it is more sensitive to outliers and errors.
- The median does not use all the values but, in general, is less sensitive to outliers and errors.

- **Mode:** It is the value that appears a larger number of times, i.e., is the value with a larger absolute (or relative) frequency.
- A variable may have more than one mode.
- If we work with grouped data, then we use the midpoint values on the previous formulae. We are thus committing an approximation error.
- Finally, note that the mean and the median (and the next measures) are defined only for quantitative variables.

6. Measures of variability

- **Variance:** It measures the average of the square of the deviations from the mean,

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{S}_X^2 = \mathrm{S}^2 = \frac{\sum_i \left(x_i - \overline{X}\right)^2}{N} = \sum_x \frac{\left(x - \overline{X}\right)^2 \cdot n(x)}{N} \\ &= \sum_x \left(x - \overline{X}\right)^2 \cdot f(x). \end{aligned}$$

- The subindex X can be omitted when we work with a single variable.
- Obviously, the variance is always non-negative.
- Note that, if we suppress the square, the previous formula is meaningless since, for every variable X, we have

$$\frac{\sum_i \left(x_i - \overline{X}\right)}{N} = 0.$$

- **Properties of the variance**

1. $S_X^2 = \overline{X^2} - \overline{X}^2$.

In words: the variance is equal to the mean of the square minus the square of the mean.

Proof:

$$S_X^2 = \frac{\sum_i \left(x_i - \overline{X}\right)^2}{N} = \frac{1}{N}\sum_i \left[x_i^2 - 2x_i\overline{X} + \overline{X}^2\right]$$

$$= \frac{1}{N}\left[\sum_i x_i^2 - 2\overline{X}\sum_i x_i + N\overline{X}^2\right] = \frac{\sum_i x_i^2}{N} - 2\overline{X}\left(\frac{\sum_i x_i}{N}\right) + \overline{X}^2$$

$$= \overline{X^2} - 2\cdot\overline{X}\cdot\overline{X} + \overline{X}^2 = \overline{X^2} - \overline{X}^2 \ . \qquad Q.E.D.$$

- Thus,

$$S_X^2 = \overline{X^2} - \overline{X}^2 = \frac{\sum_i x_i^2}{N} - \overline{X}^2 = \sum_x \frac{x^2 \cdot n(x)}{N} - \overline{X}^2 = \sum_x x^2 \cdot f(x) - \overline{X}^2.$$

2. $S_{kX}^2 = k^2 S_X^2$ or $\text{Var}(kX) = k^2 \text{Var}(X)$, where k is a scalar.

Proof:

$$S_{kX}^2 = \frac{\sum_i \left(kx_i - \overline{kX}\right)^2}{N} = \frac{\sum_i \left(kx_i - k\overline{X}\right)^2}{N} = \frac{k^2 \sum_i \left(x_i - \overline{X}\right)^2}{N} = k^2 S_X^2 \ .$$

Q.E.D.

- **Standard deviation:**

$$S_X = S = \left(S_X^2\right)^{1/2} \equiv +\sqrt{\text{Var}(X)}.$$

- Note that

$$S_{kX} = kS_X \quad \text{if} \quad k \geq 0.$$

- **Average absolute deviation:**

$$\frac{\sum_i \left|x_i - \overline{X}\right|}{N} = \sum_x \frac{\left|x - \overline{X}\right| \cdot n(x)}{N} = \sum_x \left|x - \overline{X}\right| \cdot f(x).$$

- The average absolute deviation is not differentiable with respect to the value x_i when $x_i = \overline{X}$.

- **Coefficient of variation (or variation coefficient):**

$$\mathrm{CV}_X = \mathrm{CV} = \frac{S_X}{\left|\overline{X}\right|}, \quad \text{when } \left|\overline{X}\right| \neq 0.$$

- Note that

$$\mathrm{CV}_{kX} = \frac{S_{kX}}{\left|\overline{kX}\right|} = \frac{kS_X}{\left|k\overline{X}\right|} = \frac{kS_X}{k\left|\overline{X}\right|} = \frac{S_X}{\left|\overline{X}\right|} = \mathrm{CV}_X \quad \text{if } k > 0,$$

  so that the coefficient of variation is immune to the units of measurement.

- The **range** of the variable X is the difference between the largest and the smallest value taken by the variable.

- The **p% percentile:** We order all the values of the variable X taken by the N individuals from the smallest to the largest. We re-index the values accordingly so that

$$x_1 \leq x_2 \leq x_3 \leq ... \leq x_{N-1} \leq x_N.$$

We compute the integer j (between 1 and N) nearest to the value

$$\frac{p \cdot N}{100} + \frac{1}{2}. \qquad (\star)$$

Then, the $p\%$ percentile is

$$x_{p\%} = x_j,$$

with $p \in [0, 100]$. That is, $x_{p\%}$ is the value of the variable for which a $p\%$ of the individuals exhibits a lower value of the variable.

- If there are two integers j and $j+1$ that are at the same distance from $(\star)$, then the $p\%$ percentile is

$$x_{p\%} = \frac{x_j + x_{j+1}}{2}.$$

- Note that the range equals to $x_{100\%} - x_{0\%}$.
- The **interquartile range** is equal to $x_{75\%} - x_{25\%}$.
- The median is equal to $x_{50\%}$.

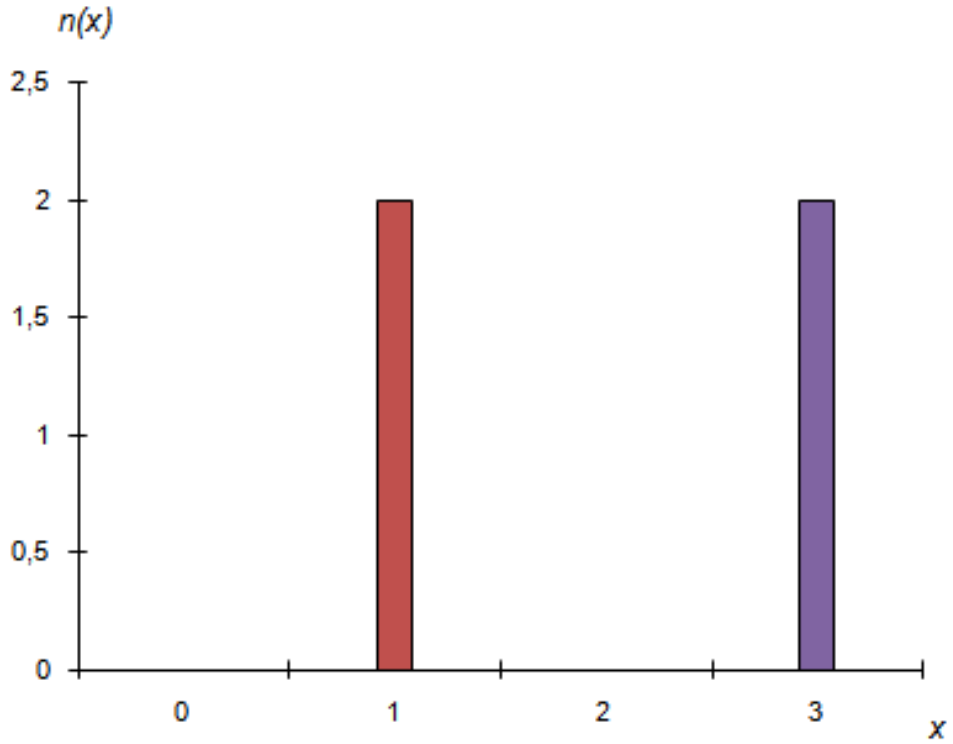
7. Other measures summarizing the shape of a distribution

- **Coefficient of Asymmetry:**

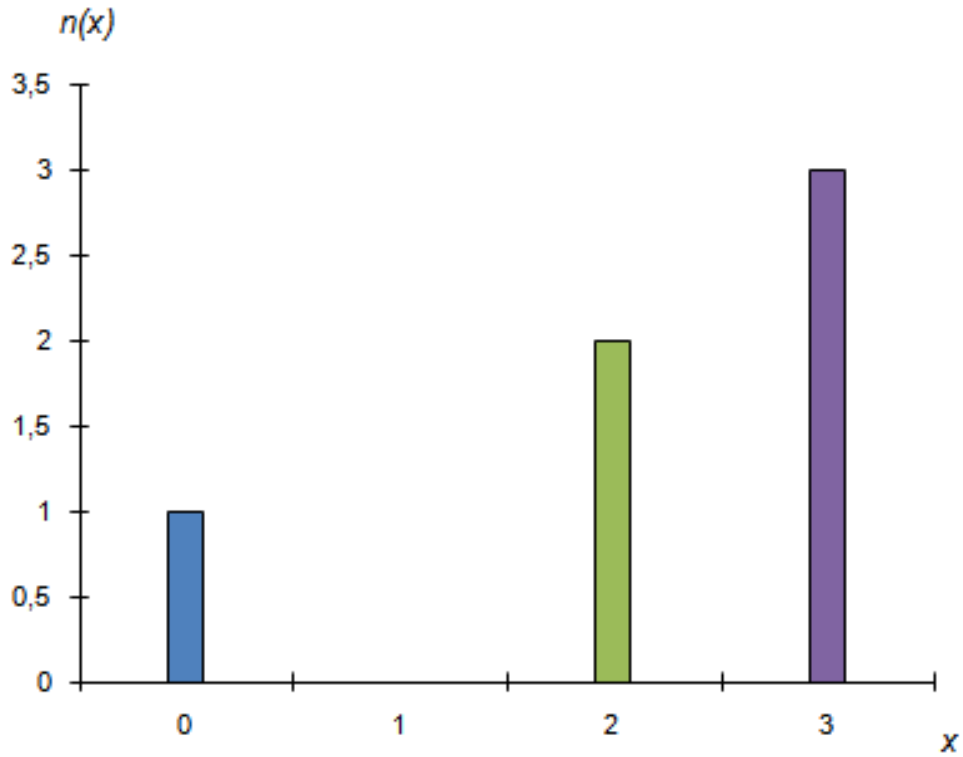
$$\text{CA}_X = \frac{\sum_i \left(x_i - \overline{X}\right)^3}{N \cdot S_X^3} = \frac{\sum_x \left(x - \overline{X}\right)^3 \cdot n(x)}{N \cdot S_X^3} = \frac{\sum_x \left(x - \overline{X}\right)^3 \cdot f(x)}{S_X^3},$$

for $S_X^3 \neq 0$.

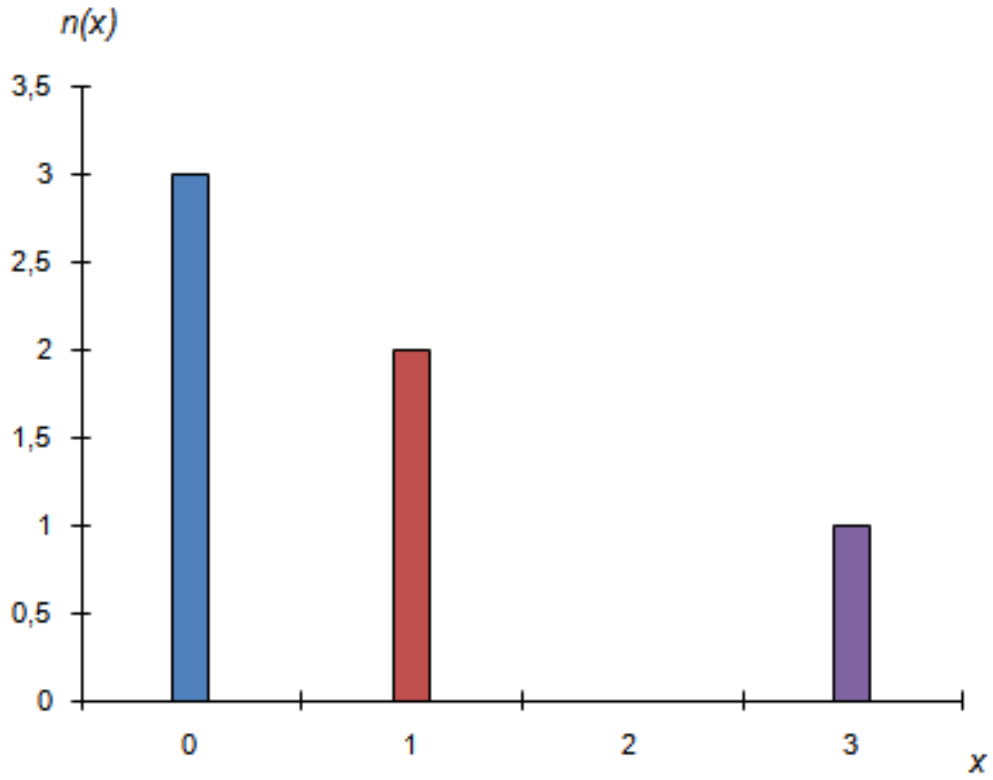
- *Examples:* $\mathrm{CA}_X = 0$. The distribution of x is symmetric.



- $\mathrm{CA}_X < 0$. The left tail is the longest.



- $\mathrm{CA}_X > 0$. The right tail is the longest

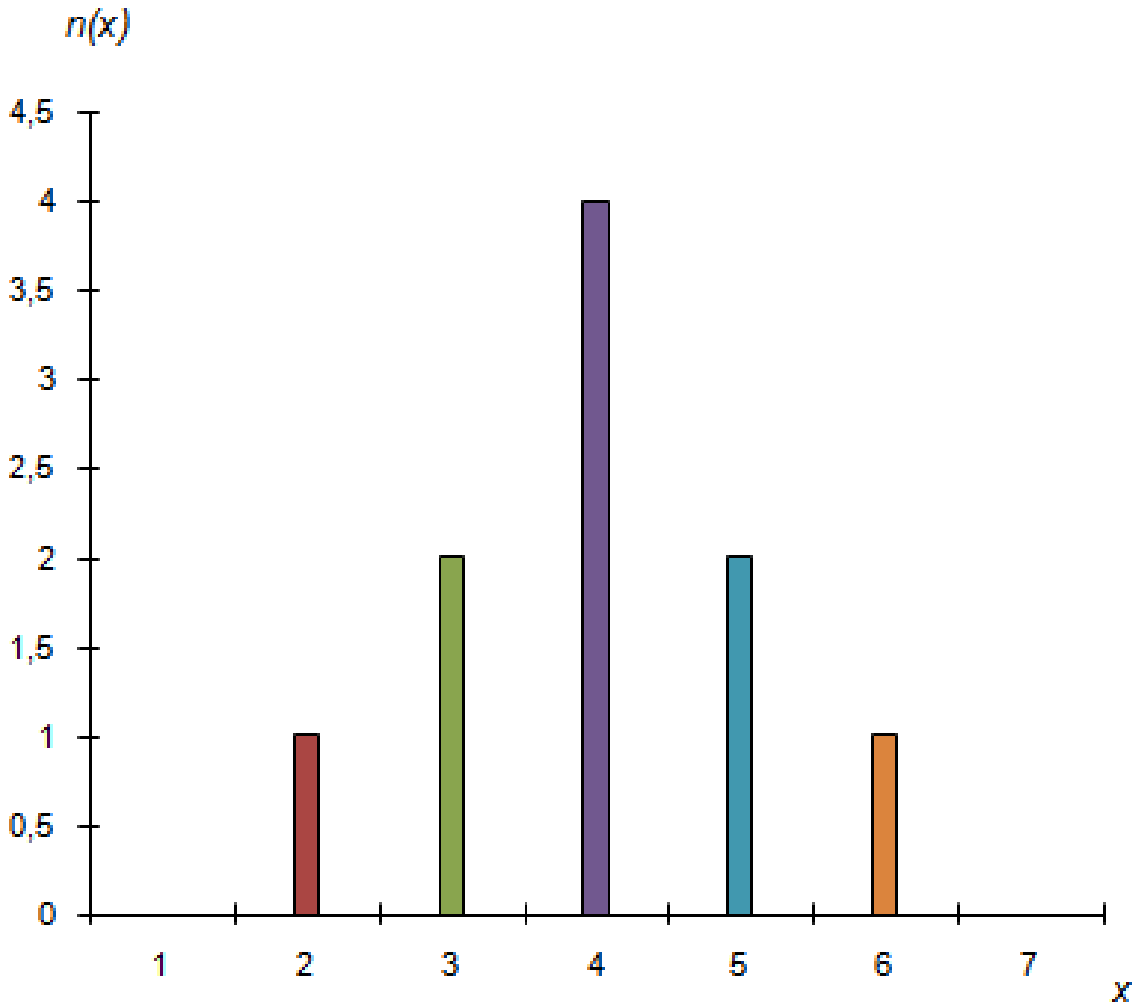


- The **coefficient of Kurtosis** measures the thickness of the tails of the distribution and is given by

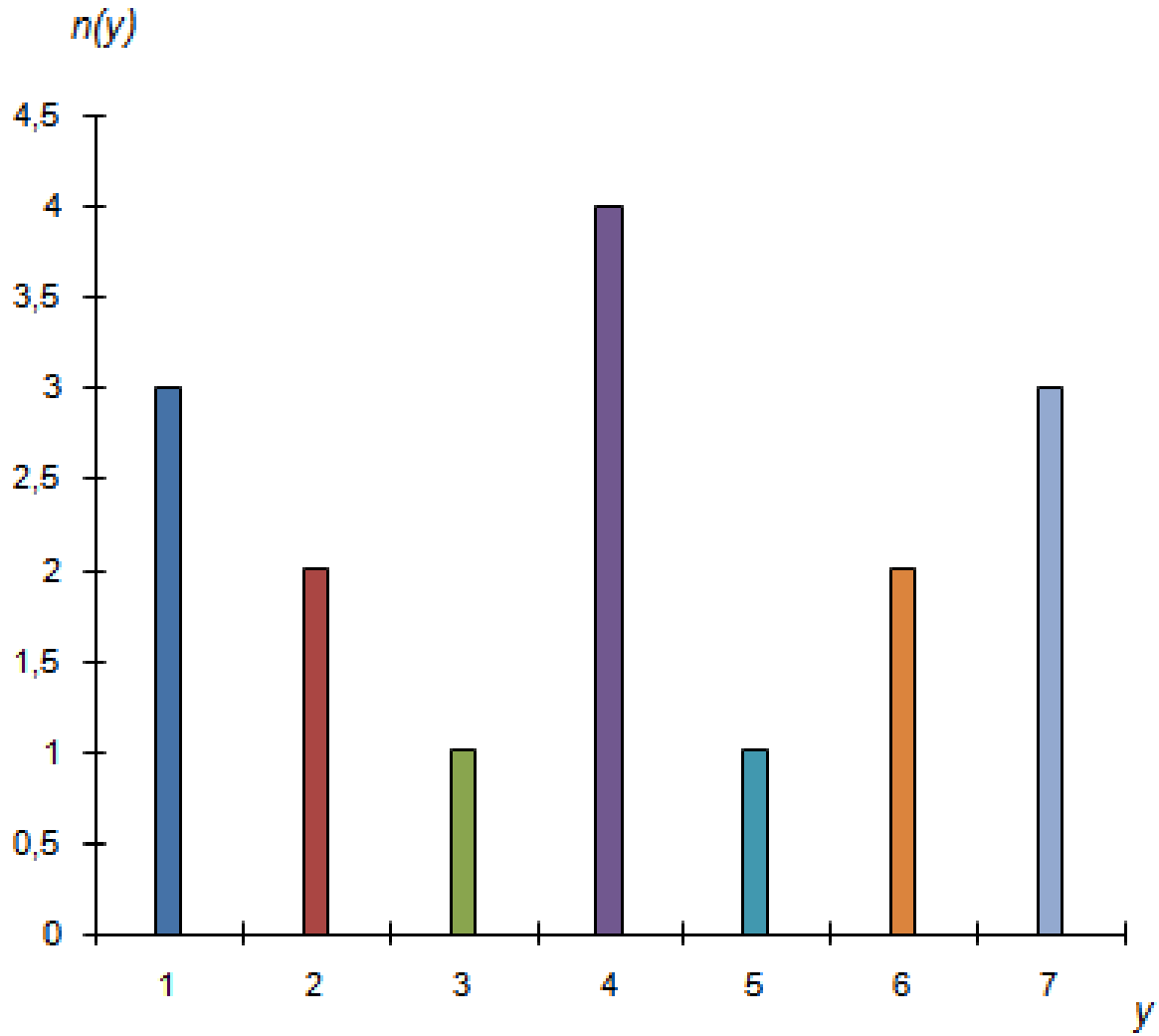
$$\mathrm{CK}_X = \frac{\sum_i \left(x_i - \overline{X}\right)^4}{N \cdot \mathrm{S}_X^4} = \frac{\sum_x \left(x - \overline{X}\right)^4 \cdot n(x)}{N \cdot \mathrm{S}_X^4} = \frac{\sum_x \left(x - \overline{X}\right)^4 \cdot f(x)}{\mathrm{S}_X^4},$$

for $\mathrm{S}_X^4 > 0$.

- *Examples:* Distribution of X



- Distribution of Y



- Then, $\mathrm{CK}_X < \mathrm{CK}_Y$ since the distribution of Y has thicker tails than the distribution of X. In other words, the distribution of X has thinner tails than the distribution of Y.

8. Multivariate frequency distributions

- The multivariate frequency distribution (or joint distribution) gives us the distribution of several variables.
- For instance, the joint distribution of absolute frequencies of two variables X and Y gives us the number of times that each pair of values (x, y) corresponding to the pair of variables (X, Y) appears in the data.
- *Example:* Consider a population consisting of 100 father/son pairs and consider two qualitative variables: $X =$ color of the parent's eyes, $Y =$ color of the son's eyes. This table summarizes the joint distribution of absolute frequencies of these two variables:

		Father's color		
		Light	Dark	
Son's color	Light	25	8	33
	Dark	12	55	67
		37	63	N =100

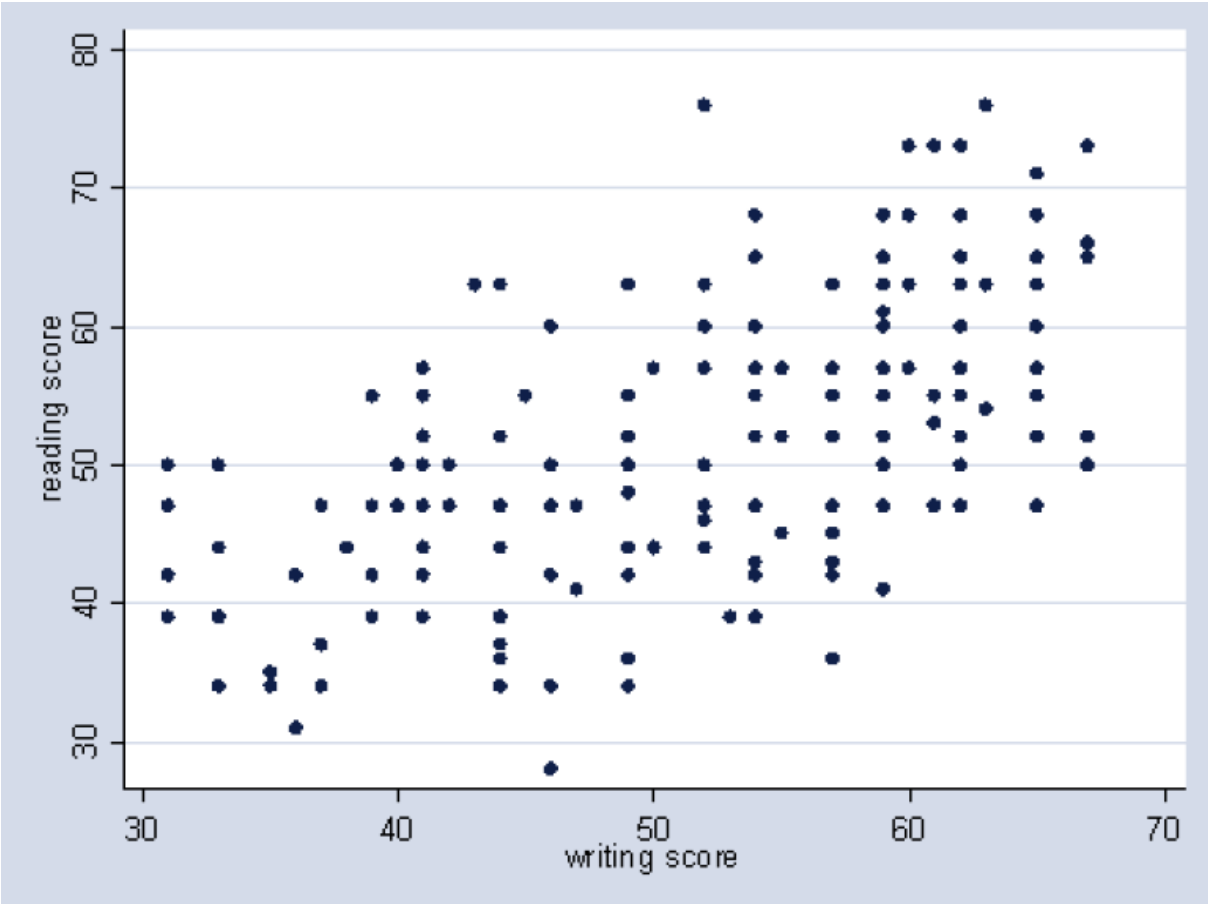
- *Another example:* Consider the quantitative variables X and Y. We have N observations for the values of these two variables.

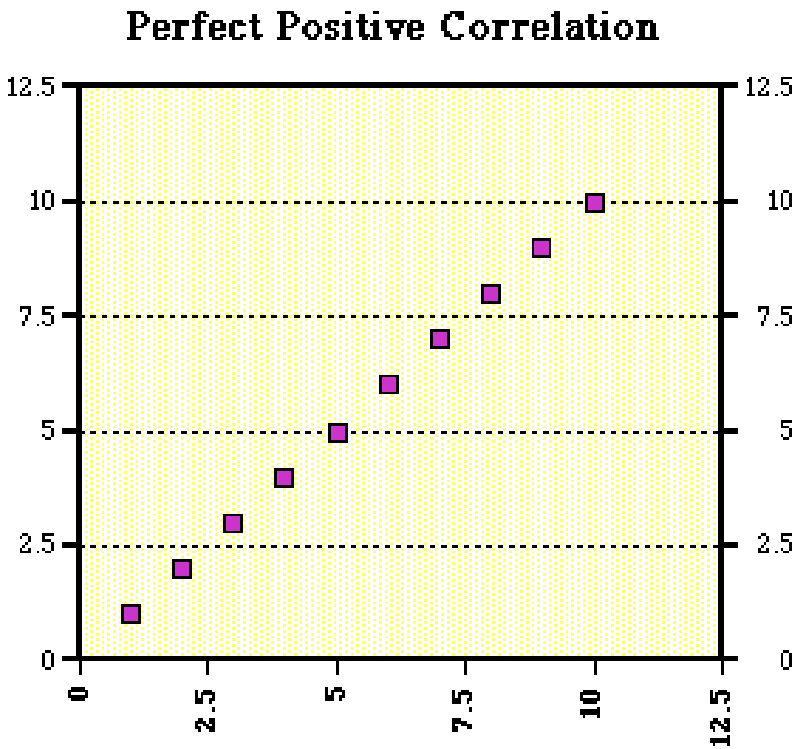
$$X : x_1, x_2, x_3, ..., x_N$$

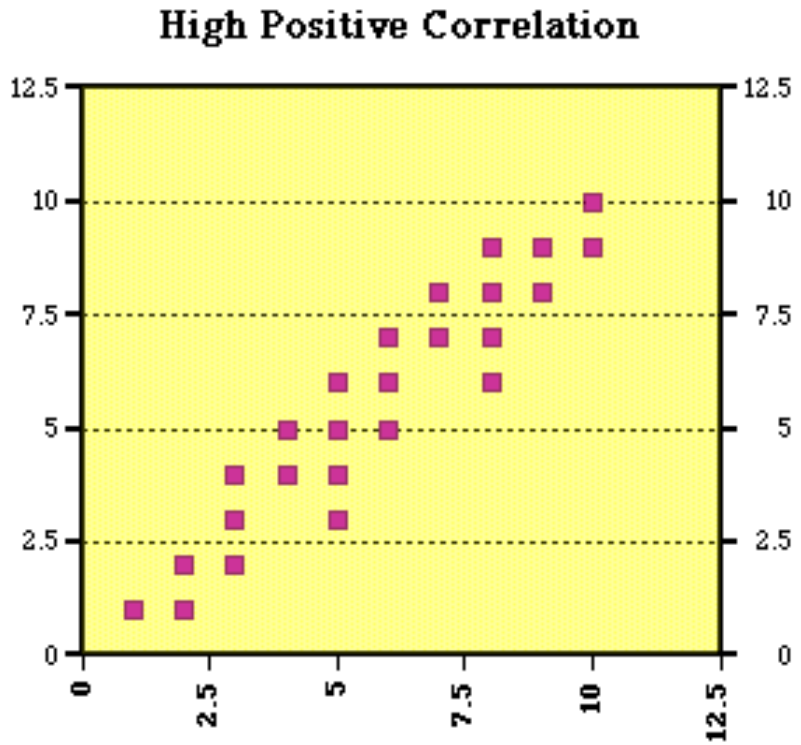
$$Y : y_1, y_2, y_3, ..., y_N$$

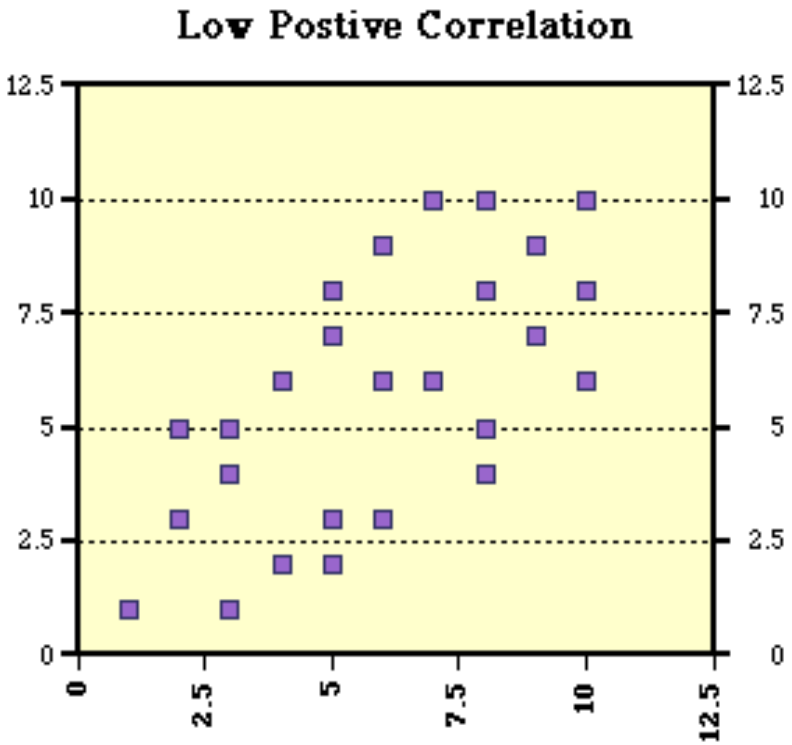
- The pair (x_i, y_i) gives the values of the variables X and Y for individual i.

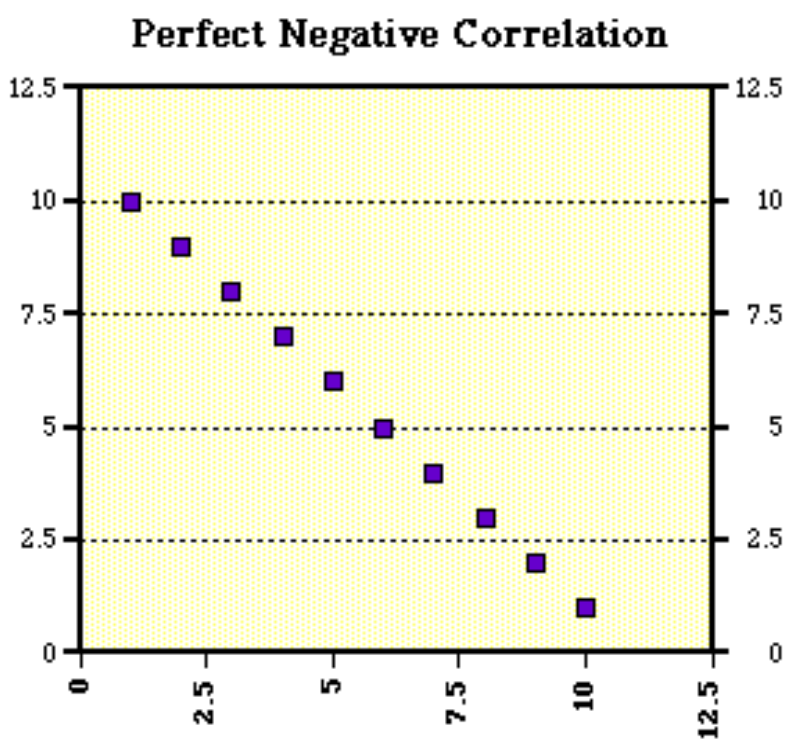
We can summarize these values in a **scatter plot,** which gives us an idea of the type of association (or correlation) between two quantitative variables:

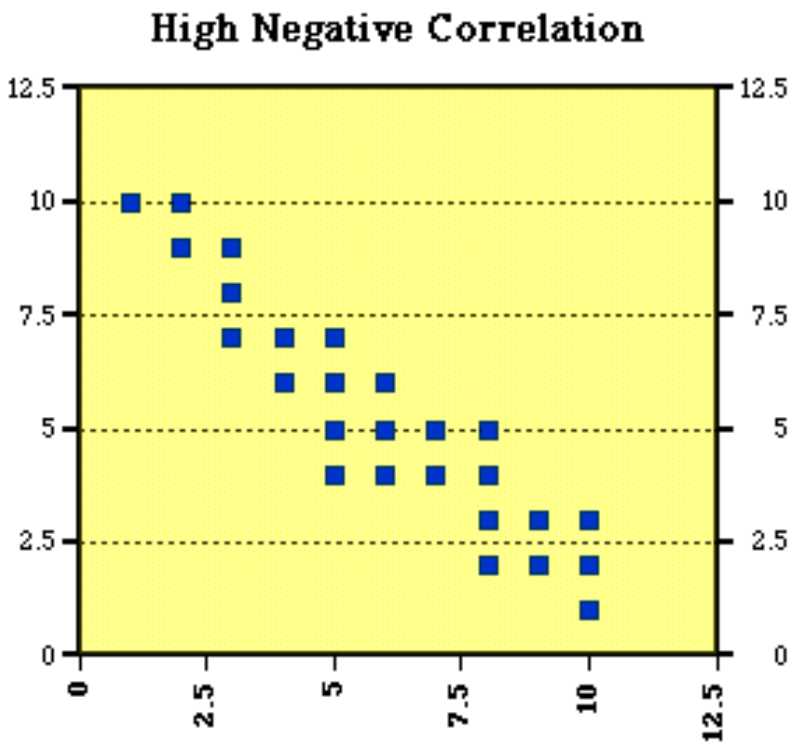


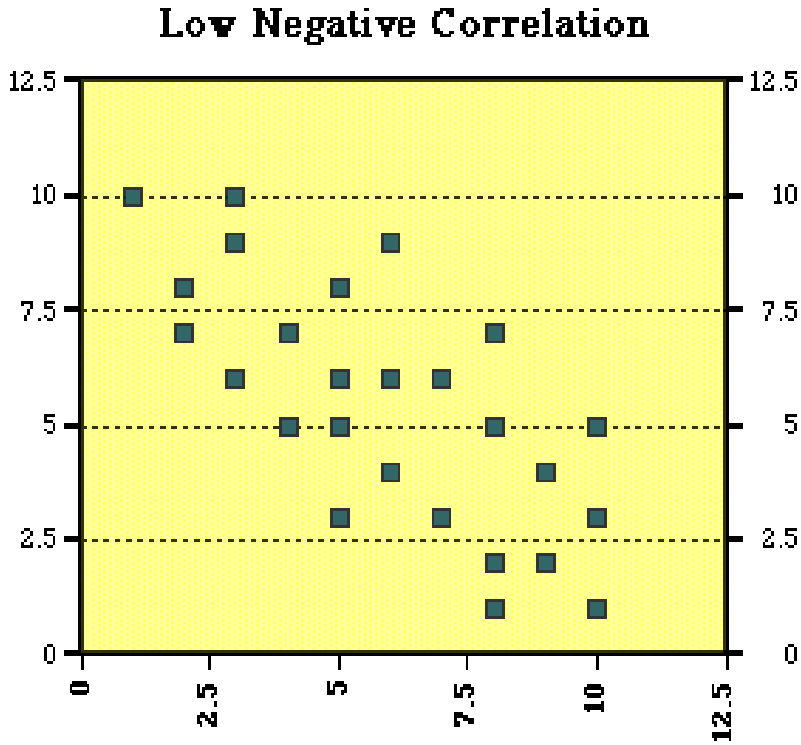

Perfect Positive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5

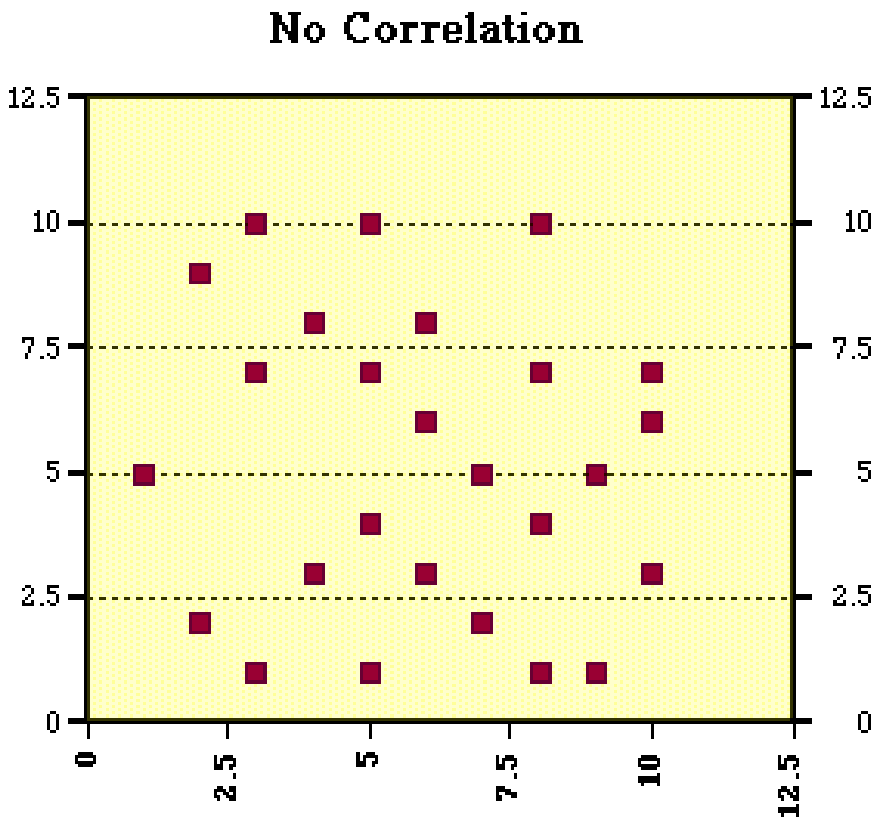

High Positive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Low Postive Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Perfect Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


High Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


Low Negative Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5


No Correlation
12.5
10
7.5
5
2.5
0
0
2.5
5
7.5
10
12.5

- *Another example:* Consider the variables
  $X =$ average monthly temperature in $^{\circ}C$ of a machine.
  $Y =$ number of breakdowns of the same machine in a month.

- We have 100 observations (months) for the values of these two variables.

$$X : x_1, x_2, x_3, ..., x_{100}$$

$$Y : y_1, y_2, y_3, ..., y_{100}$$

- For the variable X (which is continuous) we work with 3 intervals or classes. The variable Y is discrete and, in the data, only takes on the values 2, 3, 4, 5.

- The following table summarizes the distribution of absolute frequencies:

$Y \setminus X$	$(110-130]$ $120°$	$(130-150]$ $140°$	$(150-170]$ $160°$	
2	20	15	10	45
3	12	7	5	24
4	4	10	2	16
5	0	5	10	15
	36	37	27	100

- The absolute frequency $n_{X,Y}(x,y)$ of the pair (x,y) is the number of times that this pair appears in the data. Example: $n_{X,Y}(140,4) = 10$.
- Recall that the joint distribution of absolute frequencies of the variables X and Y is the function $n_{X,Y}(\cdot,\cdot)$ defined for the Cartesian product of the set of values taken by each of these variables.
- Note that $\sum_x \sum_y n_{X,Y}(x,y) = N$.

- The relative frequency $f_{X,Y}(x,y)$ of the pair (x,y) is the fraction (or percentage) of times that this pair appears in the data:

$$f_{X,Y}(x,y) = \frac{n_{X,Y}(x,y)}{N}.$$

- Example: $f_{X,Y}(140,4) = 0.1$.
- The following table summarizes the distribution of relative frequencies:

$Y \setminus X$	$(110-130]$ 120°	$(130-150]$ 140°	$(150-170]$ 160°	
2	0.20	0.15	0.10	0.45
3	0.12	0.07	0.05	0.24
4	0.04	0.10	0.02	0.16
5	0	0.05	0.10	0.15
	0.36	0.37	0.27	1

- Note that $\sum_x \sum_y f_{X,Y}(x,y) = 1$.

9. Marginal and conditional frequencies

- The **distribution of (absolute / relative) marginal frequencies of the variable** X is the frequency distribution of this variable with independency of the values taken by the other variables.
- The absolute marginal frequency $n_X(x)$ of the value x taken by the variable X is

$$n_X(x) = \sum_y n_{X,Y}(x,y).$$

- The relative marginal frequency $f_X(x)$ of the value x taken by the variable X is

$$f_X(x) = \frac{n_X(x)}{N} = \frac{\sum_y n_{X,Y}(x,y)}{N} = \sum_y \frac{n_{X,Y}(x,y)}{N} = \sum_y f_{X,Y}(x,y).$$

- Example: $n_Y(3) = 24$ and $f_Y(3) = 0.24$.

- Note that

$$\overline{X} = \frac{\sum_i x_i}{N} = \sum_x \frac{x \cdot n_X(x)}{N} = \sum_x \frac{x \cdot \sum_y n_{X,Y}(x,y)}{N}$$

$$= \sum_x \frac{\sum_y x \cdot n_{X,Y}(x,y)}{N} = \frac{\sum_x \sum_y x \cdot n_{X,Y}(x,y)}{N}.$$

or

$$\overline{X} = \sum_x x \cdot f_X(x) = \sum_x x \cdot \sum_y f_{X,Y}(x,y) = \sum_x \sum_y x \cdot f_{X,Y}(x,y),$$

and similarly for the formulae for the variance and other measures.

- The **distribution of conditional frequencies of the variable X given $Y = y$** is the relative frequency distribution of the variable X for all the observations where $Y = y$.
- The conditional frequency $f_{X|Y}(x|y)$ of the value x taken by the variable X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{n_{X,Y}(x,y)}{n_Y(y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

- Note that

$$f_{Y|X}(y|x) = \frac{n_{X,Y}(x,y)}{n_X(x)} = \frac{f_{X,Y}(x,y)}{f_X(x)} \neq \frac{f_{X,Y}(x,y)}{f_Y(y)} = f_{X|Y}(x|y).$$

- Example:

$$f_{X|Y}(160|3) = \frac{n_{X,Y}(160,3)}{n_Y(3)} = \frac{5}{24} = \frac{f_{X,Y}(160,3)}{f_Y(3)} = \frac{0.05}{0.24} = 0.2083,$$

whereas

$$f_{Y|X}(3|160) = \frac{n_{X,Y}(160,3)}{n_X(160)} = \frac{5}{27} = \frac{f_{X,Y}(160,3)}{f_X(160)} = \frac{0.05}{0.27} = 0.1852.$$

10. Covariance and correlation coefficient

- Consider the quantitative variables X and Y. We have N observations for the values of these two variables,

$$X : x_1, x_2, x_3, ..., x_N$$

$$Y : y_1, y_2, y_3, ..., y_N$$

- The **covariance** between X and Y measures the strength of the association between these two variables and is given by

$$\begin{aligned} S_{X,Y} = \text{Cov}(X, Y) &= \frac{\sum_i (x_i - \overline{X}) \cdot (y_i - \overline{Y})}{N} \\ &= \frac{\sum_x \sum_y (x - \overline{X}) \cdot (y - \overline{Y}) \cdot n_{X,Y}(x, y)}{N} \\ &= \sum_x \sum_y (x - \overline{X}) \cdot (y - \overline{Y}) \cdot f_{X,Y}(x, y). \end{aligned}$$

- Note that the covariance can be positive, negative or equal to zero.
- **Properties of the covariance**

  1. $S_{X,X} = S_X^2$.

  2. $S_{X,Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y}$.

- *In words:* the covariance is equal to the mean of the product minus the product of the means.

Proof:

$$S_{X,Y} = \frac{\sum_i \left(x_i - \overline{X}\right) \cdot \left(y_i - \overline{Y}\right)}{N} = \frac{1}{N} \sum_i \left[x_i \cdot y_i - x_i \cdot \overline{Y} - y_i \cdot \overline{X} + \overline{X} \cdot \overline{Y}\right]$$

$$= \frac{1}{N} \left[\sum_i x_i \cdot y_i - \overline{Y} \sum_i x_i - \overline{X} \sum_i y_i + N \cdot \overline{X} \cdot \overline{Y}\right]$$

$$= \frac{\sum_i x_i \cdot y_i}{N} - \overline{Y} \left(\frac{\sum_i x_i}{N}\right) - \overline{X} \left(\frac{\sum_i y_i}{N}\right) + \overline{X} \cdot \overline{Y}$$

$$= \overline{X \cdot Y} - \overline{Y} \cdot \overline{X} - \overline{X} \cdot \overline{Y} + \overline{X} \cdot \overline{Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y} \; . \qquad Q.E.D.$$

- Thus,

$$S_{X,Y} = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y} = \frac{\sum_i x_i \cdot y_i}{N} - \overline{X} \cdot \overline{Y}$$

$$= \sum_x \sum_y \frac{x \cdot y \cdot n_{X,Y}(x,y)}{N} - \overline{X} \cdot \overline{Y} = \sum_x \sum_y x \cdot y \cdot f_{X,Y}(x,y) - \overline{X} \cdot \overline{Y}.$$

3. $S_{\alpha X,\beta Y} = \alpha \cdot \beta \cdot S_{X,Y}$, where α and β are scalars.

Proof:

$$S_{\alpha X,\beta Y} = \frac{\sum_i \left(\alpha x_i - \overline{\alpha X}\right)\left(\beta y_i - \overline{\beta Y}\right)}{N} = \frac{\sum_i \left(\alpha x_i - \alpha \overline{X}\right)\left(\beta y_i - \beta \overline{Y}\right)}{N}$$

$$= \frac{\alpha \cdot \beta \cdot \sum_i \left(x_i - \overline{X}\right) \cdot \left(y_i - \overline{Y}\right)}{N} = \alpha \cdot \beta \cdot S_{X,Y} \ . \qquad Q.E.D.$$

4. $S_{X,Y} = S_{Y,X}$.

- The **coefficient of correlation** $r_{X,Y}$ between the variables X and Y is given by

$$r_{X,Y} = \frac{S_{X,Y}}{S_X \cdot S_Y} \quad \text{when } S_X > 0 \text{ and } S_Y > 0.$$

- Note that $r_{\alpha X,\beta Y} = r_{X,Y}$ if $\alpha > 0$ and $\beta > 0$ so that the coefficient of correlation is immune to the units of measurement.

- Another interesting property of the coefficient of correlation is that $|r_{X,Y}| \leq 1$, that is, $-1 \leq r_{X,Y} \leq 1$ (see the proof in the handout).

- In the temperature-breakdowns example we can compute (please do it!) the following:

$$\overline{X} = 138.2, \quad \overline{Y} = 3.01,$$

$$S_X = 15.77 > S_Y = 1.1, \text{ and } CV_X = 0.11 < CV_Y = 0.37,$$

$$S_{X,Y} = 5.62 > 0, \; r_{X,Y} = 0.32 > 0.$$

11. Mean and variance of linear combinations of variables

- Let us consider the variables X_1 and X_2,

$$X_1 : x_{11}, x_{12}, ..., x_{1N}$$

$$X_2 : x_{21}, x_{22}, ..., x_{2N}$$

and consider the scalars α_1 and α_2. Then, the values taken by the variable

$$Z = \alpha_1 X_1 + \alpha_2 X_2$$

are

$$Z : \underbrace{\alpha_1 x_{11} + \alpha_2 x_{21}}_{z_1}, \underbrace{\alpha_1 x_{12} + \alpha_2 x_{22}}_{z_2}, ..., \underbrace{\alpha_1 x_{1N} + \alpha_2 x_{2N}}_{z_N}$$

- The mean of the variable Z is

$$\overline{Z} = \frac{\sum_i z_i}{N} = \frac{\sum_i (\alpha_1 x_{1i} + \alpha_2 x_{2i})}{N} = \frac{\sum_i \alpha_1 x_{1i} + \sum_i \alpha_2 x_{2i}}{N}$$

$$= \frac{\alpha_1 \sum_i x_{1i} + \alpha_2 \sum_i x_{2i}}{N} = \frac{\alpha_1 \sum_i x_{1i}}{N} + \frac{\alpha_2 \sum_i x_{2i}}{N}$$

$$= \alpha_1 \cdot \left(\frac{\sum_i x_{1i}}{N} \right) + \alpha_2 \cdot \left(\frac{\sum_i x_{2i}}{N} \right) = \alpha_1 \overline{X}_1 + \alpha_2 \overline{X}_2 \ .$$

- *In words:* the mean of a linear combination of variables is equal to the linear combination of means.
- In particular, if $Z = X + Y$, then

$$\overline{Z} = \overline{X} + \overline{Y} \ .$$

The variance of the variable Z is

$$S_Z^2 = \frac{\sum_i \left(z_i - \overline{Z}\right)^2}{N} = \frac{\sum_i \left[\alpha_1 x_{1i} + \alpha_2 x_{2i} - \left(\alpha_1 \overline{X}_1 + \alpha_2 \overline{X}_2\right)\right]^2}{N}$$

$$= \frac{\sum_i \left[\alpha_1 \left(x_{1i} - \overline{X}_1\right) + \alpha_2 \left(x_{2i} - \overline{X}_2\right)\right]^2}{N}$$

$$= \frac{\sum_i \left[\alpha_1^2 \left(x_{1i} - \overline{X}_1\right)^2 + \alpha_2^2 \left(x_{2i} - \overline{X}_2\right)^2 + 2\alpha_1\alpha_2 \left(x_{1i} - \overline{X}_1\right)\left(x_{2i} - \overline{X}_2\right)\right]}{N}$$

$$= \alpha_1^2 \cdot \left[\frac{\sum_i \left(x_{1i} - \overline{X}_1\right)^2}{N}\right] + \alpha_2^2 \cdot \left[\frac{\sum_i \left(x_{2i} - \overline{X}_2\right)^2}{N}\right]$$

$$+2\alpha_1\alpha_2 \left[\frac{\sum_i \left(x_{1i} - \overline{X}_1\right)\left(x_{2i} - \overline{X}_2\right)}{N}\right] = \alpha_1^2 \cdot S_{X_1}^2 + \alpha_2^2 \cdot S_{X_2}^2 + 2\alpha_1\alpha_2 S_{X_1,X_2}.$$

- In particular, if $Z = X + Y$, then

$$S_Z^2 = S_X^2 + S_Y^2 + 2S_{X,Y} \ .$$

12. The mean vector and the variance covariance matrix

- Consider the following (column) vector of M variables:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{pmatrix}.$$

- The mean vector (or vector of means) of the vector of variables X is

$$\overline{X} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_M \end{pmatrix},$$

where $\overline{X}_j$ is the mean of the variable X_j.

- The variance covariance matrix (or covariance matrix) of the vector X of variables is the following $M \times M$ matrix

$$\mathrm{S} = \begin{pmatrix} \mathrm{S}_1^2 & \mathrm{S}_{12} & \cdots & \mathrm{S}_{1M} \\ \mathrm{S}_{21} & \mathrm{S}_2^2 & \cdots & \mathrm{S}_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{S}_{M1} & \mathrm{S}_{M2} & & \mathrm{S}_M^2 \end{pmatrix},$$

  where $\mathrm{S}_{jq} = \mathrm{S}_{X_j,X_q}$ and $\mathrm{S}_j^2 = \mathrm{S}_{X_j}^2 = \mathrm{S}_{X_j,X_j} = \mathrm{S}_{jj}$.

- Since $\mathrm{S}_{jq} = \mathrm{S}_{qj}$, for all pairs (j, q), the matrix S is symmetric.

- Define the vector of scalars

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \end{pmatrix}.$$

- Define the variable Z as a linear combination of the variables appearing in X,

$$Z = \sum_j \alpha_j X_j = \alpha^\intercal X,$$

where $^\intercal$ denotes the transpose.

- Then,

$$\overline{Z} = \sum_j \alpha_j \overline{X}_j = \alpha^\intercal \overline{X}$$

and

$$S_Z^2 = \sum_j \alpha_j^2 S_j^2 + 2 \sum_{\substack{q \\ j<q}} \sum_j \alpha_j \alpha_q S_{jq} = \sum_j \alpha_j^2 S_j^2 + \sum_{\substack{q \\ j \neq q}} \sum_j \alpha_j \alpha_q S_{jq}$$

$$= \sum_j \sum_q \alpha_j \alpha_q S_{jq} = \alpha^\intercal S \alpha, \quad \text{where}$$

- $\sum\limits_{\substack{q \\ j<q}} \sum\limits_j$ means $\sum\limits_{q=j+1}^{M} \sum\limits_{j=1}^{M-1}$,
- $\sum\limits_{\substack{q \\ j \neq q}} \sum\limits_j$ means $\sum\limits_{\substack{q=1 \\ q \neq j}}^{M} \sum\limits_{j=1}^{M}$,
- $\sum\limits_j \sum\limits_q$ means $\sum\limits_{j=1}^{M} \sum\limits_{q=1}^{M}$.

- Note that $S_Z^2 = \alpha^\intercal S \alpha \geq 0$ for all vector of scalars α. Therefore, the covariance matrix is (symmetric) positive semi-definite. This implies that the determinant of the matrix S is non-negative.

The Correlation Coefficient $\mathrm{r}_{X,Y}$

The correlation coefficient $\mathrm{r}_{X,Y}$ between two variables X and Y satisfies $-1 \leq \mathrm{r}_{X,Y} \leq 1$ or, equivalently, $|\mathrm{r}_{X,Y}| \leq 1$.

Proof. Consider the symmetric 2×2 variance covariance matrix of the vector of variables (X, Y),

$$\mathrm{S} = \begin{pmatrix} \mathrm{S}_X^2 & \mathrm{S}_{X,Y} \\ \\ \mathrm{S}_{X,Y} & \mathrm{S}_Y^2 \end{pmatrix}.$$

Since for every vector of scalars,

$$\alpha = \begin{pmatrix} \alpha_1 \\ \\ \alpha_2 \end{pmatrix},$$

it is true that $\mathrm{Var}(\alpha_1 X + \alpha_2 Y) = \alpha^\top \mathrm{S} \alpha \geq 0$, it follows that the matrix S is symmetric positive semi-definite. This means that the determinants of all the main minors of S are non-negative, that is, $\mathrm{S}_X^2 \geq 0$, $\mathrm{S}_Y^2 \geq 0$, and $\det(\mathrm{S}) \geq 0$. Note that

$$\det(\mathrm{S}) = \mathrm{S}_X^2 \mathrm{S}_Y^2 - (\mathrm{S}_{X,Y})^2 = \mathrm{S}_X^2 \mathrm{S}_Y^2 \left[1 - \left(\frac{\mathrm{S}_{X,Y}}{\mathrm{S}_X \mathrm{S}_Y} \right)^2 \right] = \mathrm{S}_X^2 \mathrm{S}_Y^2 \left[1 - (\mathrm{r}_{X,Y})^2 \right] \geq 0.$$

Since $\mathrm{S}_X^2 \mathrm{S}_Y^2 > 0$, we have that $\det(\mathrm{S}) \geq 0$ if and only if $(\mathrm{r}_{X,Y})^2 \leq 1$. That is, it must hold that $|\mathrm{r}_{X,Y}| \leq 1$ or, equivalently $-1 \leq \mathrm{r}_{X,Y} \leq 1$. *Q.E.D.*

Exercises. Probability and Statistics. IDEA.
Descriptive Statistics

1. (a) Compute the mean, the median and the standard deviation of the following data in cm: 28, 22, 35, 42, 44, 53, 58, 41, 40, 32, 31, 38, 37, 61, 25, 35.

   (b) Classify the previous data into 5 classes (or intervals) with a length of 10 cm each ($(20-30]$; $(30-40;]$; etc.) and compute the previous characteristic measures using the formulae for grouped data.

   (c) Draw the histogram for these grouped data.

2. Find the value of a minimizing $\sum_{i=1}^{n}(x_i - a)^2$. Discuss.

3. Prove that the arithmetic mean of the variable Z obtained by adding the data of two variables, X and Y, equals the sum of the arithmetic means of these variables.

4. Prove that, if we construct a variable Z by mixing n_1 values of X and n_2 values of Y, the mean of Z is

$$\bar{Z} = \left(\frac{n_1}{n_1+n_2}\right)\bar{X} + \left(\frac{n_2}{n_1+n_2}\right)\bar{Y},$$

   where $\bar{X}$ and $\bar{Y}$ are the means of the two initial variables.

5. If $Z = X + Y$, prove that the variance of Z could be greater or smaller than the sum of the variances of the summands.

6. In 1879 Michelson obtained the following values for the speed of the light in the air (we provide the results subtracting 299.000 from the original data, in km/sec, to make easier the computations): 850, 740, 900, 1070, 930, 850, 950, 980, 980, 880, 1000, 980, 930, 650, 760.

   In 1882 Newcomb, using another procedure, obtained (subtracting again 299.000): 883, 816, 778, 796, 682, 711, 611, 599, 1051, 781, 578, 796, 774, 820, 772.

   (a) Compute the mean and the standard deviations.

   (b) What conclusions could you infer from (a)?

7. The geometric mean of the variable X is defined as

$$G = (x_1 \cdot x_2 \cdot .... \cdot x_n)^{1/n}$$

   and the harmonic mean of X as

$$H = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}\right)^{-1}.$$

You are asked

(a) to find the relationship between the geometric mean and the mean of the logarithm of the original variable X;

(b) the same between H and the mean of Y, where $Y = X^{-1}$.

8. Prove that, if there exists a exact linear relationship between two variables, $Y = a + bX$, with $b > 0$, then their correlation coefficient is equal to one.

9. Compute the variance and the correlation coefficient of the following data:

$$X: \quad 2 \quad 1 \quad 2 \quad 1 \quad 4$$

$$Y: \quad 8 \quad 9 \quad 8 \quad 5 \quad 10$$

10. Consider the following joint distribution of absolute frequencies $n(x, y)$ of the variables $X =$ age and $Y =$ weight of a group of 42 secondary school boys. There are only two ages: 13 and 14 years. The weights (in kilograms) are grouped into three intervals (or classes): (35,45], (45,55] and (55,65], which are parametrized by their corresponding midpoints (40, 50 and 60).

		$Y =$ weight		
		40	50	60
$X =$ age	13	4	7	7
	14	3	8	13

You are asked to

(a) compute the distributions of marginal relative frequencies of age, $f_X(x)$ for $x = 13, 14,$ and weight, $f_Y(y)$ for $y = 40, 50, 60$;

(b) compute the means of age $\bar{X}$ and weight $\bar{Y}$;

(c) compute the variances of age S_X^2 and weight S_Y^2;

(d) compute the covariance between age and weight S_{XY} and the corresponding correlation coefficient r_{XY};

(e) compute the distribution of conditional (relative) frequencies of age given that the weight is 60, $f_{X|Y}\,(x\,|60)$ for $x = 13, 14$.

11. Let the variable X be the number of theater attendances and Y the number of cinema attendances. Consider the following table of monthly relative frequencies of theater and cinema attendances for a group of retired people:

		X		
		0	1	2
Y	1	0.41	0.05	0
	2	0.19	0.06	0.02
	3	0.10	0.05	0.02
	4	0.02	0.07	0.01

You are asked to

(a) find the distribution of marginal relative frequencies of the number of cinema attendances;

(b) find the distribution of conditional (relative) frequencies of the number of cinema attendances for the retired people who have not gone to the theater;

(c) compute the means and the standard deviations of the variables X and Y.

(d) compute the covariance between the variables X and Y.

12. The distribution of the number of accidents caused by 705 bus drivers in the last 4 years has been the following:

Number of accidents	n
0	114
1	157
2	158
3	115
4	78
5	44
6	21
7	7
8	6
9	1
10	3
11	1

(a) What type of variable are we dealing with? (discrete or continuous, qualitative or quantitative).

(b) Complete the frequency table by adding the relative frequencies f, the absolute cumulative frequencies N, and the relative cumulative frequencies F.

(c) Draw the bar diagram (or chart) of absolute frequencies.

(d) Compute the mode, the mean and the median.

(e) Find the 50%, 75% and 40% percentiles; $x_{50\%}$, $x_{75\%}$ and $x_{40\%}$.

13. An exporting firm sells its product to three countries (China, Japan and Korea). The variable X represents the number of commercial representatives in each country and the variable Y represents the revenues from each country (measured in units of one hundred thousand euros). The following table summarizes the information:

	X	Y
China	3	7
Japan	4	9
Korea	5	7

(a) Compute the variances of X and Y, $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$.

(b) Compute the covariance between X and Y, $\mathrm{Cov}(X, Y)$.

(c) Find the conditional (relative) frequencies of X given $Y = 7$, $f_{X|Y}(x\,|7)$ for $x = 3, 4, 5$, and the conditional (relative) frequencies of X given $Y = 9$, $f_{X|Y}(x\,|9)$ for $x = 3, 4, 5$.

14. The following observations correspond to the market value of two given shares of stock during 10 days:

    Share of stock A: 1.25, 1.34, 1.02, 1.01, 0.98, 1.12, 1.40, 1.23 , 1.10, 1.02
    Share of stock B: 1.40, 0.98, 1.32, 1.20, 0.97, 1.24, 1.10, 0.89, 1.36, 1.01

    (a) What is the mean value of each share of stock during the observed period?

    (b) What is the median of the values of each share of stock during the observed period?

    (c) From the point of view of a potential investor, and according to these observations, which of the two shares of stock is less risky (that is, which one has smaller variance)?

15. The variable X represents the number of years that a former student of Economics and Business of the UAB remained unemployed after completing her degree. The variable Y represents the number of questions of the final exam of Statistics that the same former student answered correctly. Consider the following table of relative frequencies of X and Y from the population of former students:

		X = 0	X = 1
Y	0	0,01	0,28
	1	0,07	0,17
	2	0,13	0,08
	3	0,10	0,01
	4	0,15	0,00

    You are asked to:

    (a) Find the distribution of marginal relative frequencies of the number of correct answers in the exam.

    (b) Find the distribution of conditional frequencies of the number of correct answers for the former students who remained unemployed during one year.

    (c) Compute the mean and the standard deviation of the variables X and Y.

    (d) Compute the covariance and the correlation coefficient between X and Y. What conclusion do you reach about the relationship between the number of correct answers and the number of years unemployed?

1. Probability

1.1. Random experiments and sample spaces

- **Random experiment:** experiment, trial, or observation that can be repeated under the same conditions and whose outcomes are uncertain but observable.
- **Sample space**: The set Ω of possible outcomes of a random experiment.
- **Sample point (or "state of the nature")**: An outcome of a random experiment, $\omega \in \Omega$.
- **Example 1:** Random experiment: Roll a dice. $\Omega_1 = \{1, 2, 3, 4, 5, 6\}$.
- **Example 2:** Random experiment: Flip a coin. $\Omega_2 = \{H, T\}$.
- **Example 3:** Random experiment: Roll a dice and flip a coin.

$$\Omega_3 = \Omega_1 \times \Omega_2 = \{(1, H), (2, H), (3, H), (4, H), (5, H), (6, H),$$
$$(1, T), (2, T), (3, T), (4, T), (5, T), (6, T)\}.$$

- **Example 4:** Random experiment: You pick an individual from a large population and observe his/her wealth measured in cents.

$$\Omega_4 = \left\{ \begin{array}{l} \text{the set of all integers comprised between} \\ \text{the minimum and the maximum wealth} \end{array} \right\}.$$

- A set Ω is discrete or countable if there exists an injective (or one-to-one) function from Ω to the set of natural numbers $\mathbb{N}$.
- A set Ω is continuous if it is not discrete.
- Sample spaces can be either **discrete** (or countable) or **continuous**.
- Discrete sample spaces can be either **finite** or (countable) **infinite**.

1.2. Composition of experiments and combinatorics

- **Fundamental theorem of combinatorics.** Consider an experiment (or operation) that consists of K sub-experiments (or steps) and each step j can be made in n_j different ways, $j = 1, 2, ..., K$, then the whole operation can be made in $\prod_{j=1}^{K} n_j = n_1 \cdot n_2 \cdot ... \cdot n_K$ different ways.

- Sometimes the experiment consists of selecting permutations (or orders or arrangements) of n objects.

- The number of **permutations** P_n of n different objects is given by

$$P_n = n! \equiv n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 3 \cdot 2 \cdot 1,$$

where n is a natural number $(n = 0, 1, 2, ...)$ and the symbol "!" denotes the factorial.

- **Variations** (or permutations) of r objects taken from a set of n different objects ($n \geq r$): V_n^r.

- Here only r objects are taken from the set of n different objects (with $n \geq r$) and, thus, two variations may differ because their objects are different or because they are arranged in a different order.

- The number of variations of r objects taken from a set of n different objects is given by

$$V_n^r = \underbrace{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+1)}_{r \text{ elements}} = \frac{n!}{(n-r)!},$$

  where n and r are natural numbers with $n \geq r$.

- **Example:** A club has 24 members. Its executive committee is formed by 4 members of the club: the president, the vice-president, the treasurer and the secretary. Then, this executive committee can be formed in $24 \cdot 23 \cdot 22 \cdot 21 = 255\,024$ different ways.

- **Permutations with repeated objects.**

  The number of permutations of n objects with n_1 of type 1, n_2 of type 2,..., n_K of type K, with $\sum_{j=1}^{K} n_j = n$, is given by

$$\frac{n!}{n_1!n_2!...n_K!} = \frac{n!}{\prod_{j=1}^{K} n_j!}.$$

- **Example:** With the letters appearing in the word 'Tallahassee' we can form

$$\frac{11!}{1!3!2!1!2!2!} = \frac{39\,916\,800}{48} = 831\,600$$

  different words.

- Number of **combinations** of r objects taken from a set of n different objects ($n \geq r$), C_n^r:
  - number of ways we can select r objects from a set of n different objects (we do not care here about the order but only about the objects we pick); or
  - number of subsets of r elements from a set of n elements.
- The number of combinations is given by the following combinatorial number:

$$C_n^r = \binom{n}{r} \equiv \frac{n!}{r!(n-r)!} = \frac{V_n^r}{r!},$$

  where n and r are natural numbers with $n \geq r$.
- The combinatorial number $\binom{n}{r}$ also gives us the number of permutations of n objects with r of type 1 and $(n-r)$ of type 2.

- Number of ways of partitioning a set with n elements into K subsets of $n_1, n_2, ..., n_K$ elements, with $\sum_{j=1}^{K} n_j = n$:

$$\binom{n}{n_1, n_2, ..., n_K} \equiv \frac{n!}{n_1! n_2! ... n_K!},$$

which coincides with the formula of the number of permutations with repeated objects.

- **Properties of combinatorial numbers:**

1.

$$\binom{n}{r} = \binom{n}{n-r}.$$

Proof:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n!}{(n-r)!r!} = \binom{n}{n-r}.$$

2.

$$0! = 1.$$

Proof:

$$\binom{n}{n} = 1 = \frac{n!}{n!0!} = \frac{1}{0!} \Longrightarrow 0! = 1.$$

- The combinatorial number $\binom{n}{r}$ is also a coefficient in the polynomial expansion of a binomial power (Newton's binomial theorem):

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} y^r.$$

- Moreover,

$$\left(\sum_{j=1}^{K} x_j \right)^n =$$

$$\sum_{r_1=0}^{n} \sum_{r_2=0}^{n} \dots \sum_{r_K=0}^{n} \binom{n}{r_1, r_2, \dots, r_K} \left(x_1^{r_1} \cdot x_2^{r_2} \cdot \dots \cdot x_K^{r_K} \right), \text{ with } \sum_{j=1}^{K} r_j = n.$$

- **Example.** Let $(x+y+z)^6$. Then, the value of the coefficient c in the term cx^3yz^2 of the expansion is

$$c = \binom{6}{3,1,2} = \frac{6!}{3!1!2!} = 60.$$

1.3. sigma-algebras, events, and measurable spaces.

- B^c denotes the complement of B relative to Ω, that is, $B^c = \{\omega \in \Omega \,|\omega \notin B\}$
- $A \cup B = \{\omega \in \Omega \,|\omega \in A, \text{ or } \omega \in B, \text{ or both}\}$.
- $A \cap B = \{\omega \in \Omega \,|\omega \in A \text{ and } \omega \in B\}$.
- $A \backslash B = \{\omega \in \Omega \,|\omega \in A \text{ and } \omega \notin B\}$. Therefore, $B^c = \Omega \backslash B$.

- **Definition.** A σ-algebra (or σ-field) on Ω is a collection $\mathcal{F}$ of subsets of Ω, such that

  1. $\emptyset \in \mathcal{F}$, $\Omega \in \mathcal{F}$.

  2. If $B \in \mathcal{F}$, then $B^c \in \mathcal{F}$.

  3. If $\{B_1, B_2, ...\}$ is a countable (finite or infinite) collection of elements of $\mathcal{F}$, then $\bigcup_i B_i \in \mathcal{F}$.

- *Note:* If $\{B_1, B_2, ...\}$ is a countable (finite or infinite) collection of elements of the σ-algebra $\mathcal{F}$, then $\bigcap_i B_i \in \mathcal{F}$.

- **Proof.** Note that, if $\{B_1, B_2, ...\}$ is a countable collection of elements of the σ-algebra $\mathcal{F}$, then $\{B_1^c, B_2^c, ...\}$ is also a countable collection of elements of $\mathcal{F}$. Therefore,

$$\bigcup_i B_i^c \in \mathcal{F} \Longrightarrow \left(\bigcup_i B_i^c \right)^c \in \mathcal{F} \Longrightarrow \bigcap_i B_i \in \mathcal{F}. \quad Q.E.D. \qquad (\star)$$

- Moreover, part 3 of the definition of σ-algebra could be replaced by the following:

  3'. If $\{B_1, B_2, ...\}$ is a countable (finite or infinite) collection of elements of $\mathcal{F}$, then $\bigcap_i B_i \in \mathcal{F}$.

  This replacement is justified by just interchanging the set operations $\bigcup_i$ and $\bigcap_i$ in $(\star)$.

- In the definition of **algebra** on Ω we replace the word "countable" in (3) by "finite".
- Therefore, a σ-algebra is an algebra, but the converse is not true.
- The elements of a σ-algebra are called measurable sets, $B \in \mathcal{F}$.
- If Ω is a sample space and we define a σ-algebra $\mathcal{F}$ on it, the elements of $\mathcal{F}$ are called "events".
- B is an event if the question "Does $\omega \in B$?" has a definite "yes" or "not" answer after the experiment has been performed and delivered the outcome ω, for all outcomes $\omega \in \Omega$.

- The smallest σ-algebra on Ω is $\{\emptyset, \Omega\}$, whereas the largest is the collection of all subsets of Ω (the power set, denoted by 2^{Ω}).
- Let the number of elements (or cardinality) of Ω be $\#\Omega = n$. The number of elements of the power set 2^{Ω} is

$$\#2^{\Omega} = \binom{n}{0} + \binom{n}{1} + ... + \binom{n}{n} = (1+1)^n = 2^n.$$

- **Example:** We roll a dice and flip a coin. This random experiment has 12 sample points (or outcomes):

$$\Omega = \{(1,H),(2,H),(3,H),(4,H),(5,H),(6,H),$$
$$(1,T),(2,T),(3,T),(4,T),(5,T),(6,T)\}.$$

- Examples of events in the power set 2^{Ω}:

  $A = \{(2,H),(4,H),(6,H)\} \leftarrow$Even number of points and Head

  $B = \{(1,H),(2,H),(3,H),(4,H),(5,H),(6,H)\} \leftarrow$Head

  $C = \{(2,H),(4,H)\} \leftarrow$Number of points even no larger than 4 and Head

- The power set 2^{Ω} contains $2^{12} = 4096$ different events.

- **Definition.** Let $S = \{A_1, A_2, ...\}$ be a collection of arbitrary subsets of Ω. The σ-algebra $\sigma(S)$ generated by S is the smallest σ-algebra containing S.
- **Note:** $\{\emptyset, \Omega\} \subset \sigma(S) \subset 2^{\Omega}$.
- **Example.** Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $S = \{\{1\}, \{1, 3, 4\}\}$. Then,

$$\sigma(S) = \{\Omega, \emptyset, \{1\}, \{1, 3, 4\}, \{2, 3, 4, 5, 6\}, \{2, 5, 6\}, \{1, 2, 5, 6\}, \{3, 4\}\}.$$

- Usually, we consider σ-algebras generated by partitions of Ω.
- **Definition.** A partition of the set Ω is a collection of subsets of Ω, $\{B_1, B_2, ...\}$ such that

  (a) $B_i \neq \emptyset$, for all i,

  (b) $B_i \bigcap B_j = \emptyset$, for all pairs (i, j) with $i \neq j$,

  (c) $\bigcup_i B_i = \Omega$.

- **Example.** Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $S = \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$ is a partition of Ω. Then,

  $$\sigma(S) = \{\Omega, \emptyset, \{1, 2\}, \{3, 4\}, \{5, 6\}, \{1, 2, 3, 4\}, \{1, 2, 5, 6\}, \{3, 4, 5, 6\}\}.$$

- If $\mathcal{F}_1 \subset \mathcal{F}_2$, then we say that the σ-algebra $\mathcal{F}_2$ is finer than $\mathcal{F}_1$ (or that $\mathcal{F}_1$ is coarser than $\mathcal{F}_2$).

- $\{\emptyset, \Omega\}$ is the coarsest σ-algebra on Ω, whereas the power set 2^Ω is the finest σ-algebra on Ω.

- **Definition.** The pair $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-algebra on Ω is called a measurable space.

1.4. Measure, probability, and probability spaces.

- **Definition.** A measure μ on the measurable space $(\Omega, \mathcal{F})$ is a set function on $\mathcal{F}$ taking values on the set of non-negative extended real numbers,

$$\mu : \mathcal{F} \longrightarrow [0, \infty] \equiv \overline{\mathbb{R}}_+,$$

such that

1. If $\{B_1, B_2, ...\}$ is a countable (finite or infinite) collection of mutually disjoint measurable sets (elements of $\mathcal{F}$), then

$$\mu\left(\bigcup_i B_i\right) = \sum_i \mu\left(B_i\right) \qquad \text{(countable additivity)}.$$

2. $\mu(\emptyset) = 0$.

- Property 2 is redundant if at least one of the measurable sets has finite measure.

- A signed measure on the measurable space $(\Omega, \mathcal{F})$ is a countably additive set function on $\mathcal{F}$ taking values on the set of extended real numbers $\overline{\mathbb{R}} \equiv [-\infty, \infty]$ such that $\mu(\emptyset) = 0$.
- Therefore, a measure is a signed measure that takes values on $\overline{\mathbb{R}}_+$.

- **Definition.** The triple $(\Omega, \mathcal{F}, \mu)$, where $(\Omega, \mathcal{F})$ is a measurable space and μ is a measure on it, is called a measure space.
- **Definition.** A probability measure (or "probability") P on $(\Omega, \mathcal{F})$, where Ω is a sample space and $\mathcal{F}$ is a σ-algebra of events, is a measure such that $P(\Omega) = 1$.
- **Definition.** The triple $(\Omega, \mathcal{F}, P)$, where P is a probability on $(\Omega, \mathcal{F})$ is called a probability space.
- *Note:* We only assign probability to events.

- Usually, if the sample space is discrete and the σ-algebra of events is the power set 2^{Ω}, we assign probabilities to all the sample points and, using the countable additivity property, we construct a probability on the measurable space $\left(\Omega, 2^{\Omega}\right)$.
- **Proposition.** Consider the probability space $\left(\Omega, 2^{\Omega}, P\right)$, where $\Omega = \{\omega_1, \omega_2, ...\}$ is a discrete sample space, then

$$P(B) = \sum_{\omega_i \in B} P\{\omega_i\}.$$

- **Proposition (Laplace).** Consider the probability space $\left(\Omega, 2^{\Omega}, P\right)$. If $\#\Omega = N$, $\#B = n$, and all the sample points are equally likely, then

$$P(B) = \frac{n}{N} \quad \text{(Laplace formula)}.$$

- **Proof:** Observe that $\Omega = \bigcup_{i=1}^{N} \{\omega_i\}$. Therefore,

$P(\Omega) = \sum_{i=1}^{N} P\{\omega_i\}$. Since all the sample points are equally likely, $P\{\omega_i\} = p$ for all i. Then,

$$1 = P(\Omega) = \sum_{i=1}^{N} p = Np,$$

which implies that $p = \frac{1}{N}$. Moreover,

$$P(B) = \sum_{\omega_i \in B} \underbrace{P\{\omega_i\}}_{=p} = \sum_{\omega_i \in B} \frac{1}{N} = \frac{n}{N}$$

since the event B has n elements. $Q.E.D.$

Pierre Simon de Laplace (1749 - 1827)

- **Example:** We roll a dice. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^{\Omega}$, $A = \{1, 2, 3\}$, $B = \{3, 4\}$, $C = \{4, 5\}$, and $D = \{4\}$.
- Then, $P(A) = 3/6 = 1/2$, $P(B) = 2/6 = 1/3$, $P(C) = 2/6 = 1/3$ and $P(D) = 1/6$.
- $A \cup C = \{1, 2, 3, 4, 5\}$ so that $P(A \cup C) = 5/6$.
- Note also that $A \cap C = \emptyset$ so that

$$P(A \cup C) = P(A) + P(C) = \frac{3}{6} + \frac{2}{6} = \frac{5}{6}.$$

- **Note on the relationship between probability and frequency:** If we randomly pick an object (or individual) from a population having a variable X distributed according to the relative frequency f_X, then

$$P\{X \in B\} = \sum_{x \in B} f_X(x),$$

where $\{X \in B\}$ is the event where the variable X of the object we pick takes a value belonging to the set B.

- **Notation:**

$$P(A \cap B) = P(A, B).$$

1.5. Properties of probability measures.

1. $P(A^c) = 1 - P(A)$.
2. If $A \subset B$, then $P(A) \leq P(B)$.
3. $0 \leq P(A) \leq 1$.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
5.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C). \end{aligned}$$

- Properties 2, 4, and 5 also hold for a general measure μ and Property 1 then becomes $\mu(A^c) = \mu(\Omega) - \mu(A)$, while Property 3 becomes $0 \leq \mu(A) \leq \mu(\Omega)$.

- All the previous rules of probability have to be satisfied both when the probability is objective (i.e., based on scientific predictions) and when the probability is subjective (i.e., based on beliefs).
- **Example:** We roll a dice. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^{\Omega}$, $A = \{1, 2, 3\}$ and $B = \{3, 4\}$.
- Then, $P(A) = 3/6 = 1/2$ and $P(B) = 2/6 = 1/3$.
- $A \cup B = \{1, 2, 3, 4\}$ and $A \cap B = \{3\}$.
- Then, $P(A \cup B) = 4/6 = 2/3$ and $P(A \cap B) = 1/6$.
- Note that

$$P(A \cup B) = \frac{4}{6} = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6}.$$

1.6. Conditional probability.

- **Definition.** If A and B are events in the sample space Ω and $P(A) \neq (>)0$, the conditional probability of B given A is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

- Therefore, if A and B are events in the sample space Ω and $P(A) \neq (>)0$, then

$$P(A \cap B) = P(A) \cdot P(B|A).$$

- **Proposition.** If $P\left(A_1 \cap A_2 \cap ... \cap A_{n-1}\right) > 0$, then

$$P\left(A_1 \cap A_2 \cap ... \cap A_n\right) =$$

$$P\left(A_1\right) \cdot P\left(A_2 \left| A_1\right.\right) \cdot P\left(A_3 \left| A_1 \cap A_2\right.\right) \cdot ... \cdot P\left(A_n \left| A_1 \cap A_2 \cap ... \cap A_{n-1}\right.\right).$$

- **Proof.**

$$P\left(A_1\right) \cdot P\left(A_2 \left| A_1\right.\right) \cdot P\left(A_3 \left| A_1 \cap A_2\right.\right) \cdot ... \cdot P\left(A_n \left| A_1 \cap A_2 \cap ... \cap A_{n-1}\right.\right)$$

$$= P\left(A_1\right) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdot ... \cdot \frac{P(A_1 \cap A_2 \cap ... \cap A_n)}{P(A_1 \cap A_2 \cap ... \cap A_{n-1})}$$

$$= P(A_1 \cap A_2 \cap ... \cap A_n). \qquad Q.E.D.$$

Note: If we randomly pick an object (or individual) from a population having the variables X and Y distributed according to the relative joint frequency $f_{X,Y}$ and, thus, with the relative marginal frequencies f_X and f_Y, and the conditional frequencies $f_{X|Y}$ and $f_{Y|X}$, then

$$P\{X \in B, Y \in C\} = \sum_{x \in B} \sum_{y \in C} f_{X,Y}(x, y) \Longrightarrow$$

$$P\{X = x, Y = y\} = f_{X,Y}(x, y),$$

$$P\{X \in B\} = \sum_{x \in B} f_X(x) \implies P\{X = x\} = f_X(x),$$

$$P\{Y \in C\} = \sum_{y \in C} f_Y(y) \implies P\{Y = y\} = f_Y(y),$$

$$P\{X \in B \,|\, Y \in C\} = \frac{P\{X \in B, Y \in C\}}{P\{Y \in C\}} = \frac{\sum_{x \in B} \sum_{y \in C} f_{X,Y}(x,y)}{\sum_{y \in C} f_Y(y)}$$

and

$$P\{Y \in C \,|\, X \in B\} = \frac{P\{X \in B, Y \in C\}}{P\{X \in B\}} = \frac{\sum_{x \in B} \sum_{y \in C} f_{X,Y}(x,y)}{\sum_{x \in B} f_X(x)}$$

$\Longrightarrow$

$$P\{X \in B \,|\, Y{=}y\} = \sum_{x \in B} f_{X|Y}(x \,|\, y), P\{Y \in C \,|\, X{=}x\} = \sum_{y \in C} f_{Y|X}(y \,|\, x)$$

and

$$P\{X = x \,|\, Y = y\} = f_{X|Y}(x \,|\, y), \quad P\{Y = y \,|\, X = x\} = f_{Y|X}(y \,|\, x),$$

where $\{X \in B\}$ is the event where the variable X of the object we pick takes a value belonging to the set B and $\{Y \in C\}$ is the event where the variable Y of the object we pick takes a value belonging to the set C.

- **Example.** We extract two cards with "no replacement" from a deck of poker cards. The probability that the two cards will be aces is

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1) = \frac{4}{52} \cdot \frac{3}{51} = 0.0045.$$

- Assume now that the extractions are made "with replacement", that is, the cards are introduced back in the deck after each extraction. The probability that the two cards so extracted will be aces is

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1) = \frac{4}{52} \cdot \frac{4}{52} = 0.0059.$$

- **Example:** We extract three cards with "no replacement" from a deck of poker cards. The probability that all the three cards will be aces is

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2)$$
$$= \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = 0.000181.$$

1.7. Independent events.

- **Definition.** The events A and B are independent if

$$P(A \cap B) = P(A) \cdot P(B).$$

- **Proposition.** Let $P(A) > 0$. The events A and B are independent if and only if

$$P(B) = P(B|A).$$

- **Proof.** Since $P(A \cap B) = P(A) \cdot P(B|A)$ if $P(A) > 0$ and the events A and B are independent (i.e., $P(A \cap B) = P(A) \cdot P(B)$), we immediately obtain the desired result. $Q.E.D.$

Example. We toss 3 coins (or the same coin three times). This random experiment has 8 equally likely outcomes,

$$\Omega = \{(HHH), (HHT), (HTH), (THH), (TTH), (THT), (HTT), (TTT)\}$$

so that the probability of each outcome is $1/8$.

- Consider the following events when $\mathcal{F} = 2^{\Omega}$:

  $A = \{(HHH), (HHT)\}, \leftarrow$ {Head in the first two coins}

  $B = \{(HHT), (HTT), (THT), (TTT)\}, \leftarrow$ {Tail in the third coin}

  $C = \{(HTT), (THT), (TTH)\} \leftarrow$ {Exactly two tails}

- $A \cap B = \{(HHT)\}$ and $B \cap C = \{(HTT), (THT)\}$.

- $P(A) = 2/8 = 1/4$, $P(B) = 4/8 = 1/2$, $P(C) = 3/8$, $P(A \cap B) = 1/8$, and $P(B \cap C) = 2/8 = 1/4$.
- The events A and B are independent since

$$P(A \cap B) = \frac{1}{8} = P(A) \cdot P(B) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}.$$

- The events B and C are not independent since

$$P(B \cap C) = \frac{1}{4} \neq P(B) \cdot P(C) = \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16}.$$

- **Proposition.** If A and B are independent events, then

  (a) A^c and B are independent.

  (b) A^c and B^c are independent.

- **Proof. (a)** Observe that

$$B = \Omega \cap B = (A \cup A^c) \cap B = (A \cap B) \cup (A^c \cap B).$$

Since $(A \cap B)$ and $(A^c \cap B)$ are disjoint events,

$$P(B) = P\left[(A \cap B) \cup (A^c \cap B)\right] = P(A \cap B) + P(A^c \cap B),$$

which is equivalent to

$$P(A^c \cap B) = P(B) - P(A \cap B).$$

Therefore, since $P(A^c) = 1 - P(A)$ and the events A and B are independent (i.e., $P(A \cap B) = P(A) \cdot P(B)$), we have

$$\begin{aligned} P(A^c \cap B) &= P(B) - P(A \cap B) = P(B) - P(A) \cdot P(B) \\ &= P(B)\left[1 - P(A)\right] = P(B) \cdot P(A^c), \end{aligned}$$

which proves the independency between A^c and B.

- **(b)** Obvious from (a). $Q.E.D.$

- **Definition.** The events in the collection $S = \{A_1, A_2, ...\}$ are independent if the probability of every finite intersection of events in S equals the product of their respective probabilities.

1.8. Theorem of total probability.

- **Theorem of Total Probability.** Let $\{B_1, B_2, ...\}$ be a countable (finite or infinite) collection of events that constitutes a partition of the sample space Ω and assume that $P(B_i) > 0$ for all i. Then,

$$P(A) = \sum_i P(B_i) \cdot P\left(A \,|B_i\right), \quad \text{for every event } A.$$

- **Proof.** Since

$$A = A \cap \Omega = A \cap \left(\bigcup_i B_i \right) = \bigcup_i (A \cap B_i)$$

and the events in the countable collection $\{A \cap B_1, A \cap B_2, ...\}$ are disjoint, we get

$$P\left(A\right) = P\left(\bigcup_i (A \cap B_i) \right) = \sum_i P(A \cap B_i) = \sum_i P(B_i) \cdot P\left(A \,|B_i\right),$$

where the last equality follows since $P(A \cap B_i) = P(B_i) \cdot P\left(A \,|B_i\right)$ if $P(B_i) > 0$. *Q.E.D.*

- **Example.** We extract two cards with "no replacement" from a deck of poker cards. The probability that the second card is an ace is

$$P(A_2) = P(A_1) \cdot P(A_2 | A_1) + P(A_1^c) \cdot P(A_2 | A_1^c)$$

$$= \left(\frac{4}{52} \cdot \frac{3}{51}\right) + \left(\frac{48}{52} \cdot \frac{4}{51}\right) = \frac{4}{52}.$$

- Assume now that the extractions are made "with replacement", that is, the cards are introduced back in the deck after each extraction. The probability that the second card is an ace is, obviously,

$$P(A_2) = P(A_1) \cdot P(A_2 | A_1) + P(A_1^c) \cdot P(A_2 | A_1^c)$$

$$= \left(\frac{4}{52} \cdot \frac{4}{52}\right) + \left(\frac{48}{52} \cdot \frac{4}{52}\right) = \frac{4}{52}.$$

- Note that, if the extractions are made with "no replacement", then the events A_1 and A_2 are not independent since

$$P(A_1 \cap A_2) = \frac{4}{52} \cdot \frac{3}{51} \neq P(A_1) \cdot P(A_2) = \frac{4}{52} \cdot \frac{4}{52}.$$

- If the extractions are made "with replacement", then the events A_1 and A_2 are independent since

$$P(A_1 \cap A_2) = \frac{4}{52} \cdot \frac{4}{52} = P(A_1) \cdot P(A_2) = \frac{4}{52} \cdot \frac{4}{52}.$$

- The theorem of total probability can be modified to apply for intersections of events as follows:
- **Proposition.** Let $\{B_1, B_2, ...\}$ be a countable (finite or infinite) collection of events that constitutes a partition of the event B and assume that $P(B_i) > 0$ for all i. Then,

$$P(A \cap B) = \sum_i P(B_i) \cdot P\left(A \middle| B_i\right), \quad \text{for every event } A.$$

- **Proof.** Since $\bigcup_i B_i = B$, we have that $\bigcup_i (A \cap B_i) = A \cap B$, where the events in the countable collection $\{A \cap B_1, A \cap B_2, ...\}$ are disjoint. Thus,

$$P(A \cap B) = P\left(\bigcup_i (A \cap B_i)\right) = \sum_i P\left(A \cap B_i\right)$$
$$= \sum_i P(B_i) \cdot P\left(A \middle| B_i\right). \qquad Q.E.D.$$

- If we make $B = \Omega$ in the previous proposition, then we recover the original theorem of total probability since $P(A \cap \Omega) = P(A)$.

1.9. Bayes' theorem.

Thomas Bayes (1702 - 1761)

- **Bayes' theorem.** Let $\{B_1, B_2, ...\}$ be a countable (finite or infinite) collection of events that constitutes a partition of the sample space Ω and assume that $P(B_i) > 0$ for all i. Then, for every event A such that $P(A) > 0$,

$$P(B_j|A) = \frac{P(B_j) \cdot P(A|B_j)}{\sum_i P(B_i) \cdot P(A|B_i)}, \text{ for all } j = 1, 2, ...$$

- **Proof.** Obvious since

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} \text{ when } P(A) > 0,$$

and $P(A \cap B_j) = P(B_j) \cdot P(A|B_j)$ if $P(B_j) > 0$, while $P(A) = \sum_i P(B_i) \cdot P(A|B_i)$ if $P(B_i) > 0$ for all i, as follows from the theorem of total probability. *Q.E.D.*

- **Note:** $P(B_j)$ is the prior probability of the event B_j, $P(A|B_j)$ is the conditional probability of the event A given B_j (also called the "likelihood" of A when B_j occurs), and $P(B_j|A)$ is the posterior probability of the event B_j given A, for $j = 1, 2, ...$
- **Example:** We want to select a candidate for a job and we observe the result of a test undertaken by the candidate.
- $\{g, b\} \equiv \{\text{candidate is good for a job, candidate is bad for a job}\}$, constitutes a partition of Ω. Thus, $b = g^c$.
- $\{p, f\} \equiv \{\text{candidate has passed a test, candidate has failed a test}\}$.
- *Prior probabilities:* $P(g) = 0.25$ so that $P(b) = 0.75$.
- *Likelihoods:* $P(p|g) = 0.99$ so that $P(f|g) = 1 - P(p|g) = 0.01$,

  and

  $P(f|b) = 0.83$ so that $P(p|b) = 1 - P(f|b) = 0.17$.

- Then,

$$P(g\,|p\,) = \frac{\overbrace{P(g)\cdot P(p\,|g\,)}^{P(p\cap g)}}{\underbrace{P(g)\cdot P(p\,|g\,)+P(b)\cdot P(p\,|b\,)}_{P(p)}}$$

$$= \frac{\overbrace{0.25\cdot 0.99}^{0.2475}}{\underbrace{(0.25\cdot 0.99)+(0.75\cdot 0.17)}_{=0.375}} = 0.66.$$

- Note that $P(b\,|p\,) = 1 - P(g\,|p\,) = 0.34$.
- Moreover, $P(p) = 0.375$ so that $P(f\,) = 0.625$.
- Finally, $P(p \cap g) = P\{\text{pass and good}\} = 0.2475$.

Exercises. Probability and Statistics. IDEA.
1. Probability

1. Prove that, for any positive integer (or natural number) n larger or equal than 2 and $r = 1, 2, ..., n-1$,

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}.$$

Hint: Use the fact that

$$(1+y)^n = (1+y)(1+y)^{n-1} = (1+y)^{n-1} + y(1+y)^{n-1}.$$

2. Prove that

$$\sum_{r=0}^{k} \binom{m}{r}\binom{n}{k-r} = \binom{m+n}{k}.$$

Hint: Use the fact that

$$(1+y)^{m+n} = (1+y)^m (1+y)^n.$$

3. Suppose that we are concerned with the completion of a highway construction job, which may be delayed because of a strike. Suppose, furthermore, that the probabilities are 0.60 that there will be a strike, 0.85 that the job will be completed on time if there is no strike, and 0.35 that the job will be completed on time if there is a strike. What is the probability that the job will be completed on time?

4. A strictly positive integer I is selected, with $P\{I = n\} = \left(\frac{1}{2}\right)^n$, $n = 1, 2, ...$ If I takes the value n, a coin with probability e^{-n} of heads is tossed once.

   (a) Find the probability that the resulting toss is a head.

   (b) Find the conditional probability of $\{I = 5\}$ given that we know that the resulting toss has been a tail.

5. The members of a consulting firm rent cars from three rental agencies: 60% from agency 1, 30% from agency 2, and 10% from agency 3. If 9% of the cars from agency 1 need a tune-up, 20% of the cars from agency 2 need a tune-up, and 6% of the cars from agency 3 need a tune-up,

   (a) what is the probability that a rental car delivered to the firm will need a tune-up?

   (b) if a rental car delivered to the firm needs a tune-up, what is the probability that it came from the rental agency 2?

6. Let us define a generalized combinatorial number as

$$\binom{r}{x} = \frac{\prod_{i=0}^{x-1}(r-i)}{x!},$$

for any real number r and any natural number x. Moreover, by definition, we have

$$\binom{r}{0} = 1,$$

for any real number r.

Prove that,

(a) If n is a natural number such that $n \geq x$, then we obtain the traditional formula for a combinatorial number,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

(b) If n is a natural number such that $n < x$,

$$\binom{n}{x} = 0.$$

(c) For any real number r and any natural number x,

$$\binom{-r}{x} = (-1)^x \binom{r+x-1}{x}.$$

(d) If n is a strictly positive natural number,

$$\binom{-n}{x} = (-1)^x \binom{n+x-1}{n-1}.$$

(e) For any natural number x,

$$\binom{-1}{x} = (-1)^x \quad \text{and} \quad \binom{-2}{x} = (-1)^x(1+x).$$

(f) If r and z are two real numbers with $|z| < 1$,

$$(1+z)^r = \sum_{x=0}^{\infty} \binom{r}{x} z^x.$$

(g) For any real number z and any natural number n,

$$(1+z)^n = \sum_{x=0}^{n} \binom{n}{x} z^x.$$

7. In a scientific meeting of 100 people, 60 of them speak English only, 30 speak French only, and the remaining 10 speak both languages. Compute the probability that two randomly selected participants will be able to understand each other.

8. Ten individuals are randomly ordered and all possible orderings are equally likely. Find the probability that two given individuals be contiguous if the 10 individuals are ordered: a) in a row, b) in a circle.

9. We have 11 urns, which are numbered from 2 to 12. The composition of the urns is the following:

number of the urn	2	3	4	5	6	7	8	9	10	11	12
number of white balls	0	1	2	3	4	5	4	3	2	1	0
number of black balls	1	1	1	1	1	1	1	1	1	1	1

   We roll two dices simultaneously. If the sum of points is k, we make an extraction with replacement of a ball from the urn k. If the extracted ball is black, the game is over. If the ball is white, we roll again the dices and repeat the whole process. Compute the probability of making at least 3 successive extractions.

10. There are n balls in a box. Each ball is either black or red and we assume that the $n+1$ different possible compositions of the box are equally likely.

    (a) We randomly pick one ball and it turns out to be red. What is the probability that k out of the n balls in the box were red?

    (b) Assume that $n = 6$ and that we randomly pick two balls with replacement (i.e., we put back in the box the picked ball after each extraction) and it turns out that both balls are red. What is the probability that 5 out of the 6 balls were red in the composition of the box?

    (c) Assume now that $n = 7$ and that we extract two balls with no replacement from the box. What is the probability that both balls will be black? What is the probability that the first extracted ball will be black and the second red?

11. We have two decks of Spanish cards (with 48 cards where 12 of them are figures). We randomly extract a card from the first deck and we insert it in the second deck. Then, we roll a dice. If we get one dot, we randomly extract a card from the first deck; if we get two dots, we randomly extract a card from the second deck; otherwise, we merge the two decks and we randomly extract a card from the resulting single deck. Find the probability of extracting a figure in this second extraction.

12. The Chicago Bulls and the Golden State Warriors are playing each other in the NBA (National Basketball Association) final playoffs. In this set of games, the first team to have won four games is declared to be the (world) champion.

(a) Assume that the Bulls are slightly better than the Warriors, such that

$$P\{\text{Bulls beat the Warriors in the first game}\} = 0.6$$

and that the outcome of a game affects slightly the probability of the outcome on the following game so that

$$P\{\text{Bulls win a game if they have won the previous game}\} = 0.7$$

but

$$P\{\text{Bulls win a game if they have lost the previous game}\} = 0.5.$$

(i) What is the probability that the Warriors will win a game if they have lost the previous game?

(ii) What is the probability that the Bulls will win the playoffs in only four games?

(iii) What is the probability that the Warriors will win the playoffs in only four games?

(iv) What is the probability that the Warriors have won the first game if we know that they have lost the second game?

(b) If the two teams were evenly matched, that is,

$$P\{\text{Bulls beat the Warriors in the first game}\} = 0.5$$

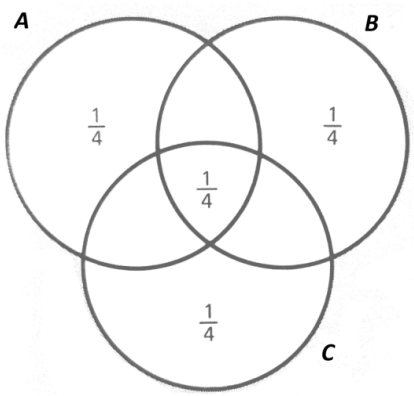
and the probability of winning a game did not depend upon the outcome of the previous game, would it be more likely or less likely that the series would end in just four games than under the conditions specified in (a)?

13. In a group of 20 students of the IDEA program, 12 of them have passed the Micro exam only, 6 have passed the Macro exam only, and the remaining two students have passed both the Micro and the Macro exams. We pick randomly three students from this group.

(a) Compute the probability that all three students have passed the same exam.

(b) If we know that the three students we have picked have passed the same exam, what is the probability that the three have passed the Micro exam.

14. Consider the events A, B, and C on the probability space $(\Omega, \mathcal{F}, P)$. Assume that $P(A) = 1/2$, $P(B) = 1/2$, $P(C) = 1/2$, $P(A \cap B) = 1/4$, $P(A \cap C) = 1/4$, $P(B \cap C) = 1/4$, and $P(A \cap B \cap C) = 1/4$. The following figure summarizes these probabilities:

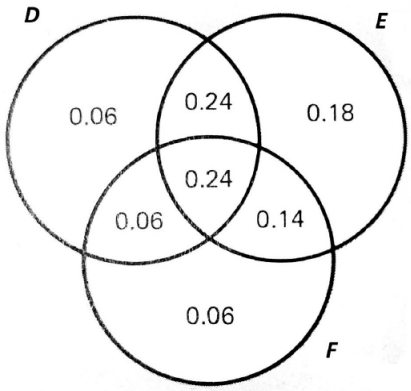


(a) Show that A and B are independent, A and C are independent, and B and C are independent, but A, B, and C are not independent. *Note:* this is an example of three events that are pairwise independent without being independent.

(b) Prove that $P(A \cup B \cup C) = 1$.

(c) Find $P\left((A \cap B) \cap C^c\right)$, where C^c is the complement of the event C. Note that the event $(A \cap B) \cap C^c$ contains all the sample points belonging to both A and B but not belonging to C. That is, $(A \cap B) \cap C^c = (A \cap B) \setminus C$.

Consider now the events D, E, and F on the probability space $(\Omega, \mathcal{F}, P)$. Assume that $P(D) = 0.6$, $P(E) = 0.8$, $P(F) = 0.5$, $P(D \cap E) = 0.48$, $P(D \cap F) = 0.3$, $P(E \cap F) = 0.38$, and $P(D \cap E \cap F) = 0.24$. The following figure summarizes these probabilities:



(d) Are D and E are independent? Are D and F independent? Are E and F independent? Are D, E, and F independent? Show that $P(D \cap E \cap F) = P(D) \cdot P(E) \cdot P(F)$. *Note:* this is an example of three events whose probability of their intersection is equal to the product of their respective probabilities. However, they are not independent since they are not all pairwise independent.

(e) Find $P\left((D \cap E) \cap F^c\right)$, where F^c is the complement of the event F. Note that $(D \cap E) \cap F^c = (D \cap E) \setminus F$.

(f) Find $P\left((D \cup E \cup F)^c\right)$, where $(D \cup E \cup F)^c$ is the complement of the event $D \cup E \cup F$.

15. Prove the following generalization of the theorem of total probability:

    Theorem of total conditional probability. If A and C are events of the measurable space $(\Omega, \mathcal{F})$ and $\{B_1, B_2, ...\}$ is a discrete collection of events that constitutes a partition of the sample space Ω, with $P(B_i \cap C) > 0$ for $i = 1, 2, ...$, then

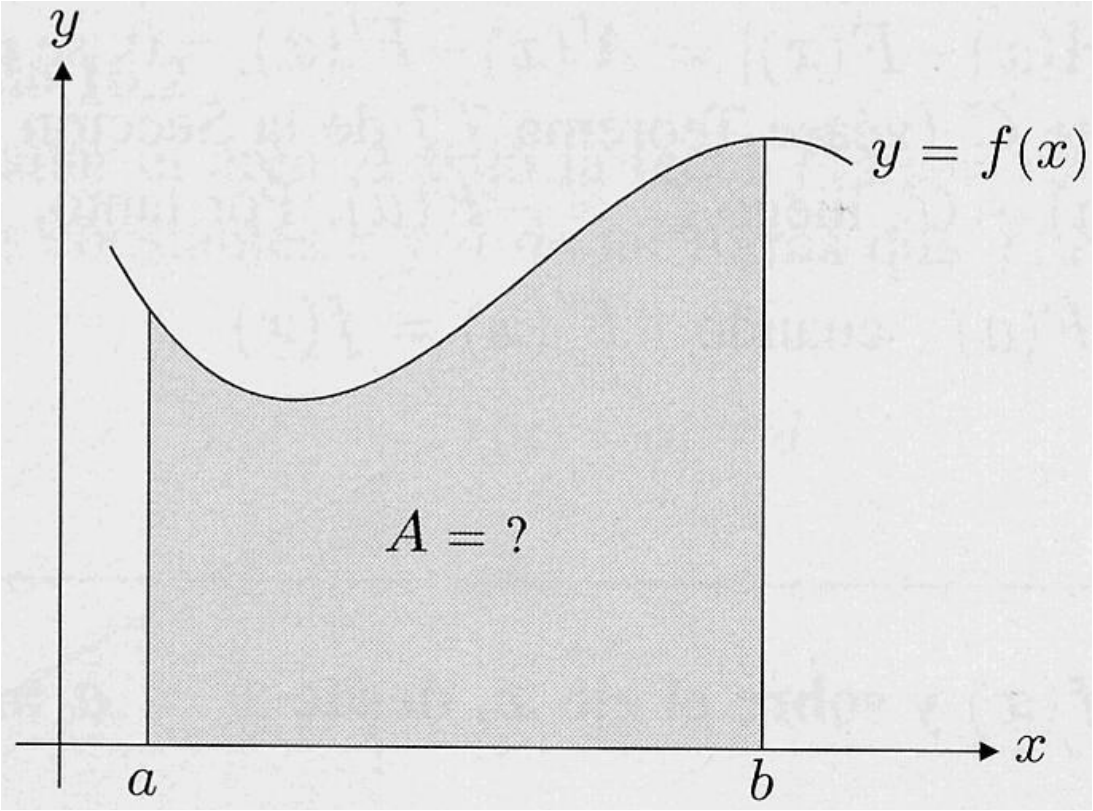
    $$P(A|C) = \sum_i P(B_i|C) \cdot P(A|B_i \cap C) \quad \text{for all } A \in \mathcal{F} \text{ and } C \in \mathcal{F}.$$

16. Consider a group of n individuals and a task that has to be done. Nobody wants to do this task. The "drawing straws" method is used to select the individual who is going to do it. Thus, individuals are randomly ordered and the last individual becomes the group leader. Then, the group leader takes n straws and ensures that one of them is shorter than the others. The leader then grabs all of the straws in her fist, such that all of them appear to be of the same length. The group leader offers the clenched fist to the group. Each member of the group draws sequentially a straw from the fist of the group leader. When an individual draws the shortest straw he must perform the task and the game finishes. Obviously, if nobody has picked the shortest straw, then this straw is the one left in the leader's fist and, then, the leader has to do the task. Does the order in which individuals pick the straw affect their probability of picking the shortest straw? *Hint:* Use the theorem of Exercise 15.

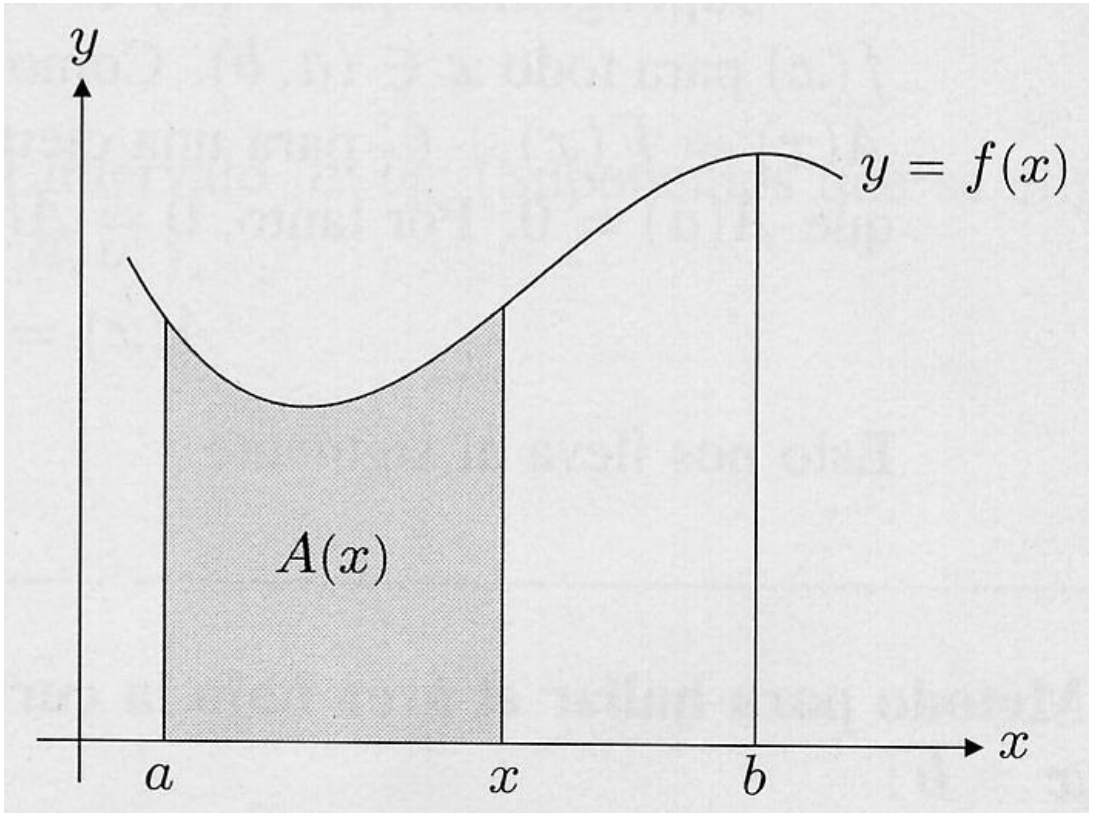
A Primer on Integration

1. The fundamental theorem of calculus

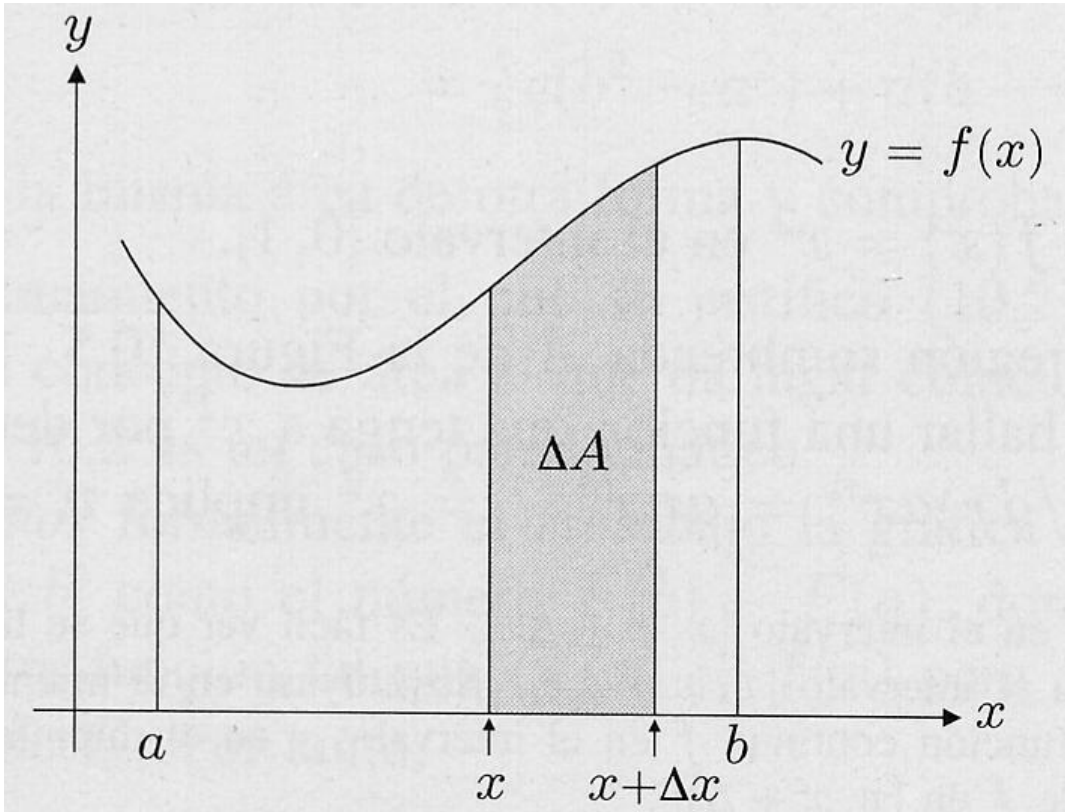
Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a continuous and non-negative function ($f \geq 0$). We want to find the area A between the graph of the function and the horizontal axis on the interval $[a, b]$.



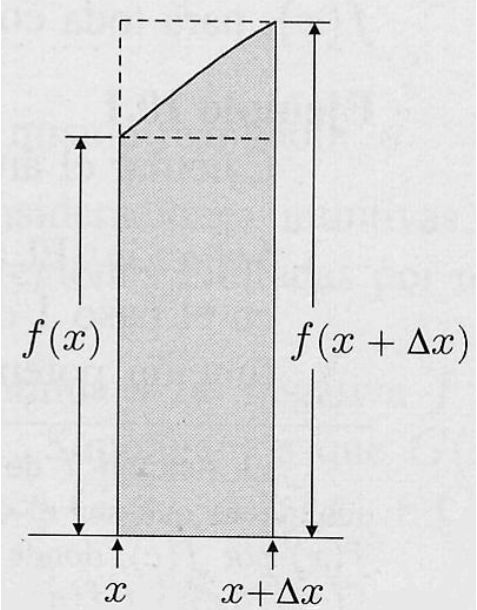
Let us fix the value a and make the area a function of the upper endpoint of the interval, $A(x)$,



Let us increase the value x by a small (infinitessimal) amount Δx. The increase in the area is $\Delta A = A(x + \Delta x) - A(x)$.



- Let us have a closer look to the increase of the area, ΔA,



- We see that

$$f(x)\cdot\Delta x \leq A(x+\Delta x)-A(x) \leq f(x+\Delta x)\cdot\Delta x, \quad \text{if } f \text{ is increasing,}$$

whereas

$$f(x)\cdot\Delta x \geq A(x+\Delta x)-A(x) \geq f(x+\Delta x)\cdot\Delta x, \quad \text{if } f \text{ is decreasing.}$$

- Note that for Δx sufficiently small, f is either increasing or decreasing on the interval $[x, x+\Delta x]$.

- Divide the three terms in the previous expression by Δx

$$f(x) \underset{\geq}{\leq} \frac{A(x+\Delta x) - A(x)}{\Delta x} \underset{\geq}{\leq} f(x+\Delta x)$$

and take the limit when $\Delta x \longrightarrow 0$,

$$\lim_{\Delta x \to 0} f(x) \underset{\geq}{\leq} \underbrace{\lim_{\Delta x \to 0} \frac{A(x+\Delta x) - A(x)}{\Delta x}}_{A'(x)} \underset{\geq}{\leq} \lim_{\Delta x \to 0} f(x+\Delta x).$$

- Since f is continuous it holds that $\lim_{\Delta x \to 0} f(x+\Delta x) = f(x)$.

Therefore, we get

$$f(x) \underset{\geq}{\leq} A'(x) \underset{\geq}{\leq} f(x)$$

$$\Longrightarrow A'(x) = f(x). \qquad \textbf{(Fundamental theorem of calculus I)}$$

- Therefore, the derivative of the area $A(x)$ at the point x is equal to the function f evaluated at the point x, and this is true for all x.
- This means that the problem of finding the area on the interval $[a, x]$ is equivalent to the problem of finding a function F such that $F' = f$.
- **Definition.** F is a "primitive", "antiderivative" or "indefinite integral" of the function f if $F' = f$. We write the indefinite integral of f as

$$\int f(x)dx.$$

- Hence, the area $A(\cdot)$ (as a function of the upper endpoint x) is a "primitive", "antiderivative" or "indefinite integral" of the function f.

- **Lemma 1.** If F is a primitive of f then $G = F + C$, where C is a constant (or scalar), is also a primitive of f.
- **Lemma 2.** If F and G are primitives of f then $G(x) - F(x) = C$, for all x.

- Hence,

$$A(x) = F(x) + C, \quad \text{where } F' = f.$$

- Note that $A(a) = 0$ so that

$$0 = A(a) = F(a) + C \Longrightarrow C = -F(a).$$

and, thus,

$$A(x) \quad = \quad F(x) - F(a), \quad \text{where } F' = f.$$

(Fundamental theorem of calculus II)

- In particular, the area A between the graph of the continuous and non-negative function f and the horizontal axis on the interval $[a, b]$ is

$$A = F(b) - F(a), \quad \text{where } F' = f.$$

- **Definition.** The definite integral of the continuous function f on the interval $[a, b]$ is

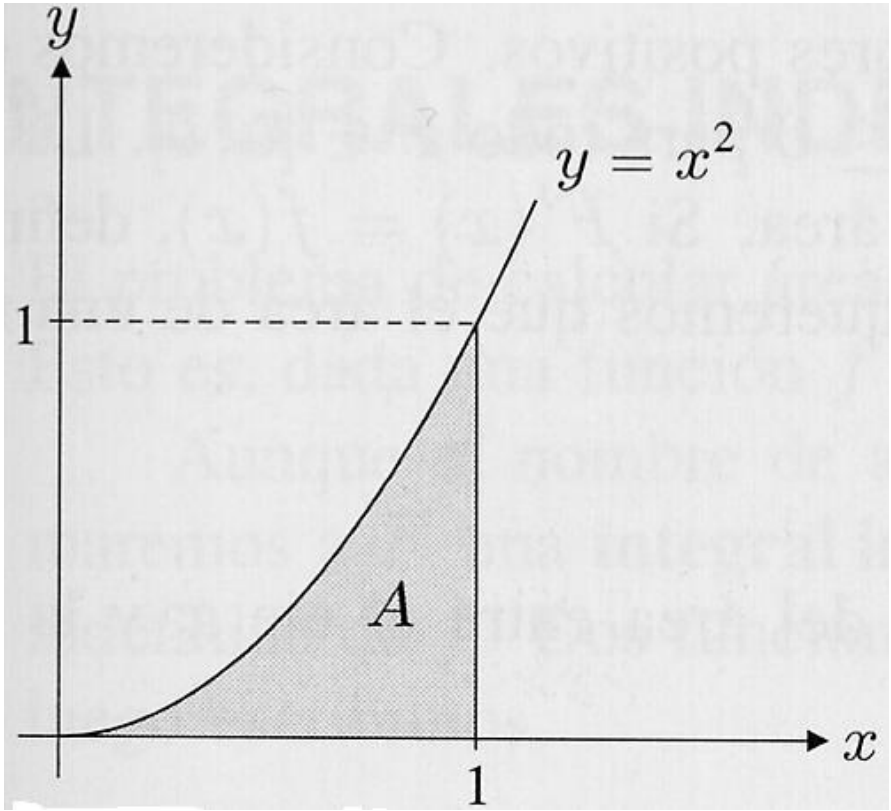
$$\int_a^b f(x)dx = F(b) - F(a) \equiv [F(x)]_a^b \equiv F(x)|_a^b, \quad \text{where } F' = f.$$

- Thus, when f is continuous and non-negative on the interval $[a, b]$, the area A between the graph of f and the horizontal axis on the interval $[a, b]$ is

$$A = \int_a^b f(x)dx.$$

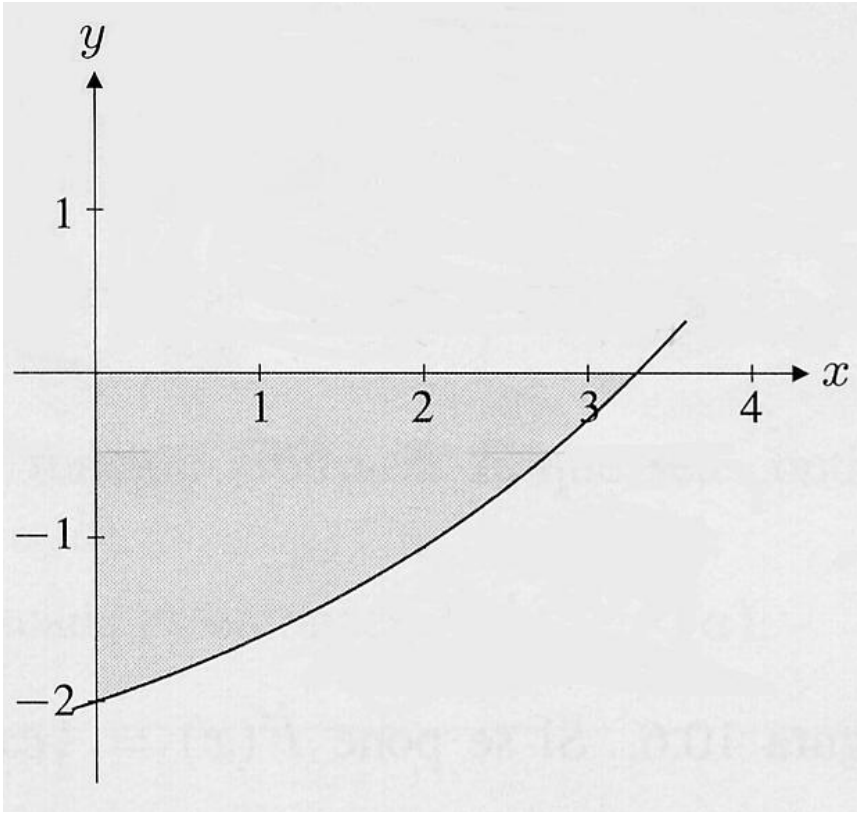
- This is the concept of integral due to Newton and Leibniz. Other more general concepts of integral are those of Riemann and of Lebesgue.

Example:



$$A = \int_0^1 x^2 dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}.$$

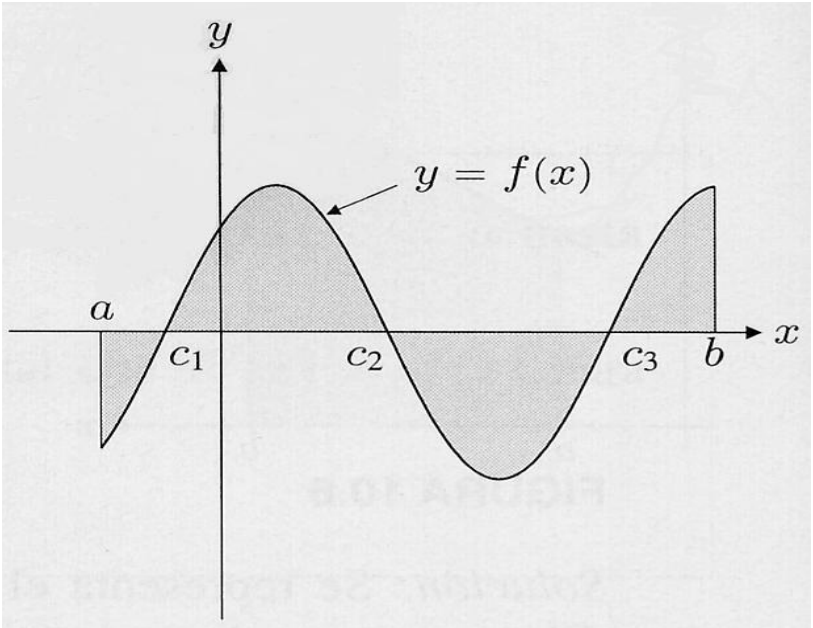
If the function f were negative,



then the area A of the shaded region will be

$$A = \int_a^b [-f(x)]\,dx = -\underbrace{\int_a^b f(x)dx}_{\text{negative}} \geq 0 \ .$$

If we want to find he area A between the graph of a continuous function f and the horizontal axis on the interval $[a, b]$, where the function takes both positive and negative values on that interval,



then the area of the shaded region will be

$$A = \underbrace{-\int_a^{c_1} f(x)dx}_{\text{negative}} + \underbrace{\int_{c_1}^{c_2} f(x)dx}_{\text{positive}} \underbrace{-\int_{c_2}^{c_3} f(x)dx}_{\text{negative}} + \underbrace{\int_{c_3}^{b} f(x)dx}_{\text{positive}} \geq 0.$$

2. The indefinite integral and its properties

- Recall that the indefinite integral of a continuous function f is

$$\int f(x)dx = F(x) + C, \quad \text{where } F' = f \text{ and } C \text{ is a constant.}$$

- **Properties of the indefinite integral of a continuous function:**
- **1.**

$$\int cf(x)dx = c\int f(x)dx, \quad \text{where } c \text{ is a constant.}$$

- **2.**

$$\int [f(x) + g(x)]\, dx = \int f(x)dx + \int g(x)dx.$$

Summing up:

$$\int \left(\sum_{n=1}^{N} \left[c_n f_n(x) \right] \right) dx = \sum_{n=1}^{N} \left[c_n \int f_n(x) dx \right], \text{ where } \{c_n\}_{n=1}^{N} \text{ are scalars.}$$

3. The definite integral and its properties

- Recall that the definite integral of the continuous function f is

$$\int_a^b f(x)dx = F(b) - F(a), \quad \text{where } F' = f.$$

- **Note:**

$$\int_a^b f(x)dx = \int_a^b f(z)dz = F(b) - F(a), \quad \text{where } F' = f.$$

- That is, the variable appearing as the argument of the function f is a "mute" variable.

Properties of the definite integral of a continuous function:

- **1.**
$$\int_a^b f(x)dx = -\int_b^a f(x)dx.$$

- **2.**
$$\int_a^a f(x)dx = 0.$$

- **3.**
$$\int_a^b cf(x)dx = c\int_a^b f(x)dx, \quad \text{where } c \text{ is a constant.}$$

- **4.**
$$\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)dx + \int_a^b g(x)dx.$$

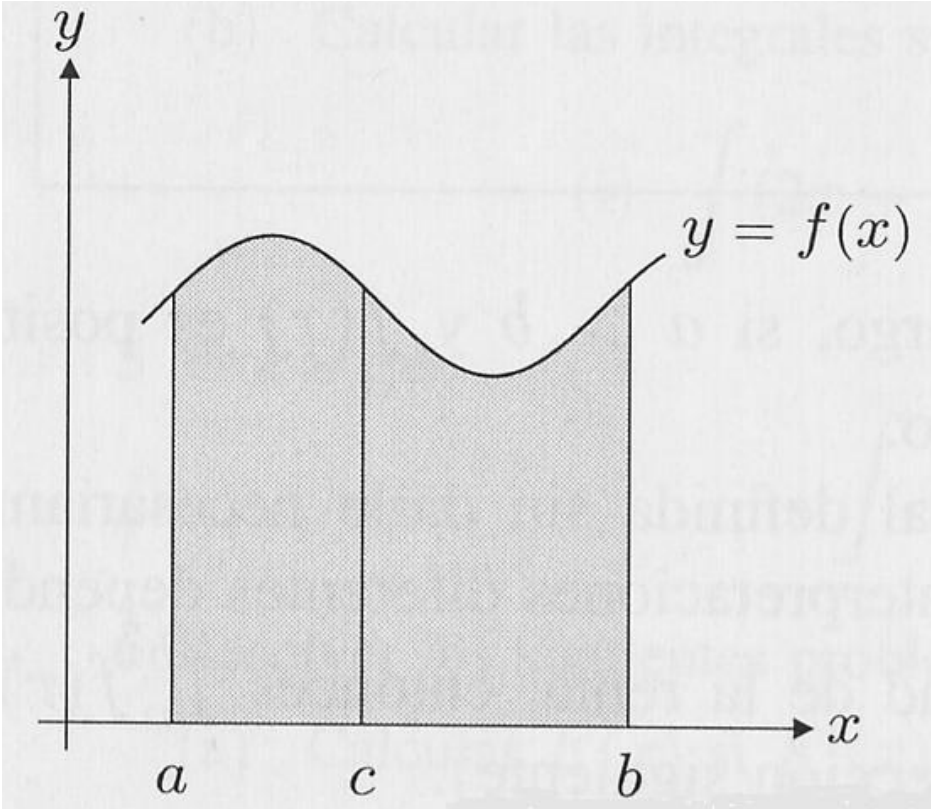
- **3 and 4 imply that**

$$\int_a^b \left(\sum_{n=1}^{N} \left[c_n f_n(x) \right] \right) dx = \sum_{n=1}^{N} \left[c_n \int_a^b f_n(x) dx \right],$$

where $\{c_n\}_{n=1}^{N}$ are scalars.

- **5.**

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx.$$



- **6.**

$$\frac{d}{dx}\int_a^x f(z)dz = f(x)$$

and

$$\frac{d}{dx}\int_x^b f(z)dz = -f(x).$$

- **7.** Let $b > a$. If $f(x) \geq g(x)$ for all $x \in [a, b]$, then

$$\int_a^b f(x)dx \geq \int_a^b g(x)dx.$$

- **8.**

$$\left| \int_a^b f(x)dx \right| \leq \int_a^b |f(x)|\, dx \quad \text{for } b > a.$$

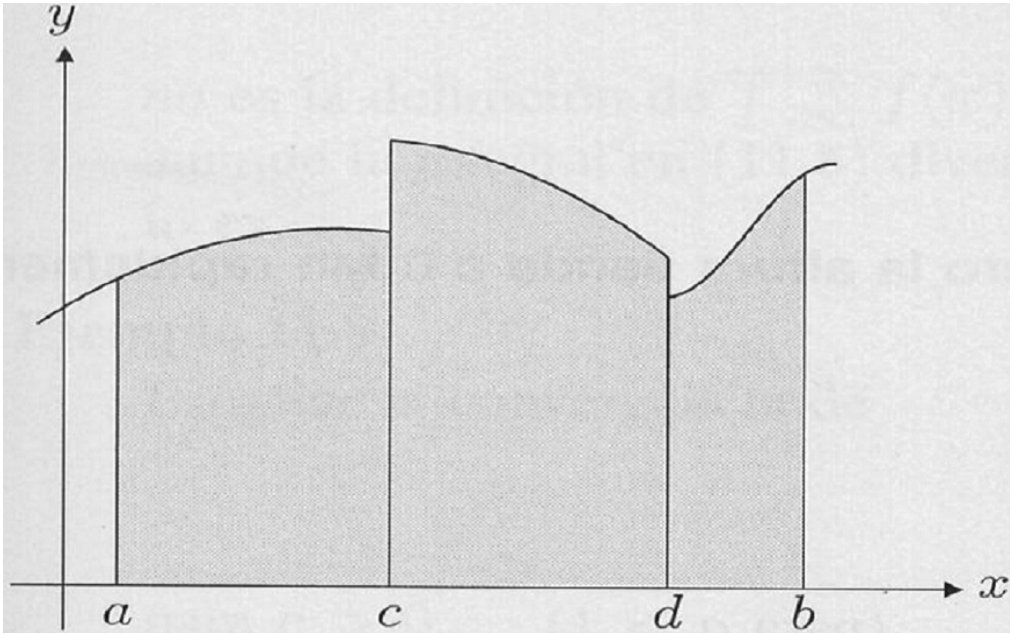
- **9. Cauchy-Schwarz Inequality:**

$$\left[\int_a^b |f(x)g(x)|\, dx \right]^2 \leq \left[\int_a^b [f(x)]^2\, dx \right] \cdot \left[\int_a^b [g(x)]^2\, dx \right]$$

or

$$\int_a^b |f(x)g(x)|\, dx \leq \left[\int_a^b [f(x)]^2\, dx \right]^{1/2} \cdot \left[\int_a^b [g(x)]^2\, dx \right]^{1/2} \text{ for } b > a.$$

- If the function f has a countable number of discontinuities on the interval $[a, b]$, then we perform the integral for each subinterval where the function is continuous and then we sum all the resulting integrals to obtain the integral on the whole interval $[a, b]$.



$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^d f(x)dx + \int_d^b f(x)dx$$

- Note that it does not matter if we count a point twice since the integral on a single point is zero by property 2.

4. Differentiation of integrals

- **On the differentiability and continuity of the integral**

$$H(z) \equiv \int_a^z f(t)dt, \quad \text{where } z \in [a, b].$$

- **1.** If f is continuous at x, then H is differentiable at x and $H'(x) = f(x)$.
- **2.** If f is discontinuous at x, then H is not differentiable at x.
- **3.** The function H is continuous on $[a, b]$,

$$\lim_{z \to x} H(z) = \lim_{z \to x} \int_a^z f(t)dt = \int_a^x f(t)dt = H(x), \quad \text{for all } x \in [a, b].$$

- Let $f(x,y)$ be a function such that the partial derivative $\dfrac{\partial f(x,y)}{\partial y}$ exists and is continuous. Then,

$$\frac{d}{dy}\int_a^b f(x,y)dx = \int_a^b \frac{\partial f(x,y)}{\partial y}dx.$$

Therefore, one may interchange the integral and partial differential operators.

- **Proof.**

$$\frac{d}{dy}\int_a^b f(x,y)dx = \lim_{h\to 0}\frac{\int_a^b f(x,y+h)dx - \int_a^b f(x,y)dx}{h} =$$

$$\lim_{h\to 0}\frac{\int_a^b \left[f(x,y+h)-f(x,y)\right]dx}{h} =$$

$$\int_a^b \lim_{h\to 0}\left[\frac{f(x,y+h)-f(x,y)}{h}\right]dx = \int_a^b \frac{\partial f(x,y)}{\partial y}dx.$$

- **Leibniz rule.** Let $f(x,y)$ be a function such that the partial derivative $\dfrac{\partial f(x,y)}{\partial y}$ exists and is continuous, and $a(y)$ and $b(y)$ be differentiable functions. Then,

$$\frac{d}{dy}\int_{a(y)}^{b(y)} f(x,y)dx =$$

$$\int_{a(y)}^{b(y)} \frac{\partial f(x,y)}{\partial y}dx + f(b(y),y)\cdot b'(y) - f(a(y),y)\cdot a'(y).$$

- **Proof.** Since the variable y appears thrice in the integral $\int_{a(y)}^{b(y)} f(x,y)dx$, we apply the chain rule to obtain

$$\frac{d}{dy}\int_{a(y)}^{b(y)} f(x,y)dx = \int_{a(y)}^{b(y)} \frac{\partial f(x,y)}{\partial y}dx + \left[\frac{d}{db}\int_{a(y)}^{b(y)} f(x,y)dx\right]\cdot b'(y)$$

$$+ \left[\frac{d}{da}\int_{a(y)}^{b(y)} f(x,y)dx\right]\cdot a'(y) =$$

$$\int_{a(y)}^{b(y)} \frac{\partial f(x,y)}{\partial y}dx + f(b(y),y)\cdot b'(y) - f(a(y),y)\cdot a'(y).$$

5. Integration rules: immediate integrals and integration by parts

- **Immediate integrals:**

$f(x)$	$\int f(x)dx = F(x) + C$
0	C (an arbitrary constant)
x^n, with $n \neq -1$	$\frac{x^{n+1}}{n+1} + C$
$1 \left(= x^0\right)$	$x + C$

$f(x)$	$\int f(x)dx = F(x) + C$
$\frac{1}{x}\left(= x^{-1}\right)$, with $x \neq 0$	$\ln \lvert x \rvert + C$
$f'(x) \cdot [f(x)]^n$, with $n \neq -1$	$\frac{[f(x)]^{n+1}}{n+1} + C$
$\frac{f'(x)}{f(x)}\left(= f'(x) \cdot [f(x)]^{-1}\right)$, with $f(x) \neq 0$	$\ln \lvert f(x) \rvert + C$

$f(x)$	$\int f(x)dx = F(x) + C$
e^x	$e^x + C$
$f'(x) \cdot e^{f(x)}$	$e^{f(x)} + C$
a^x with $a > 0$	$\frac{a^x}{\ln a} + C$

$f(x)$	$\int f(x)dx = F(x) + C$
$\sin x$	$-\cos x + C$
$\cos x$	$\sin x + C$
$\tan x$, with $\cos x \neq 0$	$-\ln \lvert\cos x\rvert + C$
$\ln x$, with $x > 0$	$x\left[(\ln x) - 1\right] + C$

- **Examples:**
- **1.**

$$\int x \cdot \underbrace{\left(x^2+4\right)}_{f(x)}^{1/2} dx = \int \frac{1}{2} \cdot \underbrace{2x}_{f'(x)} \underbrace{\left(x^2+4\right)^{1/2}}_{[f(x)]^{1/2}} dx$$

$$= \frac{1}{2} \int \underbrace{2x}_{f'(x)} \cdot \underbrace{\left(x^2+4\right)^{1/2}}_{[f(x)]^{1/2}} dx = \frac{1}{2}\left[\frac{\left(x^2+4\right)^{3/2}}{3/2}\right] + C$$

$$= \frac{1}{3}\left(x^2+4\right)^{3/2} + C.$$

- **2.**

$$\int (\tan x)\, dx = \int \frac{\sin x}{\underbrace{\cos x}_{f(x)}} dx = \int -\left(\frac{-\sin x}{\cos x}\right) dx = -\int \frac{\overbrace{-\sin x}^{f'(x)}}{\underbrace{\cos x}_{f(x)}} dx$$

$$= -\ln|\cos x| + C, \quad \text{if } \cos x \neq 0.$$

- **3.**

$$\int e^{\overbrace{3x+2}^{f(x)}} dx = \int \frac{1}{3} \cdot \underbrace{3}_{f'(x)} e^{\overbrace{3x+2}^{f(x)}} dx = \frac{1}{3} \int \underbrace{3}_{f'(x)} e^{\overbrace{3x+2}^{f(x)}} dx$$

$$= \frac{1}{3} e^{3x+2} + C.$$

- **Integration by parts:**
- Let F and G be the primitives of the continuous functions f and g, respectively. Then,

$$\frac{d\left[F(x)\cdot G(x)\right]}{dx} = \underbrace{f(x)}_{F'(x)}\cdot G(x) + F(x)\cdot \underbrace{g(x)}_{G'(x)}.$$

- Therefore, computing the indefinite integral of both sides we get

$$F(x)\cdot G(x) + C = \int f(x)\cdot G(x)dx + \int F(x)\cdot g(x)dx$$

or

$$\int F(x)\cdot g(x)dx = F(x)\cdot G(x) - \int f(x)\cdot G(x)dx + C.$$

- For definite integrals we have

$$\int_a^b F(x) \cdot g(x) dx = [F(x) \cdot G(x)]_a^b - \int_a^b f(x) \cdot G(x) dx.$$

- **Example:** Let $x > 0$,

$$\int (\ln x)\, dx = \int \left[\underbrace{(\ln x)}_{F(x)} \cdot \underbrace{1}_{g(x)} \right] dx = ...$$

- Make $F(x) = \ln x$ and $g(x) = 1$ so that $f(x) = \dfrac{1}{x}$ and $G(x) = x$.
-

$$... = \underbrace{(\ln x)}_{F(x)} \cdot \underbrace{x}_{G(x)} - \int \underbrace{\frac{1}{x}}_{f(x)} \cdot \underbrace{x}_{G(x)} dx + C = (\ln x) \cdot x - \int 1 dx + C$$

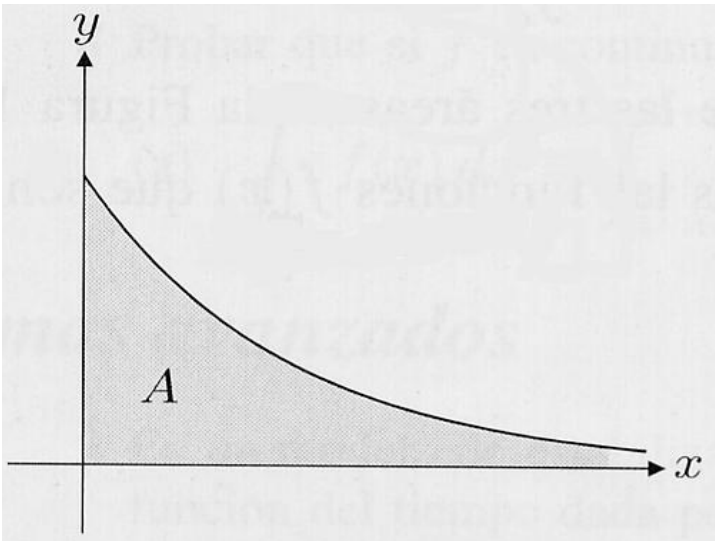
$$= (\ln x) \cdot x - x + C = x\left[(\ln x) - 1\right] + C.$$

- Moreover, the corresponding definite integral is

$$\int_a^b (\ln x)\, dx = b\left[(\ln b) - 1\right] - a\left[(\ln a) - 1\right].$$

6. Improper integrals

- So far we have looked at integrals on a closed interval $[a, b]$, where a and b are real numbers.
- Let us consider integrals of continuous functions on non-closed intervals. These integrals are called **improper**.



- Integral on the interval $[a, \infty)$:

$$\int_a^{\infty} f(x)dx \equiv \lim_{b \to \infty} \int_a^b f(x)dx.$$

- Integral on the interval $(\infty, b]$:

$$\int_{-\infty}^{b} f(x)dx \equiv \lim_{a \to -\infty} \int_{a}^{b} f(x)dx.$$

- Integral on the interval $(-\infty, \infty)$:

$$\int_{-\infty}^{\infty} f(x)dx \equiv \lim_{a \to -\infty} \int_{a}^{0} f(x)dx + \lim_{b \to \infty} \int_{0}^{b} f(x)dx.$$

- Integral on the right-semiclosed interval $(a, b]$:

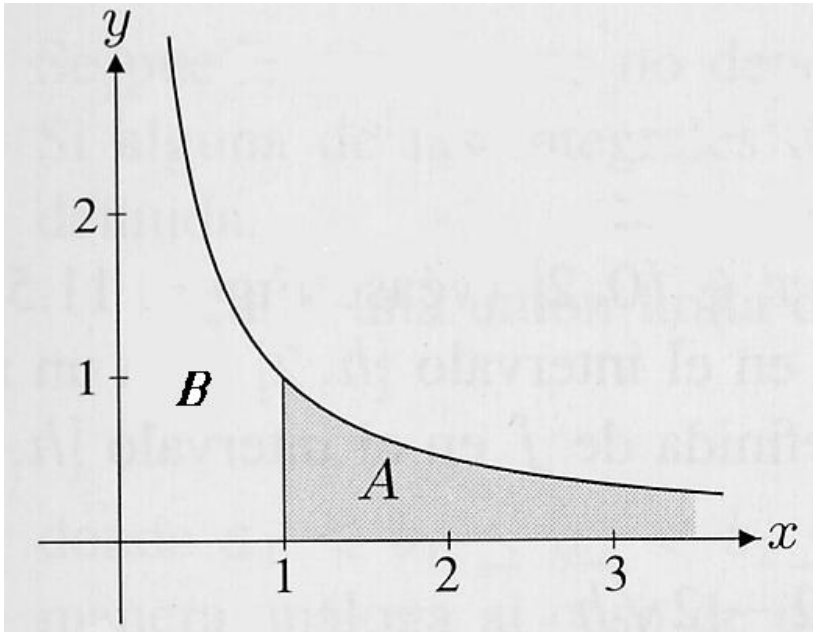
$$\int_{a^+}^{b} f(x)dx \equiv \lim_{z \to a^+} \int_{z}^{b} f(x)dx, \quad \text{where } z > a.$$

- Integral on the left-semiclosed interval $[a, b)$:

$$\int_{a}^{b-} f(x)dx \equiv \lim_{z \to b^-} \int_{a}^{z} f(x)dx, \quad \text{where } z < b.$$

- Integral on the open interval (a, b) :

$$\int_{a^+}^{b-} f(x)dx \equiv \underbrace{\lim_{z \to a^+} \int_z^c f(x)dx}_{B} + \underbrace{\lim_{z \to b^-} \int_c^z f(x)dx}_{A}, \quad \text{with } c \in (a, b) .$$



- All the previous limits might fail to exist (i.e., they could be equal to $\infty - \infty$) or be equal to $\pm\infty$. In the latter case, we say that the improper integral "diverges".
- **Examples:**
- **1.**

$$\int_{0^+}^{1} \frac{1}{x} dx = \lim_{z \to 0^+} \left[\ln |x| \right]_z^1 = \ln 1 - \lim_{z \to 0^+} (\ln |x|) = 0 - (-\infty) = \infty,$$

  so that the previous improper integral diverges.
- **2.**

$$\int_{1}^{\infty} \frac{1}{x} dx = \lim_{b \to \infty} \left[\ln |x| \right]_1^b = \lim_{b \to \infty} (\ln |x|) - \ln 1 = \infty - 0 = \infty,$$

  so that the previous improper integral diverges.
- **3.**

$$\int_{1}^{\infty} \frac{1}{x^2} dx = \lim_{b \to \infty} \left[-\frac{1}{x} \right]_1^b = \lim_{b \to \infty} \left(-\frac{1}{b} \right) - \left(-\frac{1}{1} \right) = 0 + 1 = 1.$$

7. Economic applications of integration

- **Investment and the stock of capital.**
- Under discrete time, $t = 0, 1, 2, ...$
- If K_t is the stock of capital at the beginning of period t, and I_t is the amount of investment during period t, we have

$$K_{t+1} = K_t + I_t \quad \text{or} \quad K_{t+1} - K_t = I_t.$$

- Therefore,

$$K_t = K_0 + \sum_{s=0}^{t-1} I_s.$$

- Under continuous time, $t \in [0, \infty)$.
- If $K(t)$ is the stock of capital at period t and $I(t)$ is the instantaneous amount of investment during the time interval $(t, t + \Delta t)$, where Δt is infinitessimal, we have

$$K(t + \Delta t) - K(t) = I(t) \cdot \Delta t$$

or

$$\frac{K(t + \Delta t) - K(t)}{\Delta t} = I(t),$$

so that

$$\lim_{\Delta t \to 0} \frac{K(t + \Delta t) - K(t)}{\Delta t} = K'(t) = I(t).$$

and, hence,

$$K(t) - K(0) = \int_0^t I(s)ds \quad \text{or} \quad K(t) = K(0) + \int_0^t I(s)ds.$$

- **Present value of income flows.**
- Under discrete time, $t = 0, 1, 2, ...$
- If $r \geq 0$ is the interest rate per period and $W_t > 0$ is the wealth at the beginning of period t, we have

$$W_{t+1} = (1 + r) W_t \quad \text{or} \quad W_{t+1} - W_t = rW_t.$$

- Therefore,

$$W_t = (1 + r)^t W_0 \quad \text{or} \quad W_0 = \frac{W_t}{(1 + r)^t},$$

where the latter equality means that the present value (at date $t = 0$) of having W_t euros at the future date t is $\frac{W_t}{(1 + r)^t}$ euros.

- An implication of the previous formula is that the present value PV of a stream of income $\{y_t\}_{t=0}^{T}$ is

$$PV = \sum_{t=0}^{T} \frac{y_t}{(1+r)^t}.$$

- If the time horizon of the stream is infinite, $T \longrightarrow \infty$, then

$$PV = \sum_{t=0}^{\infty} \frac{y_t}{(1+r)^t},$$

where the previous sum could diverge.

- Under continuous time, $t \in [0, \infty)$.

- If $r \geq 0$ is the instantaneous interest rate, $W(t)$ is the wealth at date t, and Δt is an infinitessimal time length, then we have

$$W(t + \Delta t) - W(t) = rW(t)\Delta t$$

or

$$\frac{W(t + \Delta t) - W(t)}{\Delta t} = rW(t).$$

- Thus,

$$\lim_{\Delta t \to 0} \frac{W(t + \Delta t) - W(t)}{\Delta t} = W'(t) = rW(t),$$

which can be written as

$$\frac{W'(t)}{W(t)} = r.$$

- Finding the definite integral in both sides, we have

$$\int_0^t \frac{W'(s)}{W(s)} ds = \int_0^t r ds$$

so that

$$[\ln W(s)]_0^t = [rs]_0^t .$$

Thus,

$$\ln W(t) - \ln W(0) = \ln \left(\frac{W(t)}{W(0)} \right) = rt,$$

which implies that

$$\frac{W(t)}{W(0)} = e^{rt} \quad \text{or} \quad W(t) = W(0)e^{rt} \quad \text{or} \quad W(0) = \frac{W(t)}{e^{rt}} = W(t)e^{-rt},$$

where the last equality says that the present value (at date $t = 0$) of having $W(t)$ euros at the future date t is $W(t)e^{-rt}$ euros.

- An implication of the previous formula is that the present value PV of a stream of income $y(t)$ for $t \in [0, T]$ is

$$PV = \int_0^T y(t)e^{-rt}dt.$$

- If the time horizon of the stream is infinite, $T \longrightarrow \infty$, then

$$PV = \int_0^\infty y(t)e^{-rt}dt,$$

where the previous improper integral may diverge.

- Note that, if r_c is the instantaneous interest rate in continuous time and r_d is the interest rate per period in discrete time, then

$$1 + r_d = e^{r_c}$$

so that

$$r_c = \ln(1 + r_d) \quad \text{or} \quad r_d = e^{r_c} - 1.$$

8. Integration with respect to several variables

- Assume that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuous function.
- The definite integral over a rectangle $A = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_{n-1}, b_{n-1}] \times [a_n, b_n]$ is

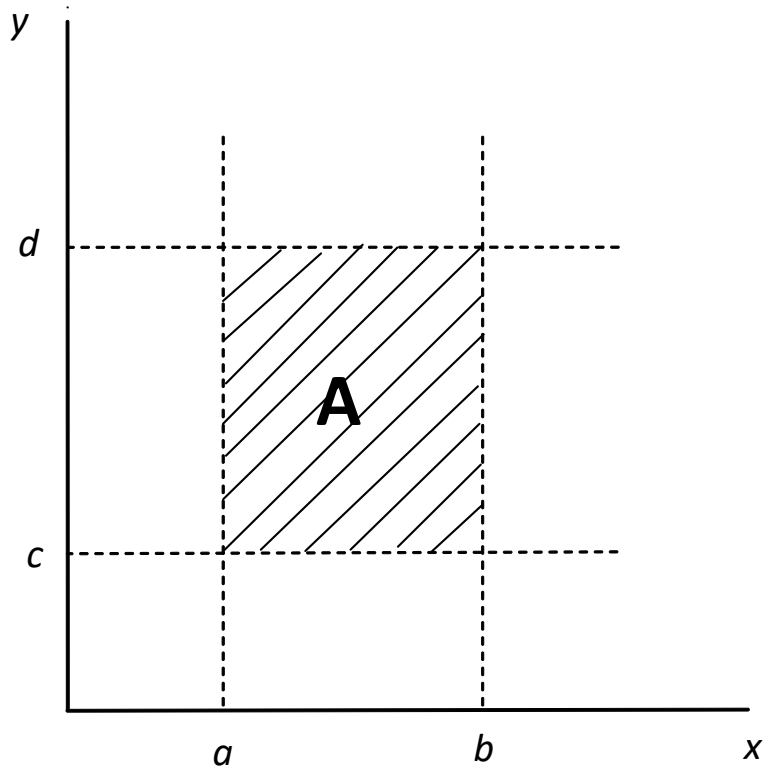
$$\int_A f(x_1, x_2, ..., x_{n-1}, x_n)\, d(x_1, x_2, ..., x_{n-1}, x_n) =$$

$$\int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} .... \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x_1, x_2, ..., x_{n-1}, x_n)\, dx_1 dx_2 ... dx_{n-1} dx_n =$$

$$\int_{a_n}^{b_n} \left[\int_{a_{n-1}}^{b_{n-1}} .... \left[\int_{a_2}^{b_2} \left[\int_{a_1}^{b_1} f(x_1, x_2, ..., x_{n-1}, x_n)\, dx_1 \right] dx_2 \right] ... dx_{n-1} \right] dx_n,$$

with $b_i > a_i$, $i = 1, 2, ...n$.

Example:



$$\int_A f(x,y)\, d(x,y), \quad \text{where } f(x,y) = x \cdot y \quad \text{and} \quad A = [a,b] \times [c,d].$$

$$\int_A f(x,y)\, d(x,y) = \int_c^d \int_a^b xy dx dy = \int_c^d \left[\int_a^b xy dx \right] dy$$

$$= \int_c^d y \left[\int_a^b x dx \right] dy = \int_c^d y \left[\frac{x^2}{2} \right]_a^b dy$$

$$= \int_c^d y \left(\frac{b^2}{2} - \frac{a^2}{2} \right) dy = \left(\frac{b^2}{2} - \frac{a^2}{2} \right) \int_c^d y dy = \left(\frac{b^2}{2} - \frac{a^2}{2} \right) \cdot \left[\frac{y^2}{2} \right]_c^d$$

$$= \left(\frac{b^2}{2} - \frac{a^2}{2} \right) \cdot \left(\frac{d^2}{2} - \frac{c^2}{2} \right) = \frac{\left(b^2 - a^2\right) \cdot \left(d^2 - c^2\right)}{4}.$$

Properties of the multiple integral (or integral with respect to several variables) over a rectangle:

- **1.**

$$\int_{a_n}^{b_n} \dots \int_{a_2}^{b_2} \int_{a_1}^{b_1} \underbrace{f_1(x_1) \cdot f_2(x_2) \cdot \dots \cdot f_n(x_n)}_{f(x_1,x_2,\dots,x_n)} dx_1 dx_2 \dots dx_n =$$

$$\left[\int_{a_1}^{b_1} f_1(x_1)\, dx_1\right] \cdot \left[\int_{a_2}^{b_2} f_2(x_2)\, dx_2\right] \cdot \dots \cdot \left[\int_{a_n}^{b_n} f_n(x_n)\, dx_n\right].$$

- **2.** Fubini's theorem:

$$\int_c^d \int_a^b f(x,y)\, dx dy = \int_a^b \int_c^d f(x,y)\, dy dx.$$

- **3.**

$$\frac{\partial^n}{\partial x_1 \partial x_2 ... \partial x_n} \int_{a_n}^{x_n} ... \int_{a_2}^{x_2} \int_{a_1}^{x_1} f\left(t_1, t_2, ..., t_{n-1}, t_n\right) dt_1 dt_2 ... dt_n$$

$$= f\left(x_1, x_2, ..., x_{n-1}, x_n\right).$$

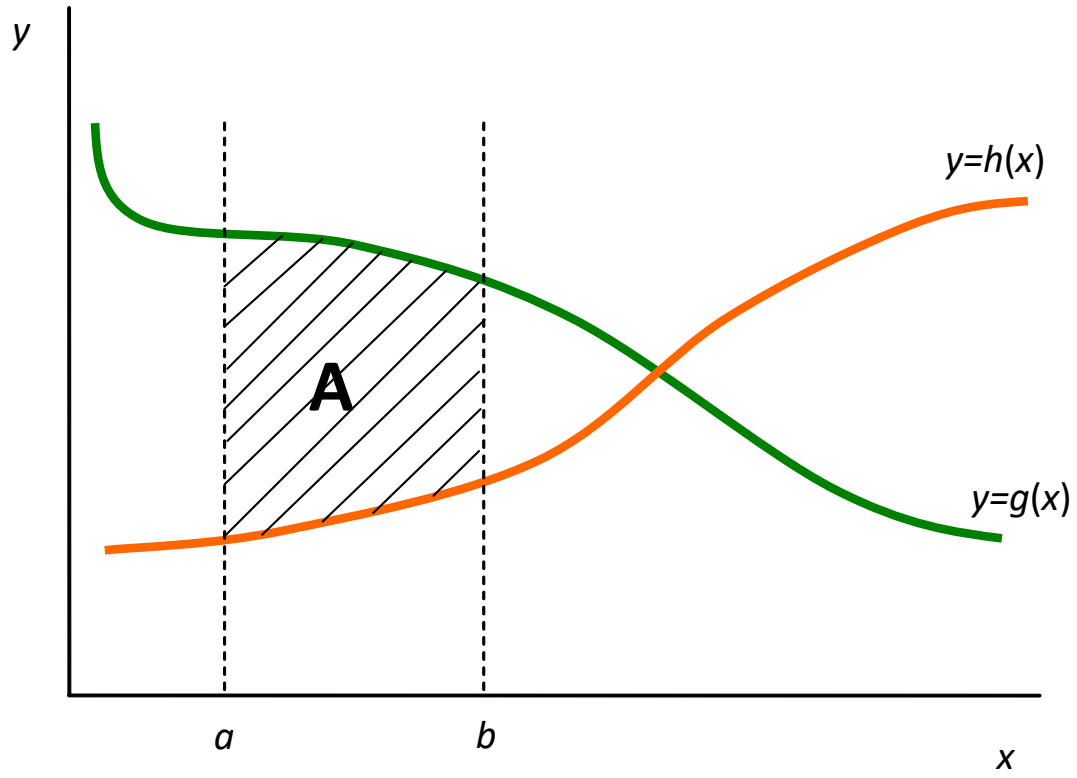
- **Integral over a non-rectangular region:**

$$\int_A f\left(x_1, x_2, ..., x_{n-1}, x_n\right) d\left(x_1, x_2, ..., x_{n-1}, x_n\right),$$

where

$$A \neq [a_1, b_1] \times [a_2, b_2] \times ... \times [a_{n-1}, b_{n-1}] \times [a_n, b_n].$$

- **Example:**
- Consider the following non-rectangular region A :



$$\int_A f(x,y)\, d(x,y) = \int_a^b \int_{h(x)}^{g(x)} f(x,y)\, dydx$$

$$= \int_a^b \left[\int_{h(x)}^{g(x)} f(x,y)\, dy \right] dx.$$

Exercises. Probability and Statistics. IDEA.
A Primer on Integration

1. Let f be a continuous function on $[a,b]$. Prove the following integral version of the mean value theorem:

$$\text{There exists an } \varepsilon \in [a,b] \text{ such that } \int_a^b f(x)dx = f(\varepsilon)(b-a).$$

2. Let us assume that f is a continuous function on $[a,b]$ with $f(x) \geq 0$ for all $x \in [a,b]$. Prove that

$$\int_a^b f(x)dx = 0$$

if and only if $f(x) = 0$ for all $x \in [a,b]$.

3. Prove the Cauchy-Schwarz inequality: Let f and g continuous functions on $[a,b]$, then it holds that

$$\left(\int_a^b |f(x)g(x)|\,dx\right)^2 \leq \left(\int_a^b [f(x)]^2\,dx\right)\cdot\left(\int_a^b [g(x)]^2\,dx\right).$$

4. Taking into account that the term being integrated is a square and, thus, it is positive, the following argument must be erroneous. Explain why.

$$\int_0^2 \frac{1}{(x-1)^2}dx = \left[-\frac{1}{x-1}\right]_0^2 = -1-1 = -2.$$

5. Let f and g be two continuous and differentiable functions on the real line. Find the derivative of $h(x)$ and the second derivative of $k(x)$, where

$$\text{(a) } h(x) = x^2\int_0^x f(t)g(t)dt \qquad \text{(b) } k(x) = e^{\int_0^x f(t)g(t)dt}.$$

6. Find the derivative of $G(x)$ in the following cases:

$$\text{(a) } G(x) = \int_1^e \ln(xt)\,dt$$

$$\text{(b) } G(x) = \int_0^{2x} t^3 dt$$

$$\text{(c) } G(x) = \int_{x^2}^x \frac{\sin(xt)}{t}\,dt \qquad (x>0).$$

7. **.** Compute the following definite integrals:

$$\text{(a) } \int_0^{\ln 2} 8e^{2x+3}dx \qquad \text{(b) } \int_0^{\pi/2} \frac{\sin x}{1+\cos x}dx \qquad \text{(c) } \int_1^2 x^2\ln x\,dx.$$

8. Find the following primitives (or indefinite integrals or antiderivatives):

(a) $\int x \ln x \, dx$ (b) $\int \frac{1}{\tan x} dx$ (c) $\int \sin^3 x \, dx$

(d) $\int \frac{\sqrt{x^2+4}+x}{\sqrt{x^2+4}} dx$ (e) $\int x^2 e^{2x} dx$ (f) $\int e^{3x} \cos 2x \, dx$

9. Compute the following integrals (some of them could be divergent or fail to exist):

(a) $\int_{-\infty}^{\infty} \frac{x}{1+x^2} \, dx$ (b) $\int_{0^+}^{1} x^{-2/3} dx$

(c) $\int_{-\infty}^{0} e^{rx} dx$ (analyze the result for all r) (d) $\int_{1}^{5} \frac{1}{x-2} \, dx$.

10. We have seen that the present value of an infinite income stream (or flow) $y(t)$ discounted at a constant interest rate $r > 0$ is given by $\int_0^{\infty} y(t) e^{-rt} dt$. Let us assume that the income stream $y(t)$ has the functional form $y(t) = A\,(2^t)$, that is, the income in $t = 0$ is A and it doubles in each time unit. For which values of the interest rate the present value of this income flow is finite? Find the present value in this case.

11. Compute the following double integrals on the specified domains:

(a) $\int\limits_D (2x + 3y + 4) \; d\,(x, y)$, where $D = [0, 1] \times [0, 2]$, that is, $0 \leq x \leq 1$ and $0 \leq y \leq 2$.

(b) $\int\limits_C \frac{1}{y} \; d\,(x, y)$, where C is the region of $\mathbb{R}^2$ such that $0 \leq x \leq y, \; 0 \leq y \leq 1, \; x + y \geq 1/2$.

(c) $\int\limits_B xy^2 \; d\,(x, y)$, where B is the region of $\mathbb{R}^2$ such that $0 \leq x \leq 1, \; x \leq y \leq x^2 + 1$.

12. It is known that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

and

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}.$$

Find $\int_0^{\infty} e^{-x^2} dx$ and $\int_0^{\infty} e^{-\frac{1}{2}x^2} dx$.

Measure Theory

Additional reference:

Kirman, A.P. (1982). Measure theory with applications to economics. Chapter 5 in *Handbook of Mathematical Economics*, ed. K.J. Arrow and M. Intriligator, Vol. 1, 159–209.

1. Lebesgue measure

Henri L. Lebesgue (1875 - 1941)

Émile Borel (1871 - 1956)

- Let I be an interval in $\mathbb{R}$. It may be open: (a, b), closed: $[a, b]$, right-semiclosed: $(a, b]$, or left-semiclosed: $[a, b)$, with $a < b$.
- The length of I is
$$\ell(I) = b - a.$$
- **Definition.** The σ-algebra $\mathcal{B}$ of the Borel sets in $\mathbb{R}$ is the σ-algebra generated by the collection of all the intervals of $\mathbb{R}$. In other words, $\mathcal{B}$ is the smallest σ-algebra containing all the intervals of $\mathbb{R}$.
- *Note:* A single point in $\mathbb{R}$ is a Borel set since
$$\{a\} = \bigcap_{n=1}^{\infty} \left[a, a + \frac{1}{n}\right].$$

- Moreover, $(-\infty, b) = \bigcup_{n=1}^{\infty} (a-n, b)$,

  $(a, \infty) = \bigcup_{n=1}^{\infty} (a, b+n)$, $(-\infty, b] = \bigcup_{n=1}^{\infty} (a-n, b]$,

  $[a, \infty) = \bigcup_{n=1}^{\infty} [a, b+n)$, etc.

- Not all subsets of $\mathbb{R}$ are Borel sets.

- **Definition.** Let B be a Borel set in $\mathbb{R}$, $B \in \mathcal{B}$. The **Lebesgue measure** of the Borel set B is

$$\mu(B) = \inf_{C} \sum_{k} \ell(I_k),$$

  where C is a countable union of intervals in $\mathbb{R}$ covering B, i.e., $B \subset C = \bigcup_{k} I_k$, where $\{I_1, I_2, ...\}$ is a countable collection of intervals in $\mathbb{R}$.

- Let $I \subset \mathbb{R}^n$ be an interval (or rectangle or box), i.e.,

$$I = \left\{ x \in \mathbb{R}^n \left| a_1 \underset{(<)}{\leq} x_1 \underset{(<)}{\leq} b_1, a_2 \underset{(<)}{\leq} x_2 \underset{(<)}{\leq} b_2, ..., a_n \underset{(<)}{\leq} x_n \underset{(<)}{\leq} b_n \right. \right\}.$$

- The length (or the area or the volume) of I is $\ell(I) = \prod_{i=1}^{n} (b_i - a_i)$.
- **Definition.** The σ-algebra $\mathcal{B}(\mathbb{R}^n)$ of the Borel sets in $\mathbb{R}^n$ is the σ-algebra generated by the collection of all the intervals (or rectangles or boxes) of $\mathbb{R}^n$. In other words, $\mathcal{B}(\mathbb{R}^n)$ is the smallest σ-algebra containing all the intervals of $\mathbb{R}^n$.
- **Definition.** Let B be a Borel set in $\mathbb{R}^n$, $B \in \mathcal{B}(\mathbb{R}^n)$. The **Lebesgue measure** of the Borel set B is

$$\mu(B) = \inf_{C} \sum_{k} \ell(I_k),$$

where C is a countable union of intervals (or rectangles or boxes) in $\mathbb{R}^n$ covering B, i.e., $B \subset C = \bigcup_k I_k$, where $\{I_1, I_2, ...\}$ is a countable collection of intervals (or rectangles or boxes) in $\mathbb{R}^n$.

- *Note:* The σ-algebra $\mathcal{B}(\mathbb{R}^n)$ of the Borel sets in $\mathbb{R}^n$ is sometimes also defined as the σ-algebra generated by the collection of all the open (or all the closed) sets of $\mathbb{R}^n$.
- **Proposition.** If $A = \{x_1, x_2, ...\}$ is a countable subset of $\mathbb{R}$, then $\mu(A) = 0$.
- **Proof.**

$$0 \leq \mu(A) \leq \inf_{\varepsilon > 0} \sum_k \ell\left[x_k - \varepsilon, x_k + \varepsilon\right] = 0.$$

- *Example:* Let $\mathbb{Q}$ be the set of rational numbers in $[0, 1]$, then $\mu(\mathbb{Q}) = 0$ since $\mathbb{Q}$ is countable. Moreover, the Lebesgue measure of the set $\mathbb{Q}^c$ of irrational numbers in $[0, 1]$ is $\mu(\mathbb{Q}^c) = 1$.
- *Note:* If $A = \{x_1, x_2, ...\}$ is a countable subset of $\mathbb{R}^n$, then $\mu(A) = 0$.
- The Lebesgue measure is the unique measure on the Borel sets of $\mathbb{R}^n$ with the property that the measure of any interval (or rectangle or box) equals its length (or area or volume).

2. Finite and sigma-finite measures

- **Definition.** A measure on $(\Omega, \mathcal{F})$ is finite if $\mu(\Omega) < \infty$.
- The Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ is not finite.
- A probability measure on $(\Omega, \mathcal{F})$ is finite since $P(\Omega) = 1 < \infty$.
- **Definition.** A measure on $(\Omega, \mathcal{F})$ is σ-finite if $\Omega = \bigcup_i A_i$, where $\{A_1, A_2, ...\}$ is a countable collection of elements of $\mathcal{F}$ with $\mu(A_i) < \infty$ for all i.
- *Examples:*
- (1) The Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ is σ-finite.
- (2) Counting measure: $\mu(A) = \#A$. Obviously, the counting measure on $(\mathbb{R}, \mathcal{B})$ is not σ-finite.
- (3) The Lebesgue measure extended on $\left(\overline{\mathbb{R}}, \mathcal{B}\left(\overline{\mathbb{R}}\right)\right)$ is not σ-finite. Note that $\mathcal{B}\left(\overline{\mathbb{R}}\right)$ is the σ-algebra generated by the intervals in $\overline{\mathbb{R}}$.

3. Lebesgue-Stieltjes measures

- **Definition.** A Lebesgue-Stieltjes measure on $\mathbb{R}^n$ is a measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that $\mu(I) < \infty$ for every bounded interval (or rectangle or box) I in $\mathbb{R}^n$, $n = 1, 2, ...$
- The Lebesgue measure is a particular case of Lebesgue-Stieltjes measure.
- A Lebesgue-Stieltjes measure on $\mathbb{R}^n$ is σ-finite (but not necessarily finite).

4. Distribution functions

- **Definition.** A distribution function $F : \mathbb{R} \longrightarrow \mathbb{R}$ is a (weakly) increasing and right-continuous function.
- Increasing: $a < b \Longrightarrow F(a) \leq F(b)$.
- Right-continuous at x_0: $\lim_{x \to x_0^+} F(x) \equiv F\left(x_0^+\right) = F\left(x_0\right)$, where $x > x_0$.
- A function is right-continuous if it is right-continuous at all points of its domain.

- **Proposition.**

- (a) Let μ be a Lebesgue-Stieltjes measure on $\mathbb{R}$. Let F be defined up to an additive constant by $F(b) - F(a) = \mu\,(a, b]$ for all $a < b$. Then, F is a distribution function.

- (b) Let $F : \mathbb{R} \longrightarrow \mathbb{R}$ be a distribution function. Let μ be a measure on $(\mathbb{R}, \mathcal{B})$ such that $\mu\,(a, b] = F(b) - F(a)$, for all $a < b$, and $\mu(B) = \inf_{C} \sum_{k} \mu\left(\hat{I}_k\right)$ for all $B \in \mathcal{B}$, where C is a countable union of right-semiclosed intervals in $\mathbb{R}$ covering B, i.e., $B \subset C = \bigcup_{k} \hat{I}_k$, where $\left\{\hat{I}_1, \hat{I}_2, ...\right\}$ is a countable collection of right-semiclosed intervals in $\mathbb{R}$. Then, μ is a Lebesgue-Stieltjes measure on $\mathbb{R}$.

- *Note:*

$$\lim_{x \to x_0^+} F(x) - F(x_0) = \lim_{x \to x_0^+} \mu(x_0, x] = \mu(\emptyset) = 0,$$

$$F(x_0) - \lim_{x \to x_0^-} F(x) = \lim_{x \to x_0^-} \mu(x, x_0] = \mu\{x_0\}.$$

- Thus, the distribution function $F : \mathbb{R} \longrightarrow \mathbb{R}$ is discontinuous at x_0 if and only if $\mu\{x_0\} > 0$.

- The σ-algebra generated by the right-semiclosed intervals is the σ-algebra of the Borel sets. To see this, observe that

$$(a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \tfrac{1}{n}\right], \ (-\infty, b) = \bigcup_{n=1}^{\infty} (a - n, b),$$

$$(a, \infty) = \bigcup_{n=1}^{\infty} (a, b + n), \quad [a, b] = \{(-\infty, a) \cup (b, \infty)\}^c,$$

$$[a, b) = (c, b) \cap [a, d], \quad (-\infty, b] = \bigcup_{n=1}^{\infty} (a - n, b],$$

$$[a, \infty) = \bigcup_{n=1}^{\infty} [a, b + n), \quad \{a\} = \bigcap_{n=1}^{\infty} \left[a, a + \tfrac{1}{n}\right], \quad \text{etc.}$$

- Usually in Statistics, when μ is a finite measure (for instance, when μ is a probability) on $(\mathbb{R}, \mathcal{B})$, we can make the following normalization without loss of generality:

$$\lim_{x \to -\infty} F(x) \equiv F(-\infty) = 0$$

so that

$$F(x) = F(x) - F(-\infty) = \mu(-\infty, x].$$

- Recall that the Lebesgue measure is a Lebesgue-Stieltjes measure. The distribution function associated with the Lebesgue measure is $F(x) = x + K$, where K is an arbitrary constant. However, in this case, $F(-\infty) = -\infty$ since the Lebesgue measure is not a finite measure.
- Clearly, the Lebesgue measure satisfies

$$\mu(a, b] = F(b) - F(a) = b - a.$$

Lebesgue-Stieltjes measure of intervals:

$$\mu\{a\} = F(a) - \underbrace{F(a^-)}_{\lim_{x \to a^-} F(x)}.$$

$$\mu(a,b] = F(b) - F(a).$$

$$\begin{aligned} \mu[a,b] &= \mu(a,b] + \mu\{a\} = F(b) - F(a) + F(a) - F(a^-) \\ &= F(b) - F(a^-). \end{aligned}$$

$$\begin{aligned} \mu(a,b) &= \mu(a,b] - \mu\{b\} = F(b) - F(a) - F(b) + F(b^-) \\ &= F(b^-) - F(a). \end{aligned}$$

$$\begin{aligned} \mu[a,b) &= \mu(a,b) + \mu\{a\} = F(b^-) - F(a) + F(a) - F(a^-) \\ &= F(b^-) - F(a^-). \end{aligned}$$

- **Definition.** A measure on $(\Omega, \mathcal{F})$ is concentrated on $B \in \mathcal{F}$ if $\mu(B^c) = 0$.
- **Example:** Assume a Lebesgue-Stieltjes measure concentrated on the set of rational numbers in the interval $[a, b]$ such that $0 < \mu[a, b] < \infty$. Then, the corresponding distribution function $F : \mathbb{R} \longrightarrow \mathbb{R}$ is continuous at each irrational point of $[a, b]$, the function F must have discontinuity points on the interval $[a, b]$ and, thus, these discontinuities must occur at rational points.

- **Extension to** $\left(\mathbb{R}^n, \mathcal{B}\left(\mathbb{R}^n\right)\right)$:
- Right-semiclosed interval (or rectangle or box) $(a, b]$ in $\mathbb{R}^n$:

$$(a, b] = \left\{x \in \mathbb{R}^n \,|a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2, ..., a_n < x_n \leq b_n\right\}.$$

- **Definition.** A distribution function $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a (weakly) increasing and right-continuous function.
- Increasing: $a < b \Longrightarrow F(a) \leq F(b)$, where a and b belong to $\mathbb{R}^n$.
- Right-continuous at x_0: $\lim_{x \to x_0^+} F(x) \equiv F\left(x_0^+\right) = F\left(x_0\right)$, where $x > x_0 \in \mathbb{R}^n$.
- **Notation.** Let $x = (x_1, ..., x_n)^\intercal$ and $y = (y_1, ..., y_n)^\intercal$ belong to $\mathbb{R}^n$.

  $x = y$ means that $x_i = y_i$ for $i = 1, ..., n$.

  $x \geqq y$ means that $x_i \geq y_i$ for $i = 1, ..., n$.

  $x > y$ means that $x_i \geq y_i$ for $i = 1, ..., n$ and $x \neq y$.

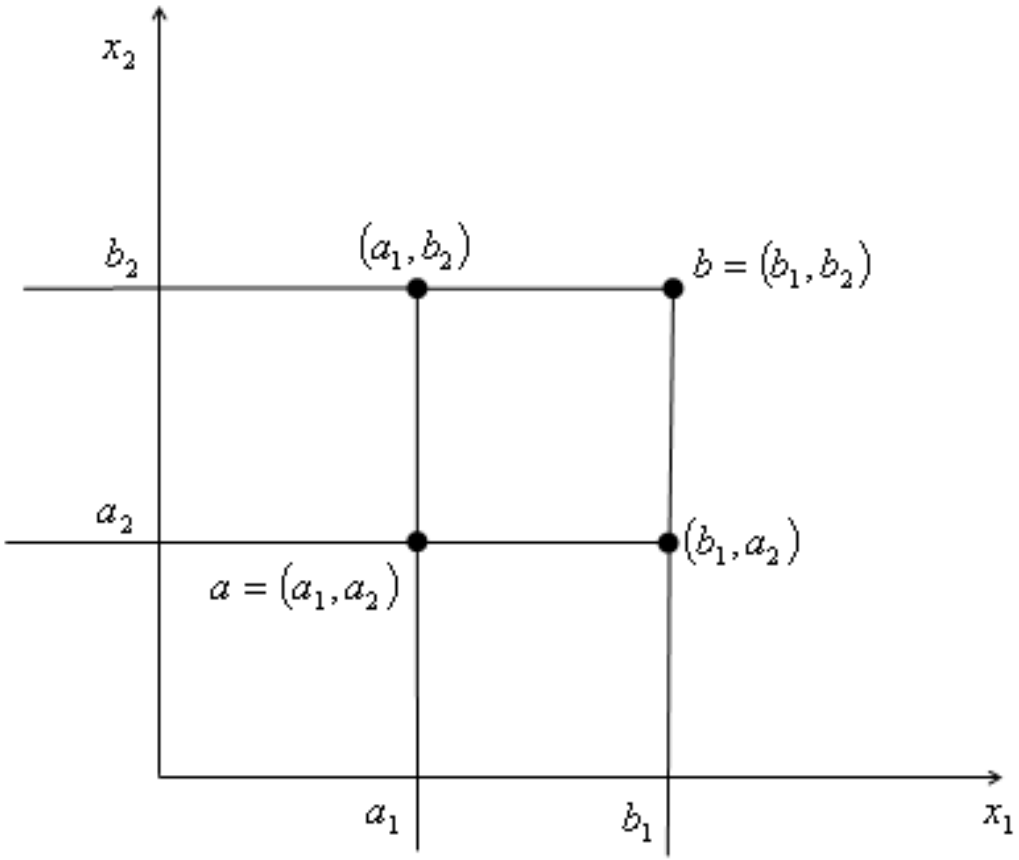
  $x \gg y$ means that $x_i > y_i$ for $i = 1, ..., n$.
- The same relationships hold for row vectors.

- **From μ to F :**
- Assume that μ is a finite measure on $\left(\mathbb{R}^n, \mathcal{B}\left(\mathbb{R}^n\right)\right)$. Define the corresponding distribution F as follows:

$$F(x) = \mu(-\infty, x] = \mu\underbrace{\{\omega \in \mathbb{R}^n \,|-\infty < \omega_i \leq x_i,\ i = 1, ..., n\}}_{(-\infty, x] \subset \mathbb{R}^n}.$$

- Note that $F(x) \longrightarrow 0$ if at least one of the components x_i of $x \in \mathbb{R}^n$ tends to $-\infty$.
- Moreover $F(x) \longrightarrow \mu\left(\mathbb{R}^n\right)$ if all the components x_i, $i = 1, ..., n$, of $x \in \mathbb{R}^n$ tend to ∞.
- Obviously, if μ is a probability on $\left(\mathbb{R}^n, \mathcal{B}\left(\mathbb{R}^n\right)\right)$ then $F(x) \longrightarrow 1 = \mu\left(\mathbb{R}^n\right)$ if all the components x_i, $i = 1, ..., n$, of $x \in \mathbb{R}^n$ tend to ∞.

- **From F to μ :**
- $n = 2$



$$\mu(a, b] = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2).$$

- $n = 3$

$$\begin{aligned}\mu(a, b] &= F(b_1, b_2, b_3) - F(a_1, b_2, b_3) - F(b_1, a_2, b_3) \\ &\quad - F(b_1, b_2, a_3) + F(b_1, a_2, a_3) + F(a_1, b_2, a_3) \\ &\quad + F(a_1, a_2, b_3) - F(a_1, a_2, a_3).\end{aligned}$$

- We extend the formula to all the Borel sets of $\mathbb{R}^n$ as follows:

$$\mu(B) = \inf_C \sum_k \mu\left(\hat{I}_k\right) \text{ for all } B \in \mathcal{B},$$

where C is a countable union of right-semiclosed intervals (or rectangles or boxes) in $\mathbb{R}^n$ covering B, i.e., $B \subset C = \bigcup_k \hat{I}_k$, where $\left\{\hat{I}_1, \hat{I}_2, ...\right\}$ is a countable collection of right-semiclosed intervals (or rectangles or boxes) in $\mathbb{R}^n$.

5. Measurable and Borel measurable functions

- **Definition.** Consider the two measurable spaces $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$. The function $f : \Omega \longrightarrow \Omega'$ is measurable if $f^{-1}(A) \in \mathcal{F}$ for each $A \in \mathcal{F}'$.
- Notation for measurable function: $f : (\Omega, \mathcal{F}) \longrightarrow (\Omega', \mathcal{F}')$.
- $f : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a (real-valued) Borel measurable function.
- $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is an (extended real-valued) Borel measurable function.
- $f : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ is a (real vector-valued) Borel measurable function.
- $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}^n, \mathcal{B}\right)$ is an (extended real vector-valued) Borel measurable function.

- Recall that we can write $\mathcal{B}\left(\overline{\mathbb{R}}\right)$, $\mathcal{B}\left(\mathbb{R}^n\right)$, $\mathcal{B}\left(\overline{\mathbb{R}}^n\right)$.
- Note that $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is Borel measurable iff

$$\{\omega \mid f(\omega) < c\} \in \mathcal{F}\ , \quad \text{for all } c \in \overline{\mathbb{R}}.$$

- or iff $\{\omega \mid f(\omega) > c\} \in \mathcal{F}$, or iff $\{\omega \mid f(\omega) \leq c\} \in \mathcal{F}$, or iff $\{\omega \mid f(\omega) \geq c\} \in \mathcal{F}$, for all $c \in \overline{\mathbb{R}}$.

6. Properties of measurable functions.

- **1.** If $f_n : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$, for all n, and $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ (i.e., $f_n(\omega) \to f(\omega)$) for all $\omega \in \Omega$, then $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$. (The pointwise limit of Borel measurable functions is Borel measurable).
- **2.** If $f_1 : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ and $f_2 : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ are Borel measurable functions, so are $f_1 + f_2$, $f_1 - f_2$, $f_1 \cdot f_2$, and f_1 / f_2 (assuming these are well defined, i.e., we do not add $+\infty$ and $-\infty$, divide by 0, or divide ∞ by ∞).
- *Note:* $\infty - \infty$, $\frac{a}{0}$ for all $a \in \mathbb{R}$, and $\frac{\infty}{\infty}$ are not well-defined (or are "indeterminate" or "indefinite") in $\overline{\mathbb{R}}$.

- **3.** If $f : (\Omega_1, \mathcal{F}_1) \longrightarrow (\Omega_2, \mathcal{F}_2)$ and $g : (\Omega_2, \mathcal{F}_2) \longrightarrow (\Omega_3, \mathcal{F}_3)$, then $g(f) : (\Omega_1, \mathcal{F}_1) \longrightarrow (\Omega_3, \mathcal{F}_3)$. (The composition of measurable functions is measurable).
- **4.** $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}^n, \mathcal{B}\right)$ if and only if $f_i : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$, for $i = 1, ..., n$, where $f = (f_1, ..., f_n)$.
- *Note:* $f_i = p_i(f)$, where the projection maps p_i, for $i = 1, ..., n$, satisfy $p_i : \left(\overline{\mathbb{R}}^n, \mathcal{B}\right) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$.

- **Definition.** The Borel measurable function $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is simple if it takes on only finitely many distinct values $\{y_1, y_2, \ldots, y_n\}$.
- Therefore, $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is simple if it can be written as

$$f(\omega) = \sum_{i=1}^{n} y_i \mathbb{I}_{A_i}(\omega), \quad \text{for all } \omega \in \Omega,$$

where $f(\omega) = y_i$ if $\omega \in A_i$ and $\mathbb{I}_{A_i}$ is the indicator function of A_i. That is,

$$\mathbb{I}_A(\omega) = \begin{cases} 1 \text{ if } \omega \in A \\ \\ 0 \text{ if } \omega \notin A. \end{cases}$$

- Note that $\{A_i\}_{i=1}^{n}$ is a collection of sets in $\mathcal{F}$ that form a partition of Ω.

- **5.** Every non-negative Borel measurable function $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is the limit of an increasing sequence of non-negative, finite-valued, simple functions f_n.
- **6.** Every Borel measurable function $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is the limit of a sequence of finite-valued simple functions f_n, with $|f_n| \leq |f|$ for all n.

7. Integral with respect to a measure and its properties

Arithmetic of $\overline{\mathbb{R}}$: if $a \in \mathbb{R}$, then

- $a + \infty = \infty$,
- $a - \infty = -\infty$,
- $a/\infty = a/(-\infty) = 0$,
- $a \cdot \infty = \infty$ if $a > 0$,
- $a \cdot \infty = -\infty$ if $a < 0$,
- $0 \cdot \infty = 0 \cdot (-\infty) = 0$,
- $\infty + \infty = \infty$,
- $-\infty - \infty = -\infty$.
- Recall that $\infty - \infty$, $\frac{a}{0}$ for all $a \in \mathbb{R}$, and $\frac{\infty}{\infty}$ are not well-defined (or are "indeterminate" or "indefinite") in $\overline{\mathbb{R}}$.

- **Definition of integral with respect to a measure μ on $(\Omega, \mathcal{F})$.**
- If $h : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is a simple Borel measurable function, we define

$$\int_{\Omega} h d\mu = \sum_{i=1}^{n} y_i \mu\left(A_i\right), \quad \text{where } h(\omega) = y_i \text{ if } \omega \in A_i, \tag{1}$$

  provided $+\infty$ and $-\infty$ do not appear in the sum.
- If $h : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is a non-negative Borel measurable function, we define

$$\int_{\Omega} h d\mu = \sup_{s} \int_{\Omega} s d\mu, \tag{2}$$

  where s is a simple Borel measurable function with $0 \leq s \leq h$.
- *Note:* the integral in (1) may be equal to $\pm\infty$, whereas that in (2) may be equal to $+\infty$

- If $h : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is a Borel measurable function, let

$$h^+ = \max(h, 0) \geq 0, \quad \text{(the positive part of } h\text{)}$$

$$h^- = \max(-h, 0) \geq 0, \quad \text{(the negative part of } h\text{)}$$

  and observe that $h = h^+ - h^-$.

- We define the integral of h w.r.t. μ as follows:

$$\int_\Omega h d\mu = \int_\Omega h^+ d\mu - \int_\Omega h^- d\mu,$$

  provided it is not of the form $\infty - \infty$.

- We could write $\int_\Omega h(\omega) d\mu(\omega)$ or $\int_\Omega h(\omega) \mu(d\omega)$.

- **Definition:** The function h is said to be μ-integrable (or integrable w.r.t. μ) if $\int_\Omega h d\mu$ is finite.
- If $\int_\Omega h d\mu$ is of the form $\infty - \infty$, we say that the integral $\int_\Omega h d\mu$ does not exist. Therefore, the condition of μ-integrability requires the previous existence of the integral.
- From now on, when we write an integral, it should be understood that the integral exists.
- Sometimes $\int_\Omega h d\mu$ is called "abstract Lebesgue integral".
- We define

$$\int_B h(\omega) d\mu(\omega) = \int_\Omega h(\omega) \mathbb{I}_B(\omega) d\mu(\omega), \quad \text{where } B \in \mathcal{F}.$$

- Let $f : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$. The Lebesgue integral of f is the integral of f with respect to Lebesgue measure and can be written as

$$\int_B f(x_1, ..., x_n) d(x_1, x_2, .., x_n), \quad \text{for } B \in \mathcal{B}(\mathbb{R}^n).$$

- **Properties of the integral** (for extended real-valued Borel measurable functions)**:**
- **1.**
$$\int_{\Omega} chd\mu = c \int_{\Omega} hd\mu, \quad \text{for any constant } c.$$
- **2.** If $g \geq h$, then
$$\int_{\Omega} gd\mu \geq \int_{\Omega} hd\mu.$$
- **3.**
$$\left| \int_{\Omega} hd\mu \right| \leq \int_{\Omega} |h| \, d\mu.$$

- **4.** If $\int_{\Omega} h d\mu$ exists, so does $\int_{B} h d\mu$ for all $B \in \mathcal{F}$.
- **5.** Define $\lambda(B) = \int_{B} h d\mu$ for all $B \in \mathcal{F}$. Then, λ is a signed measure on $(\Omega, \mathcal{F})$.
- **Proof.** Let $\{B_i\}$ be a countable collection of disjoint measurable sets, then

$$\lambda\left(\bigcup_i B_i\right) = \int_{\bigcup_i B_i} h d\mu = \sum_i \left(\int_{B_i} h d\mu\right) = \sum_i \lambda\left(B_i\right).$$

Moreover, $\lambda(\emptyset) = \int_{\{\emptyset\}} h d\mu = 0$. *Q.E.D.*

- **6.**

$$\int_{\Omega} \left(\sum_i f_i \right) d\mu = \sum_i \left(\int_{\Omega} f_i d\mu \right).$$

- **7.** h is integrable if and only if $|h|$ is integrable.
- Note that $|h| = h^+ + h^-$.
- **8.** If $|g| \leq h$ and h is integrable, then g is integrable.

- **Definition:** Let f and g be Borel measurable functions, then $f = g$ almost everywhere with respect to μ (or a.e. $[\mu]$) if

$$\mu\{f \neq g\} \equiv \mu\{\omega \in \Omega \mid f(\omega) \neq g(\omega)\} = 0.$$

- **9.** If $f = 0$ a.e. with respect to μ, then $\int_\Omega f d\mu = 0$.
- **10.** If $f = g$ a.e. $[\mu]$, then

$$\int_\Omega f d\mu = \int_\Omega g d\mu.$$

- **11.** If $h \geq 0$ and $\int_\Omega h d\mu = 0$, then $h = 0$ a.e. (with respect to μ).
- **12.** If h is integrable (with respect to μ), then h is finite a.e. (with respect to μ).
- **13.**

$$\mu(B) = \int_B 1 d\mu \equiv \int_B d\mu \equiv \int_\Omega \mathbb{I}_B(\omega) d\mu(\omega) \text{ for all } B \in \mathcal{F}.$$

8. Convergence of integrals

- **Definition (convergence almost everywhere)**. Let $f_1, f_2, ...,$ and f be extended real-valued Borel measurable functions on $(\Omega, \mathcal{F}, \mu)$. The sequence of functions $\{f_n\}$ converges to f almost everywhere with respect to the measure μ (or $f = \lim_{n\to\infty} f_n$ a.e. $[\mu]$ or $f_n \longrightarrow f$ a.e. $[\mu]$) if

$$\mu\left\{\omega \in \Omega \;\middle|\; \lim_{n\to\infty} f_n(\omega) \neq f(\omega)\right\} = 0.$$

- **Fatou's lemma.** Let $f_1, f_2, ...$ be a sequence of non-negative a.e. $[\mu]$, extended real-valued Borel measurable functions on $(\Omega, \mathcal{F}, \mu)$ and

$$f = \liminf_{n\to\infty} f_n \equiv \lim_{n\to\infty}\left(\inf_{k\geq n} f_k\right) \quad \text{a.e. } [\mu].$$

Then the function f is Borel measurable and

$$\int_\Omega f d\mu \leq \liminf_{n\to\infty} \int_\Omega f_n d\mu,$$

where the integrals may be infinite.

- **Monotone Convergence Theorem (Levi).** Let $f_1, f_2, ...$ be a non-decreasing a.e. $[\mu]$ sequence of extended real-valued Borel measurable functions on $(\Omega, \mathcal{F}, \mu)$ and $f = \lim_{n\to\infty} f_n$ a.e. $[\mu]$ (or $f_n \longrightarrow f$ a.e. $[\mu]$). Then f is Borel measurable and

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \int_\Omega f d\mu \quad \left(\text{or } \int_\Omega f_n d\mu \longrightarrow \int_\Omega f d\mu \right),$$

  where the integrals may be infinite.

- **Dominated Convergence Theorem (Lebesgue).** Let $f_1, f_2, ..., f$, and g be extended real-valued Borel measurable functions on $(\Omega, \mathcal{F}, \mu)$ with $|f_n| \leq g$ a.e. $[\mu]$ for all n, where g is μ-integrable, and $f = \lim_{n\to\infty} f_n$ a.e. $[\mu]$ (or $f_n \longrightarrow f$ a.e. $[\mu]$). Then f is μ-integrable and

$$\lim_{n\to\infty} \int_\Omega f_n d\mu = \int_\Omega f d\mu \quad \left(\text{or } \int_\Omega f_n d\mu \longrightarrow \int_\Omega f d\mu \right).$$

- Thus, under the assumptions of the previous two theorems, "the limit of integrals is equal to the integral of the limit."

9. The Lebesgue and the Lebesgue-Stieltjes integrals and their relation with the Riemann and the Riemann-Stieltjes integrals.

- Let $f : (\mathbb{R}, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$
- Riemann integral of the function f on $[a, b]$:

$$\int_a^b f(x)dx.$$

- Lebesgue integral of the function f on $[a, b]$:

$$\int_{[a,b]} f(x)dx.$$

- *Note:*

$$\int_{[a,b]} f(x)dx = \int_{(a,b)} f(x)dx = \int_{(a,b]} f(x)dx = \int_{[a,b)} f(x)dx.$$

- Riemann-Stieltjes (or just Stieltjes) integral of the function f on $(a, b]$:

$$\int_a^b f(x)dF(x),$$

where F is a distribution function (i.e., an increasing and right-continuous function) on $\mathbb{R}$.

- Lebesgue-Stieltjes integral of the function f on $(a, b]$:

$$\int_{(a,b]} fdF \quad \text{or} \quad \int_{(a,b]} f(x)dF(x),$$

where we are integrating with respect to the Lebesgue-Stieltjes measure μ on $(a, b]$ associated with the distribution function F.

- Note that

$$\int_{[a,b]} fdF = \int_{(a,b]} fdF + f(a)\mu\{a\} = \int_{(a,b]} fdF + f(a)\left[F(a) - F(a^-)\right].$$

- See the handout for the relation between these integrals.
- The function defined on $[0,1]$ by $f(x) = 1$ if x is irrational, and $f(x) = 0$ if x is rational, is the standard example of a function that is Lebesgue integrable but not Riemann integrable. Its Lebesgue integral is $\int_{[0,1]} f dx = 1$.

10. Absolute continuity of measures

- **Definition.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $h : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ a Borel measurable function such that $\int_\Omega h d\mu$ exists. Then the countably additive set function (or signed measure),

$$\lambda(A) = \int_A h d\mu, \quad \text{for all } A \in \mathcal{F},$$

  is called the indefinite integral of h (with respect to μ).

- If μ is Lebesgue measure and $A = [a, x]$, then

$$\lambda[a, x] = \int_{[a,x]} h(t) dt$$

  is the familiar indefinite integral of calculus.

- **Definition.** If μ is a measure on $(\Omega, \mathcal{F})$ and λ is a signed-measure on $(\Omega, \mathcal{F})$, we say that λ is absolutely continuous with respect to μ, $\lambda \ll \mu$, if $\mu(A) = 0$ implies that $\lambda(A) = 0$.

- **Proposition.** If λ is an indefinite integral of the Borel measurable function h w.r.t. μ, i.e.,

$$\lambda(A) = \int_A h d\mu, \quad \text{for all } A \in \mathcal{F},$$

then $\lambda \ll \mu$.

11. Radon-Nikodym theorem.

- **Radon-Nikodym Theorem.** Let μ be a σ-finite measure on $(\Omega, \mathcal{F})$ and λ be a signed-measure on $(\Omega, \mathcal{F})$. Assume that $\lambda \ll \mu$. Then, there exists a function $g : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ such that

$$\lambda(A) = \int_A g d\mu, \quad \text{for all } A \in \mathcal{F}.$$

  This Borel measurable function g is called the Radon-Nikodym derivative of λ w.r.t. μ, and is written $g = d\lambda/d\mu$. Finally, if h is another such function, then $g = h$ a.e. $[\mu]$.

- Therefore, if μ is σ-finite, $\lambda \ll \mu$ if and only if λ is an indefinite integral of some Borel measurable function h w.r.t. μ.

- If the signed-measure λ is finite then the function g should be finite a.e. $[\mu]$. Therefore, in this case, we can assume without loss of generality that the existing function g is finite, $g : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$.

- *Note:* The assumption that μ is σ-finite is crucial.
- **Example:** Let μ be the counting measure on $(\mathbb{R}, \mathcal{B})$ and λ be the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$.
- Obviously, $\mu(B) = 0$ implies that $\lambda(B) = 0$, i.e., $\lambda \ll \mu$.
- Assume that

$$\lambda(A) = \int_A g d\mu, \quad \text{for all } A \in \mathcal{B},$$

for some function $g : (\mathbb{R}, \mathcal{B}) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$. Take $A = \{x\}$ so that $\lambda\{x\} = g(x)$. Since $\lambda\{x\} = 0$ for all $x \in \mathbb{R}$, it holds that $g(x) = 0$ for all $x \in \mathbb{R}$. Therefore, we obtain that $\lambda(A) = 0$ for all $A \in \mathcal{B}$, which is a contradiction.

- In this example the Radon-Nikodym theorem does not apply since the counting measure μ on $(\mathbb{R}, \mathcal{B})$ is not σ-finite.

12. Absolutely continuous functions

- **Definition.** The function $f : [a, b] \longrightarrow \mathbb{R}$ is absolutely continuous on $[a, b]$ if there exists a Borel measurable function $g : ([a, b], \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$, which is integrable w.r.t. Lebesgue measure, such that

$$f(x) - f(a) = \int_{[a,x]} g(t)dt, \quad \text{for } a \leq x \leq b.$$

- **Note:** In the previous definition we could have $g : ([a, b], \mathcal{B}) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$, but g should be finite a.e. on $[a, b]$ w.r.t. Lebesgue measure since f is a real-valued function.

- **Theorem.** Let the function f be absolutely continuous on $[a, b]$, i.e.,

$$f(x) - f(a) = \int_{[a,x]} g(t)dt, \quad \text{for } a \leq x \leq b, \qquad (\star)$$

for some Borel measurable function g defined on $[a, b]$, then $f' = g$ a.e. w.r.t. Lebesgue measure.

- Therefore, f is absolutely continuous on $[a, b]$ if and only if it is the integral of its derivative (a.e. w.r.t. Lebesgue measure).
- Clearly, if f is absolutely continuous on $[a, b]$, then

$$\lim_{h \to 0} f(x+h) - f(x) = \lim_{h \to 0} \int_{[a, x+h]} g(t)dt - \int_{[a,x]} g(t)dt$$

$$= \lim_{h \to 0} \int_{[x, x+h]} g(t)dt = \int_{\{x\}} g(t)dt = 0 \quad \text{for all } x \in [a, b].$$

- Therefore, absolute continuity implies continuity but the converse is not true (counterexample: the Cantor function).
- *Note:* In $(\star)$ if g is continuous at x then f is differentiable at x and $f'(x) = g(x)$.

13. Product measures

- **Definition:** Consider the measurable spaces $(\Omega_i, \mathcal{F}_i)$, $i = 1, ..., n$, and let

$$\Omega = \Omega_1 \times \Omega_2 \times ... \times \Omega_n = \prod_{i=1}^{n} \Omega_i.$$

A measurable rectangle in the product space Ω is a set

$$A = A_1 \times A_2 \times ... \times A_n = \prod_{i=1}^{n} A_i,$$

where $A_i \in \mathcal{F}_i$, for $i = 1, ..., n$.

- **Definition:** The σ-algebra on Ω generated by the measurable rectangles, $\sigma\left(\left\{\prod_{i=1}^{n} A_i \middle| A_i \in \mathcal{F}_i\right\}\right)$, is called the product σ-algebra and is denoted by

$$\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes ... \otimes \mathcal{F}_n = \bigotimes_{i=1}^{n} \mathcal{F}_i.$$

- Note that the product σ-algebra is not the Cartesian product of the $\mathcal{F}_i$'s, $\mathcal{F}_1 \times \mathcal{F}_2 \times ... \times \mathcal{F}_n = \prod_{i=1}^{n} \mathcal{F}_i$. The Cartesian product of the $\mathcal{F}_i$'s is the set of measurable rectangles. Therefore, $\bigotimes_{i=1}^{n} \mathcal{F}_i = \sigma\left(\prod_{i=1}^{n} \mathcal{F}_i\right)$.

- The σ-algebra $\mathcal{B}(\mathbb{R}^n)$ of the Borel sets of $\mathbb{R}^n$ is the product σ-algebra generated by the measurable rectangles $A = A_1 \times A_2 \times ... \times A_n$, where A_i is a Borel set of $\mathbb{R}$ for $i = 1, ..., n$, that is, $\mathcal{B}(\mathbb{R}^n) = \bigotimes_{i=1}^{n} \mathcal{B}_i = \sigma\left(\prod_{i=1}^{n} \mathcal{B}_i\right)$.
- **Assumption for the next two theorems.** Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ be a measure space where μ_i is σ-finite for $i = 1, 2$ and let $\Omega = \Omega_1 \times \Omega_2$ and $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$.

Product Measure Theorem. There is a unique measure μ on $(\Omega, \mathcal{F})$ such that

$$\mu\left(A_1 \times A_2\right) = \mu_1(A_1) \cdot \mu_2(A_2), \quad \text{for all } A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2.$$

Furthermore, μ is σ-finite and, if μ_1 and μ_2 are probability measures, then μ is also a probability measure. The measure μ is called the product measure of μ_1 and μ_2 ($\mu = \mu_1 \times \mu_2$) and is given by

$$\mu(B) = \int_{\Omega_1} \mu_2\left(B\left(\omega_1\right)\right) d\mu_1\left(\omega_1\right) = \int_{\Omega_2} \mu_1\left(B\left(\omega_2\right)\right) d\mu_2\left(\omega_2\right),$$

$$\text{for all } B \in \mathcal{F},$$

where

$$B\left(\omega_1\right) = \left\{\omega_2 \in \Omega_2 \,|(\omega_1, \omega_2) \in B\right\}, \quad \text{(section of } \omega_1\text{)},$$

$$B\left(\omega_2\right) = \left\{\omega_1 \in \Omega_1 \,|(\omega_1, \omega_2) \in B\right\}, \quad \text{(section of } \omega_2\text{)}.$$

14. Fubini's theorem

- Recall that $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$, and $\mu = \mu_1 \times \mu_2$.
- **Fubini's Theorem.** If $f : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is a Borel measurable function such that $\int_\Omega f d\mu$ exists, then

$$\int_\Omega f(\omega_1, \omega_2)\, d\mu(\omega_1, \omega_2) = \int_{\Omega_1} \int_{\Omega_2} f(\omega_1, \omega_2)\, d\mu_2(\omega_2)\, d\mu_1(\omega_1)$$

$$= \int_{\Omega_2} \int_{\Omega_1} f(\omega_1, \omega_2)\, d\mu_1(\omega_1)\, d\mu_2(\omega_2).$$

- Observe that $\mu(B) = \int_\Omega \mathbb{I}_B d\mu$. Therefore, if we make $f = \mathbb{I}_B$ in Fubini's theorem, we obtain the Product Measure theorem since

$$\mu(B) = \int_\Omega \mathbb{I}_B(\omega_1, \omega_2)\, d\mu(\omega_1, \omega_2)$$

$$= \int_{\Omega_1} \int_{\Omega_2} \mathbb{I}_B(\omega_1, \omega_2)\, d\mu_2(\omega_2)\, d\mu_1(\omega_1)$$

$$= \int_{\Omega_1} \int_{\Omega_2} \mathbb{I}_{B(\omega_1)}(\omega_2)\, d\mu_2(\omega_2)\, d\mu_1(\omega_1) = \int_{\Omega_1} \mu_2(B(\omega_1))\, d\mu_1(\omega_1).$$

- Similarly,

$$\mu(B) = \int_\Omega \mathbb{I}_B(\omega_1, \omega_2)\, d\mu(\omega_1, \omega_2)$$

$$= \int_{\Omega_2} \int_{\Omega_1} \mathbb{I}_B(\omega_1, \omega_2)\, d\mu_1(\omega_1)\, d\mu_2(\omega_2)$$

$$= \int_{\Omega_2} \int_{\Omega_1} \mathbb{I}_{B(\omega_2)}(\omega_1)\, d\mu_1(\omega_1)\, d\mu_2(\omega_2) = \int_{\Omega_2} \mu_1(B(\omega_2))\, d\mu_2(\omega_2).$$

- Note also that

$$\int_B f(\omega_1,\omega_2)\, d\mu(\omega_1,\omega_2) = \int_\Omega f(\omega_1,\omega_2)\, \mathbb{I}_B(\omega_1,\omega_2) d\mu(\omega_1,\omega_2)$$

$$= \int_{\Omega_1}\int_{\Omega_2} f(\omega_1,\omega_2)\, \mathbb{I}_B(\omega_1,\omega_2) d\mu_2(\omega_2)\, d\mu_1(\omega_1)$$

$$= \int_{\Omega_1}\int_{\Omega_2} f(\omega_1,\omega_2)\, \mathbb{I}_{B(\omega_1)}(\omega_2) d\mu_2(\omega_2)\, d\mu_1(\omega_1)$$

or

$$\int_B f(\omega_1,\omega_2)\, d\mu(\omega_1,\omega_2) = \int_\Omega f(\omega_1,\omega_2)\, \mathbb{I}_B(\omega_1,\omega_2) d\mu(\omega_1,\omega_2)$$

$$= \int_{\Omega_2}\int_{\Omega_1} f(\omega_1,\omega_2)\, \mathbb{I}_B(\omega_1,\omega_2) d\mu_1(\omega_1)\, d\mu_2(\omega_2)$$

$$= \int_{\Omega_2}\int_{\Omega_1} f(\omega_1,\omega_2)\, \mathbb{I}_{B(\omega_2)}(\omega_1) d\mu_1(\omega_1)\, d\mu_2(\omega_2)\,.$$

- Both the Product Measure theorem and Fubini's theorem can be extended to n factors with $\Omega = \Omega_1 \times \Omega_2 \times ... \times \Omega_n \equiv \prod_{i=1}^{n} \Omega_i$,

  $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes ... \otimes \mathcal{F}_n \equiv \bigotimes_{i=1}^{n} \mathcal{F}i$, and

  $\mu = \mu_1 \times \mu_2 \times ... \times \mu_n \equiv \prod_{i=1}^{n} \mu_i$.

- The Lebesgue measure on $\left(\mathbb{R}^n, \mathcal{B}\left(\mathbb{R}^n\right)\right)$ is the product measure of n Lebesgue measures on $\left(\mathbb{R}, \mathcal{B}\right)$. Thus,

$$\int_{\mathbb{R}^2} f\left(x_1, x_2\right) d\left(x_1, x_2\right) = \int_{\mathbb{R}} \int_{\mathbb{R}} f\left(x_1, x_2\right) dx_2 dx_1$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} f\left(x_1, x_2\right) dx_1 dx_2.$$

- See the handout for an application of Fubini's theorem.

Riemann and Lebesgue Integrals

Let $[a,b]$ be a bounded closed interval of real numbers and let f be a bounded real-valued function defined on $[a,b]$, fixed throughout the discussion. If $P = \{(x_{i-1}, x_i]\}_{i=1}^n$ is a subinterval partition of $(a,b]$ with $x_{i-1} < x_i$ for $i = 1, ..., n$, $x_0 = a$, and $x_n = b$, we may construct the upper and lower sums of f relative to P (also called upper and lower Darboux sums) as follows.

Let

$$M_i = \sup\{f(y) \,|y \in (x_{i-1}, x_i]\}, \quad i = 1, ..., n,$$
$$m_i = \inf\{f(y) \,|y \in (x_{i-1}, x_i]\}, \quad i = 1, ..., n,$$

and define step functions α and β on $[a,b]$, called the upper and lower functions corresponding to P, by

$$\alpha(x) = M_i \quad \text{if} \quad x \in (x_{i-1}, x_i], \quad i = 1, ..., n,$$

$$\beta(x) = m_i \quad \text{if} \quad x \in (x_{i-1}, x_i], \quad i = 1, ..., n,$$

and both $\alpha(a)$ and $\beta(a)$ may be chosen arbitrarily. The upper and lower Darboux sums of f relative to P are given by

$$U(P) = \sum_{i=1}^{n} M_i (x_i - x_{i-1}),$$

$$L(P) = \sum_{i=1}^{n} m_i (x_i - x_{i-1}).$$

Consider the measure space $(\Omega, \mathcal{F}, \mu)$, where $\Omega = [a,b]$, $\mathcal{F} = \mathcal{B}([a,b])$, the Borel sets of $[a,b]$, and $\mu =$ Lebesgue measure. Since α and β are simple functions, we have that the Darboux sums are equal to the following Lebesgue integrals:

$$U(P) = \int_{[a,b]} \alpha dx,$$
$$L(P) = \int_{[a,b]} \beta dx.$$

Observe that the arbitrary values of $\alpha(a)$ and $\beta(a)$ do not affect the values of the Darboux sums.

Now let $\{P_k\}_{k=1}^{\infty}$ be a sequence of subinterval partitions of $(a,b]$ such that P_{k+1} is a refinement of P_k for each k (i.e., the intervals in P_{k+1} are subintervals of the intervals in P_k) and the length of the largest subinterval

of P_k approaches 0 as $k \to \infty$. If α_k and β_k are the upper and lower functions corresponding to P_k, then

$$\alpha_1 \geq \alpha_2 \geq ... \geq f \geq ... \geq \beta_2 \geq \beta_1.$$

Thus, α_k and β_k tend to limit functions α and β. If $|f|$ is bounded by B, then all $|\alpha_k|$ and $|\beta_k|$ are bounded by B as well, and the function that is constant at B is Lebesgue integrable on $[a, b]$ since

$$\mu\,[a, b] = b - a < \infty$$

and, hence,

$$\int_{[a,b]} B dx = B\,(b - a) < \infty.$$

By the dominated convergence theorem,

$$\lim_{k\to\infty} U(P_k) = \lim_{k\to\infty} \int_{[a,b]} \alpha_k dx = \int_{[a,b]} \alpha dx,$$

$$\lim_{k\to\infty} L(P_k) = \lim_{k\to\infty} \int_{[a,b]} \beta_k dx = \int_{[a,b]} \beta dx.$$

The integrals $\int_{[a,b]} \alpha dx$ and $\int_{[a,b]} \beta dx$ are called, respectively, the upper and lower Darboux integrals of f for the sequence $\{P_k\}_{k=1}^{\infty}$ of subinterval partitions.

Note that

$$f \text{ is continuous at } x \quad \text{if and only if} \quad \alpha(x) = f(x) = \beta(x),$$

as follows from a standard $\varepsilon - \delta$ argument.

If $\lim_{k\to\infty} U(P_k) = \lim_{k\to\infty} L(P_k) =$ a finite number r, independent of the particular sequence of subinterval partitions, f is said to be Riemann integrable on $[a, b]$, and the number $r \equiv \int_a^b f dx$ is said to be (the value of) the Riemann integral of f on $[a, b]$. The above argument shows that f is Riemann integrable iff

$$\int_{[a,b]} \alpha dx = \int_{[a,b]} \beta dx,$$

independent of the particular sequence of subinterval partitions. Moreover, if f is Riemann integrable, then

$$\int_a^b f dx = \int_{[a,b]} \alpha dx = \int_{[a,b]} \beta dx = \int_{[a,b]} f dx.$$

From the previous discussion we have the following:

Proposition 1. Let f be a bounded real-valued function on $[a,b]$.

(a) The function f is Riemann integrable on $[a,b]$ if and only if f is continuous almost everywhere on $[a,b]$ with respect to Lebesgue measure.

(b) If f is Riemann integrable on $[a,b]$, then f is integrable with respect to Lebesgue measure on $[a,b]$, and the two integrals are equal, $\int_{[a,b]} f dx = \int_a^b f dx$.

The previous proposition gives us in part (a) the precise criterion for Riemann integrability. From its part (b) we see that Lebesgue integration is more general than Riemann integration as Riemann integrability implies Lebesgue integrability (but the converse is not necessarily true).

Riemann-Stieltjes (or Stieltjes) and Lebesgue-Stieltjes Integrals

Let f be a bounded real-valued function defined on the bounded interval $[a,b]$ and F is a distribution function (i.e., an increasing and right-continuous function) on $\mathbb{R}$ that induces the corresponding Lebesgue-Stieltjes measure μ on the Borel sets of $(a,b]$. Define M_i, m_i, α, and β as before and take

$$U(P) = \sum_{i=1}^{n} M_i \left(F\left(x_i\right) - F\left(x_{i-1}\right)\right) = \int_{(a,b]} \alpha d\mu \equiv \int_{(a,b]} \alpha dF,$$

$$L(P) = \sum_{i=1}^{n} m_i \left(F\left(x_i\right) - F\left(x_{i-1}\right)\right) = \int_{(a,b]} \beta d\mu \equiv \int_{(a,b]} \beta dF.$$

Let $\{P_k\}_{k=1}^{\infty}$ be a sequence of subinterval partitions of $(a,b]$ such that the length of the largest subinterval of P_k approaches 0 as $k \to \infty$ and P_{k+1} is a refinement of P_k for each k, and α_k and β_k be the upper and lower functions corresponding to P_k. Then,

$$\lim_{k\to\infty} U(P_k) = \int_{(a,b]} \alpha dF,$$

$$\lim_{k\to\infty} L(P_k) = \int_{(a,b]} \beta dF,$$

where $\alpha = \lim_{k\to\infty} \alpha_k$ and $\beta = \lim_{k\to\infty} \beta_k$. If $U(P_k)$ and $L(P_k)$ approach the same limit s independent of the particular sequence of subinterval partitions, this number $s \equiv \int_a^b f dF$ is called the Riemann-Stieltjes integral of f with respect to F on $(a,b]$, and f is said to be Riemann-Stieltjes integrable with respect to F on $(a,b]$.

Proposition 2. Let f be a bounded real-valued function on $[a,b]$ and F be a distribution function on $\mathbb{R}$.

(a) The function f is Riemann-Stieltjes integrable with respect to F on $(a,b]$ if and only if f is continuous almost everywhere on $(a,b]$ with respect to the Lebesgue-Stieltjes measure μ on $(a,b]$ induced by F.

(b) If f is Riemann-Stieltjes integrable with respect to F on $(a,b]$, then f is integrable with respect to the Lebesgue-Stieltjes measure μ on $(a,b]$ induced by F, and the two integrals are equal, $\int_{(a,b]} f dF = \int_a^b f dF$.

An application of Fubini's theorem

Let $\Omega_i = \mathbb{R}$, $i = 1, 2$, and $\Omega = \Omega_1 \times \Omega_2$. Therefore, $\Omega = \mathbb{R}^2$.
Let $\mu_i =$ Lebesgue measure on $(\mathbb{R}, \mathcal{B})$, $i = 1, 2$, where $\mathcal{B}$ is the $\sigma-$algebra of the Borel sets in $\mathbb{R}$.
Let $\mathcal{B} \otimes \mathcal{B} = \mathcal{B}(\mathbb{R}^2)$.
Let $\mu = \mu_1 \times \mu_2$. Therefore, μ is the Lebesgue measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$.
Define the Borel measurable function $f : (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \longrightarrow (\mathbb{R}, \mathcal{B})$,

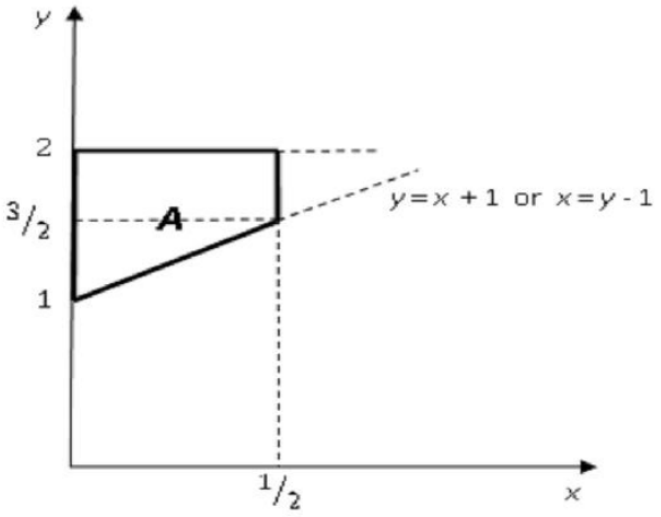
$$f(x, y) = \frac{3}{5}x(y + x).$$

We want to compute the following Lebesgue integral:

$$\int_A f d\mu = \int_A f(x, y) d(x, y),$$

where the Borel set A is given by

$$A = \left\{(x, y) \in \mathbb{R}^2 \mid 0 \leq x < 1/2\,, x + 1 < y \leq 2\right\}.$$



Then,

$$\int_A f(x, y) d(x, y) = \int_{\mathbb{R}^2} f(x, y) \mathbb{I}_A(x, y) d(x, y)$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) \mathbb{I}_A(x, y) dy dx = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) \mathbb{I}_{A(x)}(y) dy dx,$$

where the second equality follows from Fubini's theorem and

$$A(x) = \begin{cases} \{y \in \mathbb{R} \mid y \in (x + 1, 2]\} & \text{for } x \in [0, 1/2) \\ \\ \emptyset & \text{for } x \notin [0, 1/2)\,. \end{cases}$$

Hence,

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)\mathbb{I}_{A(x)}(y)dydx = \int_{[0,1/2)}\int_{(x+1,2]} f(x,y)dydx + \underbrace{\int_{[0,1/2)^c}\int_{\{\emptyset\}} f(x,y)dydx}_{=0}$$

$$= \int_0^{1/2}\int_{x+1}^{2} \frac{3}{5}x(y+x)dydx = \frac{63}{640},$$

where the second equality follows from the continuity of the function f, which allows us to use Riemann integration to compute the Lebesgue integral.

Alternatively, we could integrate first with respect to the variable x and then with respect to the variable y according to Fubini's theorem. Thus,

$$\int_A f(x,y)d(x,y) = \int_{\mathbb{R}^2} f(x,y)\mathbb{I}_A(x,y)d(x,y)$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)\mathbb{I}_A(x,y)dxdy = \int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)\mathbb{I}_{A(y)}(x)dxdy,$$

where

$$A(y) = \begin{cases} \{x \in \mathbb{R} \mid x \in [0, y-1)\} & \text{for } y \in (1, 3/2] \\ \{x \in \mathbb{R} \mid x \in [0, 1/2)\} & \text{for } y \in (3/2, 2] \\ \emptyset & \text{for } y \notin (1, 2]\,. \end{cases}$$

Hence,

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)\mathbb{I}_{A(y)}(x)dxdy =$$

$$\int_{(1,3/2]}\int_{[0,y-1)} f(x,y)dxdy + \int_{(3/2,2]}\int_{[0,1/2)} f(x,y)dxdy + \underbrace{\int_{(1,2]^c}\int_{\{\emptyset\}} f(x,y)dxdy}_{=0}$$

$$= \int_1^{3/2}\int_0^{y-1} \frac{3}{5}x(y+x)dxdy + \int_{3/2}^{2}\int_0^{1/2} \frac{3}{5}x(y+x)dxdy = \frac{63}{640}.$$

Exercises. Probability and Statistics. IDEA.
Measure Theory

1. Let Ω be a countably infinite set, and let $\mathcal{F}$ consist of all subsets of Ω $\left(\mathcal{F} = 2^{\Omega}\right)$. Define $\mu(A) = 0$ if A is finite, $\mu(A) = \infty$ if A is infinite.

   (a) Show that μ is finitely additive but not countably additive.

   (b) Show that Ω is the limit of an increasing sequence of sets A_n with $\mu(A_n) = 0$ for all n, but $\mu(\Omega) = \infty$.

2. Let μ be a Lebesgue-Stieltjes measure on $\mathbb{R}$ corresponding to a continuous distribution function.

   (a) If A is a countable subset of $\mathbb{R}$, show that $\mu(A) = 0$.

   (b) If $\mu(A) > 0$, must A include an open interval?

   (c) If $\mu(A) > 0$ and $\mu(A^c) = 0$, must A be dense in $\mathbb{R}$? *Note*: A set is dense in $\mathbb{R}$ if its closure is $\mathbb{R}$.

   (d) Do the answers to (b) and (c) change if μ is restricted to be Lebesgue measure?

3. The following statement is true:

   "A distribution function on $\mathbb{R}$ is monotone and thus has only countably many points of discontinuity."

   Is this also true for a distribution function on $\mathbb{R}^n$, $n > 1$?

4. If f_1, f_2,... are extended real-valued Borel measurable functions on $(\Omega, \mathcal{F})$, $n = 1, 2, ...$, show that $\sup_n f_n$ and $\inf_n f_n$ are Borel measurable.

5. Let $f = f(x, y)$ be a real-valued function of two real variables, defined for $a < y < b$, $c < x < d$. Assume that for each x, $f(x, \cdot)$ is a Borel measurable function of y, and that there is a Borel measurable function $g : (a, b) \to \mathbb{R}$ such that $|f(x, y)| \leq g(y)$ for all x, y, and $\int_{(a,b)} g(y)dy < \infty$. If $x_0 \in (c, d)$ and $\lim_{x \to x_0} f(x, y)$ exists for all $y \in (a, b)$, show that

$$\lim_{x \to x_0} \int_{(a,b)} f(x, y)dy = \int_{(a,b)} \left[\lim_{x \to x_0} f(x, y)\right] dy.$$

6. The function defined on $[0, 1]$ by $f(x) = 1$ if x is irrational, and $f(x) = 0$ if x is rational, is the standard example of a function that is Lebesgue integrable (it is 1 a.e.) but not Riemann integrable. But what is wrong with the following reasoning:

   "If we consider the behavior of f on the irrationals, f assumes the constant value 1 and is therefore continuous. Since the rationals have Lebesgue measure 0, f is therefore continuous almost everywhere and hence is Riemann integrable."

7. Give an example of a sequence of functions f_n on $[a,b]$ such that each f_n is Riemann integrable, $|f_n| \leq 1$ for all n, $f_n \to f$ everywhere, but f is not Riemann integrable.

8. Prove the Borel-Cantinelli Lemma:

   If $A_1, A_2, ... \in \mathcal{F}$ and $\sum_{n=1}^{\infty} \mu(A_n) < \infty$, then $\mu\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) = 0$.

   Note that $\omega \in \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$ iff $\omega \in A_n$ for infinitely many n.

9. Give an example of a measure μ and a nonnegative finite-valued Borel measurable function g such that the measure λ defined by $\lambda(A) = \int_A g \, d\mu$ is not σ-finite.

10. Let $f, f_1, f_2, ...$ be Borel measurable functions on $(\Omega, \mathcal{F}, \mu)$. We say that f_n converges to f in measure (or in μ-measure) if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mu\{\omega \mid |f_n(\omega) - f(\omega)| \geq \varepsilon\} = 0.$$

    If f_n converges both to f and g in measure, show that $f = g$ a.e. with respect to μ.

11. Let $f_n : (\mathbb{R}, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$, $n = 1, 2, ...$, be

$$f_n(x) = \begin{cases} 1 & \text{for } x \in [0,1] \\ \\ 0 & \text{otherwise} \end{cases}$$

    if n is odd and

$$f_n(x) = \begin{cases} 1 & \text{for } x \in (1,2] \\ \\ 0 & \text{otherwise} \end{cases}$$

    if n is even.

    Does the following equality hold:

$$\liminf_{n\to\infty} \int_{\mathbb{R}} f_n dx = \int_{\mathbb{R}} \liminf_{n\to\infty} f_n dx \ ?$$

12. Let $f_n : (\mathbb{R}, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$, $n = 1, 2, ...$, be

$$f_n(x) = \begin{cases} n & \text{for } x \in (0, 1/n] \\ \\ 0 & \text{otherwise} \end{cases}$$

    Does the following equality hold:

$$\lim_{n\to\infty} \int_{\mathbb{R}} f_n dx = \int_{\mathbb{R}} \lim_{n\to\infty} f_n dx \ ?$$

13. Let $\Omega_1 = \Omega_2 =$ the set of positive integers, $\mathcal{F}_1 = \mathcal{F}_2 =$ all subsets, $\mu_1 = \mu_2 =$ counting measure, $f(n,n) = n$, $f(n,n+1) = -n$, $n = 1, 2, ...,$ $f(i,j) = 0$ if $j \neq i$ or $j \neq i+1$. Show that $\int_{\Omega_1} \int_{\Omega_2} f \, d\mu_2 \, d\mu_1 = 0$ and $\int_{\Omega_2} \int_{\Omega_1} f \, d\mu_1 \, d\mu_2 = \infty$. *Note*: Fubini's theorem fails since the integral of f with respect to $\mu_1 \times \mu_2$ does not exist.

14. Let $\Omega_1 = \Omega_2 = \mathbb{R}$, $\mathcal{F}_1 = \mathcal{F}_2 = \mathcal{B}$, $\mu_1 =$ Lebesgue measure, $\mu_2 =$ counting measure. Let $A = \{(\omega_1, \omega_2) \,|\omega_1 = \omega_2\} \in \mathcal{F}_1 \otimes \mathcal{F}_2$. Show that

$$\int_{\Omega_1} \int_{\Omega_2} \mathbb{I}_A \, d\mu_2 \, d\mu_1 = \int_{\Omega_1} \mu_2 \left(A(\omega_1)\right) d\mu_1(\omega_1) = \infty,$$

but

$$\int_{\Omega_2} \int_{\Omega_1} \mathbb{I}_A \, d\mu_1 \, d\mu_2 = \int_{\Omega_2} \mu_1 \left(A(\omega_2)\right) d\mu_2(\omega_2) = 0.$$

Recall that $A(\omega_1) = \{\omega_2 \in \Omega_2 \,|\, (\omega_1, \omega_2) \in A\}$ and, symmetrically, $A(\omega_2) = \{\omega_1 \in \Omega_1 \,|\, (\omega_1, \omega_2) \in A\}$. *Note*: Fubini's theorem fails since μ_2 is not $\sigma-$finite.

2. Random Variables and Distributions

2.1. Random objects and random variables

- **Definition.** A random object is a measurable function $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\Omega', \mathcal{F}')$, where Ω is a sample space and $\mathcal{F}$ is the σ-algebra of events.
- **Definition.** A (real-valued) random variable is a measurable function $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$, where Ω is a sample space and $\mathcal{F}$ is the σ-algebra of events.

- Similarly, when Ω is a sample space and $\mathcal{F}$ is the σ-algebra of events,
  - $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ is an "extended (real-valued)" random variable.
  - $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ is a "(real-valued) random vector" or a "(real-valued) multivariate random variable".
  - $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow \left(\overline{\mathbb{R}}^n, \mathcal{B}\right)$ is an "extended (real-valued) random vector" or a "extended (real-valued) multivariate random variable".
- A random vector is just a vector of random variables:

$$\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) \quad \text{or} \quad \tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)^\top$$

2.2. Probability distributions

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.
- **Definition.** The probability distribution (or distribution) of a random object $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ is a probability measure $P_{\tilde{x}}$ on $(\Omega', \mathcal{F}')$ defined by

$$P_{\tilde{x}}(B) = P\left(\tilde{x}^{-1}(B)\right) \quad \text{for all } B \in \mathcal{F}'$$

or

$$P_{\tilde{x}}(B) = P\left\{\omega \in \Omega \mid \tilde{x}(\omega) \in B\right\} = P\left\{\tilde{x} \in B\right\} \quad \text{for all } B \in \mathcal{F}'.$$

- Obviously,

$$P_{\tilde{x}}(B) = \int_B 1 dP_{\tilde{x}} \equiv \int_B dP_{\tilde{x}} \equiv \int_{\Omega'} \mathbb{I}_B(x) dP_{\tilde{x}}(x) \quad \text{for all } B \in \mathcal{F}'$$

or

$$P_{\tilde{x}}(B) = \int\limits_{\tilde{x}^{-1}(B)} 1 dP \equiv \int\limits_{\tilde{x}^{-1}(B)} dP \equiv \int_{\Omega} \mathbb{I}_{\tilde{x}^{-1}(B)}(\omega) dP(\omega) \text{ for all } B \in \mathcal{F}'.$$

- **Example:** We roll a balanced dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and consider the random variable $\tilde{x} : (\Omega, 2^{\Omega}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ defined as

$$\tilde{x}(\omega) = \begin{cases} 1 & \text{if } \omega = 1, 2, 3, 4 \\ \\ 7 & \text{if } \omega = 5, 6. \end{cases}$$

- The induced probability $P_{\tilde{x}}$ on $(\mathbb{R}, \mathcal{B})$ (or distribution of $\tilde{x}$) satisfies

$$\begin{array}{ll} P_{\tilde{x}}\{1\} = P\{1, 2, 3, 4\} = 2/3, & P_{\tilde{x}}\{7\} = P\{5, 6\} = 1/3, \\ P_{\tilde{x}}\{12\} = P(\emptyset) = 0, & P_{\tilde{x}}(-3, 1) = P(\emptyset) = 0, \\ P_{\tilde{x}}[-3, 1] = P\{1, 2, 3, 4\} = 2/3, & P_{\tilde{x}}[5, 8] = P\{7\} = 1/3, \\ P_{\tilde{x}}\left[\pi, \sqrt{13}\right] = P(\emptyset) = 0, & P_{\tilde{x}}(-\infty, 12] = P(\Omega) = 1, \\ P_{\tilde{x}}[10, \infty) = P(\emptyset) = 0, & P_{\tilde{x}}(1, \infty) = P\{5, 6\} = 1/3, \\ P_{\tilde{x}}(-\infty, 2] = P\{1, 2, 3, 4\} = 2/3, & \text{etc.} \end{array}$$

- Moreover, using the properties of the probability, we obtain the distribution for all Borel sets in $\mathbb{R}$.

- **Definition.** The support $supp\left(P_{\tilde{x}}\right)$ of the distribution of the random vector $\tilde{x}:\left(\Omega,\mathcal{F}\right)\longrightarrow\left(\mathbb{R}^{n},\mathcal{B}\right)$ is the smallest closed subset of $\mathbb{R}^{n}$ whose complement has zero probability distribution, $P_{\tilde{x}}\left\{\left[supp\left(P_{\tilde{x}}\right)\right]^{c}\right\}=0$.

- **Definition.** Two random objects $\tilde{x}$ and $\tilde{y}$ defined on $\left(\Omega,\mathcal{F},P\right)$ and taking values on $\left(\Omega',\mathcal{F}'\right)$ are equivalent (or equal) in distribution $\left(\tilde{x}\overset{d}{=}\tilde{y}\right)$ if they have the same distribution, $P_{\tilde{x}}=P_{\tilde{y}}$.

- **Example:** We toss a balanced coin and consider $\tilde{x}:\left(\Omega,2^{\Omega},P\right)\longrightarrow\left(\mathbb{R},\mathcal{B}\right)$, and $\tilde{y}:\left(\Omega,2^{\Omega},P\right)\longrightarrow\left(\mathbb{R},\mathcal{B}\right)$ defined as

$$\tilde{x}(\omega)=\begin{cases}-1 & \text{if } \omega=H\\ \\ 1 & \text{if } \omega=T\end{cases}\qquad\text{and}\qquad\tilde{y}(\omega)=\begin{cases}-1 & \text{if } \omega=T\\ \\ 1 & \text{if } \omega=H.\end{cases}$$

Thus, $\tilde{x}\overset{d}{=}\tilde{y}$.

- An event A is sure if $A = \Omega$.
- An event A is almost sure (a.s.) if $P(A) = 1$.
- An event A is negligible if $P(A) = 0$.

- **Definition.** We say that two random objects defined on $(\Omega, \mathcal{F}, P)$ and taking values on $(\Omega', \mathcal{F}')$ are equal, $\tilde{x} = \tilde{y}$, if $\tilde{x}(\omega) = \tilde{y}(\omega)$ for all $\omega \in \Omega$.
- **Definition.** We say that two random objects defined on $(\Omega, \mathcal{F}, P)$ and taking values on $(\Omega', \mathcal{F}')$ are equal almost surely (a.s.), $\tilde{x} \overset{a.s.}{=} \tilde{y}$, if

$$P\{\tilde{x} = \tilde{y}\} = P\{\omega \in \Omega \mid \tilde{x}(\omega) = \tilde{y}(\omega)\} = 1 ,$$

  or, equivalently, if

$$P\{\tilde{x} \neq \tilde{y}\} = P\{\omega \in \Omega \mid \tilde{x}(\omega) \neq \tilde{y}(\omega)\} = 0.$$

- Note that the concept of "a.s." is the same as that of "a.e." The only difference is that "a.e." applies to functions defined on measure spaces, whereas "a.s." applies to random objects defined on probability spaces.
- Obviously, $\tilde{x} = \tilde{y} \Longrightarrow \tilde{x} \overset{a.s.}{=} \tilde{y}$. Moreover, $\tilde{x} \overset{a.s.}{=} \tilde{y} \Longrightarrow \tilde{x} \overset{d}{=} \tilde{y}$ but the converse is not true (see the previous example of tossing a balanced coin where $\tilde{x} \overset{d}{=} \tilde{y}$ but $\tilde{x} \overset{a.s.}{\neq} \tilde{y}$ since $P\{\tilde{x} \neq \tilde{y}\} = 1$).

2.3. Distribution function and quantile function of a random variable

- Note that the distribution $P_{\tilde{x}}$ of a random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a probability measure on $(\mathbb{R}, \mathcal{B})$ and, thus, is a finite measure.
- Therefore, the distribution $P_{\tilde{x}}$ of a random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a Lebesgue-Stieltjes measure on $\mathbb{R}$ satisfying $P_{\tilde{x}}(\mathbb{R}) = 1$.

- **Definition.** The (cumulative) distribution function (cdf) $F_{\tilde{x}} : \mathbb{R} \longrightarrow \mathbb{R}$ of a random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ (or of its distribution) is the distribution function associated with the distribution $P_{\tilde{x}}$, i.e.,

$$P_{\tilde{x}}(a, b] = P\{a < \tilde{x} \leq b\} = F_{\tilde{x}}(b) - F_{\tilde{x}}(a),$$

where we make the normalization $\lim_{x \to -\infty} F_{\tilde{x}}(x) = 0$.

- Therefore,

$$P_{\tilde{x}}(-\infty, x] = P\{\tilde{x} \leq x\} = F_{\tilde{x}}(x) - \lim_{x \to -\infty} F_{\tilde{x}}(x) = F_{\tilde{x}}(x).$$

- Moreover,

$$\lim_{x \to \infty} F_{\tilde{x}}(x) = P_{\tilde{x}}(-\infty, \infty) = P\{\tilde{x} \in \mathbb{R}\} = 1.$$

- Thus, the distribution function of a random variable $\tilde{x}$ is (weakly) increasing, right-continuous, and satisfies $\lim_{x \to -\infty} F_{\tilde{x}}(x) = 0$ and $\lim_{x \to \infty} F_{\tilde{x}}(x) = 1$. Therefore, we can restrict the range of $F_{\tilde{x}}$ so that $F_{\tilde{x}} : \mathbb{R} \longrightarrow [0, 1]$ without loss of generality.

- Assume that the distribution function $F_{\tilde{x}}$ of the random variable $\tilde{x}$ is strictly increasing and continuous.

- **Definition.** The quantile function (or percentile function, percent-point function, inverse cdf or inverse distribution function) $Q_{\tilde{x}} : (0,1) \longrightarrow \mathbb{R}$ of the random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ (or of its distribution function) maps the probability value $p \in (0,1)$ to the value $x \in \mathbb{R}$ such that

$$F_{\tilde{x}}(x) = P\left\{\tilde{x} \leq x\right\} = p.$$

- Thus, the quantile function $Q_{\tilde{x}}$ assigns to the probability p a threshold value x so that the probability of $\tilde{x}$ being less or equal than x is p.

- Then,

$$Q_{\tilde{x}}(p) = F_{\tilde{x}}^{-1}(p) \text{ for all } p \in (0,1),$$

so that the quantile function $Q_{\tilde{x}}$ is strictly increasing and continuous.

- In the general case of cdf's that are increasing (but not necessarily strictly increasing) and right-continuous (but not necessarily continuous), the quantile function is usually defined as

$$Q_{\tilde{x}}(p) = \inf \{x \,|\, F_{\tilde{x}}(x) \geq p\} \quad \text{for all } p \in (0,1)\,. \tag{1}$$

- Note that the quantile function is thus in general (weakly) increasing and left-continuous and satisfies the following:

$$Q_{\tilde{x}}(p) \leq x \;\; \text{if and only if} \;\; F_{\tilde{x}}(x) \geq p.$$

2.4. Discrete random variables

- **Definition.** A random variable $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$ is discrete if its range (i.e., the image of Ω) is countable or discrete (either finite or infinite). That is, a discrete random variable may take on only a countable number of distinct values.
- If Ω is discrete then $\tilde{x}$ is discrete. The converse is not true.
- Let $\{x_1, x_2, ...\}$ be the range $\tilde{x}(\Omega)$ of the discrete random variable $\tilde{x}$.
- If $\tilde{x}$ is discrete there is a countable partition $\mathcal{A} = \{A_1, A_2, ...\}$ of Ω with
$$A_i = \{\omega \in \Omega \,|\, \tilde{x}(\omega) = x_i\}, \quad \text{for all } x_i \in \tilde{x}(\Omega).$$
- Therefore, $A_i = \tilde{x}^{-1}(x_i)$, for all $x_i \in \tilde{x}(\Omega)$.
- The σ-algebra $\sigma(\mathcal{A})$ generated by the partition $\mathcal{A}$ is the smallest σ-algebra that makes the random variable $\tilde{x}$ measurable.

- The distribution of a discrete random variable $\tilde{x}$ (which is said to have a discrete distribution) satisfies:

$$P_{\tilde{x}}\left\{x_i\right\} = P\left\{\tilde{x} = x_i\right\} = P(A_i), \text{ for all } x_i \in \tilde{x}(\Omega).$$

- **Definition.** The probability mass function (pmf) - or just probability function -, $f_{\tilde{x}} : \tilde{x}(\Omega) \longrightarrow [0,1]$, of a discrete random variable $\tilde{x}$ (or of a discrete distribution $P_{\tilde{x}}$) is given by

$$f_{\tilde{x}}(x) = P\left\{\tilde{x} = x\right\} = P_{\tilde{x}}\left\{x\right\}, \quad \text{for all } x \in \tilde{x}(\Omega).$$

- Note that $f_{\tilde{x}}(x) > 0$ if and only if $x \in supp\left(P_{\tilde{x}}\right) \subset \tilde{x}(\Omega)$.

- **Properties of the probability and distribution functions of a discrete random variable:**
- **1.**

$$\sum_{x \in \tilde{x}(\Omega)} f_{\tilde{x}}(x) = 1.$$

- **2.** Any function $f : \tilde{x}(\Omega) \longrightarrow [0, 1]$, where $\tilde{x}(\Omega)$ is countable, satisfying $\sum_{x \in \tilde{x}(\Omega)} f(x) = 1$ can serve as a probability function of a discrete distribution.
- **3.**

$$F_{\tilde{x}}(x) = \sum_{t \leq x} f_{\tilde{x}}(t), \quad \text{with } t \in \tilde{x}(\Omega).$$

- **4.**

$$P_{\tilde{x}}(B) = P\{\tilde{x} \in B\} = \sum_{x \in B} f_{\tilde{x}}(x), \quad \text{for all } B \in \mathcal{B}, \quad \text{with } x \in \tilde{x}(\Omega).$$

- **5.**

$$f_{\tilde{x}}(x) = F_{\tilde{x}}(x) - \lim_{t \to x^-} F_{\tilde{x}}(t), \text{ for } x \in \tilde{x}(\Omega).$$

In particular, if the range of $\tilde{x}$ can be ordered so that $x_1 < x_2 < ... < x_{i-1} < x_i < x_{i+1} < ...$, then $f_{\tilde{x}}(x_1) = F_{\tilde{x}}(x_1)$ and $f_{\tilde{x}}(x_i) = F_{\tilde{x}}(x_i) - F_{\tilde{x}}(x_{i-1})$ for $i = 2, 3, ...$

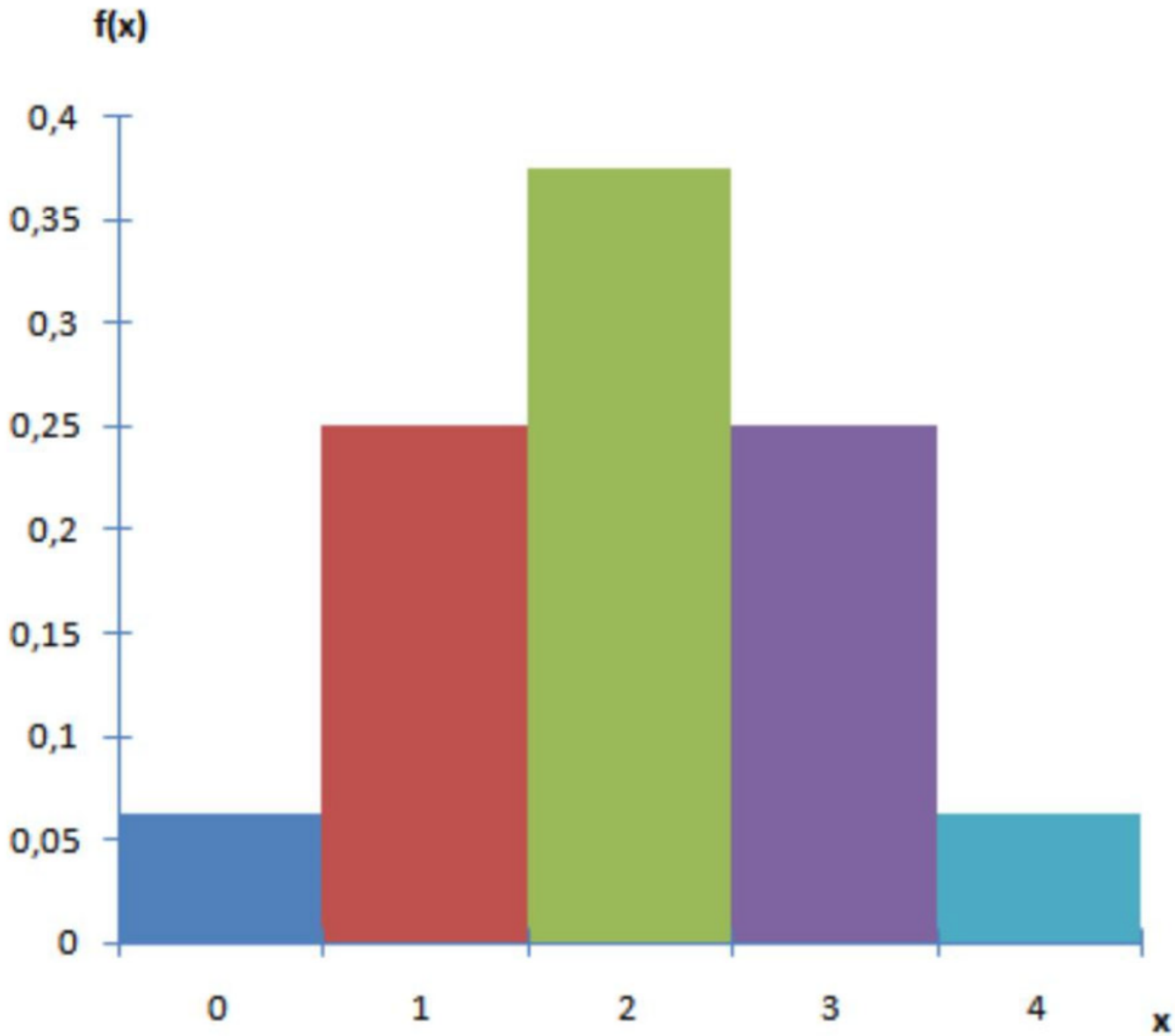
- **Example:** Let $\tilde{x}$ be the number of heads when tossing 4 coins.

$$f_{\tilde{x}}(x) = \begin{cases} 1/16 & \text{for } x = 0 \\ 4/16 & \text{for } x = 1 \\ 6/16 & \text{for } x = 2 \\ 4/16 & \text{for } x = 3 \\ 1/16 & \text{for } x = 4, \end{cases}$$

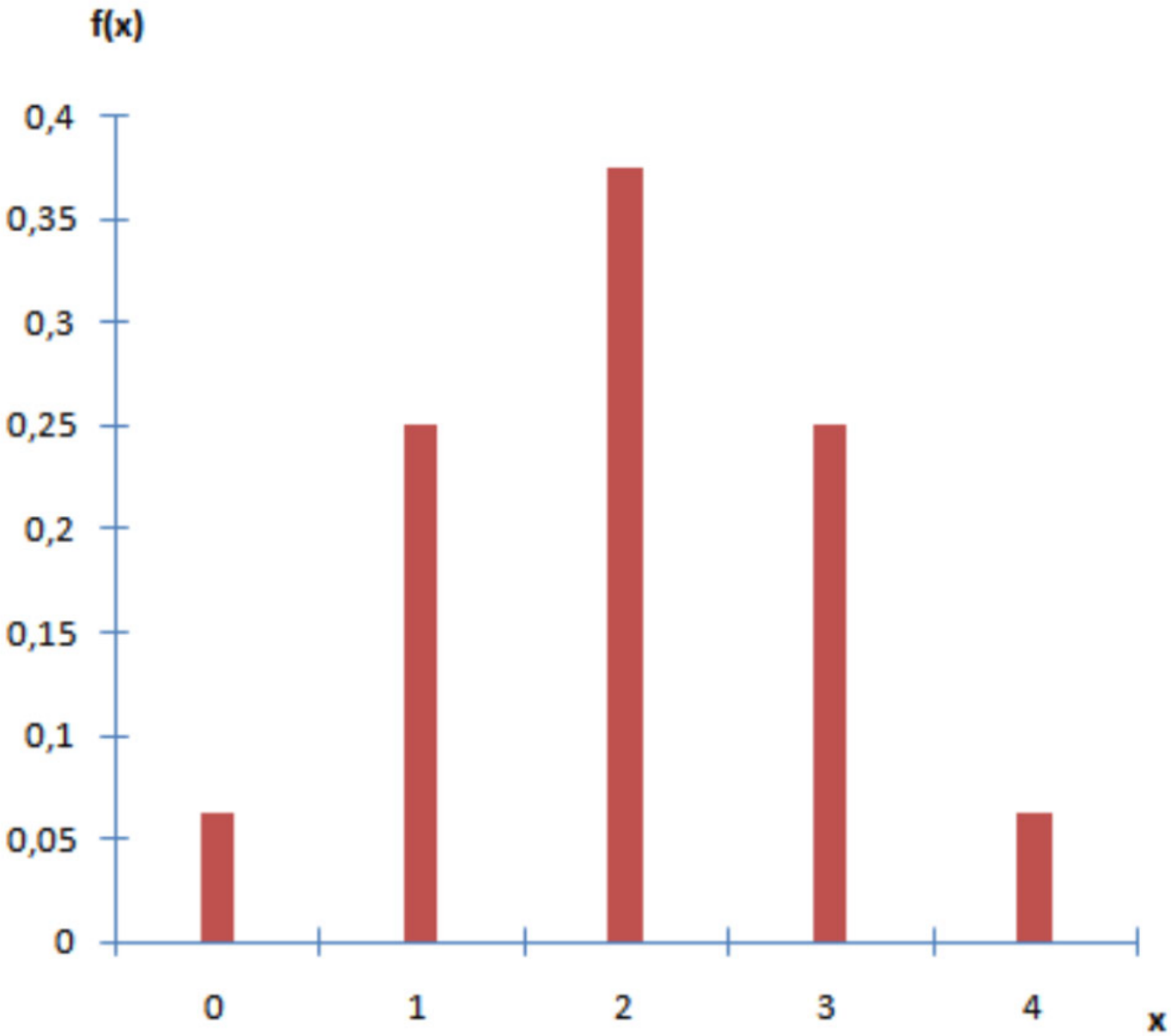
or

$$f_{\tilde{x}}(x) = \frac{1}{16}\binom{4}{x}, \quad \text{for } x = \underbrace{0, 1, 2, 3, 4}_{\tilde{x}(\Omega)}.$$

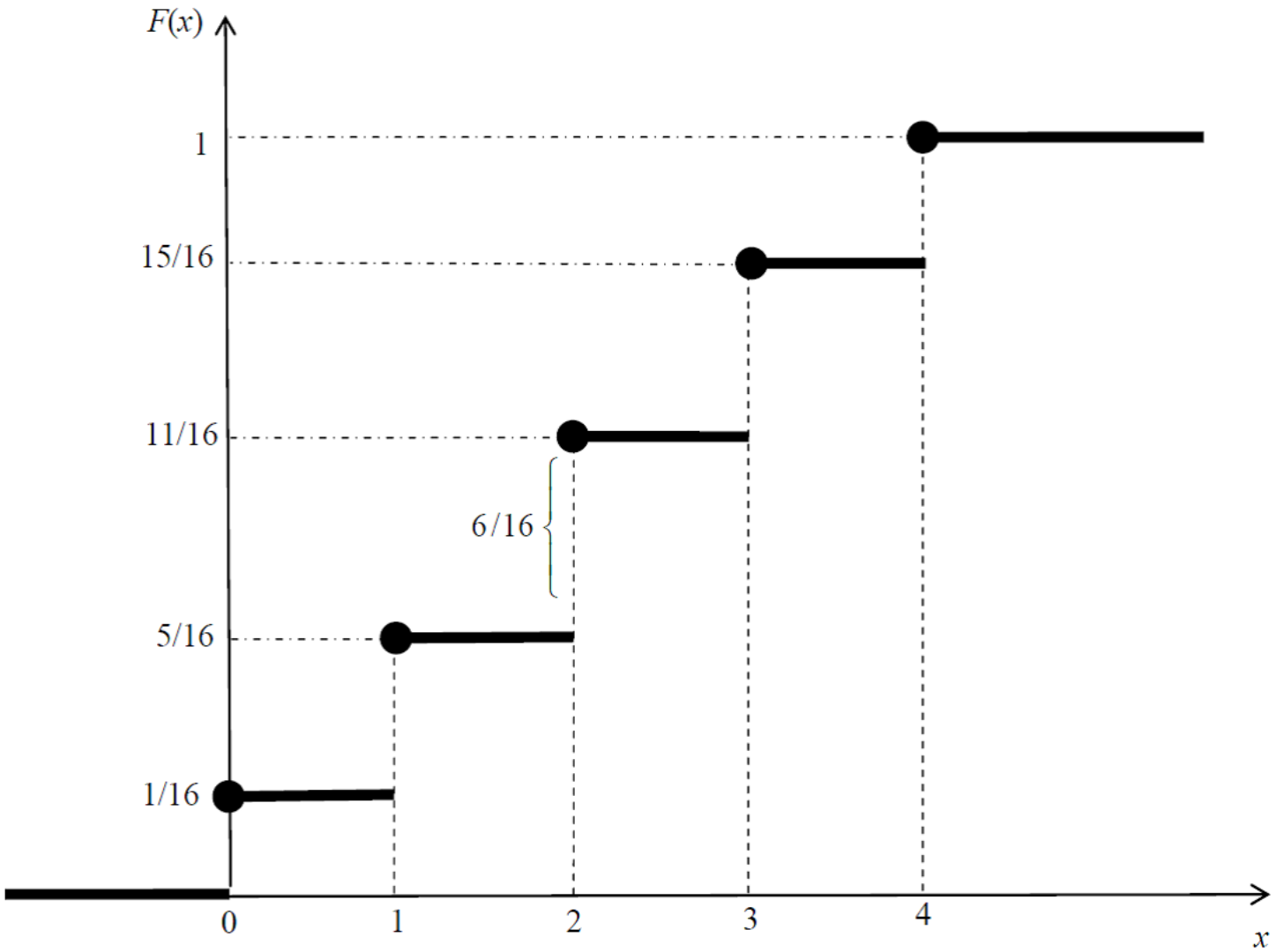
- Probability Histogram:



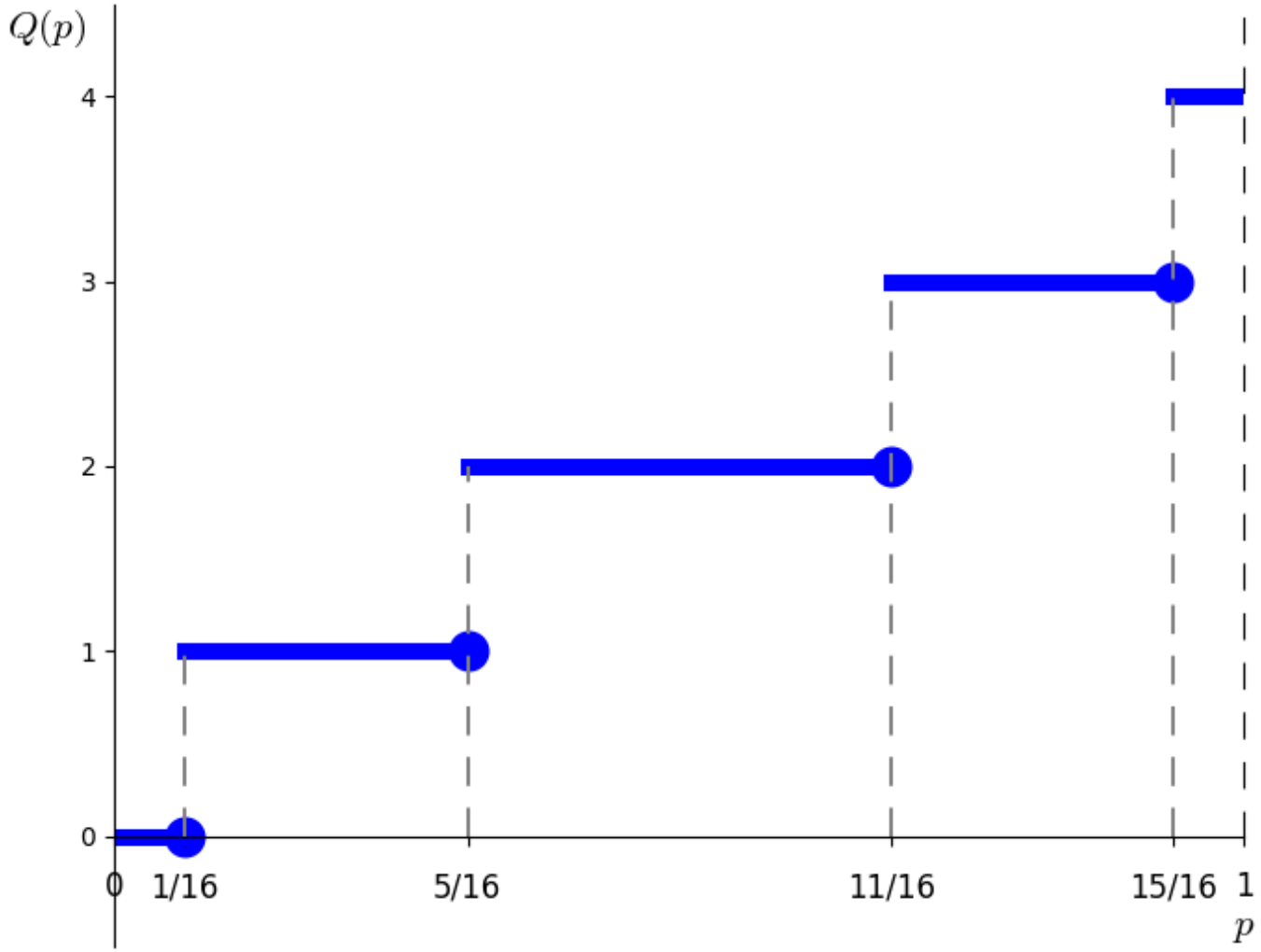
- Probability Bar Chart:



- Distribution function:



- The quantile function ($Q_{\tilde{x}}(p) = \inf\{x \,|F_{\tilde{x}}(x) \geq p\}$ for $p \in (0,1)$) of the previous distribution is weakly increasing and left-continuous:



2.5. Continuous and absolutely continuous random variables

- **Definition 1.** A random variable $\tilde{x}$ is continuous if its range $\tilde{x}(\Omega)$ is continuous.
- **Definition 2.** A random variable $\tilde{x}$ is continuous if its distribution function $F_{\tilde{x}}$ is continuous, that is, if $P_{\tilde{x}}\{x\} = P\{\tilde{x} = x\} = 0$ for all $x \in \mathbb{R}$.
- Continuity according to Definition 2 implies continuity according to Definition 1.
- **Definition.** A random variable $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$ is absolutely continuous if its distribution function $F_{\tilde{x}}$ is absolutely continuous, i.e., if there exists a Borel measurable function $f_{\tilde{x}} : (\mathbb{R}, \mathcal{B}) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$ that is integrable with respect to Lebesgue measure such that

$$F_{\tilde{x}}(x) - F_{\tilde{x}}(a) = \int_{[a,x]} f_{\tilde{x}}(t)dt, \text{ for all } a \in \mathbb{R},\ x \in \mathbb{R}, \text{ with } a \leq x.$$

- Absolute continuity implies continuity.
- Random variables that are neither discrete nor absolutely continuous are called "mixed".
- **Equivalent definition:** A random variable $\tilde{x}$ is absolutely continuous if its distribution $P_{\tilde{x}}$ is absolutely continuous with respect to Lebesgue measure.
- Therefore, thanks to the Radon-Nikodym theorem, there exists a Borel measurable function $f_{\tilde{x}} : (\mathbb{R}, \mathcal{B}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ such that

$$P_{\tilde{x}}(B) = \int_B f_{\tilde{x}}(x)dx, \text{ for all } B \in \mathcal{B}.$$

2.6. Density

- The Borel measurable function $f_{\tilde{x}} : (\mathbb{R}, \mathcal{B}) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ such that

$$P_{\tilde{x}}(B) = \int_B f_{\tilde{x}}(x)dx, \text{ for all } B \in \mathcal{B},$$

  is called the probability density function (pdf) - or density function or just "density" - of the random variable $\tilde{x}$ (or of the distribution $P_{\tilde{x}}$).

- Since $P_{\tilde{x}}(\mathbb{R}) = 1$, the density function $f_{\tilde{x}}$ is integrable with respect to Lebesgue measure on $(\mathbb{R}, \mathcal{B})$.

- Moreover, the density $f_{\tilde{x}}$ is finite a.e. with respect to Lebesgue measure on $(\mathbb{R}, \mathcal{B})$.

- The density function $f_{\tilde{x}}$ of the random variable $\tilde{x}$ is the Radon-Nikodym derivative of its distribution with respect to Lebesgue measure, $f_{\tilde{x}} = dP_{\tilde{x}} / dx$.

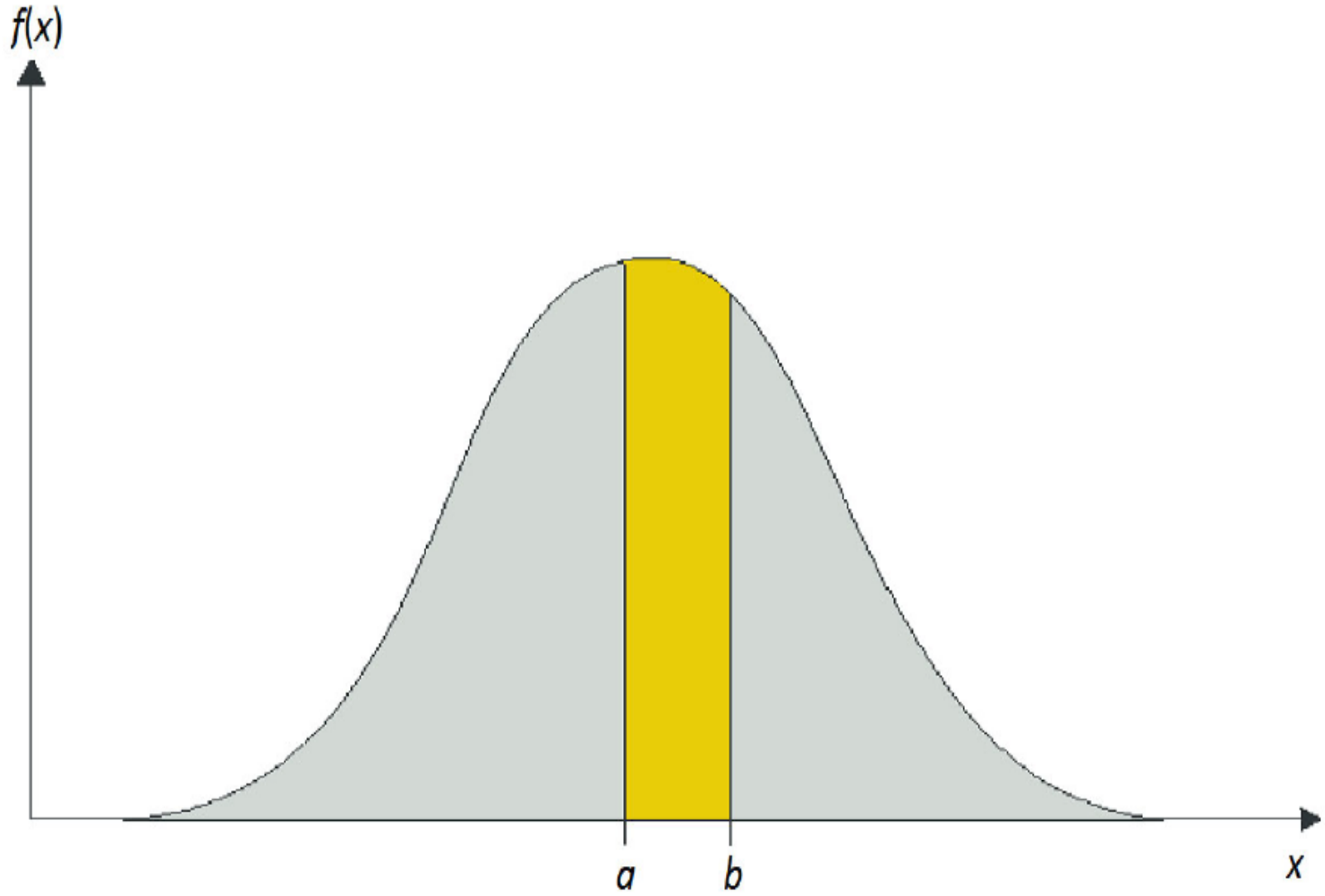
- *Note:* If $\tilde{x}$ is absolutely continuous, then

$$P_{\tilde{x}}(a,b] = P_{\tilde{x}}(a,b) = P_{\tilde{x}}[a,b] = P_{\tilde{x}}[a,b) =$$

$$F_{\tilde{x}}(b) - F_{\tilde{x}}(a) = \int_{[a,b]} f_{\tilde{x}}(x)dx.$$

- **Notation:** If the random variable $\tilde{x}$ has the distribution $P_{\tilde{x}}$, we write $\tilde{x} \sim P_{\tilde{x}}$, $\tilde{x} \sim F_{\tilde{x}}$, or $\tilde{x} \sim f_{\tilde{x}}$, where $F_{\tilde{x}}$ is the corresponding distribution function and $f_{\tilde{x}}$ is the corresponding probability or density function.

$P_{\tilde{x}}[a, b]$ is given by the area of the yellow region



- **Properties of the density:**
- **1.**
$$\int_{\mathbb{R}} f_{\tilde{x}}(x)dx = 1.$$
- **2.**
$$F_{\tilde{x}}(x) = \int_{(-\infty,x]} f_{\tilde{x}}(t)dt.$$
- **3.** Any non-negative (a.e. w.r.t. Lebesgue measure) Borel measurable function $f : (\mathbb{R}, \mathcal{B}) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$ satisfying $\int_{\mathbb{R}} f(x)dx = 1$ can serve as a density of an absolutely continuous distribution on $(\mathbb{R}, \mathcal{B})$.
- **4.** If $\tilde{x}$ is absolutely continuous, then $f_{\tilde{x}} = F'_{\tilde{x}}$ when the derivative of $F_{\tilde{x}}$ exists. Moreover, the derivative $F'_{\tilde{x}}$ exists a.e. w.r.t. Lebesgue measure. If $f_{\tilde{x}}$ is continuous at x then $F_{\tilde{x}}$ is differentiable at x and $f_{\tilde{x}}(x) = F'_{\tilde{x}}(x)$.

2.7. Random vectors

- $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B})$.
- $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ or $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)^\intercal$.
- $\tilde{x}_i = p_i(\tilde{x})$, where $p_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ is the projection to the ith coordinate.
- The distribution of the random vector $\tilde{x}$ is a probability measure on $(\mathbb{R}^n, \mathcal{B})$ given by

$$P_{\tilde{x}}(B) = P\left(\tilde{x}^{-1}(B)\right) \text{ for all } B \in \mathcal{B}(\mathbb{R}^n) .$$

- The distribution function (cdf) of the random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$, $F_{\tilde{x}} : \mathbb{R}^n \longrightarrow \mathbb{R}$, is given by

$$F_{\tilde{x}}\underbrace{(x_1, x_2, ..., x_n)}_{x \in \mathbb{R}^n} = P\left\{\underbrace{\tilde{x}_i \leq x_i, \text{ for } i = 1, 2, ..., n}_{\tilde{x} \leqq x}\right\}.$$

- The distribution function of a random vector $\tilde{x}$ is a¡ (i) (weakly) **increasing**,...
- (Increasing: $a < b \Longrightarrow F(a) \leq F(b)$, where a and b belong to $\mathbb{R}^n$.)
- (ii) **right-continuous**,..
- (Right-continuous at x_0: $\lim_{x \to x_0^+} F(x) \equiv F\left(x_0^+\right) = F\left(x_0\right)$, where $x > x_0 \in \mathbb{R}^n$)
- (iii) $F_{\tilde{x}}(x) \to 0$ if at least one of the components x_i of $x \in \mathbb{R}^n$ tends to $-\infty$, and
- (iv) $F_{\tilde{x}}(x) \to 1$ if all the components x_i, $i = 1, ..., n$, of $x \in \mathbb{R}^n$ tend to ∞.

- The random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is discrete if its range $\tilde{x}(\Omega)$ is countable (or discrete).
- The probability function (pmf), $f_{\tilde{x}} : \tilde{x}_1(\Omega) \times \tilde{x}_2(\Omega) \times ... \times \tilde{x}_n(\Omega) \longrightarrow [0, 1]$, of a discrete random vector $\tilde{x}$ is given by:

$$f_{\tilde{x}}(x) = P_{\tilde{x}}\{x\} = P\left\{\underbrace{(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)}_{\tilde{x}} = \underbrace{(x_1, x_2, ..., x_n)}_{x \in \mathbb{R}^n}\right\} =$$

$$P\{\tilde{x}_i = x_i, \text{ for } i = 1, 2, ..., n\}, \text{ for all } x \in \tilde{x}_1(\Omega) \times \tilde{x}_2(\Omega) \times ... \times \tilde{x}_n(\Omega).$$

- *Note:* $\tilde{x}(\Omega) \subset \tilde{x}_1(\Omega) \times \tilde{x}_2(\Omega) \times ... \times \tilde{x}_n(\Omega)$.

- **Properties of the probability and distribution functions of a discrete random vector:**
- **1.**

$$\sum_{x \in \tilde{x}(\Omega)} f_{\tilde{x}}(x) = 1 \quad \text{or} \quad \sum_{x \in \tilde{x}_1(\Omega) \times \tilde{x}_2(\Omega) \times ... \times \tilde{x}_n(\Omega)} f_{\tilde{x}}(x) = 1.$$

- **2.**

$$F_{\tilde{x}}(x) = \sum_{t \leqq x} f_{\tilde{x}}(t), \quad \text{with } t = (t_1, t_2, ..., t_n) \in \tilde{x}(\Omega),$$

where $t \leqq x$ means that $t_i \leq x_i$ for $i = 1, 2, ..., n$.

- **3.**

$$P_{\tilde{x}}(B) = P\{\tilde{x} \in B\} = \sum_{x \in B} f_{\tilde{x}}(x), \quad \text{for all } B \in \mathcal{B}(\mathbb{R}^n).$$

- The random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ (or its distribution) is absolutely continuous if there exists a Borel measurable function $f_{\tilde{x}} : (\mathbb{R}^n, \mathcal{B}) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$, called the density (pdf), that is integrable with respect to Lebesgue measure on $(\mathbb{R}^n, \mathcal{B})$, such that

$$P_{\tilde{x}}(B) = \int_B f_{\tilde{x}}(x_1, x_2, ..., x_n)\, d(x_1, x_2, ..., x_n), \text{ for all } B \in \mathcal{B}(\mathbb{R}^n).$$

- **Properties of the density of a random vector**:
- **1.**

$$\int_{\mathbb{R}^n} f_{\tilde{x}}(x)dx = \int_{\mathbb{R}} ... \int_{\mathbb{R}} f_{\tilde{x}}(\underbrace{x_1, ..., x_n}_{x \in \mathbb{R}^n})dx_1 ... dx_n = 1.$$

- **2.**

$$F_{\tilde{x}}(x) = \int_{(-\infty, x_n]} \int_{(-\infty, x_{n-1}]} ... \int_{(-\infty, x_1]} f_{\tilde{x}}(t_1, t_2, ..., t_n)dt_1 dt_2 ... dt_n.$$

- **3.** Any non-negative (a.e. w.r.t. Lebesgue measure) Borel measurable function $f : (\mathbb{R}^n, \mathcal{B}) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$ satisfying

$$\int_{\mathbb{R}} ... \int_{\mathbb{R}} f(x_1, x_2, ..., x_n) dx_1 dx_2, ..., dx_n = 1$$

  can serve as a density of an absolutely continuous distribution on $(\mathbb{R}^n, \mathcal{B})$.

- **4.** If the random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is absolutely continuous, then

$$f_{\tilde{x}}(x_1, x_2, ..., x_n) = \frac{\partial^n F_{\tilde{x}}(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}.$$

  when this nth crossed partial derivative of $F_{\tilde{x}}$ exists. Moreover, this derivative exists a.e. w.r.t. Lebesgue measure on $(\mathbb{R}^n, \mathcal{B})$.

2.8. Marginal distributions

- **Definition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a random vector with distribution $P_{\tilde{x}}$. The marginal distribution of $\tilde{x}_i$, for $i = 1, ..., n$, is given by

$$P_{\tilde{x}_i}(B) = P_{\tilde{x}}(\mathbb{R} \times ... \times \underset{\substack{\uparrow \\ i}}{B} \times ... \times \mathbb{R}), \quad \text{for all } B \in \mathcal{B}(\mathbb{R}).$$

- **Definition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a discrete random vector with the probability function $f_{\tilde{x}}$, the marginal probability function of $\tilde{x}_i$, for $i = 1, ..., n$, is given by

$$f_{\tilde{x}_i}(x_i) = \sum_{x_1 \in \tilde{x}_1(\Omega)} ... \sum_{x_{i-1} \in \tilde{x}_{i-1}(\Omega)} \sum_{x_{i+1} \in \tilde{x}_{i+1}(\Omega)} ... \sum_{x_n \in \tilde{x}_n(\Omega)} f_{\tilde{x}}(\underbrace{x_1, ..., x_{i-1}, x_i, x_{i+1} ..., x_n}_{x}).$$

for all $x_i \in \tilde{x}_i(\Omega)$.

- **Definition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be an absolutely continuous random vector with the density $f_{\tilde{x}}$, the marginal density of $\tilde{x}_i$, for $i = 1, ..., n$, is given by

$$f_{\tilde{x}_i}(x_i) = \int_{\mathbb{R}} ... \int_{\mathbb{R}} f_{\tilde{x}}\underbrace{(x_1, ..., x_{i-1}, x_i, x_{i+1} ..., x_n)}_{x} dx_1 ... dx_{i-1} dx_{i+1} ... dx_n,$$

for all $x_i \in \mathbb{R}$.

- *Note:* From the marginal probability or density functions we can construct the marginal distributions in the usual way.

- **Example 1:** The discrete random vector $(\tilde{x}, \tilde{y})$, where $\tilde{x}$ is the number of points when rolling a dice and $\tilde{y}$ is the number of heads when tossing a coin has a probability function $f_{\tilde{x},\tilde{y}}(x, y)$ summarized in the following table:

$y \backslash x$	1	2	3	4	5	6	$f_{\tilde{y}}(y)$
0	1/12	1/12	1/12	1/12	1/12	1/12	1/2
1	1/12	1/12	1/12	1/12	1/12	1/12	1/2
$f_{\tilde{x}}(x)$	1/6	1/6	1/6	1/6	1/6	1/6	1

- The marginal probability functions of $\tilde{x}$ and $\tilde{y}$ are summarized in the "margins".

Example 2: The absolutely continuous random vector $(\tilde{x}, \tilde{y})$ has the following density:

$$f_{\tilde{x},\tilde{y}}(x,y) = \begin{cases} \frac{2}{3}(x+2y) & \text{for } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Marginal densities:

$$f_{\tilde{x}}(x) = \int_{-\infty}^{\infty} f_{\tilde{x},\tilde{y}}(x,y)dy = \int_0^1 \frac{2}{3}(x+2y)\,dy = \frac{2}{3}\left[xy+y^2\right]_0^1$$

$$= \frac{2}{3}(x+1), \quad \text{for } 0 < x < 1.$$

- Therefore,

$$f_{\tilde{x}}(x) = \begin{cases} \frac{2}{3}(x+1) & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Similarly,

$$f_{\tilde{y}}(y) = \int_{-\infty}^{\infty} f_{\tilde{x},\tilde{y}}(x,y)dx = \int_0^1 \frac{2}{3}(x+2y)\,dx = \frac{2}{3}\left[\frac{x^2}{2}+2xy\right]_0^1$$

$$= \frac{2}{3}\left(\frac{1}{2}+2y\right) = \frac{1}{3}(1+4y), \quad \text{for } 0<y<1.$$

- Therefore,

$$f_{\tilde{y}}(y) = \begin{cases} \frac{1}{3}(1+4y) & \text{for } 0<y<1 \\ 0 & \text{otherwise.} \end{cases}$$

2.9. Independent random variables

- **Definition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a random vector defined on $(\Omega, \mathcal{F}, P)$ with the distribution $P_{\tilde{x}}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. The random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are said to be independent if, for all collections of sets $B_1, B_2, ..., B_n$ belonging to $\mathcal{B}(\mathbb{R})$, we have

$$P\{\tilde{x}_1 \in B_1, ..., \tilde{x}_n \in B_n\} = P\{\tilde{x}_1 \in B_1\} \cdot ... \cdot P\{\tilde{x}_n \in B_n\}$$

or, equivalently, if the distribution of the random vector $\tilde{x}$ is equal to the product measure of the marginal distributions,

$$P_{\tilde{x}} = P_{\tilde{x}_1} \times P_{\tilde{x}_2} \times ... \times P_{\tilde{x}_n} \equiv \prod_{i=1}^{n} P_{\tilde{x}_i},$$

that is,

$$P_{\tilde{x}}\{B_1 \times B_2 ... \times B_n\} = P_{\tilde{x}_1}(B_1) \cdot P_{\tilde{x}_2}(B_2) \cdot ... \cdot P_{\tilde{x}_n}(B_n).$$

- **Definition.** Let $\tilde{x}_i : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_i, \mathcal{F}_i)$, for $i = 1, ..., n$. The random objects $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are said to be independent if, for all sets $B_1 \in \mathcal{F}_1, ..., B_n \in \mathcal{F}_n$,

$$P\left\{\tilde{x}_1 \in B_1, ..., \tilde{x}_n \in B_n\right\} = P\left\{\tilde{x}_1 \in B_1\right\} \cdot ... \cdot P\left\{\tilde{x}_n \in B_n\right\}.$$

- **Equivalent definition.** Let $\tilde{x}_i : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_i, \mathcal{F}_i)$, for $i = 1, ..., n$. The collection of random objects $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are said to be independent if the joint distribution

$$P_{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n} : \bigotimes_{i=1}^{n} \mathcal{F}_i \longrightarrow [0, 1]$$

of these n random objects is equal to the product measure of the marginal distributions,

$$P_{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n} = \prod_{i=1}^{n} P_{\tilde{x}_i},$$

where $P_{\tilde{x}_i} : \mathcal{F}_i \longrightarrow [0, 1]$ is the marginal distribution of the random object $\tilde{x}_i$, $i = 1, ..., n$, and $\bigotimes_{i=1}^{n} \mathcal{F}_i$ is the product $\sigma-$algebra.

- **Proposition.** Let $\tilde{x}_i : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_i, \mathcal{F}_i)$, for $i = 1, ..., n$, be a collection of independent random objects and $g_i : (\Omega_i, \mathcal{F}_i) \longrightarrow (\Omega_i', \mathcal{F}_i')$, for $i = 1, ..., n$, be measurable functions. Then, the random objects $g_i(\tilde{x}_i) : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_i', \mathcal{F}_i')$, for $i = 1, ..., n$, are independent.

Proof. If

$$P\left\{\tilde{x}_1 \in B_1, ..., \tilde{x}_n \in B_n\right\} = P\left\{\tilde{x}_1 \in B_1\right\} \cdot ... \cdot P\left\{\tilde{x}_n \in B_n\right\},$$

for all sets $B_1 \in \mathcal{F}_1, ..., B_n \in \mathcal{F}_n$, then

$$P\left\{\tilde{x}_1 \in g_1^{-1}\left(B_1'\right), ..., \tilde{x}_n \in g_n^{-1}\left(B_n'\right)\right\}$$

$$= P\left\{\tilde{x}_1 \in g_1^{-1}\left(B_1'\right)\right\} \cdot ... \cdot P\left\{\tilde{x}_n \in g_n^{-1}\left(B_n'\right)\right\},$$

for all sets $B_1' \in \mathcal{F}_1', ..., B_n' \in \mathcal{F}_n'$, since $g_1^{-1}\left(B_1'\right) \in \mathcal{F}_1, ..., g_n^{-1}\left(B_n'\right) \in \mathcal{F}_n$ due to the measurability of g_i, for $i = 1, ..., n$. Therefore,

$$P\left\{g_1\left(\tilde{x}_1\right) \in B_1', ..., g_n\left(\tilde{x}_n\right) \in B_n'\right\} = P\left\{g_1\left(\tilde{x}_1\right) \in B_1'\right\} \cdot ... \cdot P\left\{g_n\left(\tilde{x}_n\right) \in B_n'\right\},$$

for all sets $B_1' \in \mathcal{F}_1', ..., B_n' \in \mathcal{F}_n'$, which proves the independency of the random objects $g_i\left(\tilde{x}_i\right) : (\Omega, \mathcal{F}, P) \longrightarrow \left(\Omega_i', \mathcal{F}_i'\right)$, for $i = 1, ..., n$. $Q.E.D.$

- **Proposition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a random vector with the distribution function $F_{\tilde{x}} : \mathbb{R}^n \longrightarrow [0,1]$ and let $F_i : \mathbb{R} \longrightarrow [0,1]$ be the marginal distribution function of $\tilde{x}_i$, for $i = 1, ..., n$. Then, the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are independent if and only if

$$F_{\tilde{x}}(x_1, ..., x_n) = F_1(x_1) \cdot F_2(x_2) \cdot ... \cdot F_n(x_n),$$

$$\text{for all } x = (x_1, ..., x_n) \in \mathbb{R}^n.$$

- **Proposition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a discrete random vector with the probability function $f_{\tilde{x}} : \tilde{x}_1(\Omega) \times ... \times \tilde{x}_n(\Omega) \longrightarrow [0, 1]$ and let $f_i : \tilde{x}_i(\Omega) \longrightarrow [0, 1]$ be the marginal probability function of $\tilde{x}_i$, for $i = 1, ..., n$. Then, the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are independent if and only if

$$f_{\tilde{x}}(x_1, ..., x_n) = f_1(x_1) \cdot f_2(x_2) \cdot ... \cdot f_n(x_n),$$

$$\text{for all } x = (x_1, ..., x_n) \in \tilde{x}_1(\Omega) \times ... \times \tilde{x}_n(\Omega).$$

- **Proposition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be an absolutely continuous random vector with the density function $f_{\tilde{x}} : \mathbb{R}^n \longrightarrow \overline{\mathbb{R}}$ and let $f_i : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$ be the marginal density function of $\tilde{x}_i$, for $i = 1, ..., n$. Then, the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are independent if and only if

$$f_{\tilde{x}}(x_1, ..., x_n) = f_1(x_1) \cdot f_2(x_2) \cdot ... \cdot f_n(x_n),$$

$$\text{for all } x = (x_1, ..., x_n) \in \mathbb{R}^n.$$

2.10. Generalized conditional probability

- Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ and let us fix the event $B \in \mathcal{F}$. From the Radon-Nikodym theorem we know that there exists a Borel measurable function $g : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}, \mathcal{B})$ such that

$$\underbrace{P\left(\{\tilde{x} \in A\} \cap B\right)}_{\lambda(A)} = \int_A g(x) dP_{\tilde{x}}(x), \text{ for all } A \in \mathcal{F}',$$

  since $\lambda \ll P_{\tilde{x}}$. The function g is called the conditional probability of B given $\tilde{x} = x$ and is written as $g(x) = P(B \,|\tilde{x} = x\,)$. The conditional probability is essentially unique for a given $B \in \mathcal{F}$ (i.e., if there exists another such function h, then $g = h$ a.e. $[P_{\tilde{x}}]$).

- Therefore,

$$P\left(\{\tilde{x} \in A\} \cap B\right) = \int_A P(B \,|\tilde{x} = x\,) dP_{\tilde{x}}(x),$$

  with $P(B \,|\tilde{x} = \cdot\,) : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}, \mathcal{B})$.

- However, sometimes the conditional probability is viewed as a measure on $(\Omega, \mathcal{F})$,

$$P(\cdot \,|\tilde{x} = x) : \mathcal{F} \longrightarrow \mathbb{R}.$$

- Moreover, if $A = \Omega'$, then

$$P\left(\left\{\tilde{x} \in \Omega'\right\} \cap B\right) = P\left(\Omega \cap B\right) = P(B) = \int_{\Omega'} P(B \,|\tilde{x} = x)\, dP_{\tilde{x}}(x),$$

  which is a generalization of the theorem of total probability.

- Note that, if $\tilde{x}$ is an absolutely continuous random variable, then $P(B \,|\tilde{x} = x)$ is a conditional probability given an event $(\{\tilde{x} = x\})$ that has zero probability!

2.11. Conditional distributions

- **Definition.** Let $(\tilde{x}, \tilde{y})$ be a vector of two random objects $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_x, \mathcal{F}_x)$ and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_y, \mathcal{F}_y)$, and let $C \in \mathcal{F}_y$ be a fixed measurable set. The conditional distribution of $\tilde{y}$ given $\tilde{x} = x$ is the Borel measurable function $P_{\tilde{y}|\tilde{x}}(C|x) : (\Omega_x, \mathcal{F}_x) \longrightarrow (\mathbb{R}, \mathcal{B})$ given by

$$P_{\tilde{y}|\tilde{x}}(C|x) = P\{\tilde{y} \in C | \tilde{x} = x\}, \text{ for all } x \in \Omega_x,$$

  which is essentially unique w.r.t. $P_{\tilde{x}}$.

- However, sometimes the conditional distribution is viewed as a measure on $(\Omega_y, \mathcal{F}_y)$,

$$P_{\tilde{y}|\tilde{x}}(\cdot|x) : \mathcal{F}_y \longrightarrow \mathbb{R}.$$

- Assume that the random vector $(\tilde{x}, \tilde{y})$ is discrete with the probability function $f_{\tilde{x},\tilde{y}} : \tilde{x}(\Omega) \times \tilde{y}(\Omega) \longrightarrow [0,1]$. Then, the conditional distribution $P_{\tilde{y}|\tilde{x}}(y|x) = P\{\tilde{y} = y | \tilde{x} = x\}$ must satisfy

$$P\{\tilde{x} \in A, \tilde{y} \in C\} = \sum_{x \in A} P_{\tilde{y}|\tilde{x}}(C|x) \underbrace{P\{\tilde{x} = x\}}_{f_{\tilde{x}}(x)}$$

$$= \sum_{x \in A} \underbrace{\sum_{y \in C} P_{\tilde{y}|\tilde{x}}(y|x)}_{P_{\tilde{y}|\tilde{x}}(C|x)} f_{\tilde{x}}(x), \text{ for all } A \in \mathcal{B}(\mathbb{R}) \text{ and } C \in \mathcal{B}(\mathbb{R}). \quad (*)$$

- Let us define the conditional distribution $P_{\tilde{y}|\tilde{x}}(y|x)$ as follows:

$$P_{\tilde{y}|\tilde{x}}(y|x) = \frac{P\{\tilde{x} = x, \tilde{y} = y\}}{P\{\tilde{x} = x\}} = \frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)} \equiv f_{\tilde{y}|\tilde{x}}(y|x),$$

for all $(x,y) \in \tilde{x}(\Omega) \times \tilde{y}(\Omega)$ with $f_{\tilde{x}}(x) > 0$.

- The function $f_{\tilde{y}|\tilde{x}}\left(\cdot\,|x\right):\tilde{y}(\Omega)\longrightarrow[0,1]$, for all $x\in\tilde{x}(\Omega)$ such that $f_{\tilde{x}}(x)>0$, is the conditional probability function of $\tilde{y}$ given $\tilde{x}=x$.
- The previous definition of the conditional probability function (or conditional distribution) of $\tilde{y}$ given $\tilde{x}=x$ is the right one since the expression $(*)$ becomes

$$P\left\{\tilde{x}\in A,\tilde{y}\in C\right\}=\sum_{x\in A}\sum_{y\in C}f_{\tilde{y}|\tilde{x}}\left(y\,|x\right)f_{\tilde{x}}(x)$$

$$=\sum_{x\in A}\sum_{y\in C}\frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)}f_{\tilde{x}}(x)=\sum_{x\in A}\sum_{y\in C}f_{\tilde{x},\tilde{y}}(x,y),$$

$$\text{for all } A\in\mathcal{B}\left(\mathbb{R}\right) \text{ and } C\in\mathcal{B}\left(\mathbb{R}\right).$$

- Assume that the random vector $(\tilde{x}, \tilde{y})$ is absolutely continuous with the density $f_{\tilde{x},\tilde{y}} : \mathbb{R}^2 \longrightarrow \overline{\mathbb{R}}$. Then, we would like to have an expression like this:

$$P\{\tilde{x} \in A, \tilde{y} \in C\} = \int_A P_{\tilde{y}|\tilde{x}}(C|x) dP_{\tilde{x}}(x)$$

$$= \int_A P_{\tilde{y}|\tilde{x}}(C|x) f_{\tilde{x}}(x) dx = \int_A \underbrace{\left[\int_C f_{\tilde{y}|\tilde{x}}(y|x)\, dy\right]}_{P_{\tilde{y}|\tilde{x}}(C|x)} f_{\tilde{x}}(x) dx, \qquad (**)$$

for all $A \in \mathcal{B}(\mathbb{R})$ and $C \in \mathcal{B}(\mathbb{R})$.

- Let us define the conditional density of $\tilde{y}$ given $\tilde{x} = x$, $f_{\tilde{y}|\tilde{x}}(\cdot\,|x) : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$, for all $x \in \mathbb{R}$ such that $f_{\tilde{x}}(x) > 0$, as follows:

$$f_{\tilde{y}|\tilde{x}}(y\,|x) = \frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)}, \quad \text{for all } (x,y) \in \mathbb{R}^2 \text{ with } f_{\tilde{x}}(x) > 0.$$

- The previous definition of the conditional density of $\tilde{y}$ given $\tilde{x} = x$ is the right one since the expression $(**)$ becomes

$$P\{\tilde{x} \in A, \tilde{y} \in C\} = \int_A \int_C f_{\tilde{y}|\tilde{x}}(y\,|x)\, f_{\tilde{x}}(x) dy dx$$

$$= \int_A \int_C \frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)} f_{\tilde{x}}(x) dy dx = \int_A \int_C f_{\tilde{x},\tilde{y}}(x,y) dy dx,$$

$$\text{for all } A \in \mathcal{B}(\mathbb{R}) \text{ and } C \in \mathcal{B}(\mathbb{R}).$$

- If the discrete (absolutely continuous) random variables $\tilde{x}$ and $\tilde{y}$ are independent then

$$f_{\tilde{y}|\tilde{x}}\left(y\,|x\right) = \frac{f_{\tilde{x},\tilde{y}}(x,y)}{f_{\tilde{x}}(x)} = \frac{f_{\tilde{x}}(x)\cdot f_{\tilde{y}}\left(y\right)}{f_{\tilde{x}}(x)} = f_{\tilde{y}}\left(y\right), \quad \text{for } f_{\tilde{x}}(x) > 0.$$

  That is, the conditional probability function (density function) is equal to the corresponding "unconditional" probability function (density function).

- In general, if the two random objects $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_x, \mathcal{F}_x)$ and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_y, \mathcal{F}_y)$ are independent then

$$P_{\tilde{y}|\tilde{x}}\left(C\,|x\right) = P_{\tilde{y}}(C), \text{ for all } C \in \mathcal{F}_y \text{ a.s. } \left[P_{\tilde{x}}\right].$$

- Note that from the conditional probability and density functions we can obtain the conditional distribution in the usual way, namely,

$$P_{\tilde{y}|\tilde{x}}\left(C\left|x\right.\right)=P\left\{\tilde{y}\in C\left|\tilde{x}=x\right.\right\}=\sum_{y\in C}f_{\tilde{y}|\tilde{x}}\left(y\left|x\right.\right),\text{ for all }C\in\mathcal{B},$$

or

$$P_{\tilde{y}|\tilde{x}}\left(C\left|x\right.\right)=P\left\{\tilde{y}\in C\left|\tilde{x}=x\right.\right\}=\int_{C}f_{\tilde{y}|\tilde{x}}\left(y\left|x\right.\right)dy,\text{ for all }C\in\mathcal{B},$$

where

$$P_{\tilde{y}|\tilde{x}}\left(C\left|\cdot\right.\right):\left(\mathbb{R},\mathcal{B}\right)\longrightarrow\left(\mathbb{R},\mathcal{B}\right)$$

or, sometimes,

$$P_{\tilde{y}|\tilde{x}}\left(\cdot\left|x\right.\right):\mathcal{B}\left(\mathbb{R}\right)\longrightarrow\mathbb{R}.$$

- Note again that, if $\tilde{x}$ is an absolutely continuous random variable, then $P_{\tilde{y}|\tilde{x}}\left(C\left|x\right.\right)$ is a conditional distribution (and, hence, a conditional probability $P\left\{\tilde{y}\in C\left|\tilde{x}=x\right.\right\}$) given the event $\left\{\tilde{x}=x\right\}$, which has zero probability! This conditional distribution is well defined when the marginal density of $\tilde{x}$ evaluated at x, $f_{\tilde{x}}(x)$, is strictly positive.

- **Example:** The absolutely continuous random vector $(\tilde{x}, \tilde{y})$ has the following density:

$$f_{\tilde{x},\tilde{y}}(x,y) = \begin{cases} \frac{2}{3}(x+2y) & \text{for } 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

- We have already proved that the marginal density of the random variable $\tilde{y}$ is

$$f_{\tilde{y}}(y) = \begin{cases} \frac{1}{3}(1+4y) & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

- Therefore, the conditional density of $\tilde{x}$ given $\tilde{y} = y$ is

$$f_{\tilde{x}|\tilde{y}}(x\,|y) = \begin{cases} \dfrac{\frac{2}{3}(x+2y)}{\frac{1}{3}(1+4y)} = \dfrac{2x+4y}{1+4y} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

  for $0 < y < 1$.

- Thus, the conditional density of $\tilde{x}$ given $\tilde{y} = 1/4$ is

$$f_{\tilde{x}|\tilde{y}}\left(x\left|\frac{1}{4}\right.\right) = \begin{cases} \dfrac{2x+1}{2} = \dfrac{1}{2}(2x+1) & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$P_{\tilde{x}|\tilde{y}}\left(\left(-\infty,\frac{1}{3}\right]\middle|\frac{1}{4}\right)=P\left\{\tilde{x}\leq\frac{1}{3}\middle|\tilde{y}=\frac{1}{4}\right\}=\int_{-\infty}^{1/3}f_{\tilde{x}|\tilde{y}}\left(x\middle|\frac{1}{4}\right)dx$$

$$=\int_{0}^{1/3}\frac{1}{2}\left(2x+1\right)dx=\frac{1}{2}\left[x^{2}+x\right]_{0}^{1/3}=\frac{1}{2}\left(\frac{1}{9}+\frac{1}{3}\right)=\frac{1}{2}\cdot\frac{4}{9}=\frac{4}{18}=\frac{2}{9},$$

while

$$P_{\tilde{x}}\left(-\infty,\frac{1}{3}\right]=\int_{-\infty}^{1/3}\underbrace{\int_{-\infty}^{\infty}f_{\tilde{x},\tilde{y}}(x,y)dy}_{f_{\tilde{x}}(x)}dx$$

$$=\int_{0}^{1/3}\underbrace{\int_{0}^{1}\frac{2}{3}\left(x+2y\right)dy}_{\frac{2}{3}(x+1)}dx=\frac{7}{27}.$$

- If we have more than 2 random variables, we can generalize the previous conditional probability and density functions.
- **Example:**

$$f_{\tilde{x}_1,\tilde{x}_3|\tilde{x}_2,\tilde{x}_4}\left(x_1, x_3 \left| x_2, x_4 \right.\right) = \frac{f_{\tilde{x}_1,\tilde{x}_2,\tilde{x}_3,\tilde{x}_4}\left(x_1, x_2, x_3, x_4\right)}{f_{\tilde{x}_2,\tilde{x}_4}\left(x_2, x_4\right)},$$

where

$$f_{\tilde{x}_2,\tilde{x}_4}\left(x_2, x_4\right) = \sum_{x_1 \in \tilde{x}_1(\Omega)} \sum_{x_3 \in \tilde{x}_3(\Omega)} f_{\tilde{x}_1,\tilde{x}_2,\tilde{x}_3,\tilde{x}_4}\left(x_1, x_2, x_3, x_4\right) > 0,$$

or

$$f_{\tilde{x}_2,\tilde{x}_4}\left(x_2, x_4\right) = \int_{\mathbb{R}} \int_{\mathbb{R}} f_{\tilde{x}_1,\tilde{x}_2,\tilde{x}_3,\tilde{x}_4}\left(x_1, x_2, x_3, x_4\right) dx_1 dx_3 > 0.$$

Exercises. Probability and Statistics. IDEA.
2. Random Variables and Distributions

1. Verify that $f(x) = \dfrac{2x}{k(k+1)}$ for $x = 1, 2, 3, ..., k$ can serve as the probability function for a random variable.

2. For what values of k can

$$f(x) = (1-k)\,k^x \quad \text{for } x = 0, 1, 2, ...$$

serve as the probability function of a random variable?

3. If the density of the random variable $\tilde{z}$ is given by

$$f(z) = \begin{cases} kze^{-z^2} & \text{for } z > 0 \\ 0 & \text{for } z \leq 0, \end{cases}$$

(a) Find the value of k;

(b) Find the distribution function of this random variable;

(c) Sketch the graphs of the density and the distribution functions;

(d) Find the quantile function of this random variable and sketch its graph.

4. If the density of the random variable $\tilde{z}$ is given by

$$f(z) = \begin{cases} -kz & \text{for } -1 < z < 0 \\ kz & \text{for } 0 \leq z < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

(a) Find the value of k;

(b) Find the distribution function of this random variable and sketch its graph;

(c) Compute $P\left\{-\dfrac{1}{2} < \tilde{z} < \dfrac{1}{2}\right\}$;

(d) Find the quantile function of this random variable and sketch its graph.

5. In a certain city the daily consumption of water (in millions of liters) is a random variable whose density is given by

$$f(x) = \begin{cases} \dfrac{1}{9}xe^{-x/3} & \text{for } x > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

What are the probabilities that on a given day

(a) the water consumption in this city is no more than 6 million liters;

(b) the water supply is inadequate if the daily capacity of this city is 9 million liters?

6. If the joint density of $\tilde{x}$ and $\tilde{y}$ is given by

$$f(x,y) = \begin{cases} \dfrac{1}{y} & \text{for } 0 < x < y,\ 0 < y < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

find the probability that the sum of the values taken on by the two random variables will exceed 1/2.

7. If $\widetilde{p}$, the price of a certain commodity (in dollars), and $\widetilde{s}$, total sales (in 10 000 units), are random variables whose distribution can be approximated with the joint density

$$f(p,s) = \begin{cases} 5pe^{-ps} & \text{for } 0.20 < p < 0.40,\ s > 0 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

find the probabilities that

(a) the price will be less than 30 cents and sales will exceed 20 000 units;

(b) the price will be between 25 cents and 30 cents and sales will be less than 10 000 units.

8. Given the joint probability function

$$f(x,y,z) = \frac{xyz}{108} \quad \text{for } x = 1,2,3;\ y = 1,2,3;\ z = 1,2,$$

find

(a) the joint marginal probability function of $\tilde{x}$ and $\tilde{y}$;

(b) the joint marginal probability function of $\tilde{x}$ and $\tilde{z}$;

(c) the marginal probability function of $\tilde{x}$;

(d) the conditional probability function of $\tilde{z}$ given $\tilde{x} = 1$ and $\tilde{y} = 2$;

(e) the joint conditional probability function of $\tilde{y}$ and $\tilde{z}$ given $\tilde{x} = 3$.

9. Check whether the random variables $\tilde{x}$ and $\tilde{y}$ are independent, if their joint probability function is given by

(a) $f(x,y) = 1/4$ for $x = -1$ and $y = -1$, $x = -1$ and $y = 1$, $x = 1$ and $y = -1$, and $x = 1$ and $y = 1$;

(b) $f(x,y) = 1/3$ for $x = 0$ and $y = 0$, $x = 0$ and $y = 1$, and $x = 1$ and $y = 1$.

10. If the joint density of $\tilde{x}$ and $\tilde{y}$ is given by

$$f(x,y) = \begin{cases} \frac{1}{4}(2x+y) & \text{for } 0 < x < 1,\ 0 < y < 2 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

find

(a) the marginal density of $\tilde{x}$;

(b) the marginal density of $\tilde{y}$;

(c) the conditional density of $\tilde{x}$ given $\tilde{y} = 1$;

(d) the conditional density of $\tilde{y}$ given $\tilde{x} = 1/4$.

11. If $\tilde{x}$ is the amount (in dollars) a salesperson spends on gasoline during a day and $\tilde{y}$ is the amount (in dollars) for which the salesperson is reimbursed, and the joint density of these random variables is given by

$$f(x,y) = \begin{cases} \frac{20-x}{25x} & \text{for } 10 < x < 20,\ \frac{x}{2} < y < x \\ \\ 0 & \text{elsewhere,} \end{cases}$$

find

(a) the marginal densities of $\tilde{x}$ and $\tilde{y}$;

(b) the conditional density of $\tilde{y}$ given $\tilde{x} = 12$;

(c) the probability that the salesperson will be reimbursed at least \$8 when spending \$12.

12. Give an example of random variables $\tilde{x}$ and $\tilde{y}$ (on the same probability space) such that $\tilde{x}$ and $\tilde{y}$ each have densities, but the random vector $(\tilde{x}, \tilde{y})$ does not.

13. (a) Let $\tilde{x}$ be an absolutely continuous random variable with density $f_{\tilde{x}}(x)$. If $\tilde{x} = x$, let $\tilde{y}$ be discrete, with the probability function $f_{\tilde{y}|\tilde{x}}(y\,|x) = P\{\tilde{y} = y \mid \tilde{x} = x\}$ specified. Show that there is a conditional density of $\tilde{x}$ given $\tilde{y}$, namely,

$$f_{\tilde{x}|\tilde{y}}(x\,|y) = \frac{f_{\tilde{x}}(x) f_{\tilde{y}|\tilde{x}}(y\,|x)}{f_{\tilde{y}}(y)},$$

where

$$f_{\tilde{y}}(y) = P\{\tilde{y} = y\} = \int_{\mathbb{R}} f_{\tilde{x}}(x) f_{\tilde{y}|\tilde{x}}(y\,|x)\,dx > 0.$$

(b) Let $\tilde{x}$ be a discrete random variable with the probability function $f_{\tilde{x}}(x)$. If $\tilde{x} = x$, let $\tilde{y}$ be absolutely continuous, with the density function $f_{\tilde{y}|\tilde{x}}(y\,|x)$ specified. Show that there is a conditional probability function of $\tilde{x}$ given $\tilde{y}$, namely,

$$f_{\tilde{x}|\tilde{y}}(x\,|y) = \frac{f_{\tilde{x}}(x) f_{\tilde{y}|\tilde{x}}(y\,|x)}{f_{\tilde{y}}(y)},$$

where

$$f_{\tilde{y}}(y) = \sum_{x \in \tilde{x}(\Omega)} f_{\tilde{x}}(x) f_{\tilde{y}|\tilde{x}}\,(y\,|x\,) > 0.$$

14. If $\tilde{x}$ is a random vector with density f, and $A = \{\tilde{x} \in B_0\}$, $B_0 \in \mathcal{B}(\mathbb{R}^n)$, show that there is a conditional density for $\tilde{x}$ given A, namely,

$$f_{\tilde{x}|A}\,(x\,|A\,) = \begin{cases} \dfrac{f(x)}{P(A)} & \text{if } x \in B_0 \\ \\ 0 & \text{if } x \notin B_0. \end{cases}$$

Note: The interpretation of the conditional density is that

$$P\,\{\tilde{x} \in B\,|A\}\,\} = \int_B f_{\tilde{x}|A}\,(x\,|A\,)\,dx, \quad B \in \mathcal{B}(\mathbb{R}^n).$$

15. Assume that the distribution of the random variable $\tilde{x}$ has the following density:

$$f(x) = \begin{cases} 1 & \text{if } x \in (0,1) \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Use the definition of conditional probability given in Chapter 1 (Probability) to compute

$$P\,\{\tilde{x} < 1/3\,|\,\tilde{x} < 1/2\}\,,\ P\,\{\tilde{x} < 3/4\,|\,\tilde{x} < 1/2\}\,,\ \text{and } P\,\{1/3 < \tilde{x} < 3/4\,|\,\tilde{x} < 1/2\}\,.$$

(b) Use the conditional density found in Exercise 14 of this list to compute the same conditional probabilities of part (a).

(c) Consider the random variable $\tilde{x}$ having the distribution $P_{\tilde{x}}$. Find the conditional probability $P\,\{\tilde{x} \in B\,|\,\tilde{x} = x\,\}$ and check that it satisfies the theorem of total probability,

$$P\,\{\tilde{x} \in B\}\,\} = \int_{\mathbb{R}} P\,\{\tilde{x} \in B\,|\,\tilde{x} = x\,\}\,dP_{\tilde{x}}(x).$$

16. If the joint density of $\tilde{x}$ and $\tilde{y}$ is given by

$$f\,(x,y) = \begin{cases} x + y & \text{for } 0 < x < 1,\ 0 < y < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

find the corresponding joint distribution function $F\,(x,y)$. Check that

$$\frac{\partial^2 F\,(x,y)}{\partial x \partial y} = f\,(x,y) \quad \text{a.e. with respect to Lebesgue measure.}$$

17. The random vector $(\tilde{x}, \tilde{y})$ is distributed according to the following joint density function:

$$f(x,y) = \begin{cases} c(x^2 - y) & \text{if } x \in (1,2)\,,\ y \in (0,1) \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of the constant c.

(b) Find the marginal density function of $\tilde{x}$, $f_{\tilde{x}}(x)$, and the marginal distribution function of $\tilde{y}$, $F_{\tilde{y}}(y)$.

(c) Are $\tilde{x}$ and $\tilde{y}$ independent random variables?

(d) Find the conditional density function of $\tilde{x}$ given $\tilde{y} = 1/4$, $f_{\tilde{x}|\tilde{y}}\left(x \,\middle|\, \frac{1}{4}\right)$.

18. The joint and marginal distribution functions of the random variables $\tilde{x}$ and $\tilde{y}$ are $F(x,y)$, $F_{\tilde{x}}(x)$ and $F_{\tilde{y}}(y)$. Let us consider the following functions:

$$F^{+}(x,y) = \min\left\{F_{\tilde{x}}(x), F_{\tilde{y}}(y)\right\},$$

and

$$F^{-}(x,y) = \max\left\{F_{\tilde{x}}(x) + F_{\tilde{y}}(y) - 1, 0\right\}.$$

(a) Prove that both $F^{+}(x,y)$ and $F^{-}(x,y)$ are distribution functions associated with probability measures on the $\sigma-$algebra $\mathcal{B}(\mathbb{R}^2)$ of the Borel sets of $\mathbb{R}^2$.

(b) Prove that the marginal distribution functions of both $F^{+}(x,y)$ and $F^{-}(x,y)$ are also $F_{\tilde{x}}(x)$ and $F_{\tilde{y}}(y)$.

(c) Prove that $F^{-}(x,y) \leq F(x,y) \leq F^{+}(x,y)$ for all $(x,y) \in \mathbb{R}^2$.

3. Expectation

3.1. Mathematical expectation

- **Definition.** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ be an extended random variable with the distribution $P_{\tilde{x}}$. The (mathematical or arithmetic) expectation (or expected value or mean) of $\tilde{x}$ (or of its distribution) is defined by

$$\mathrm{E}(\tilde{x}) = \int_{\Omega} \tilde{x}(\omega)\, dP(\omega) \quad \left(= \int_{\Omega} \tilde{x} dP\right)$$

or

$$\mathrm{E}(\tilde{x}) = \int_{\overline{\mathbb{R}}} x dP_{\tilde{x}}(x) \quad \left(= \int_{\overline{\mathbb{R}}} x dP_{\tilde{x}}\right), \qquad (\star)$$

whenever these integrals exist, i.e., whenever they are not of the form $\infty - \infty$ (indefinite).

- Obviously, if $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a (real-valued) random variable, we can replace $\overline{\mathbb{R}}$ by $\mathbb{R}$ in $(\star)$ and use the following notation:

$$\mathrm{E}(\tilde{x}) = \int_{\mathbb{R}} x dF_{\tilde{x}}(x) \quad \left(= \int_{\mathbb{R}} x dF_{\tilde{x}} = \int_{\mathbb{R}} x dP_{\tilde{x}}\right).$$

- Recall that

$$\mathrm{E}(\tilde{x}) = \mathrm{E}(\tilde{x}^+) - \mathrm{E}(\tilde{x}^-) = \int_{\{\tilde{x}\geq 0\}} \tilde{x} dP - \int_{\{\tilde{x}\leq 0\}} (-\tilde{x})\, dP.$$

- From now on, when we write an expectation, it should be understood that the expectation exists.
- If $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a discrete random variable with probability function $f_{\tilde{x}} : \tilde{x}(\Omega) \longrightarrow [0,1]$,

$$\mathrm{E}(\tilde{x}) = \sum_{x\in\tilde{x}(\Omega)} x\ f_{\tilde{x}}(x).$$

- If $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ is an absolutely continuous random variable with density $f_{\tilde{x}} : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$,

$$\mathrm{E}(\tilde{x}) = \int_{\mathbb{R}} x\ f_{\tilde{x}}(x) dx.$$

- See the handout for the proof of the previous "intuitive" formula.

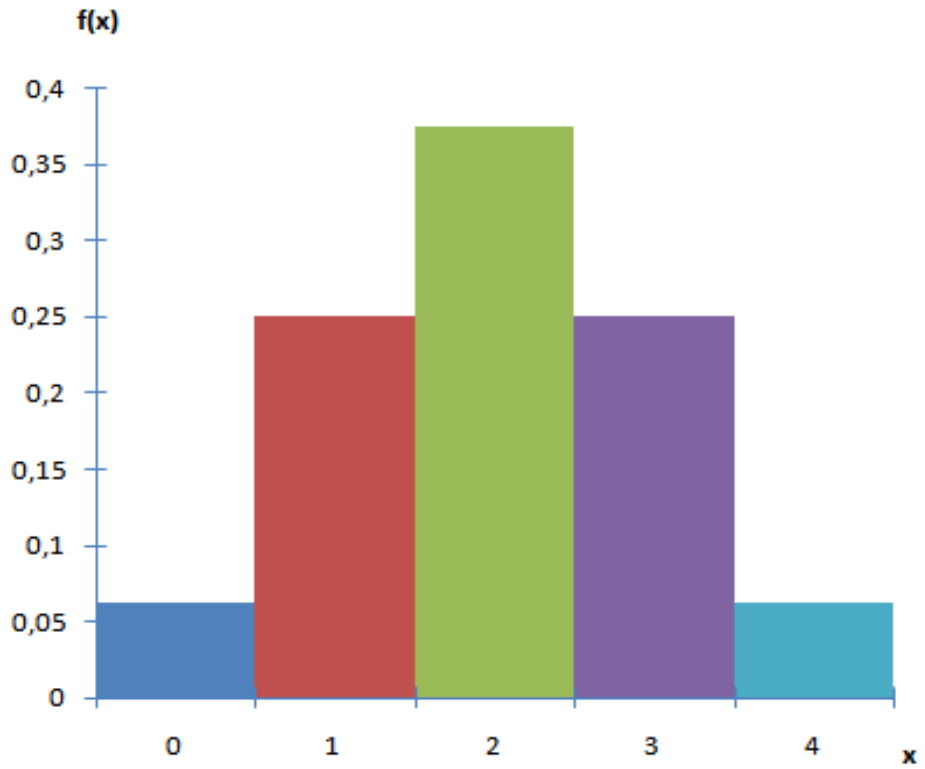
- **Example 1:**

  Let $\tilde{x}$ be the number of heads when tossing 4 coins.

$$f_{\tilde{x}}(x) = \begin{cases} 1/16 & \text{for } x = 0 \\ 4/16 & \text{for } x = 1 \\ 6/16 & \text{for } x = 2 \\ 4/16 & \text{for } x = 3 \\ 1/16 & \text{for } x = 4, \end{cases}$$

  or

$$f_{\tilde{x}}(x) = \frac{1}{16}\binom{4}{x}, \quad \text{for } x = \underbrace{0, 1, 2, 3, 4}_{\tilde{x}(\Omega)}.$$



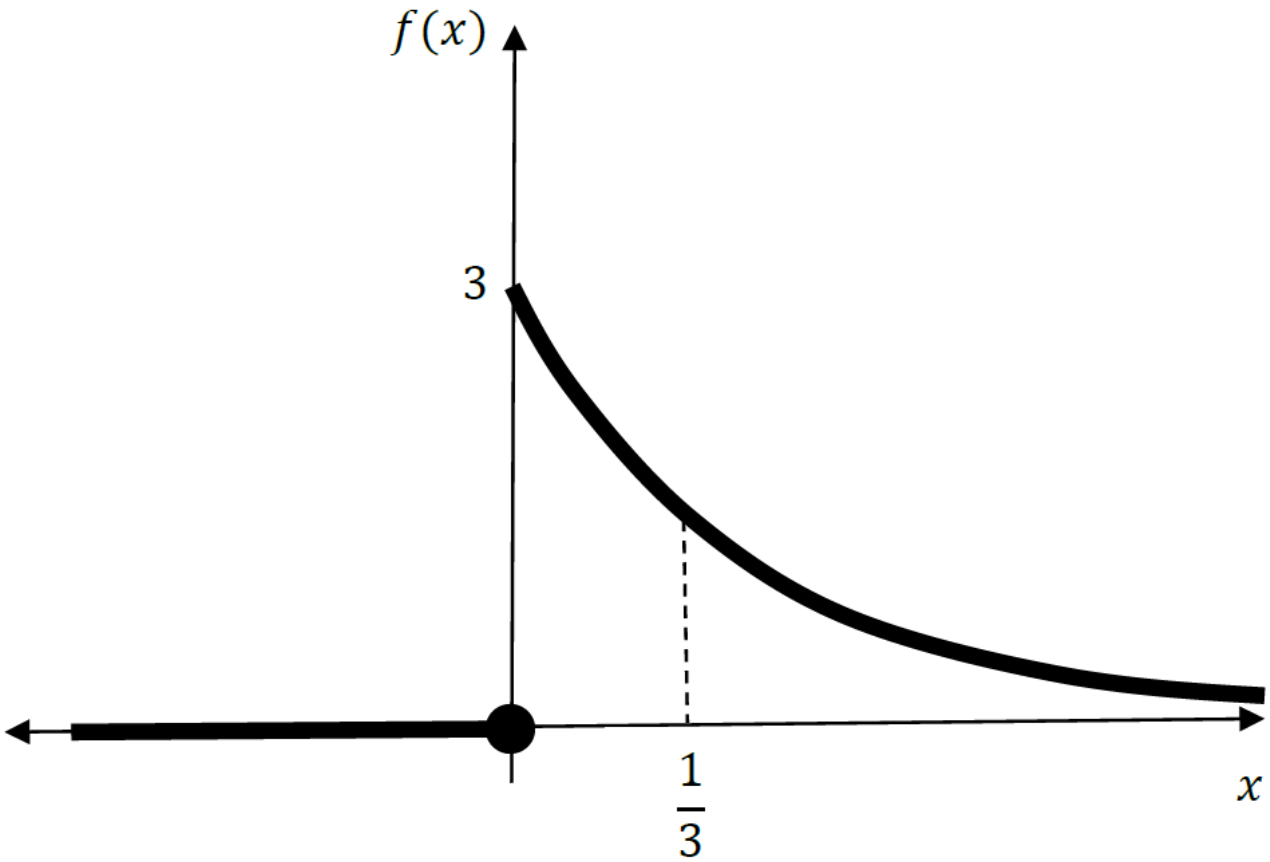
$$\mathrm{E}(\tilde{x}) = \left(0 \cdot \tfrac{1}{16}\right) + \left(1 \cdot \tfrac{4}{16}\right) + \left(2 \cdot \tfrac{6}{16}\right) + \left(3 \cdot \tfrac{4}{16}\right) + \left(4 \cdot \tfrac{1}{16}\right) = 2.$$

- The median of $\tilde{x}$ (or of its distribution) is the real number m for which $P\{\tilde{x} \leq m\} \geq 1/2$ and $P\{\tilde{x} \geq m\} \geq 1/2$. If there are two values satisfying this definition, we take the average of them. Obviously, $m = 2 = \mathrm{E}(\tilde{x})$ in this example.

- **Example 2:**

  The random variable $\tilde{x}$ has the following density:

$$f(x) = \begin{cases} 3e^{-3x} & \text{for } x > 0 \\ \\ 0 & \text{for } x \leq 0 \end{cases}$$



$$\mathrm{E}(\tilde{x}) = \int_{-\infty}^{\infty} x\, f(x)dx = \int_{-\infty}^{0} x \cdot 0 \cdot dx + \int_{0}^{\infty} \underbrace{x}_{H(x)} \underbrace{3e^{-3x}}_{g(x)} dx$$

$$= 0 + \left[H(x)G(x)\right]_0^{\infty} - \int_0^{\infty} h(x)G(x)dx$$

$$= \underbrace{\left[-xe^{-3x}\right]_0^{\infty}}_{-\lim_{x\to\infty}\left(\frac{x}{e^{3x}}\right)+0=-\lim_{x\to\infty}\left(\frac{1}{3e^{3x}}\right)=0} - \int_0^{\infty} \underbrace{1}_{h(x)} \cdot \underbrace{\left(-e^{-3x}\right)}_{G(x)} dx = \left[\frac{e^{-3x}}{-3}\right]_0^{\infty} = \frac{1}{3},$$

↑ L'Hôpital's rule

where $H' = h$ and $G' = g$.

- Since the distribution of $\tilde{x}$ is absolutely continuous in this example, its median m satisfies $P\{\tilde{x} \leq m\} = P\{\tilde{x} \geq m\}$ and, thus,

$$P\{\tilde{x} \leq m\} = F(m) = \int_0^m 3e^{-3x}dx = 1 - e^{-3m} = \frac{1}{2}$$

$$\Longrightarrow m = \frac{1}{3}\ln 2 = 0.23105 < \frac{1}{3} = \mathrm{E}(\tilde{x}).$$

3.2. Properties of mathematical expectation

- **Proposition (the law of the unconscious statistician - LOTUS).** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object with the distribution $P_{\tilde{x}}$ and let $g : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}, \mathcal{B})$ be a Borel measurable function. If $\tilde{y} = g(\tilde{x})$, then

$$\mathrm{E}(\tilde{y}) = \int_{\Omega} g(\tilde{x})\, dP = \int_{\Omega'} g(x) dP_{\tilde{x}}(x).$$

- **Corollary 1.** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be a random vector with the distribution $P_{\tilde{x}}$ and let $g : (\mathbb{R}^n, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$ be a Borel measurable function. If $\tilde{y} = g(\tilde{x})$, then

$$\mathrm{E}(\tilde{y}) = \mathrm{E}(g(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)) =$$

$$\int_{\Omega} g(\tilde{x})\, dP = \int_{\mathbb{R}^n} g(\underbrace{x_1, x_2, ..., x_n}_{x \in \mathbb{R}^n}) dP_{\tilde{x}}(x_1, x_2, ..., x_n).$$

- **Corollary 2.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be a random vector with the distribution $P_{\tilde{x}}$, then

$$\mathrm{E}(\tilde{x}_i) = \int_{\Omega} \tilde{x}_i \, dP = \int_{\mathbb{R}^n} x_i \, dP_{\tilde{x}}(x_1, x_2, ..., x_n) \quad \text{for } i = 1, 2, ..., n.$$

- **Proof:** Just make $g = p_i$ in Corollary 1, where p_i is the projection map to the ith coordinate, that is, $p_i(x_1, x_2, ..., x_n) = x_i$.

- **Corollary 3.** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ be a random variable with the distribution $P_{\tilde{x}}$ and let $g : (\mathbb{R}, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$ be a Borel measurable function. If $\tilde{y} = g(\tilde{x})$, then

$$\mathrm{E}(\tilde{y}) = \int_{\Omega} g(\tilde{x}) \, dP = \int_{\mathbb{R}} g(x) dP_{\tilde{x}}(x).$$

- **Geometric** and **Harmonic** expectation (or mean) of a non-negative extended random variable, $\tilde{x} \geq 0$:

$$\mathrm{GE}(\tilde{x}) = \exp\left(\mathrm{E}\left[\ln \tilde{x}\right]\right) \quad \text{and} \quad \mathrm{HE}(\tilde{x}) = \left(\mathrm{E}\left[\tilde{x}^{-1}\right]\right)^{-1} = \frac{1}{\mathrm{E}\left[1/\tilde{x}\right]}.$$

- If $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a discrete random variable with probability function $f_{\tilde{x}} : \tilde{x}(\Omega) \longrightarrow [0,1]$,

$$\mathrm{GE}(\tilde{x}) = \exp\left(\mathrm{E}\left[\ln \tilde{x}\right]\right) = \exp\left(\sum_{x \in \tilde{x}(\Omega)} f_{\tilde{x}}(x) \cdot \ln x\right)$$

$$= \exp\left(\sum_{x \in \tilde{x}(\Omega)} \ln\left[x^{f_{\tilde{x}}(x)}\right]\right) = \exp\left(\ln\left[\prod_{x \in \tilde{x}(\Omega)} x^{f_{\tilde{x}}(x)}\right]\right) = \prod_{x \in \tilde{x}(\Omega)} x^{f_{\tilde{x}}(x)}$$

and

$$\mathrm{HE}(\tilde{x}) = \left[\sum_{x \in \tilde{x}(\Omega)} \frac{f_{\tilde{x}}(x)}{x}\right]^{-1}.$$

- *Note:* Probability and distribution are expectations:
- Consider the measure space $(\Omega, \mathcal{F}, P)$, then

$$P(B) = \int_B dP = \int_\Omega \mathbb{I}_B dP = \mathrm{E}\left(\mathbb{I}_B\right), \text{ for all } B \in \mathcal{F}.$$

- Consider the random object $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$, then

$$P_{\tilde{x}}(A) = P\left\{\tilde{x} \in A\right\} = \int_A dP_{\tilde{x}}(x) = \int_\Omega \mathbb{I}_A\left(\tilde{x}\left(\omega\right)\right) dP(\omega)$$

$$= \int_{\Omega'} \mathbb{I}_A\left(x\right) dP_{\tilde{x}}(x) = \mathrm{E}\left[\mathbb{I}_A\left(\tilde{x}\right)\right], \text{ for all } A \in \mathcal{F}',$$

- or

$$P_{\tilde{x}}(A) = P\left(\tilde{x}^{-1}(A)\right) = \mathrm{E}\left[\mathbb{I}_{\tilde{x}^{-1}(A)}\right], \text{ for all } A \in \mathcal{F}'.$$

- **Proposition.** Let $\tilde{x}(\omega) = b$ a.s. $[P]$, where b is a constant. Then, $\mathrm{E}(\tilde{x}) \equiv \int_{\Omega} \tilde{x} dP = b$.

- **Proposition.** If c is a constant, then $\mathrm{E}(c\tilde{x}) = c\mathrm{E}(\tilde{x})$.

- **Proposition.** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object, $\{c_i\}_{i=1}^n$ a collection of constants, and $g_i : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}, \mathcal{B})$ measurable functions, for $i = 1, ..., n$. Then,

$$\mathrm{E}\left[\sum_{i=1}^{n} c_i g_i(\tilde{x})\right] = \sum_{i=1}^{n} c_i \mathrm{E}\left(g_i(\tilde{x})\right).$$

- **Corollary.** Let $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ be a collection of random variables on $(\Omega, \mathcal{F}, P)$ and $\{c_i\}_{i=1}^n$ a collection of constants. Then,

$$\mathrm{E}\left[\sum_{i=1}^{n} c_i \tilde{x}_i\right] = \sum_{i=1}^{n} c_i \mathrm{E}\left(\tilde{x}_i\right).$$

3.3. L^p spaces and the Cauchy-Schwarz inequality

- **Definition.** Assume a given measure space $(\Omega, \mathcal{F}, \mu)$ and let $f : (\Omega, \mathcal{F}, \mu) \longrightarrow (\mathbb{R}, \mathcal{B})$ and p be a non-negative real number, $p \geq 0$. We say that $f \in \mathrm{L}^p$ (or $f \in \mathrm{L}^p(\Omega, \mathcal{F}, \mu)$) if

$$\int_\Omega |f|^p \, d\mu < \infty.$$

  i.e., $|f|^p$ is μ-integrable.

- **Definition.** $||f||_p = \left[\int_\Omega |f|^p \, d\mu\right]^{1/p}$.

- **Hölder inequality:** Let $f \in \mathrm{L}^p$ and $g \in \mathrm{L}^q$ with $p > 1$, $q > 1$, and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$||fg||_1 \leq ||f||_p \cdot ||g||_q,$$

  that is,

$$\int_\Omega |fg| \, d\mu \leq \left[\int_\Omega |f|^p \, d\mu\right]^{1/p} \cdot \left[\int_\Omega |g|^q \, d\mu\right]^{1/q}.$$

- The **Cauchy-Schwarz inequality** is the **Hölder inequality** for $p = q = 2$, i.e., if $f \in \mathrm{L}^2$ and $g \in \mathrm{L}^2$ then

$$||fg||_1 \leq ||f||_2 \cdot ||g||_2 ,$$

- or

$$\int_\Omega |fg| \, d\mu \leq \left[\int_\Omega |f|^2 \, d\mu \right]^{1/2} \cdot \left[\int_\Omega |g|^2 \, d\mu \right]^{1/2} ,$$

- or

$$\int_\Omega |fg| \, d\mu \leq \left[\int_\Omega f^2 d\mu \right]^{1/2} \cdot \left[\int_\Omega g^2 d\mu \right]^{1/2} .$$

- Remember that

$$\int_\Omega fg d\mu \leq \left| \int_\Omega fg d\mu \right| \leq \int_\Omega |fg| \, d\mu .$$

Proof of the Cauchy-Schwarz inequality: For every constant K and any pair of functions f and g belonging to L^2, we have

$$\int_{\Omega} \left(K\left|f\right| + \left|g\right|\right)^2 d\mu \geq 0$$

and, thus,

$$\int_{\Omega} K^2 \left|f\right|^2 d\mu + \int_{\Omega} 2K\left|f\right|\left|g\right| d\mu + \int_{\Omega} \left|g\right|^2 d\mu \geq 0.$$

The previous inequality becomes

$$K^2 \underbrace{\int_{\Omega} f^2 d\mu}_{a} + K \cdot \underbrace{2\int_{\Omega} \left|fg\right| d\mu}_{b} + \underbrace{\int_{\Omega} g^2 d\mu}_{c} \geq 0.$$

This means that the above second order polynomial of K can have one or zero roots. That is, the discriminant $D = b^2 - 4ac$ has to be less than or equal to zero. Therefore,

$$\left[2\int_\Omega |fg|\, d\mu\right]^2 - 4\left[\int_\Omega f^2 d\mu\right]\left[\int_\Omega g^2 d\mu\right] \leq 0.$$

Rearranging the above inequality and dividing by 4, one obtains

$$\left[\int_\Omega |fg|\, d\mu\right]^2 \leq \left[\int_\Omega f^2 d\mu\right] \cdot \left[\int_\Omega g^2 d\mu\right]$$

and taking the positive square root in both sides, we obtain the desired inequality. *Q.E.D.*

3.4. Moments and its properties: mean and variance

- Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$.
- **Definition.** Let k be a natural number, $k = 0, 1, 2, ...$ The kth (non-central) moment of $\tilde{x}$ is

$$\mu'_k(\tilde{x}) = \mathrm{E}\left(\tilde{x}^k\right) = \int_\Omega \tilde{x}^k dP = \int_{\mathbb{R}} x^k dP_{\tilde{x}}(x).$$

- **Definition.** Let k be a natural number, $k = 0, 1, 2, ...$, and assume that $\mathrm{E}(\tilde{x})$ is finite. The kth central moment of $\tilde{x}$ is

$$\begin{aligned}\mu_k(\tilde{x}) &= \mathrm{E}\left[(\tilde{x} - \mathrm{E}(\tilde{x}))^k\right] \\ &= \int_\Omega (\tilde{x} - \mathrm{E}(\tilde{x}))^k dP = \int_{\mathbb{R}} (x - \mathrm{E}(\tilde{x}))^k dP_{\tilde{x}}(x).\end{aligned}$$

- Odd moments may fail to exist and, whenever they exist, they may be equal to $\pm\infty$.
- Even moments always exist but they may be equal to $+\infty$.

- Note that the kth moment of $\tilde{x}$ is finite if and only if $\tilde{x}^k$ is P-integrable. Therefore, the finiteness of $\mathrm{E}\left(\tilde{x}^k\right)$ requires the previous existence of the integral $\int_\Omega \tilde{x}^k dP$.

- Let k be a natural number, $k = 0, 1, 2, ...$ Since

$$\mathrm{E}\left(\tilde{x}^k\right) = \mathrm{E}\left[\left(\tilde{x}^k\right)^+\right] - \mathrm{E}\left[\left(\tilde{x}^k\right)^-\right]$$

and

$$\mathrm{E}\left(\left|\tilde{x}\right|^k\right) = \mathrm{E}\left(\left|\tilde{x}^k\right|\right) = \mathrm{E}\left[\left(\tilde{x}^k\right)^+\right] + \mathrm{E}\left[\left(\tilde{x}^k\right)^-\right],$$

then $\mathrm{E}\left(\tilde{x}^k\right)$ is finite if and only if $\mathrm{E}\left(\left|\tilde{x}\right|^k\right)$ is finite. Note that $\mathrm{E}\left(\left|\tilde{x}\right|^k\right)$ always exists.

- In particular, $\mathrm{E}\left(\tilde{x}\right)$ is finite if and only if $\mathrm{E}\left(\left|\tilde{x}\right|\right)$ is finite. Moreover, if $\mathrm{E}\left(\tilde{x}\right)$ does not exist (it is of the indefinite form $\infty - \infty$) or is equal to $\pm\infty$, then $\mathrm{E}\left(\left|\tilde{x}\right|\right) = \infty$.

- **Proposition.** If k is a natural number, $k = 0, 1, 2, ...$ and $\mathrm{E}\left(\tilde{x}^k\right)$ is finite, then $\mathrm{E}\left(\tilde{x}^j\right)$ is finite for all natural numbers j smaller or equal than k, $j = 0, 1, ..., k$.

- Recall that $\mathrm{E}\left(\tilde{x}^k\right)$ is finite if and only if $\mathrm{E}\left(\left|\tilde{x}\right|^k\right)$ is finite, where $k = 0, 1, 2, ...$ Therefore, the proposition says that $\tilde{x} \in \mathrm{L}^k\left(\Omega, \mathcal{F}, P\right) \Longrightarrow \tilde{x} \in \mathrm{L}^j\left(\Omega, \mathcal{F}, P\right)$ for all $j = 0, 1, ..., k$, where k is a natural number.

- The proposition follows immediately from the following lemma:

- **Lemma:** If $\tilde{x} \in \mathrm{L}^k$, where k is a non-negative real number, then $\tilde{x} \in \mathrm{L}^j$ for all real numbers $j \in [0, k]$.

- Hence, the lemma says that $\mathrm{L}^k\left(\Omega, \mathcal{F}, P\right) \subset \mathrm{L}^j\left(\Omega, \mathcal{F}, P\right)$ for $0 \leq j \leq k$. This inclusion relationship does not hold in general for measures that are not finite.

Proof of the Lemma: We will prove that $\mathrm{E}\left(|\tilde{x}|^j\right) < \infty$ for all real numbers $j \in [0, k]$ when $\mathrm{E}\left(|\tilde{x}|^k\right) < \infty$ with $k \geq 0$. The cases where either $j = 0$ or $k = 0$ are trivial so that we only consider the case where $k > 0$ and $j \in (0, k]$ as follows:

$$\mathrm{E}\left(|\tilde{x}|^j\right) = \int_\Omega |\tilde{x}|^j \, dP = \int_{\left\{|\tilde{x}|^j < 1\right\}} |\tilde{x}|^j \, dP + \int_{\left\{|\tilde{x}|^j \geq 1\right\}} |\tilde{x}|^j \, dP$$

$$\leq \int_{\left\{|\tilde{x}|^j < 1\right\}} dP + \int_{\left\{|\tilde{x}|^j \geq 1\right\}} \left(|\tilde{x}|^j\right)^{k/j} dP + \int_{\left\{|\tilde{x}|^j < 1\right\}} |\tilde{x}|^k \, dP =$$

$$\int_{\left\{|\tilde{x}|^j < 1\right\}} dP + \int_\Omega |\tilde{x}|^k \, dP = P\left\{|\tilde{x}|^j < 1\right\} + \mathrm{E}\left(|\tilde{x}|^k\right) < \infty. \qquad Q.E.D.$$

- **Note 1:** If $\mathrm{E}(\tilde{x})$ fails to exist (it is of the form $\infty - \infty$) or is equal to $\pm\infty$, then $\mathrm{E}(\tilde{x}^k) = \infty$ when k is an even strictly positive natural number, $k = 2, 4, 6, ...$
- To see this observe that in this case $\mathrm{E}(|\tilde{x}|) = \infty$ and $\mathrm{E}(\tilde{x}^k) = \mathrm{E}\left(|\tilde{x}|^k\right) \geq 0$ when k is even. According to the previous Lemma, if $\mathrm{E}(|\tilde{x}|) = \infty$ then $\mathrm{E}\left(|\tilde{x}|^k\right) = \mathrm{E}(\tilde{x}^k) = \infty$.
- **Note 2:** If $\mathrm{E}(\tilde{x})$ fails to exist (i.e., it is of the form $\infty - \infty$), then $\mathrm{E}(\tilde{x}^k)$ fails to exists when k is an odd natural number, $k = 1, 3, 5, ...$
- To see this observe that in this case, since $\tilde{x}^+ \geq 0$ and $\tilde{x}^- \geq 0$, $\mathrm{E}(\tilde{x}) = \mathrm{E}(\tilde{x}^+) - \mathrm{E}(\tilde{x}^-) = \mathrm{E}(|\tilde{x}^+|) - \mathrm{E}(|\tilde{x}^-|) = \infty - \infty$. Thus, $\mathrm{E}(|\tilde{x}^+|) = \infty$ and $\mathrm{E}(|\tilde{x}^-|) = \infty$. According to the previous Lemma, $\mathrm{E}(|\tilde{x}^+|) = \infty$ implies that $\mathrm{E}\left[(|\tilde{x}^+|)^k\right] = \mathrm{E}\left[(\tilde{x}^k)^+\right] = \infty$ and $\mathrm{E}[|\tilde{x}^-|] = \infty$ implies that $\mathrm{E}\left[(|\tilde{x}^-|)^k\right] = \mathrm{E}\left[(\tilde{x}^k)^-\right] = \infty$. Therefore, $\mathrm{E}(\tilde{x}^k) = \mathrm{E}\left[(\tilde{x}^k)^+\right] - \mathrm{E}\left[(\tilde{x}^k)^-\right]$ is of the form $\infty - \infty$.

- **Note 3:** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$ be an extended real-valued random variable. If $\mathrm{E}(\tilde{x})$ is finite, then $\tilde{x}$ is finite a.e. with respect to P.

- From now on, when we write a moment $\mu'_k(\tilde{x})$ of a (extended) random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow \left(\overline{\mathbb{R}}, \mathcal{B}\right)$, it should be understood that this moment exists and is finite, that is, $\tilde{x} \in \mathrm{L}^k(\Omega, \mathcal{F}, P)$.

- *Note:*

$$\mu_0'(\tilde{x}) = \mu_0(\tilde{x}) = \mathrm{E}(\tilde{x}^0) = \mathrm{E}(1) = 1$$

and

$$\mu_1(\tilde{x}) = \mathrm{E}[\tilde{x} - \mathrm{E}(\tilde{x})] = \mathrm{E}(\tilde{x}) - \mathrm{E}(\tilde{x}) = 0.$$

- μ_1' is the mean of $\tilde{x}$ and is denoted by μ (or $\mu_{\tilde{x}}$),

$$\mu = \mu_1' = \mathrm{E}(\tilde{x}).$$

- μ_2 is the variance of $\tilde{x}$ and is denoted by σ^2, $\sigma_{\tilde{x}}^2$ or $\mathrm{Var}(\tilde{x})$,

$$\sigma^2 = \mathrm{Var}(\tilde{x}) = \mathrm{E}\left[(\tilde{x} - \mathrm{E}(\tilde{x}))^2\right].$$

- The inverse of the variance of $\tilde{x}$ is called the precision of $\tilde{x}$,

$$\tau_{\tilde{x}} = \frac{1}{\sigma_{\tilde{x}}^2}.$$

- **Proposition.** Let $\tilde{x}(\omega) = b$ a.s. $[P]$, where b is a constant. Then $\text{Var}(\tilde{x}) = \text{Var}(b) = 0$.
- **Proposition.** If c is a constant, then $\text{Var}(c\tilde{x}) = c^2\text{Var}(\tilde{x})$.
- **Definition.** The standard deviation $\sigma_{\tilde{x}}$ (or σ) of $\tilde{x}$ is

$$\sigma_{\tilde{x}} = [\text{Var}(\tilde{x})]^{1/2}.$$

- **Proposition.** If $c > 0$ is a constant, then $\sigma_{c\tilde{x}} = c\sigma_{\tilde{x}}$.
- **Definition.** The coefficient of variation of $\tilde{x}$ is

$$\text{CV}_{\tilde{x}} = \frac{\sigma}{|\mu|}, \quad \text{with } \mu \neq 0.$$

- **Proposition.** If c is a constant, then $\text{CV}_{c\tilde{x}} = \text{CV}_{\tilde{x}}$.

- **Proposition.**

$$\text{Var}(\tilde{x}) = \text{E}\left(\tilde{x}^2\right) - \left[\text{E}\left(\tilde{x}\right)\right]^2$$

or

$$\sigma^2 \equiv \mu_2 = \mu_2' - \left(\mu_1'\right)^2 \equiv \mu_2' - \mu^2.$$

- **Proof.**

$$\sigma^2 = \text{E}\left[\left(\tilde{x} - \mu\right)^2\right] = \text{E}\left[\tilde{x}^2 - 2\tilde{x}\mu + \mu^2\right] = \text{E}\left(\tilde{x}^2\right) - 2\mu\text{E}\left(\tilde{x}\right) + \mu^2$$

$$= \text{E}\left(\tilde{x}^2\right) - 2\mu^2 + \mu^2 = \text{E}\left(\tilde{x}^2\right) - \mu^2. \qquad Q.E.D$$

- In general: If k is a natural number, we get the following formula for the kth central moment by just making the binomial expansion:

$$\mathrm{E}\left[\left(\tilde{x}-\mathrm{E}\left(\tilde{x}\right)\right)^{k}\right]=\sum_{n=0}^{k}\binom{k}{n}\mathrm{E}\left(\tilde{x}^{n}\right)\left[-\mathrm{E}\left(\tilde{x}\right)\right]^{k-n}$$

or

$$\mu_{k}=\sum_{n=0}^{k}\binom{k}{n}\mu_{n}^{\prime}\left[-\mu\right]^{k-n}.$$

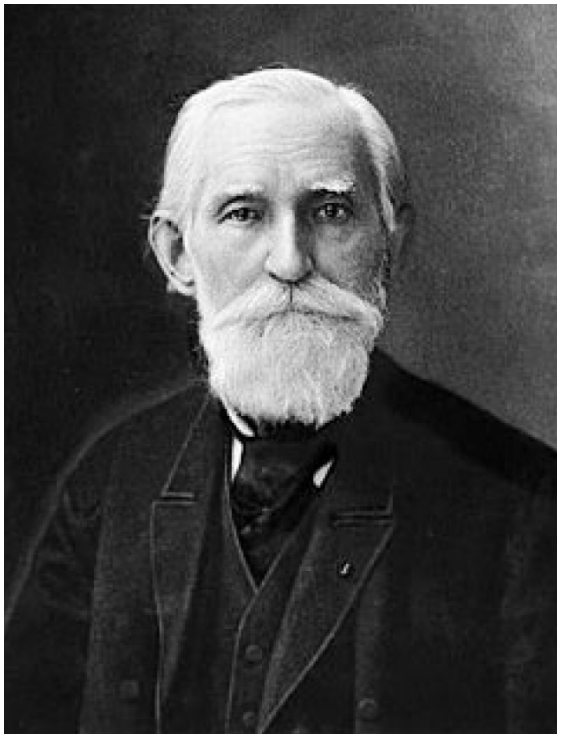
- Therefore, the central moment μ_k is finite if and only if the (non-central) moment $\mu_k^{\prime}$ is finite.
- **Definition.** The coefficient of skewness or asymmetry (for $\sigma^3 \neq 0$) is

$$\mathrm{CA}=\frac{\mu_{3}}{\sigma^{3}}.$$

- **Definition.** The coefficient of kurtosis (for $\sigma^4 > 0$) is

$$\mathrm{CK}=\frac{\mu_{4}}{\sigma^{4}}.$$

3.5. Chevyshev's inequality



Pafnuty Chebyshev (1821 -1894)

- **Theorem (Chevyshev).** Assume that the random variable $\tilde{x}$ has the finite mean μ and the finite variance $\sigma^2 > 0$. Then, for any $k \in (0, \infty)$,

$$P\left\{|\tilde{x} - \mu| < k\sigma\right\} \geq 1 - \frac{1}{k^2}, \qquad \text{(Chevyshev's inequality)}$$

or, equivalently,

$$P\left\{|\tilde{x} - \mu| \geq k\sigma\right\} \leq \frac{1}{k^2}. \qquad \text{(Chevyshev's inequality)}$$

- Making the change of variable $k\sigma = \varepsilon > 0$, Chevyshev's inequality becomes

$$P\left\{|\tilde{x} - \mu| < \varepsilon\right\} \geq 1 - \frac{\sigma^2}{\varepsilon^2}, \qquad \text{(Chevyshev's inequality)}$$

or, equivalently,

$$P\left\{|\tilde{x} - \mu| \geq \varepsilon\right\} \leq \frac{\sigma^2}{\varepsilon^2}. \qquad \text{(Chevyshev's inequality)}$$

- The Chevyshev's theorem says that a random variable with small variance is likely to take on values close to its mean.

Proof of the Theorem: We know from the definition of variance that

$$\sigma^2 = \mathrm{E}\left[(\tilde{x}-\mu)^2\right] = \int_{\mathbb{R}} (x-\mu)^2 \, dP_{\tilde{x}}(x)$$

and, dividing the integral into three parts, we get

$$\sigma^2 = \int_{(-\infty,\mu-k\sigma]} (x-\mu)^2 \, dP_{\tilde{x}}$$

$$+\int_{(\mu-k\sigma,\mu+k\sigma)} (x-\mu)^2 \, dP_{\tilde{x}} + \int_{[\mu+k\sigma,\infty)} (x-\mu)^2 \, dP_{\tilde{x}}.$$

Since $(x-\mu)^2$ is non-negative, the three integrals are also non-negative. Therefore, we can form the inequality

$$\sigma^2 \geq \int_{(-\infty,\mu-k\sigma]} (x-\mu)^2 \, dP_{\tilde{x}}(x) + \int_{[\mu+k\sigma,\infty)} (x-\mu)^2 \, dP_{\tilde{x}}(x).$$

by deleting the second integral.

Now, since $(x-\mu)^2 \geq k^2\sigma^2$ if either $x \leq \mu - k\sigma$ or $x \geq \mu + k\sigma$, it follows that

$$\sigma^2 \geq \int_{(-\infty,\mu-k\sigma]} k^2\sigma^2 dP_{\tilde{x}}(x) + \int_{[\mu+k\sigma,\infty)} k^2\sigma^2 dP_{\tilde{x}}(x)$$

and, hence, dividing both sides by $k^2\sigma^2$, we get

$$\frac{1}{k^2} \geq \int_{(-\infty,\mu-k\sigma]} dP_{\tilde{x}}(x) + \int_{[\mu+k\sigma,\infty)} dP_{\tilde{x}}(x),$$

provided $\sigma^2 > 0$. Since the sum of the two integrals in this inequality represents the probability that $\tilde{x}$ will take on a value less than or equal to $\mu - k\sigma$ or greater than or equal to $\mu + k\sigma$, we have thus shown that

$$P\{|\tilde{x} - \mu| \geq k\sigma\} \leq \frac{1}{k^2}. \qquad Q.E.D.$$

3.6. The moment-generating function

- **Definition.** The moment-generating function (MGF), $M_{\tilde{x}}(\cdot) : B \longrightarrow \mathbb{R}$, of the random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ (or of the distribution $P_{\tilde{x}}$) is

$$M_{\tilde{x}}(t) = \mathrm{E}(e^{t\tilde{x}}) = \int_{\Omega} e^{t\tilde{x}} dP = \int_{\mathbb{R}} e^{tx} dP_{\tilde{x}}(x) \geq 0.$$

  where B is the subset of $\mathbb{R}$ for which $\mathrm{E}(e^{t\tilde{x}})$ is finite.

- *Notes:*
- (1) $M_{\tilde{x}}(0) = 1$.
- (2) $M_{\tilde{x}}(t)$ could fail to be finite for some set of values of $t \in \mathbb{R}$.
- (3) If there exists a real number a such that $P\{\tilde{x} < a\} = 0$ $(P\{\tilde{x} > a\} = 0)$, then $M_{\tilde{x}}(t)$ is finite for all real numbers $t \leq (\geq) 0$.
- (4) If there exist two real numbers a and b such that $P\{a \leq \tilde{x} \leq b\} = 1$, then $M_{\tilde{x}}(t)$ is finite for all $t \in \mathbb{R}$.

- **Definition.** A neighborhood V of a point $x \in \mathbb{R}^n$ is a subset of $\mathbb{R}^n$ that includes an open set containing the point x.
- Note that, if V is a neighborhood of the point x, then x lies in the interior of V.
- **Proposition.** If there exists a neighborhood V of 0 for which $M_{\tilde{x}}(t)$ is finite for all $t \in V$, then $M_{\tilde{x}}(t)$ is infinitely differentiable at $t = 0$, the moments μ'_r of $\tilde{x}$ are finite for $r = 0, 1, 2, ...$, and

$$\left.\frac{d^r M_{\tilde{x}}(t)}{dt^r}\right|_{t=0} = \mu'_r, \quad \text{for } r = 0, 1, 2, ...$$

- **Proof.** Since, from the Taylor's expansion of e^y at $y = 0$, we have

$$e^y = \sum_{r=0}^{\infty} \frac{y^r}{r!},$$

it follows that

$$e^{tx} = \sum_{r=0}^{\infty} \frac{(tx)^r}{r!},$$

and, thus,

$$M_{\tilde{x}}(t) = \int_{\mathbb{R}} \overbrace{\left[\sum_{r=0}^{\infty} \frac{(tx)^r}{r!}\right]}^{e^{tx}} dP_{\tilde{x}}(x) = \sum_{r=0}^{\infty} \left[\int_{\mathbb{R}} \frac{(tx)^r}{r!} dP_{\tilde{x}}(x)\right] =$$

$$\sum_{r=0}^{\infty} \frac{t^r}{r!} \left(\underbrace{\int_{\mathbb{R}} x^r dP_{\tilde{x}}(x)}_{\mu'_r}\right) = 1 + \mu'_1 t + \mu'_2 \frac{t^2}{2!} + \mu'_3 \frac{t^3}{3!} + ... + \mu'_r \frac{t^r}{r!} + ...$$

Since $M_{\tilde{x}}(t) < \infty$ for all t in a neighborhood of $t = 0$, it holds that μ'_r is finite for $r = 0, 1, 2, ...$ and, thus, $M_{\tilde{x}}(t)$ is infinitely differentiable at $t = 0$. Finally,

$$\left.\frac{d^r M_{\tilde{x}}(t)}{dt^r}\right|_{t=0} = \left.\frac{d^r \mathrm{E}(e^{t\tilde{x}})}{dt^r}\right|_{t=0} = \mathrm{E}(\tilde{x}^r e^{t\tilde{x}})\big|_{t=0} = \mathrm{E}(\tilde{x}^r) = \mu'_r. \quad Q.E.D.$$

- Therefore, if some moments of a random variable $\tilde{x}$ do not exist (they are of the indefinite form $\infty - \infty$) or are not finite, then $M_{\tilde{x}}(t)$ fails to be finite in a neighborhood of $t = 0$ (and, thus, it fails to be infinitely differentiable at $t = 0$).
- However, even if a random variable has all its moments finite, the MGF may fail to be finite in a neighborhood of $t = 0$ (this is the case for the log-normal distribution).
- Since

$$M''_{\tilde{x}}(t) \equiv \frac{d^2 M_{\tilde{x}}(t)}{dt^2} = \mathrm{E}\underbrace{\left(\tilde{x}^2 e^{t\tilde{x}}\right)}_{\geq 0} \geq 0,$$

  the moment-generating function $M_{\tilde{x}}(t)$ is convex.
- Moreover, since

$$M'_{\tilde{x}}(t)\big|_{t=0} \equiv \left.\frac{dM_{\tilde{x}}(t)}{dt}\right|_{t=0} = \mathrm{E}\left(\tilde{x}\right),$$

  the slope of the moment-generating function $M_{\tilde{x}}(t)$ at $t = 0$ is the mean of $\tilde{x}$.

- **Examples:**

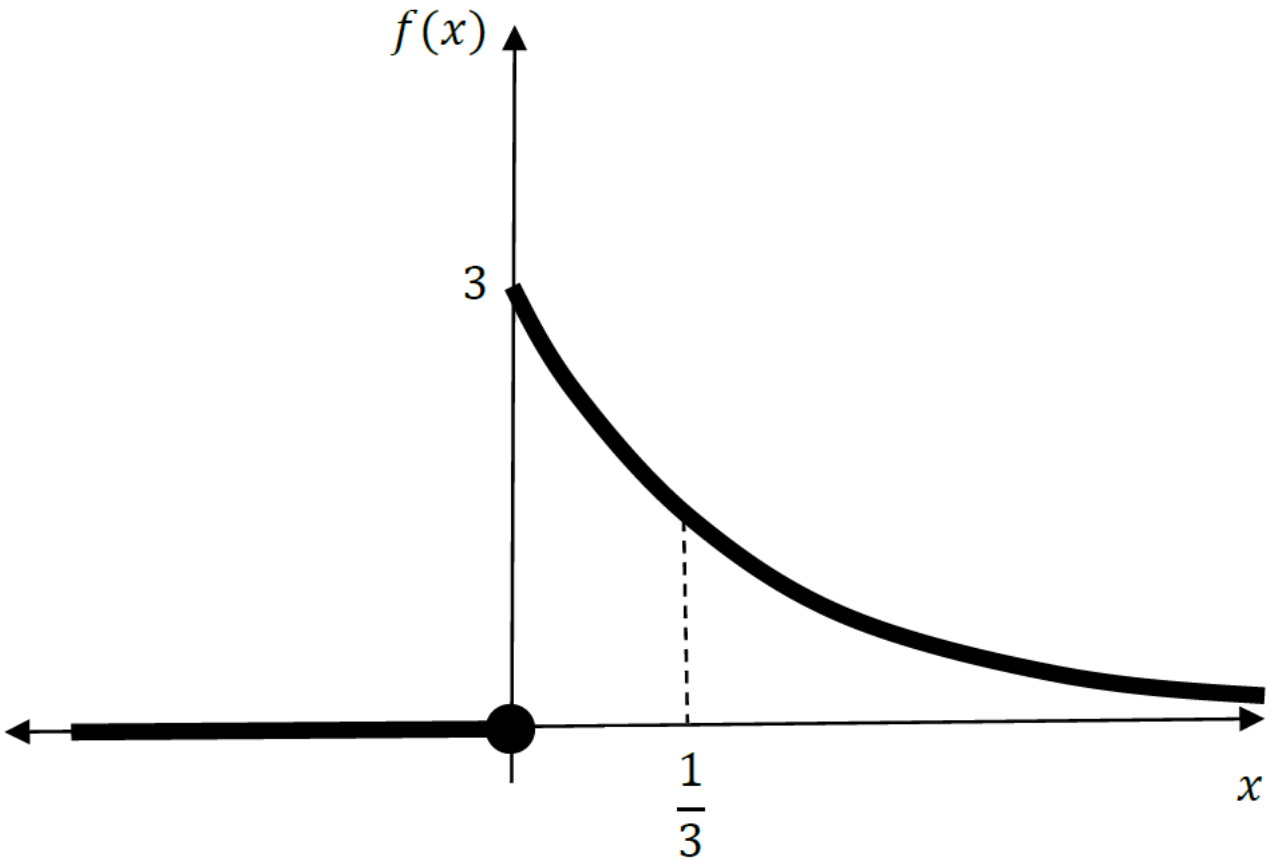
- The log-normal and F distributions have positive densities only for positive values. The Pareto distribution has positive density only for values larger than a strictly positive threshold. The Pareto, F, and log-normal distributions exhibit MGF's that are finite if and only if $t \leq 0$. The Student's t and Cauchy distributions have positive densities everywhere and exhibit MGF's that are not well defined (i.e., not finite) for all $t \neq 0$.

- The log-normal distribution has all its moments finite. The Pareto, F, and Student's t distributions could have some moments μ'_r, with $r \geq 1$, finite (depending on its parameters) but μ'_r for r sufficiently high becomes either indefinite (for the odd moments) or infinite (for the even moments). Finally, the Cauchy distribution has all its moments μ'_r, with $r \geq 1$, either indefinite (for the odd moments) or infinite (for the even moments).

- **Example:**

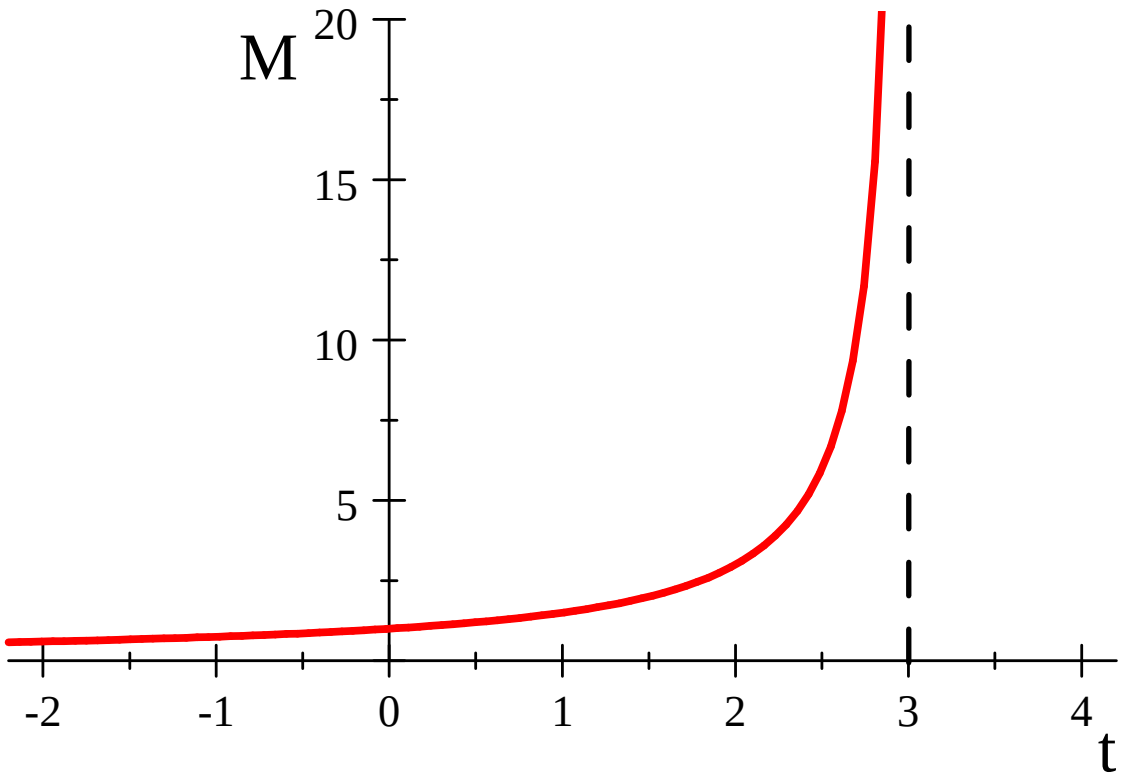
  Consider the absolutely continuous random variable $\tilde{x}$ with the density

$$f(x) = \begin{cases} 3e^{-3x} & \text{for } x > 0 \\ \\ 0 & \text{for } x \leq 0 \end{cases}$$



$$M_{\tilde{x}}(t) = \mathrm{E}(e^{t\tilde{x}}) = \int_{\mathbb{R}} e^{tx} dP_{\tilde{x}}(x) = \int_{\mathbb{R}} e^{tx} f(x) dx = \int_0^{\infty} e^{tx} 3e^{-3x} dx =$$

$$3\int_0^{\infty} e^{(t-3)x} dx = 3\left[\frac{e^{(t-3)x}}{t-3}\right]_0^{\infty} = 3 \cdot \left(0 - \frac{1}{t-3}\right) = \frac{3}{3-t}, \quad \text{for } t < 3.$$



$$M'_{\tilde{x}}(t) = \frac{3}{(3-t)^2} \implies M'_{\tilde{x}}(0) = \mu = \mu'_1 = \mathrm{E}(\tilde{x}) = \frac{3}{9} = \frac{1}{3},$$

$$M''_{\tilde{x}}(t) = \frac{6}{(3-t)^3} \implies M''_{\tilde{x}}(0) = \mathrm{E}(\tilde{x}^2) = \mu'_2 = \frac{6}{27} = \frac{2}{9},$$

$$\mathrm{Var}(\tilde{x}) = \mu_2 = \mu'_2 - (\mu'_1)^2 = \mathrm{E}(\tilde{x}^2) - [\mathrm{E}(\tilde{x})]^2 = \frac{2}{9} - \left(\frac{1}{3}\right)^2 = \frac{1}{9}.$$

- *Note:* If $M_{\tilde{x}}(t)$ is finite for all t in a neighborhood of 0, then $d^r M_{\tilde{x}}(t) / dt^r$ is continuous at $t = 0$, for $r = 0, 1, 2, ...$ In some cases, it could be necessary to compute the following limit to find the moment μ'_r :

$$\lim_{t \to 0} \frac{d^r M_{\tilde{x}}(t)}{dt^r} = \left. \frac{d^r M_{\tilde{x}}(t)}{dt^r} \right|_{t=0} = \mu'_r \ .$$

- **Properties of the moment-generating function:**

1. Let a be a constant, then $M_{\tilde{x}+a}(t) = \mathrm{E}(e^{t(\tilde{x}+a)}) = e^{at} M_{\tilde{x}}(t)$.
2. Let b be a constant, then $M_{b\tilde{x}}(t) = \mathrm{E}(e^{tb\tilde{x}}) = M_{\tilde{x}}(bt)$.
3. $M_{\frac{\tilde{x}+a}{b}}(t) = e^{\frac{a}{b}t} M_{\tilde{x}}(\frac{t}{b})$.

- Recall that

$$e = \lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x = \sum_{r=0}^{\infty} \frac{1}{r!},$$

which follows fron Newton's binomial theorem (please, check it).

3.7. The characteristic function and the Laplace transform

- Consider the set $\mathbb{C}$ of complex numbers, where $\mathbb{C} \equiv \mathbb{R}^2$ is endowed with the following two *internal* operations:

  (1) The (standard) sum,

  $$(a, b) + (c, d) = (a + c, b + d),$$

  (2) The (somewhat strange) multiplication,

  $$(a, b) \cdot (c, d) = (ac - bd, ad + bc).$$

- The complex number (a, b) has a real component a and an imaginary component b. Therefore, the complex numbers of the form $(a, 0)$ are real, $(a, 0) = a \in \mathbb{R}$.

- Therefore, if $k \in \mathbb{R}$ then

$$k \cdot (a, b) = (k, 0) \cdot (a, b) = (ka, kb),$$

  which agrees with the standard rule of multiplication of a vector by a scalar.

- The complex number $i \equiv (0, 1)$ is such that $i^2 = i \cdot i = (0, 1) \cdot (0, 1) = (-1, 0) = -1$. That is, $\sqrt{-1} = \pm i = (0, \pm 1)$.

- Observe that

$$(a, b) = a \cdot (1, 0) + b \cdot (0, 1) = a \cdot 1 + b \cdot i = a + bi.$$

- **Definition.** The modulus of the complex number (a, b) is its distance to $(0, 0)$, i.e., $\left(a^2 + b^2\right)^{1/2}$.

- **Definition.** The characteristic function, $\varphi_{\tilde{x}}(\cdot) : \mathbb{R} \longrightarrow \mathbb{C}$, of the random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ (or of its distribution) is

$$\varphi_{\tilde{x}}(t) = \mathrm{E}(e^{it\tilde{x}}) = \int_{\mathbb{R}} e^{itx} dP_{\tilde{x}}(x) = \mathrm{E}\left[\cos(t\tilde{x}) + i\sin(t\tilde{x})\right]$$

$$= \mathrm{E}\left[\cos(t\tilde{x})\right] + i\mathrm{E}\left[\sin(t\tilde{x})\right] = \left(\mathrm{E}\left[\cos(t\tilde{x})\right], \mathrm{E}\left[\sin(t\tilde{x})\right]\right),$$

  which is a well-defined (i.e., finite) complex number for all $t \in \mathbb{R}$.

- Recall that the moment-generating function, $M_{\tilde{x}}(\cdot) : B \longrightarrow \mathbb{R}$, of the random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ (or of its distribution) is

$$M_{\tilde{x}}(t) = \mathrm{E}(e^{t\tilde{x}}) = \varphi_{\tilde{x}}(-it).$$

  where B is the subset of $\mathbb{R}$ for which $\mathrm{E}(e^{t\tilde{x}})$ is finite.

- **Definition.** The Laplace transform, $\Lambda_{\tilde{x}}(\cdot) : D \longrightarrow \mathbb{R}$, of the random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ (or of its distribution) is

$$\Lambda_{\tilde{x}}(t) = \mathrm{E}(e^{-t\tilde{x}}) = M_{\tilde{x}}(-t) = \varphi_{\tilde{x}}(it).$$

  where D is the subset of $\mathbb{R}$ for which $\mathrm{E}(e^{-t\tilde{x}})$ is finite.

- *Note:* Moment-generating functions and Laplace transforms may be not well-defined around $t = 0$ (i.e., they may be not finite).

- **Proposition.** Let $P_{\tilde{x}}$ and $P_{\tilde{y}}$ be the distributions of the random variables $\tilde{x}$ and $\tilde{y}$, respectively. Then, $P_{\tilde{x}} = P_{\tilde{y}}$ if and only if $\varphi_{\tilde{x}} = \varphi_{\tilde{y}}$.
  - or iff $M_{\tilde{x}} = M_{\tilde{y}}$, provided they are well-defined (i.e., finite) in a neighborhood of $t = 0$,
  - or iff $\Lambda_{\tilde{x}} = \Lambda_{\tilde{y}}$, provided they are well-defined (i.e., finite) in a neighborhood of $t = 0$.

- **Lévy's inversion formula:**

$$P_{\tilde{x}}(a,b) + \frac{1}{2}P_{\tilde{x}}\{a,b\} = \underbrace{F_{\tilde{x}}(b) - F_{\tilde{x}}(a)}_{P_{\tilde{x}}(a,b]} - \frac{1}{2}\left[P_{\tilde{x}}\{b\} - P_{\tilde{x}}\{a\}\right]$$

$$= \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{itb}}{it}\varphi_{\tilde{x}}(t)dt.$$

- If the characteristic function $\varphi_{\tilde{x}}$ is Lebesgue integrable on $\mathbb{R}$, then $\tilde{x}$ is absolutely continuous with the density

$$f_{\tilde{x}}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-itx}}{it}\varphi_{\tilde{x}}(t)dt.$$

- The Fourier transform of the distribution $P_{\tilde{x}}$ is $\mathfrak{F}_{\tilde{x}}(t) = \mathrm{E}(e^{-it\tilde{x}}) = \varphi_{\tilde{x}}(-t)$ or $\mathfrak{F}_{\tilde{x}}(t) = \mathrm{E}(e^{-2\pi it\tilde{x}}) = \varphi_{\tilde{x}}(-2\pi t)$, whereas the Fourier transform of the function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is

$$\widehat{f}(t) = \int_{\mathbb{R}} e^{-2\pi itx} f(x)dx.$$

3.8. Product moments and its properties: covariance and correlation coefficient.

- $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^n, \mathcal{B})$.
- **Definition.** The $(r_1, r_2, ..., r_n)$ product moment of the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ is

$$\mu'_{r_1, r_2, ..., r_n}(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = \mathrm{E}\left(\tilde{x}_1^{r_1} \cdot \tilde{x}_2^{r_2} \cdot ... \cdot \tilde{x}_n^{r_n}\right)$$

$$= \int_{\Omega} \left(\tilde{x}_1^{r_1} \cdot \tilde{x}_2^{r_2} \cdot ... \cdot \tilde{x}_n^{r_n}\right) dP = \int_{\mathbb{R}^n} \left(x_1^{r_1} \cdot x_2^{r_2} \cdot ... \cdot x_n^{r_n}\right) dP_{\tilde{x}}(x_1, ..., x_n).$$

- **Definition.** The $(r_1, r_2, ..., r_n)$ product central moment of the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ is

$$\mu_{r_1, r_2, ..., r_n}(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) =$$

$$\mathrm{E}\left[(\tilde{x}_1 - \mathrm{E}(\tilde{x}_1))^{r_1} \cdot (\tilde{x}_2 - \mathrm{E}(\tilde{x}_2))^{r_2} \cdot ... \cdot (\tilde{x}_n - \mathrm{E}(\tilde{x}_n))^{r_n}\right] =$$

$$\int_{\Omega} (\tilde{x}_1 - \mathrm{E}(\tilde{x}_1))^{r_1} \cdot (\tilde{x}_2 - \mathrm{E}(\tilde{x}_2))^{r_2} \cdot ... \cdot (\tilde{x}_n - \mathrm{E}(\tilde{x}_n))^{r_n} \, dP =$$

$$\int_{\mathbb{R}^n} (x_1 - \mathrm{E}(\tilde{x}_1))^{r_1} \cdot (x_2 - \mathrm{E}(\tilde{x}_2))^{r_2} \cdot ... \cdot (x_n - \mathrm{E}(\tilde{x}_n))^{r_n} \, dP_{\tilde{x}}(x_1, ..., x_n).$$

- *Note:*

$$\mu'_r(\tilde{x}_i) = \mu'_{0,0,\ldots,\underset{\uparrow}{r},\ldots,0}(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) = \mathrm{E}(\tilde{x}_i^r),$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i$$

and

$$\mu_r(\tilde{x}_i) = \mu_{0,0,\ldots,\underset{\uparrow}{r},\ldots,0}(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) = \mathrm{E}\left[(\tilde{x}_i - \mathrm{E}(\tilde{x}_i))^r\right].$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i$$

- **Definition.** The covariance $\sigma_{\tilde{x},\tilde{y}}$ (or $\mathrm{Cov}(\tilde{x},\tilde{y})$) between (or of) $\tilde{x}$ and $\tilde{y}$ is

$$\sigma_{\tilde{x},\tilde{y}} = \mu_{1,1}(\tilde{x},\tilde{y}) = \mathrm{E}\left[(\tilde{x}-\mathrm{E}(\tilde{x}))\cdot(\tilde{y}-\mathrm{E}(\tilde{y}))\right] =$$

$$\int_{\Omega}(\tilde{x}-\mathrm{E}(\tilde{x}))(\tilde{y}-\mathrm{E}(\tilde{y}))\,dP = \int_{\mathbb{R}^2}(x-\mathrm{E}(\tilde{x}))(y-\mathrm{E}(\tilde{y}))\,dP_{\tilde{x},\tilde{y}}(x,y).$$

- *Note:* $\sigma_{\tilde{x},\tilde{x}} = \sigma^2_{\tilde{x}}$ and $\sigma_{\tilde{x},\tilde{y}} = \sigma_{\tilde{y},\tilde{x}}$.

- **Proposition.**

$$\sigma_{\tilde{x},\tilde{y}} = \mu'_{1,1}(\tilde{x},\tilde{y}) - \mu_{\tilde{x}}\mu_{\tilde{y}}$$

or

$$\text{Cov}(\tilde{x},\tilde{y}) = \text{E}(\tilde{x}\cdot\tilde{y}) - \text{E}(\tilde{x})\cdot\text{E}(\tilde{y}).$$

- **Proof.**

$$\sigma_{\tilde{x},\tilde{y}} = \text{E}\left[(\tilde{x}-\mu_{\tilde{x}})\cdot\left(\tilde{y}-\mu_{\tilde{y}}\right)\right] = \text{E}\left(\tilde{x}\tilde{y} - \tilde{x}\mu_{\tilde{y}} - \tilde{y}\mu_{\tilde{x}} + \mu_{\tilde{x}}\mu_{\tilde{y}}\right)$$

$$= \text{E}(\tilde{x}\tilde{y}) - \mu_{\tilde{x}}\mu_{\tilde{y}} - \mu_{\tilde{y}}\mu_{\tilde{x}} + \mu_{\tilde{x}}\mu_{\tilde{y}} = \text{E}(\tilde{x}\cdot\tilde{y}) - \mu_{\tilde{x}}\mu_{\tilde{y}}. \quad Q.E.D.$$

- **Definition.** The coefficient of correlation between $\tilde{x}$ and $\tilde{y}$ is

$$\rho_{\tilde{x},\tilde{y}} = \frac{\sigma_{\tilde{x},\tilde{y}}}{\sigma_{\tilde{x}} \cdot \sigma_{\tilde{y}}}, \quad \text{with } \sigma_{\tilde{x}} > 0 \text{ and } \sigma_{\tilde{y}} > 0.$$

- **Proposition.** $\rho_{\tilde{x},\tilde{y}} \in [-1, 1]$ or $\left|\rho_{\tilde{x},\tilde{y}}\right| \leq 1$.

- **Proof.** From the Cauchy-Schwarz inequality,

$$\left|\int_{\Omega} (\tilde{x} - \mathrm{E}(\tilde{x}))\,(\tilde{y} - \mathrm{E}(\tilde{y}))\,dP\right| \leq \int_{\Omega} |(\tilde{x} - \mathrm{E}(\tilde{x}))\,(\tilde{y} - \mathrm{E}(\tilde{y}))|\,dP$$

$$\leq \left[\int_{\Omega} (\tilde{x} - \mathrm{E}(\tilde{x}))^2\,dP\right]^{1/2} \left[\int_{\Omega} (\tilde{y} - \mathrm{E}(\tilde{y}))^2\,dP\right]^{1/2}.$$

That is,

$$|\sigma_{\tilde{x},\tilde{y}}| \leq \sigma_{\tilde{x}} \cdot \sigma_{\tilde{y}} \Leftrightarrow -\sigma_{\tilde{x}} \cdot \sigma_{\tilde{y}} \leq \sigma_{\tilde{x},\tilde{y}} \leq \sigma_{\tilde{x}} \cdot \sigma_{\tilde{y}} \Leftrightarrow -1 \leq \rho_{\tilde{x},\tilde{y}} \leq 1.$$

Q.E.D.

- **Proposition.** Let a and b be constants. Then

$$\sigma_{a\tilde{x},b\tilde{y}} = ab\sigma_{\tilde{x},\tilde{y}}.$$

Moreover, if $a > 0$ and $b > 0$, then

$$\rho_{a\tilde{x},b\tilde{y}} = \rho_{\tilde{x},\tilde{y}}.$$

- **Proposition.** Let $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ be independent random variables on $(\Omega, \mathcal{F}, P)$ belonging to L^1, then

$$\mathrm{E}\left(\tilde{x}_1 \cdot \tilde{x}_2 \cdot ... \cdot \tilde{x}_n\right) = \mathrm{E}\left(\tilde{x}_1\right) \cdot \mathrm{E}\left(\tilde{x}_2\right) \cdot ... \cdot \mathrm{E}\left(\tilde{x}_n\right).$$

- **Proof.** Since, from independency, $P_{\tilde{x}} = P_{\tilde{x}_1} \times P_{\tilde{x}_2} \times ... \times P_{\tilde{x}_n} = \prod_{i=n}^{n} P_{\tilde{x}_i}$, we can use Fubini's theorem so that

$$\mathrm{E}\left(\tilde{x}_1 \cdot \tilde{x}_2 \cdot ... \cdot \tilde{x}_n\right) = \int_{\mathbb{R}^n} \left(x_1 \cdot x_2 \cdot ... \cdot x_n\right) dP_{\tilde{x}}\left(x_1, ..., x_n\right)$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} ... \int_{\mathbb{R}} \left(x_1 \cdot x_2 \cdot ... \cdot x_n\right) dP_{\tilde{x}_1}\left(x_1\right) dP_{\tilde{x}_2}\left(x_2\right) ... dP_{\tilde{x}_n}\left(x_n\right)$$

$$= \left[\int_{\mathbb{R}} x_1 dP_{\tilde{x}_1}\left(x_1\right)\right] \cdot \left[\int_{\mathbb{R}} x_2 dP_{\tilde{x}_2}\left(x_2\right)\right] \cdot ... \cdot \left[\int_{\mathbb{R}} x_n dP_{\tilde{x}_n}\left(x_n\right)\right]$$

$$= \mathrm{E}\left(\tilde{x}_1\right) \cdot \mathrm{E}\left(\tilde{x}_2\right) \cdot ... \cdot \mathrm{E}\left(\tilde{x}_n\right). \qquad \textit{Q.E.D.}$$

- **Corollary.** If the random variables $\tilde{x}$ and $\tilde{y}$ are independent, then

$$\mathrm{Cov}\left(\tilde{x}, \tilde{y}\right) = 0,$$

  i.e., $\tilde{x}$ and $\tilde{y}$ are "uncorrelated".

- **Proof.**

$$\mathrm{Cov}\left(\tilde{x}, \tilde{y}\right) = \mathrm{E}\left(\tilde{x} \cdot \tilde{y}\right) - \mathrm{E}(\tilde{x}) \cdot \mathrm{E}(\tilde{y}) = \mathrm{E}(\tilde{x}) \cdot \mathrm{E}(\tilde{y}) - \mathrm{E}(\tilde{x}) \cdot \mathrm{E}(\tilde{y}) = 0,$$

  where the second equality holds because $\mathrm{E}\left(\tilde{x} \cdot \tilde{y}\right) = \mathrm{E}(\tilde{x}) \cdot \mathrm{E}(\tilde{y})$ when $\tilde{x}$ and $\tilde{y}$ are independent. *Q.E.D.*

- Independency implies $\mathrm{Cov}\left(\tilde{x}, \tilde{y}\right) = 0$, but the converse is not true.

- **Example:** Consider the following table summarizing the probability function $f_{\tilde{x},\tilde{y}}(x,y)$ of the random variables $\tilde{x}$ and $\tilde{y}$**:**

$y \setminus x$	-1	0	1	$f_{\tilde{y}}(y)$
-1	$1/6$	$1/3$	$1/6$	$2/3$
1	$1/6$	0	$1/6$	$1/3$
$f_{\tilde{x}}(x)$	$1/3$	$1/3$	$1/3$	1

- We have that $\mathrm{E}(\tilde{x}) = 0$, $\mathrm{E}(\tilde{y}) = -1/3$, and $\mathrm{E}(\tilde{x} \cdot \tilde{y}) = 0$.
- Thus,

$$\sigma_{\tilde{x},\tilde{y}} = \mathrm{E}(\tilde{x} \cdot \tilde{y}) - \mathrm{E}(\tilde{x}) \cdot \mathrm{E}(\tilde{y}) = 0.$$

- However, $\tilde{x}$ and $\tilde{y}$ are not independent since it is not true that $f_{(\tilde{x},\tilde{y})}(x,y) = f_{\tilde{x}}(x) \cdot f_{\tilde{y}}(y)$ for all $(x,y) \in \tilde{x}(\Omega) \times \tilde{y}(\Omega)$. For instance,

$$f(-1,-1) = \frac{1}{6} \neq f_{\tilde{x}}(-1) \cdot f_{\tilde{y}}(-1) = \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}.$$

3.9. Mean and variance of linear combinations of random variables: the mean vector and the covariance matrix

Proposition. Let $\{\tilde{x}_i\}_{i=1}^n$ be a collection of random variables on $(\Omega, \mathcal{F}, P)$, $\{c_i\}_{i=1}^n$ a collection of constants, and b a constant. Then,

$$\text{Var}\left(\left[\sum_{i=1}^{n} c_i \tilde{x}_i\right] + b\right) = \sum_{i=1}^{n} c_i^2 \text{Var}\left(\tilde{x}_i\right) + 2 \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} c_i c_j \text{Cov}\left(\tilde{x}_i, \tilde{x}_j\right)$$

$$= \sum_{i=1}^{n} c_i^2 \text{Var}\left(\tilde{x}_i\right) + \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{i=1}^{n} c_i c_j \text{Cov}\left(\tilde{x}_i, \tilde{x}_j\right) = \sum_{j=1}^{n} \sum_{i=1}^{n} c_i c_j \text{Cov}\left(\tilde{x}_i, \tilde{x}_j\right).$$

Proof:

$$\text{Var}\left(\left[\sum_{i=1}^{n} c_i \tilde{x}_i\right] + b\right) = \text{E}\left(\left[\left(\sum_{i=1}^{n} c_i \tilde{x}_i\right) + b - \text{E}\left(\left[\sum_{i=1}^{n} c_i \tilde{x}_i\right] + b\right)\right]^2\right)$$

$$= \text{E}\left(\left[\left(\sum_{i=1}^{n} c_i \tilde{x}_i\right) + b - \left(\sum_{i=1}^{n} c_i \text{E}\left(\tilde{x}_i\right)\right) - b\right]^2\right)$$

$$= \text{E}\left(\left[\sum_{i=1}^{n} c_i \left(\tilde{x}_i - \text{E}\left(\tilde{x}_i\right)\right)\right]^2\right)$$

$$= \text{E}\left[\sum_{i=1}^{n} c_i^2 \left(\tilde{x}_i - \text{E}\left(\tilde{x}_i\right)\right)^2 + 2 \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} c_i c_j \left(\tilde{x}_i - \text{E}\left(\tilde{x}_i\right)\right)\left(\tilde{x}_j - \text{E}\left(\tilde{x}_j\right)\right)\right]$$

$$= \sum_{i=1}^{n} c_i^2 \text{E}\left[\left(\tilde{x}_i - \text{E}\left(\tilde{x}_i\right)\right)^2\right] + 2 \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} c_i c_j \text{E}\left[\left(\tilde{x}_i - \text{E}\left(\tilde{x}_i\right)\right)\left(\tilde{x}_j - \text{E}\left(\tilde{x}_j\right)\right)\right]$$

$$= \sum_{i=1}^{n} c_i^2 \text{Var}\left(\tilde{x}_i\right) + 2 \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} c_i c_j \text{Cov}\left(\tilde{x}_i, \tilde{x}_j\right). \qquad Q.E.D.$$

- **Corollary.** If $\{\tilde{x}_i\}_{i=1}^n$ is a collection of independent random variables, then

$$\text{Var}\left(\left[\sum_{i=1}^n c_i \tilde{x}_i\right] + b\right) = \sum_{i=1}^n c_i^2 \text{Var}\left(\tilde{x}_i\right).$$

- **Corollary.** If the random variables $\tilde{x}$ and $\tilde{y}$ are independent, then

$$\text{Var}\left(\tilde{x} + \tilde{y}\right) = \text{Var}\left(\tilde{x}\right) + \text{Var}\left(\tilde{y}\right).$$

- But, in general,

$$\text{Var}\left(\tilde{x} + \tilde{y}\right) = \text{Var}\left(\tilde{x}\right) + \text{Var}\left(\tilde{y}\right) + 2\text{Cov}\left(\tilde{x}, \tilde{y}\right).$$

- Let $c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$ be a vector of scalars and $\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix}$.
- $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$, with $\mu_i = \mathrm{E}(\tilde{x}_i)$, is the vector of means.
- $$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & .. & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & .. & \sigma_{2n} \\ .. & .. & .. & .. \\ \sigma_{n1} & \sigma_{n2} & .. & \sigma_{nn} \end{pmatrix}_{n \times n},$$

  with $\sigma_{ij} = \mathrm{Cov}(\tilde{x}_i, \tilde{x}_j)$, is the covariance matrix (or variance covariance matrix).
- *Note:* Σ is symmetric ($\sigma_{ij} = \sigma_{ji}$) and the elements along the diagonal are the variances ($\sigma_{ii} = \sigma_i^2$).

- Then,

$$\mathrm{E}(c^{\intercal}\tilde{x}) = \mathrm{E}\left(\sum_{i=1}^{n} c_i \tilde{x}_i\right) = c^{\intercal}\mu$$

and

$$\mathrm{Var}(c^{\intercal}\tilde{x}) = \mathrm{Var}\left(\sum_{i=1}^{n} c_i \tilde{x}_i\right) = c^{\intercal}\Sigma c.$$

- *Note:* Since $\mathrm{Var}(c^{\intercal}\tilde{x}) = c^{\intercal}\Sigma c \geq 0$ for all $c \in \mathbb{R}^n$, then the matrix Σ is symmetric positive semi-definite, which implies that $\det(\Sigma) \geq 0$.

- **Proposition.**

$$\mathrm{Cov}\left(\tilde{y}, \sum_{i=1}^{n} c_i \tilde{x}_i\right) \equiv \mathrm{Cov}\left(\tilde{y}, c^{\intercal}\tilde{x}\right) = \sum_{i=1}^{n} c_i \mathrm{Cov}\left(\tilde{y}, \tilde{x}_i\right).$$

- **Proof.** Exercise.

3.10. Multivariate moment-generating functions

- **Definition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be a random vector. The multivariate moment-generating function, $M_{\tilde{x}}(\cdot) : B \longrightarrow \mathbb{R}$, of the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ is

$$M_{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n}(t_1, t_2, ..., t_n) = \mathrm{E}(e^{t_1\tilde{x}_1+t_2\tilde{x}_2+...+t_n\tilde{x}_n})$$
$$= \int_\Omega e^{(t_1\tilde{x}_1+t_2\tilde{x}_2+...+t_n\tilde{x}_n)} dP = \int_{\mathbb{R}^n} e^{(t_1x_1+t_2x_2+...+t_nx_n)} dP_{\tilde{x}}(x_1, ..., x_n),$$

  where B is the subset of $\mathbb{R}^n$ for which $\mathrm{E}(e^{t_1\tilde{x}_1+t_2\tilde{x}_2+...+t_n\tilde{x}_n})$ is finite.

- **Proposition.** If there exists a neighborhood V of $(0, 0, ..., 0)$ for which $M_{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n}(t_1, t_2, ..., t_n)$ is finite for all vectors $(t_1, t_2, ..., t_n) \in V$, then $M_{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n}(t_1, t_2, ..., t_n)$ is infinitely differentiable at $(t_1, t_2, ..., t_n) = (0, 0, ..., 0)$, all the product moments $\mu'_{r_1,r_2,...,r_n}(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ are finite and

$$\left.\frac{\partial^{r_1+r_2+...+r_n} M_{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n}(t_1, t_2, ..., t_n)}{\partial t_1^{r_1} \partial t_2^{r_2} ... \partial t_n^{r_n}}\right|_{(t_1,t_2,...,t_n)=(0,0,...,0)}$$
$$= \mu'_{r_1,r_2,...,r_n}(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n).$$

- **Proposition.** If the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are independent, then

$$M_{\tilde{x}_1+\tilde{x}_2+...+\tilde{x}_n}(t) = \prod_{i=1}^{n} M_{\tilde{x}_i}(t).$$

- **Proof.**

$$\mathrm{E}\left(e^{t(\tilde{x}_1+\tilde{x}_2+...+\tilde{x}_n)}\right) = \mathrm{E}\left(\prod_{i=1}^{n} e^{t\tilde{x}_i}\right) = \prod_{i=1}^{n} \mathrm{E}\left(e^{t\tilde{x}_i}\right). \qquad Q.E.D.$$

- **Proposition.** If the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are independent, then

$$M_{\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n}(t_1, t_2, ..., t_n) = \prod_{i=1}^{n} M_{\tilde{x}_i}(t_i).$$

Proof.

$$\mathrm{E}(e^{t_1\tilde{x}_1+t_2\tilde{x}_2+...+t_n\tilde{x}_n}) = \mathrm{E}\left(\prod_{i=1}^{n} e^{t_i\tilde{x}_i}\right) = \prod_{i=1}^{n} \mathrm{E}\left(e^{t_i\tilde{x}_i}\right). \qquad Q.E.D.$$

3.11. Conditional expectation

- **Definition.** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ a random variable. The conditional expectation of $\tilde{y}$ given $\tilde{x} = x$ is a (essentially unique w.r.t. $P_{\tilde{x}}$) Borel measurable function $\mathrm{E}(\tilde{y} \,|\tilde{x} = \cdot\,) : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}, \mathcal{B})$ given by

$$\mathrm{E}(\tilde{y} \,|\tilde{x} = x\,) = \int_{\Omega} \tilde{y} dP(\omega \,|\tilde{x} = x\,) = \int_{\mathbb{R}} y dP_{\tilde{y}|\tilde{x}}(y \,|x\,), \quad \text{for all } x \in \Omega'.$$

- Note that the previous conditional expectation can be seen as a random variable from $(\Omega', \mathcal{F}')$ to $(\mathbb{R}, \mathcal{B})$.
- All the properties of the (unconditional) expectation of $\tilde{y}$ also hold a.s. $[P_{\tilde{x}}]$ for the conditional expectation of $\tilde{y}$ given $\tilde{x} = x$.

- If the random vector $(\tilde{x}, \tilde{y})$ is discrete, then the conditional expectation of $\tilde{y}$ given $\tilde{x} = x$ becomes

$$\mathrm{E}(\tilde{y} \,|\tilde{x} = x\,) = \sum_{y \in \tilde{y}(\Omega)} y f_{\tilde{y}|\tilde{x}}(y \,|x\,), \text{ for all } x \text{ with } f_{\tilde{x}}(x) > 0.$$

- If the random vector $(\tilde{x}, \tilde{y})$ is absolutely continuous, then the conditional expectation of $\tilde{y}$ given $\tilde{x} = x$ becomes

$$\mathrm{E}(\tilde{y} \,|\tilde{x} = x\,) = \int_{\mathbb{R}} y f_{\tilde{y}|\tilde{x}}(y \,|x\,) dy, \text{ for all } x \text{ with } f_{\tilde{x}}(x) > 0.$$

- We can define conditional moments since moments are just expectations. For instance, the conditional variance of $\tilde{y}$ given $\tilde{x} = x$ is

$$\begin{aligned}\mathrm{Var}(\tilde{y} \,|\tilde{x} = x\,) &= \mathrm{E}\left(\left[\tilde{y} - \mathrm{E}\left(\tilde{y} \,|\tilde{x} = x\,\right)\right]^2 |\tilde{x} = x\right) \\ &= \mathrm{E}\left(\tilde{y}^2 \,|\tilde{x} = x\right) - \left[\mathrm{E}\left(\tilde{y} \,|\tilde{x} = x\,\right)\right]^2, \quad \text{for all } x \in \Omega'.\end{aligned}$$

- If the discrete (absolutely continuous) random variables $\tilde{x}$ and $\tilde{y}$ are independent then

$$\mathrm{E}(\tilde{y}\,|\tilde{x}=x\,)=\mathrm{E}(\tilde{y}),\quad \text{for } f_{\tilde{x}}(x)>0.$$

  That is, the conditional expectation is equal to the corresponding unconditional expectation.

- In general, if the random object $\tilde{x}:(\Omega,\mathcal{F},P)\longrightarrow(\Omega',\mathcal{F}')$ and the random variable $\tilde{y}:(\Omega,\mathcal{F},P)\longrightarrow(\mathbb{R},\mathcal{B})$ are independent then

$$\mathrm{E}(\tilde{y}\,|\tilde{x}=x\,)=\mathrm{E}(\tilde{y}),\quad \text{a.s. } [P_{\tilde{x}}]\,.$$

- Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega'', \mathcal{F}'')$.
- Conditional probability is a particular case of conditional expectation:

$$P(B \,|\tilde{x} = x) = \mathrm{E}(\mathbb{I}_B \,|\tilde{x} = x), \text{ for all } B \in \mathcal{F}.$$

- The conditional distribution is also a particular case of conditional expectation:

$$P_{\tilde{y}}(A \,|\tilde{x} = x) = \mathrm{E}\left(\mathbb{I}_A(\tilde{y}) \,|\tilde{x} = x\right), \text{ for all } A \in \mathcal{F}'',$$

or

$$P_{\tilde{y}}(A \,|\tilde{x} = x) = P\left(\tilde{y}^{-1}(A) \,|\tilde{x} = x\right) = \mathrm{E}\left(\mathbb{I}_{\tilde{y}^{-1}(A)} \,|\tilde{x} = x\right),$$

for all $A \in \mathcal{F}''$.

- We can define the conditional expectation of the random variable $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ given the event $B \in \mathcal{F}$ as

$$\mathrm{E}(\tilde{y} \,|B) = \mathrm{E}(\tilde{y} \,|\mathbb{I}_B = 1).$$

- **Theorem of total expectation (Adam's law).** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$. Then,

$$\mathrm{E}(\tilde{y}) = \int_{\Omega'} \mathrm{E}(\tilde{y} \,|\tilde{x} = x\,) dP_{\tilde{x}}(x).$$

- If the random vector $(\tilde{x}, \tilde{y})$ is discrete,

$$\int_{\mathbb{R}} \mathrm{E}(\tilde{y} \,|\tilde{x} = x\,) dP_{\tilde{x}}(x) = \sum_{x \in \tilde{x}(\Omega)} \left[\sum_{y \in \tilde{y}(\Omega)} y f_{\tilde{y}|\tilde{x}}(y \,|x\,) \right] f_{\tilde{x}}(x)$$

$$= \sum_{x \in \tilde{x}(\Omega)} \sum_{y \in \tilde{y}(\Omega)} y f_{\tilde{x},\tilde{y}}(x, y) = \sum_{y \in \tilde{y}(\Omega)} y \left[\sum_{x \in \tilde{x}(\Omega)} f_{\tilde{x},\tilde{y}}(x, y) \right]$$

$$= \sum_{y \in \tilde{y}(\Omega)} y f_{\tilde{y}}(y) = \mathrm{E}(\tilde{y}).$$

- If the random vector $(\tilde{x}, \tilde{y})$ is absolutely continuous,

$$\int_{\mathbb{R}} \mathrm{E}(\tilde{y} \,|\tilde{x} = x\,) dP_{\tilde{x}}(x) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} y f_{\tilde{y}|\tilde{x}}(y\,|x\,) dy \right] f_{\tilde{x}}(x) dx$$

$$= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} y f_{\tilde{x},\tilde{y}}(x, y) dy \right] dx = \int_{\mathbb{R}} \left[y \int_{\mathbb{R}} f_{\tilde{x},\tilde{y}}(x, y) dx \right] dy$$

$$= \int_{\mathbb{R}} y f_{\tilde{y}}(y) dy = \mathrm{E}(\tilde{y}).$$

3.12. Conditional expectation given a sigma-algebra

- **Definition.** Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ a random variable. The conditional expectation of $\tilde{y}$ given $\tilde{x}$ is a (essentially unique w.r.t. P) Borel measurable function $\mathrm{E}(\tilde{y} \,|\tilde{x} = \tilde{x}(\cdot)) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$ (or $\mathrm{E}(\tilde{y} \,|\tilde{x})(\cdot)$) given by

$$\mathrm{E}(\tilde{y} \,|\tilde{x})(\omega) \equiv \mathrm{E}(\tilde{y} \,|\tilde{x} = \tilde{x}(\omega))$$

$$= \int_{\Omega} \tilde{y}(\omega') dP(\omega' \,|\tilde{x} = \tilde{x}(\omega)) = \int_{\mathbb{R}} y dP_{\tilde{y}|\tilde{x}}(y \,|\tilde{x}(\omega)), \quad \text{for all } \omega \in \Omega.$$

- Note that the previous conditional expectation can be seen as a random variable from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$.

- Moreover, if $\mathrm{E}(\tilde{y} \,|\tilde{x} = x) \equiv h(x)$ a.s. $[P_{\tilde{x}}]$, where $h : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}, \mathcal{B})$, then $\mathrm{E}(\tilde{y} \,|\tilde{x}) = h(\tilde{x})$, a.s. $[P]$ where $h(\tilde{x}) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$, since $h(\tilde{x}(\omega)) = \mathrm{E}(\tilde{y} \,|\tilde{x} = \tilde{x}(\omega))$.

- **Definition.** Let $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\Omega', \mathcal{F}')$ be a random object. The σ-algebra induced (or generated) on Ω by $\tilde{x}$ is

$$\mathcal{F}(\tilde{x}) = \left\{ \tilde{x}^{-1}(A) \;\middle|\; A \in \mathcal{F}' \right\},$$

that is, $\mathcal{F}(\tilde{x})$ contains the pre-images of all the measurable sets in $\mathcal{F}'$.

- The σ-algebra $\mathcal{F}(\tilde{x})$ is the smallest (i.e., the coarsest) σ-algebra that makes the random object $\tilde{x}$ measurable, $\tilde{x} : (\Omega, \mathcal{F}(\tilde{x})) \longrightarrow (\Omega', \mathcal{F}')$. Hence, $\mathcal{F}(\tilde{x}) \subset \mathcal{F}$.
- Recall that, if the random variable $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$ is discrete, there is a countable partition $\mathcal{A} = \{A_1, A_2, ...\}$ of Ω with $A_n = \{\omega \in \Omega \mid \tilde{x}(\omega) = x_n\}$, for all $x_n \in \tilde{x}(\Omega)$. Then, the σ-algebra $\mathcal{F}(\tilde{x})$ induced on Ω by $\tilde{x}$ is the σ-algebra $\sigma(\mathcal{A})$ generated by the partition $\mathcal{A}$, i.e., the smallest σ-algebra containing $\mathcal{A}$.
- *Note:* If we have two random objects $\tilde{x}_1 : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_1, \mathcal{F}_1)$ and $\tilde{x}_2 : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega_2, \mathcal{F}_2)$ such that $\mathcal{F}(\tilde{x}_1) = \mathcal{F}(\tilde{x}_2)$, then

$$\mathrm{E}(\tilde{y} \,|\tilde{x}_1) = \mathrm{E}(\tilde{y} \,|\tilde{x}_2) \quad \text{a.s. } [P].$$

- Therefore, we can define the conditional expectation of the random variable $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ given the σ-algebra $\mathcal{G}$ on Ω as the (essentially unique w.r.t. P) Borel measurable function $\mathrm{E}(\tilde{y} \,|\mathcal{G}\,)(\cdot) : (\Omega, \mathcal{G}) \longrightarrow (\mathbb{R}, \mathcal{B})$ given by $\mathrm{E}(\tilde{y} \,|\mathcal{G}\,) = \mathrm{E}(\tilde{y} \,|\tilde{x}\,)$, where $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\Omega', \mathcal{F}')$ is any random object such that $\mathcal{G} = \mathcal{F}\,(\tilde{x})$.
- In particular, we can choose $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\Omega, \mathcal{G})$ to be the identity map, $\tilde{x}(\omega) = \omega$ for all $\omega \in \Omega$.
- Note that the conditional expectation of the random variable $\tilde{y}$ given the σ-algebra $\mathcal{G}$ can be seen as a random variable from $(\Omega, \mathcal{G})$ to $(\mathbb{R}, \mathcal{B})$.
- *Note:* $\mathrm{E}(\tilde{y} \,|\tilde{x}\,) = \mathrm{E}(\tilde{y} \,|\mathcal{F}\,(\tilde{x})\,)$ a.s. $[P]$.

- **Definition.** The random object $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\Omega', \mathcal{F}')$ is $\mathcal{G}$-measurable if $\tilde{x}^{-1}(A) \in \mathcal{G}$ for all $A \in \mathcal{F}'$. Obviously, we must have that $\mathcal{G} \subset \mathcal{F}$.
- **Proposition.** If $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\Omega', \mathcal{F}')$ is a random object, then $\tilde{x}$ is $\mathcal{G}$-measurable for all $\mathcal{G}$ such that $\mathcal{F}(\tilde{x}) \subset \mathcal{G} \subset \mathcal{F}$.
- **Definition.** Let $\mathcal{H}$ and $\mathcal{G}$ be two σ-algebras on the set Ω with $\mathcal{H} \subset \mathcal{G}$. Then, we say that $\mathcal{H}$ is coarser than $\mathcal{G}$ or, equivalently, that $\mathcal{G}$ is finer than $\mathcal{H}$.
- **Definition.** Let H and G be partitions of the sample space Ω. Then, we say that H is coarser than G or, equivalently, that G is finer than H, if every element of G is a subset of an element of H.

Example A:

- $\Omega = \{\omega_{1,}\omega_2, \omega_3, \omega_4, \omega_5\}$
- $H = \{\{\omega_{1,}\omega_2, \omega_3, \omega_4, \omega_5\}\} = \{\Omega\}$
- $G = \{\{\omega_{1,}\omega_2, \omega_3\}, \{\omega_4, \omega_5\}\},$
- $F = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}\}.$
- Observe that F is finer than G, while G is finer than H.

$$\{\omega_1\} \subset \{\omega_{1,}\omega_2, \omega_3\} \subset \Omega,$$

$$\{\omega_2\} \subset \{\omega_{1,}\omega_2, \omega_3\} \subset \Omega,$$

$$\{\omega_3\} \subset \{\omega_{1,}\omega_2, \omega_3\} \subset \Omega,$$

$$\{\omega_4\} \subset \{\omega_{4,}\omega_5\} \subset \Omega,$$

$$\{\omega_5\} \subset \{\omega_{4,}\omega_5\} \subset \Omega.$$

- Let $\mathcal{G}$ be the σ-algebra generated by the partition G, i.e., $\mathcal{G}$ is the smallest σ-algebra containing G.
- Similarly, in our example, $\mathcal{H}$ is the σ-algebra generated by H, while $\mathcal{F}$ is the σ-algebra generated by F.
- Obviously, if a countable partition H is coarser (resp. finer) than the countable partition G, then the σ-algebra $\mathcal{H}$ generated by H is coarser (rep. finer) than the σ-algebra $\mathcal{G}$ generated by G, i.e., $\mathcal{H} \subset \mathcal{G}$ (resp. $\mathcal{G} \subset \mathcal{H}$).
- *Note:* Not all σ-algebras are generated by partitions. For instance, there is no partition of $\mathbb{R}$ that generates the σ-algebra $\mathcal{B}$ of Borel sets in $\mathbb{R}$.

- **Example B:** Consider $\Omega = \mathbb{R}$ and the following uncountable partition of $\mathbb{R}$:

$$P = \{\{x\} \mid x \in \mathbb{R}\}.$$

The σ-algebra generated by P is

$$\mathcal{P} = \{A \subset \mathbb{R} \mid \text{either } A \text{ or } A^c \text{ are countable}\}.$$

Consider now the following countable partition of $\mathbb{R}$:

$$Q = \{(-\infty, 0), [0, \infty)\}.$$

The σ-algebra generated by Q is

$$\mathcal{Q} = \{\varnothing, \mathbb{R}, (-\infty, 0), [0, \infty)\}.$$

We see that P is finer than Q since, for all $x \in \mathbb{R}$, either $\{x\} \subset (-\infty, 0)$ or $\{x\} \subset [0, \infty)$. However, it is not true that $\mathcal{Q} \subset \mathcal{P}$ since neither $(-\infty, 0)$ nor $[0, \infty)$ are countable or have countable complements.

- In the previous Example A, we have $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$.
- $\mathcal{H} = \{\varnothing, \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}\}$,
- $\mathcal{G} = \{\varnothing, \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_4, \omega_5\}\}$,

  and
- $\mathcal{F} = \{\varnothing, \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_4, \omega_5\}, \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_1, \omega_4\}, \{\omega_1, \omega_5\}, \{\omega_2, \omega_3\}, \{\omega_2, \omega_4\}, \{\omega_2, \omega_5\}, \{\omega_3, \omega_4\}, \{\omega_3, \omega_5\}, \{\omega_1, \omega_2, \omega_4\}, \{\omega_1, \omega_2, \omega_5\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_1, \omega_3, \omega_5\}, \{\omega_1, \omega_4, \omega_5\}, \{\omega_2, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_5\}, \{\omega_3, \omega_4, \omega_5\}, \{\omega_1, \omega_2, \omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3, \omega_5\}, \{\omega_1, \omega_2, \omega_4, \omega_5\}, \{\omega_1, \omega_3, \omega_4, \omega_5\}, \{\omega_2, \omega_3, \omega_4, \omega_5\}\}$.
- In this example, $\mathcal{F}$ is the power set 2^{Ω} (the finest σ-algebra on Ω) since F is the finest partition of Ω.
- Note that $\#\mathcal{F} = 2^5 = 32$.
- Moreover, in this example, $\mathcal{H}$ is the coarsest σ-algebra on Ω, $\mathcal{H} = \{\varnothing, \Omega\}$, as H is the coarsest partition of Ω.

- **Example:** Consider the probability space $\left(\Omega, 2^{\Omega}, P\right)$, where the sample space is $\Omega = \{a, b, c\}$ and the probability P satisfies $P\{a\} = 1/6$, $P\{b\} = 2/3 = 4/6$, and $P\{c\} = 1/6$. Let $\tilde{x} : \left(\Omega, 2^{\Omega}, P\right) \longrightarrow (\mathbb{R}, \mathcal{B})$ be a random variable defined as follows:

$$\tilde{x}(\omega) = \begin{cases} 1 & \text{if } \omega = a \\ \\ 2 & \text{if } \omega \in \{b, c\}. \end{cases}$$

- The σ-algebra $\mathcal{F}(\tilde{x})$ induced on Ω by the random variable $\tilde{x}$ is

$$\mathcal{F}(\tilde{x}) = \{\emptyset, \Omega, \{a\}, \{b, c\}\} \subset 2^{\Omega}.$$

- Then,

$$\mathrm{E}\left(\tilde{x} \,|\mathcal{F}(\tilde{x})\right) = \tilde{x}(\omega) = \begin{cases} 1 & \text{if } \omega = a \\ \\ 2 & \text{if } \omega \in \{b, c\}. \end{cases}$$

- Consider the following partition of the sample space Ω: $G = \{\{a, b\}, \{c\}\}$, and let $\mathcal{G}$ be the σ-algebra generated by the partition G,
$$\mathcal{G} = \{\emptyset, \Omega, \{a, b\}, \{c\}\} \subset 2^{\Omega}.$$
- Since
$$P\{\tilde{x} = 1 \,|\{a, b\}\} = P(\{a\} \,|\{a, b\}) = \frac{P(\{a\} \cap \{a,b\})}{P\{a,b\}} = \frac{P\{a\}}{P\{a,b\}} = \frac{1/6}{5/6} = \frac{1}{5},$$
$$\begin{aligned} P\{\tilde{x} = 2 \,|\{a, b\}\} &= P(\{b, c\} \,|\{a, b\}) = \frac{P(\{b,c\} \cap \{a,b\})}{P\{a,b\}} \\ &= \frac{P(b)}{P\{a,b\}} = \frac{4/6}{5/6} = \frac{4}{5}, \end{aligned}$$
$$P\{\tilde{x} = 1 \,|\{c\}\} = P(\{a\} \,|\{c\}) = \frac{P(\emptyset)}{P\{c\}} = \frac{0}{1/6} = 0,$$
$$P\{\tilde{x} = 2 \,|\{c\}\} = P(\{b, c\} \,|\{c\}) = \frac{P(\{b,c\} \cap \{c\})}{P\{c\}} = \frac{P\{c\}}{P\{c\}} = \frac{1/6}{1/6} = 1,$$
- then
$$\mathrm{E}(\tilde{x} \,|\mathcal{G}) = \begin{cases} 1 \cdot \frac{1}{5} + 2 \cdot \frac{4}{5} = \frac{9}{5} & \text{if } \omega \in \{a, b\} \\ 1 \cdot 0 + 2 \cdot 1 = 2 & \text{if } \omega = c. \end{cases}$$

- Let $\mathcal{H}$ be the coarsest σ-algebra on Ω, $\mathcal{H} = \{\emptyset, \Omega\}$. Since

$$P\{\tilde{x} = 1\} = P\{a\} = \frac{1}{6},$$

$$P\{\tilde{x} = 2\} = P\{b, c\} = \frac{5}{6},$$

then

$$\mathrm{E}(\tilde{x}\,|\mathcal{H}) = \mathrm{E}(\tilde{x}) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{5}{6} = \frac{11}{6} \quad \text{for } \omega \in \{a, b, c\} = \Omega.$$

- Finally, note that, since

$$P\{a, b\} = \frac{5}{6},$$

$$P\{c\} = \frac{1}{6},$$

then

$$\mathrm{E}(\mathrm{E}(\tilde{x}\,|\mathcal{G})) = \mathrm{E}(\mathrm{E}(\tilde{x}\,|\mathcal{G})\,|\mathcal{H}) = \frac{9}{5} \cdot \frac{5}{6} + 2 \cdot \frac{1}{6} = \frac{11}{6} = \mathrm{E}(\tilde{x}\,|\mathcal{H}) = \mathrm{E}(\tilde{x}).$$

Properties of the conditional expectation given a σ-algebra:

1. (Theorem of total expectation). Let $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ and $\mathcal{G} \subset \mathcal{F}$, then $\mathrm{E}\left(\mathrm{E}\left(\tilde{y} \,|\mathcal{G}\right)\right) = \mathrm{E}\left(\tilde{y}\right)$. In particular, if $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$, then

$$\mathrm{E}\left(\mathrm{E}\left(\tilde{y} \,|\tilde{x}\right)\right) = \mathrm{E}\left(\tilde{y}\right). \qquad \text{(Adam's Law)}$$

2. If $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ are independent random variables, then

$$\mathrm{E}\left(\tilde{y} \,|\tilde{x}\right) = \mathrm{E}\left(\tilde{y}\right) \quad \text{a.s.} \quad [P].$$

3. Let $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ and $\mathcal{F}\left(\tilde{y}\right) \subset \mathcal{G} \subset \mathcal{F}$, then $\mathrm{E}(\tilde{y} \,|\mathcal{G}) = \tilde{y}$ a.s. $[P]$. In particular, $\mathrm{E}(\tilde{y} \,|\mathcal{F}) = \tilde{y}$ a.s. $[P]$, and

$$\mathrm{E}(\tilde{y} \,|\tilde{y}) = \mathrm{E}(\tilde{y} \,|\mathcal{F}\left(\tilde{y}\right)) = \tilde{y} \quad \text{a.s.} \quad [P].$$

4. Let $\tilde{z} : (\Omega, \mathcal{G}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ be random variables with $\mathcal{G} \subset \mathcal{F}$, then

$$\mathrm{E}\left(\tilde{z} \cdot \tilde{y} \,|\mathcal{G}\right) = \tilde{z} \cdot \mathrm{E}\left(\tilde{y} \,|\mathcal{G}\right) \quad \text{a.s.} \quad [P].$$

In particular, $\mathrm{E}\left(\tilde{z} \cdot \tilde{y} \,|\tilde{z}\right) = \tilde{z} \cdot \mathrm{E}\left(\tilde{y} \,|\tilde{z}\right)$ a.s. $[P]$.

5. If $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ and $\mathcal{H} = \{\emptyset, \Omega\}$ is the coarsest σ-algebra on Ω, then $\mathrm{E}\left(\tilde{y} \,|\mathcal{H}\right) = \mathrm{E}\left(\tilde{y}\right)$ a.s. $[P]$.

3.13. The law of iterated expectations

Theorem. Let $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ and $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$, then

(a) $\mathrm{E}\left(\mathrm{E}(\tilde{y}\,|\mathcal{G}_1)\,|\mathcal{G}_2\right) = \mathrm{E}(\tilde{y}\,|\mathcal{G}_1)$ a.s. $[P]$.

(b) The law of iterated expectations:

$$\mathrm{E}\left(\mathrm{E}(\tilde{y}\,|\mathcal{G}_2)\,|\mathcal{G}_1\right) = \mathrm{E}(\tilde{y}\,|\mathcal{G}_1) \text{ a.s. } [P].$$

3.14. Jensen's inequality

Johan Jensen (1859 - 1925)

- **Theorem.** Let g be a Borel measurable function from the convex set $C \subset \mathbb{R}^n$ to $\mathbb{R}$. Then,

$$\mathrm{E}\left(g\left(\tilde{y}\right)\right) \underset{(\geq)}{\leq} g\left(\mathrm{E}\left(\tilde{y}\right)\right), \qquad \text{(Jensen's inequality)}$$

  for all random vectors $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ with $\tilde{y}\left(\Omega\right) \subset C$ (i.e., $\tilde{y}\left(\omega\right) \in C$ for all $\omega \in \Omega$) if and only if g is concave (convex) on C.

- **Proof: (IF)** If g is concave, there exists a real vector $\delta \in \mathbb{R}^n$ such that

$$g\left(y\right) \leq g\left(\mathrm{E}\left(\tilde{y}\right)\right) + \delta^{\top}\left(y - \mathrm{E}\left(\tilde{y}\right)\right), \text{ for all } y \in C,$$

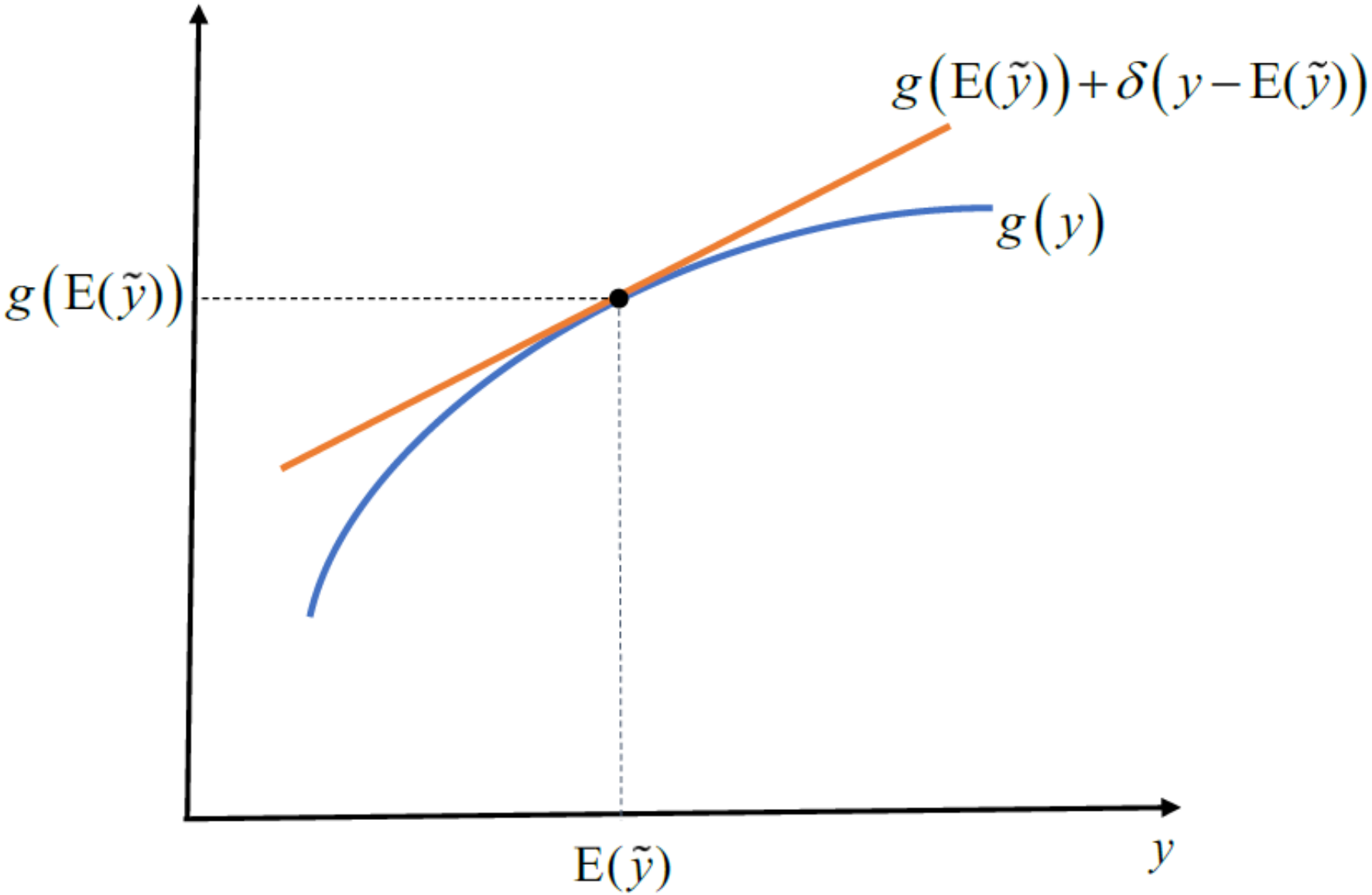
  where $\mathrm{E}\left(\tilde{y}\right) \in C \subset \mathbb{R}^n$ is the mean vector of the random vector $\tilde{y}$.

Therefore,

$$\mathrm{E}\left(g\left(\tilde{y}\right)\right) = \int_C g\left(y\right) dP_{\tilde{y}}\left(y\right) \leq \int_C \left[g\left(\mathrm{E}\left(\tilde{y}\right)\right) + \delta^{\top}\left(y - \mathrm{E}\left(\tilde{y}\right)\right)\right] dP_{\tilde{y}}\left(y\right)$$

$$= \int_C g\left(\mathrm{E}\left(\tilde{y}\right)\right) dP_{\tilde{y}}\left(y\right) + \int_C \delta^{\top}\left(y - \mathrm{E}\left(\tilde{y}\right)\right) dP_{\tilde{y}}\left(y\right)$$

$$= g\left(\mathrm{E}\left(\tilde{y}\right)\right) + \delta^{\top} \underbrace{\mathrm{E}\left[\tilde{y} - \mathrm{E}\left(\tilde{y}\right)\right]}_{=(0,0,\ldots,0)^{\top} \in \mathbb{R}^n} = g\left(\mathrm{E}\left(\tilde{y}\right)\right).$$


$g\left(\mathrm{E}(\tilde{y})\right)+\delta\left(y-\mathrm{E}(\tilde{y})\right)$
$g(y)$
$g\left(\mathrm{E}(\tilde{y})\right)$
$\mathrm{E}(\tilde{y})$
y

(ONLY IF) Consider a random vector $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ taking the values $x, y \in C$ with probability p and $(1-p)$, respectively. Then, Jensen's inequality becomes

$$pg(x) + (1-p)g(y) \leq g(px + (1-p)y),$$

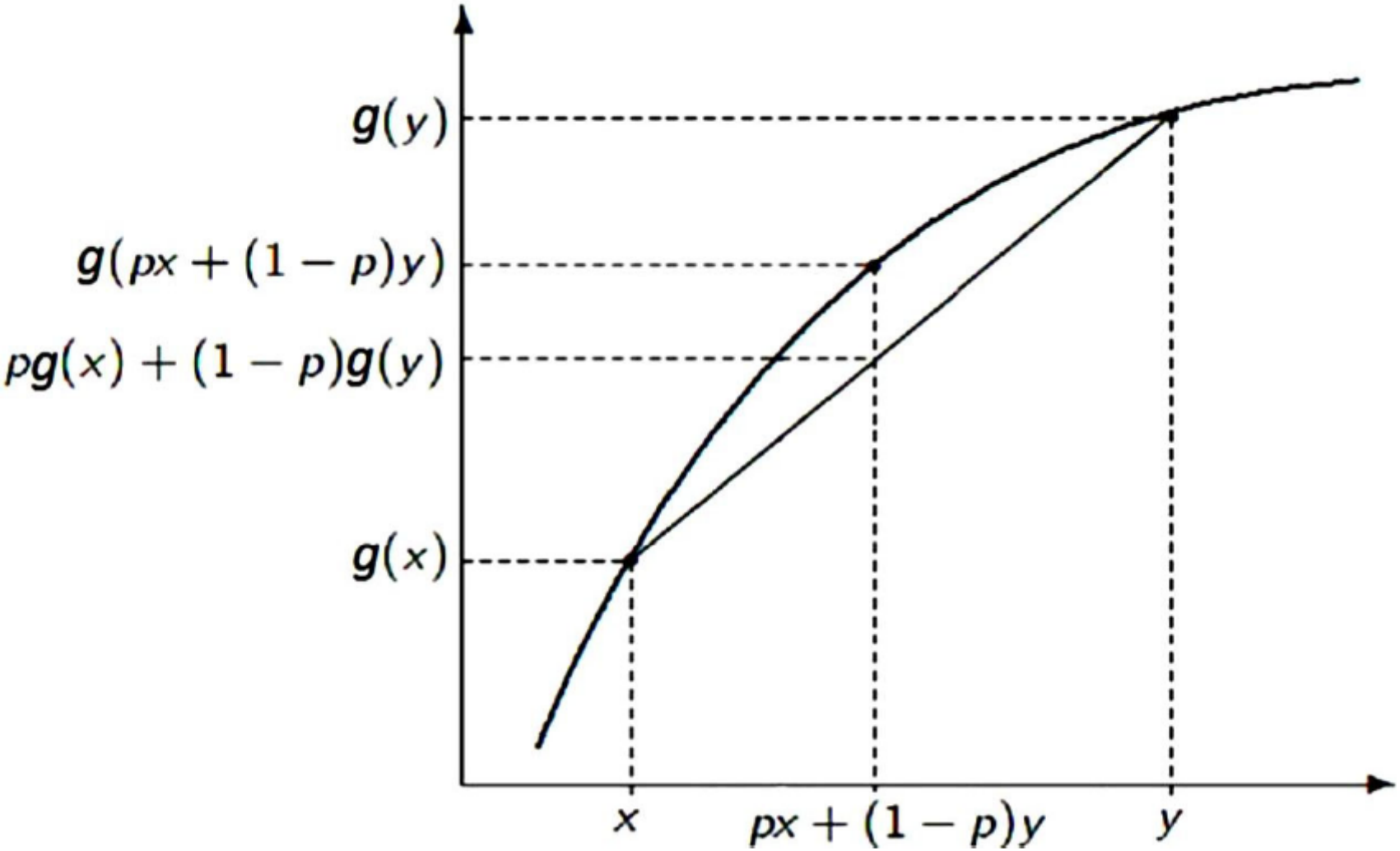
and, since this inequality holds for all values $x, y \in C$ and for all $p \in [0,1]$, we have just obtained the definition of concavity of the function g on C. $Q.E.D.$

- Jensen's inequality also holds for the conditional expectation of $\tilde{y}$ given $\tilde{x} = x$ (a.s. $[P_{\tilde{x}}]$) or given the σ-algebra $\mathcal{G}$ (a.s. $[P]$):
- For g concave (convex):

$$\mathrm{E}\left[g\left(\tilde{y}\right)|\tilde{x} = x\right] \underset{(\geq)}{\leq} g\left(\mathrm{E}\left[\tilde{y}\,|\tilde{x} = x\right]\right), \quad \text{a.s. } \left[P_{\tilde{x}}\right]$$

or

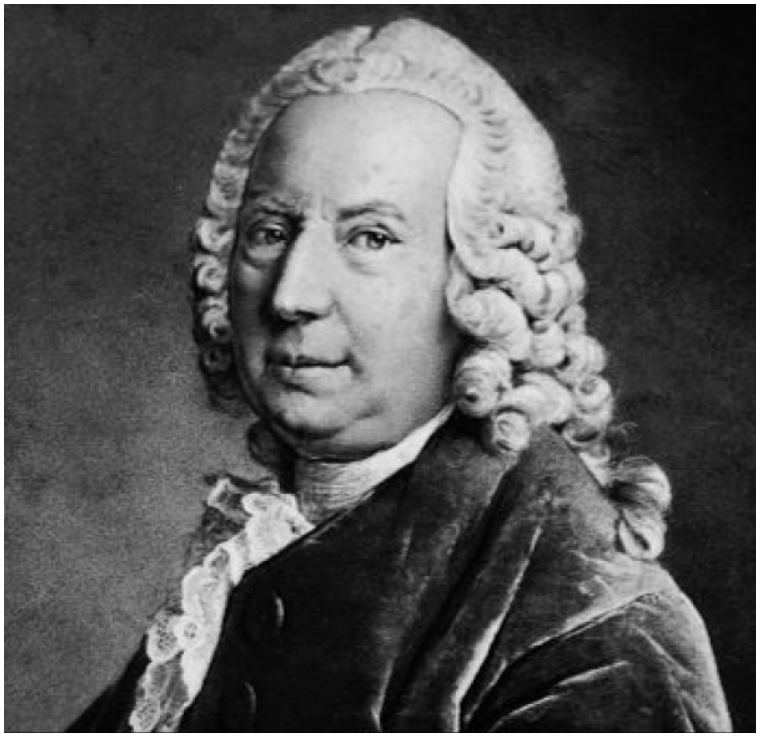
$$\mathrm{E}\left[g\left(\tilde{y}\right)|\mathcal{G}\right] \underset{(\geq)}{\leq} g\left(\mathrm{E}\left[\tilde{y}\,|\mathcal{G}\right]\right), \quad \text{a.s. } \left[P\right].$$


$g(y)$
$g(px + (1-p)y)$
$pg(x) + (1-p)g(y)$
$g(x)$
x
$px + (1-p)y$
y

- **Application to Economics:**
- **Definition.** An expected utility maximizing individual with utility function u is risk averse if

$$u\left[\mathrm{E}\left(\tilde{x}\right)\right] \geq \mathrm{E}\left[u\left(\tilde{x}\right)\right] \quad \text{for all random variables } \tilde{x}.$$

- From Jensen's inequality we can conclude that an individual is risk averse if and only if the utility function u (which is called the Bernoulli utility function) is concave.



Daniel Bernoulli (1700 - 1782)

Proof that $\mathrm{E}(\tilde{x}) = \int_{\mathbb{R}} x f_{\tilde{x}}(x)\, dx$ for absolutely continuous random variables.

Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ be an absolutely continuous random variable with density $f_{\tilde{x}}$. On the one hand, we know that

$$P_{\tilde{x}}(A) = \int_A dP_{\tilde{x}}(x) = \int_{\mathbb{R}} \mathbb{I}_A(x)\, dP_{\tilde{x}}(x). \tag{1}$$

On the other hand, from the definition of density, we have

$$P_{\tilde{x}}(A) = \int_A f_{\tilde{x}}(x)\, dx = \int_{\mathbb{R}} \mathbb{I}_A(x)\, f_{\tilde{x}}(x)\, dx, \tag{2}$$

as $\tilde{x}$ is absolutely continuous. Therefore, combining (1) and (2) we get

$$\int_A dP_{\tilde{x}}(x) = \int_A f_{\tilde{x}}(x)\, dx. \tag{3}$$

Let us compute now the integral of the non-negative simple function $s : (\mathbb{R}, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$ with respect to the probability distribution $P_{\tilde{x}}$. Recall that simple functions have the following form:

$$s(x) = \sum_{i=1}^{n} y_i \mathbb{I}_{A_i}(x), \tag{4}$$

where $\{A_1, A_2, ..., A_n\}$ is a finite partition of $\mathbb{R}$ and $s(x) = y_i$ if $x \in A_i$, $i = 1, ..., n$. Thus,

$$\begin{aligned}
\int_{\mathbb{R}} s(x)\, dP_{\tilde{x}}(x) &= \int_{\mathbb{R}} \sum_{i=1}^{n} y_i \mathbb{I}_{A_i}(x) dP_{\tilde{x}}(x) = \sum_{i=1}^{n} \int_{\mathbb{R}} y_i \mathbb{I}_{A_i}(x) dP_{\tilde{x}}(x) \\
&= \sum_{i=1}^{n} \int_{A_i} y_i dP_{\tilde{x}}(x) = \sum_{i=1}^{n} y_i \int_{A_i} dP_{\tilde{x}}(x) = \sum_{i=1}^{n} y_i \int_{A_i} f_{\tilde{x}}(x)\, dx \\
&= \sum_{i=1}^{n} \int_{A_i} s(x)\, f_{\tilde{x}}(x)\, dx = \int_{\bigcup_i A_i} s(x)\, f_{\tilde{x}}(x)\, dx = \int_{\mathbb{R}} s(x)\, f_{\tilde{x}}(x)\, dx,
\end{aligned} \tag{5}$$

where the first equality comes from (4); the second from the fact that the integral of a sum is the sum of integrals; the third from the definition of indicator function; the fourth from the fact that y_i is a constant in the Borel set A_i and, thus, it can be factorized; the fifth from (3); the sixth from the fact that $s(x) = y_i$ when $x \in A_i$; the seventh from the contable additivity property of the integral; and the eighth from the fact that $\bigcup_i A_i = \mathbb{R}$.

Let us now compute the integral $\int_{\mathbb{R}} g(x)\, dP_{\tilde{x}}(x)$, where $g : (\mathbb{R}, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$. We can use the definition of integral with respect to a measure to compute the integral of the positive and negative parts of g, g^+ and g^- (which are given by $g^+(x) = \max(g, 0) \geq 0$ and $g^- = \max(-g, 0) \geq 0$, respectively) as the supremum of the integrals of non-negative simple functions s and h such that $s \leq g^+$ and $h \leq g^-$, respectively. We thus have

$$\begin{aligned}
\mathrm{E}[g(\tilde{x})] &= \int_{\mathbb{R}} g(x)\, dP_{\tilde{x}}(x) = \int_{\mathbb{R}} g^+(x)\, dP_{\tilde{x}}(x) - \int_{\mathbb{R}} g^-(x)\, dP_{\tilde{x}}(x) \\
&= \sup_{0 \leq s \leq g^+} \int_{\mathbb{R}} s(x) dP_{\tilde{x}}(x) - \sup_{0 \leq h \leq g^-} \int_{\mathbb{R}} h(x) dP_{\tilde{x}}(x) \\
&= \sup_{0 \leq s \leq g^+} \int_{\mathbb{R}} s(x)\, f_{\tilde{x}}(x)\, dx - \sup_{0 \leq h \leq g^-} \int_{\mathbb{R}} h(x)\, f_{\tilde{x}}(x)\, dx \\
&= \int_{\mathbb{R}} g^+(x) f_{\tilde{x}}(x)\, dx - \int_{\mathbb{R}} g^-(x) f_{\tilde{x}}(x)\, dz = \int_{\mathbb{R}} g(x) f_{\tilde{x}}(x)\, dx, \qquad (6)
\end{aligned}$$

where the first equality is the definition of expectation; the second and the third come from the definition of the integral of a Borel measurable function; the fourth from (5); and the fifth and the sixth from the fact that the density $f_{\tilde{x}}$ is non-negative a.e. with respect to Lebesgue measure and from the definition of integral.

If the previous Borel measurable function g is the identity function, i.e., $g(x) = x$ for all $x \in \mathbb{R}$, then (6) becomes

$$\mathrm{E}(\tilde{x}) = \int_{\mathbb{R}} x dP_{\tilde{x}}(x) = \int_{\mathbb{R}} x f_{\tilde{x}}(x)\, dx. \qquad \textit{Q.E.D.}$$

Exercises. Probability and Statistics. IDEA.
3. Expectation

1. The density function of the absolutely continuous random variable $\tilde{x}$ is given by

$$f(x) = \begin{cases} \dfrac{1}{x \cdot (\ln 3)} & \text{for } 1 < x < 3 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

   (a) Find $\mathrm{E}(\tilde{x})$, $\mathrm{E}(\tilde{x}^2)$, and $\mathrm{E}(\tilde{x}^3)$.

   (b) Use the result of part (a) to determine the value of $\mathrm{E}(\tilde{x}^3 + 2\tilde{x}^2 - 3\tilde{x} + 1)$.

2. A casino offers a game of chance for a single player in which a fair coin is tossed at each stage. The initial stake starts at 2 dollars and is doubled every time head appears. The first time tail appears, the game ends and the player wins whatever is in the pot. Thus, the player wins 2 dollars if tail appears on the first toss, 4 dollars if head appears on the first toss and tail on the second, 8 dollars if head appears on the first two tosses and tail on the third, and so on. Mathematically, the player wins 2^x dollars, where x is the number of tosses and the probability of having x tosses is given by the probability function $f(x) = (1/2)^x$ for $x = 1, 2, 3....$ Show that $\mathrm{E}\left(2^{\tilde{x}}\right) = \infty$ so that the fair price to pay the casino for entering the game is infinite. This is the famous **St. Petersburg paradox,** according to which a player's payoff expectation is infinite if he is to receive 2^x dollars when, in a series of flips of a balanced coin, the first tail appears on the xth flip.

3. Mr. Adams and Ms. Smith are betting on repeated flips of a coin. At the start of the game Mr. Adams has a dollars and Ms. Smith has b dollars. At each flip the loser pays the winner one dollar, and the game continues until either player is "ruined". Making use of the fact that in an equitable game each player's mathematical expectation is zero, find the probability that Mr. Adams will win Ms. Smith's b dollars before he loses his a dollars.

4. Show that if $\tilde{x}$ is a random variable taking non-negative values with mean μ, then for any positive constant a,

$$P\{\tilde{x} \geq a\} \leq \frac{\mu}{a}.$$

   This inequality is called **Markov's inequality.**

5. Find the moment-generating function of the discrete random variable $\tilde{x}$, which has the probability function $f(x) = 2 \cdot (1/3)^x$ for $x = 1, 2, 3, ...$, and use it to determine the moments μ_1', μ_2', and σ^2.

6. Find the moment-generating function of the absolutely continuous random variable $\tilde{x}$ whose density is given by

$$f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ \\ 0 & \text{elsewhere} \end{cases}$$

and use it to find μ_1', μ_2' and σ^2.

7. If we let $R_{\tilde{x}}(t) = \ln M_{\tilde{x}}(t)$, show that $R_{\tilde{x}}'(0) = \mu$ and $R_{\tilde{x}}''(0) = \sigma^2$. Also, use these results to find the mean and the variance of a random variable $\tilde{x}$ having the moment-generating function

$$M_{\tilde{x}}(t) = e^{4\left(e^t - 1\right)}$$

8. Given the moment-generating function $M_{\tilde{x}}(t) = e^{3t+8t^2}$ of the random variable $\tilde{x}$, find the moment-generating function of the random variable $\tilde{z} = \frac{1}{4}(\tilde{x} - 3)$, and use it to find the mean and the variance of $\tilde{z}$.

9. If $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$ are random variables with the means $4, 9, 3$, the variances $3, 5, 7$, $\text{Cov}(\tilde{x}_1, \tilde{x}_2) = 1$, $\text{Cov}(\tilde{x}_2, \tilde{x}_3) = -2$, and $\text{Cov}(\tilde{x}_1, \tilde{x}_3) = -3$, find the mean and the variance of

   (*a*) $\tilde{y} = 2\tilde{x}_1 - 3\tilde{x}_2 + 4\tilde{x}_3$;

   (*b*) $\tilde{z} = \tilde{x}_1 + 2\tilde{x}_2 - \tilde{x}_3$.

10. If the joint density of $\tilde{x}$ and $\tilde{y}$ is given by

$$f(x, y) = \begin{cases} \frac{1}{3}(x + y) & \text{for } 0 < x < 1,\ 0 < y < 2 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

find the variance of $\tilde{w} = 3\tilde{x} + 4\tilde{y} - 5$.

11. A penny, which is unbalanced so that the probability of heads is 0.40, is tossed twice. What is the covariance between $\tilde{z}$, the number of heads obtained on the first toss, and $\tilde{w}$, the total number of heads obtained in the two tosses of the coin?

12. If getting a head is a success when we flip a coin, getting a six is a success when we roll a dice, and getting an ace is a success when we draw a card from an ordinary deck of 52 playing cards, find the mean and the standard deviation of the total number of successes when we

    (*a*) flip a balanced coin, roll a balanced dice, and then draw a card from a well-shuffled deck;

    (*b*) flip a balanced coin three times, roll a balanced dice twice, and then draw a card from a well-shuffled deck.

13. If the joint density function of two random variables $\tilde{x}$ and $\tilde{y}$ is given by

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y) & \text{for } 0 < x < 1,\ 0 < y < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

(a) find the conditional mean and the conditional variance of $\tilde{x}$ given $\tilde{y} = 1/2$;

(b) find the covariance between $\tilde{x}$ and $\tilde{y}$.

14. The random variables $\tilde{x}$ and $\tilde{y}$ have the joint density

$$f(x,y) = \begin{cases} -\frac{1}{2}\ln(x \cdot y) & \text{for } 0 < x < 1 \text{ and } 0 < y < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

Taking into account that, if $F(x) = x \cdot (\ln x) - x$, then $F'(x) = \ln x$,

(a) find the marginal density of $\tilde{x}$, $f_{\tilde{x}}(x)$.

(b) find the conditional density of $\tilde{x}$ given $\tilde{y} = y$, $f_{\tilde{x}|\tilde{y}}(x\,|y)$.

(c) are $\tilde{x}$ and $\tilde{y}$ independent?

(d) compute the expectation (or mean) of $\tilde{x}$, $\mathrm{E}(\tilde{x})$.

(e) compute the expectation of $\tilde{x}$ given $\tilde{y}$, $\mathrm{E}(\tilde{x}\,|\tilde{y})$.

(f) find $P\{\tilde{x} > 2\tilde{y}\}$.

15. (a) Let $\tilde{y}$ be an absolutely continuous random variable with density function (pdf) $g(y)$, $g : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$. If $\tilde{y} = y$, n independent observations $\tilde{x}_1, ..., \tilde{x}_n$ are taken, where each $\tilde{x}_i$ is absolutely continuous and has the same density $h(x_i; y)$, $h(\cdot; y) : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$. Indicate how to compute the conditional expectation of $\tilde{y}$ given $\tilde{x}_1 = x_1, ..., \tilde{x}_n = x_n$.

(b) Let $\tilde{y}$ be a discrete random variable with probability function (pmf) $g(y)$, $g : \tilde{y}(\Omega) \longrightarrow [0,1]$. If $\tilde{y} = y$, n independent observations $\tilde{x}_1, ..., \tilde{x}_n$ are taken, where each $\tilde{x}_i$ is discrete and has the same probability function $h(x; y)$, $h(\cdot; y) : \tilde{x}(\Omega) \longrightarrow [0,1]$. Indicate how to compute the conditional expectation of $\tilde{y}$ given $\tilde{x}_1 = x_1, ..., \tilde{x}_n = x_n$.

Note: Your answers to (a) and (b) should use only the functions $g(\cdot)$ and $h(\cdot;\cdot)$ and the values $x_1, ..., x_n$.

16. Let $\{B_1, B_2, ...\}$ be a countable collection of events with strictly positive probability that constitutes a partition of the sample space Ω. Establish the following version of the theorem of the total expectation:

$$\text{If } \mathrm{E}(\tilde{x}) \text{ exists, then } \mathrm{E}(\tilde{x}) = \sum_n P(B_n)\mathrm{E}(\tilde{x}\,|B_n).$$

Recall that $\mathrm{E}(\tilde{x}\,|B) \equiv \mathrm{E}(\tilde{x}\,|\mathbb{I}_B = 1)$, where B is an event and $\mathbb{I}_B$ is the indicator function of the set B.

17. Find the covariance of the two random variables whose joint density is given by

$$f(x,y) = \begin{cases} 2 & \text{for } x > 0,\ \ y > 0,\ \ x + y < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

18. We have a urn with 2 white balls and 3 black balls. We first extract one ball and we put it back in the urn. If the ball is white, we next extract two balls with no replacement. However, if the ball is black, we next extract only one ball. Find the moment-generating function of the number of black balls extracted in the second round.

19. Find the moment-generating function of the discrete random variable taking the values 0, 1, 2,... with the probability function

$$f(k) = P\{\tilde{x} = k\} = q^k \cdot p\,, \qquad k = 0, 1, 2, ..., \qquad q + p = 1\,, \quad q > 0\,, \quad p > 0$$

20. The moments of a random variable are $\mu_1' = 1$, $\mu_2' = 2$, and in general $\mu_k' = k$, for every strictly positive natural number k. Find its moment-generating function.

21. Is the function $g(t) = (2+t)^2$ the moment-generating function of some random variable?

22. Consider a discrete random variable $\tilde{x}$ with the following probability function:

$$f(x) = \begin{cases} p_0 & \text{for } x = 0 \\ \\ p_1 & \text{for } x = 1 \\ \\ 1 - p_0 - p_1 & \text{for } x = 2. \end{cases}$$

    (*a*) Find the moment-generating function of $\tilde{x}$.

    (*b*) Using the moment-generating function, compute $\mathrm{E}(\tilde{x})$.

    (*c*) Using the moment-generating function, compute $\mathrm{E}(\tilde{x}^2)$.

    (*d*) What is the variance of $\tilde{x}$?

23. The probability function of the discrete random variable $\tilde{x}$ is

$$f_{\tilde{x}}(x) = \frac{k}{x} \quad \text{for } x = 1, 2, 3.$$

    (*a*) The value of the constant k is 6/11. Why?

    (*b*) Find the moment-generating function of the random variable $\tilde{x}$.

    (*c*) Compute $\mathrm{E}(\tilde{x})$, $\mathrm{E}(\tilde{x}^2)$ and $\mathrm{E}(\tilde{x}^4)$ by using the moment-generating function.

    (*d*) Compute $\mathrm{Var}(\tilde{x})$.

    (*e*) Compute $\mathrm{Var}(\tilde{y})$, where $\tilde{y} = 3\tilde{x}^2 + 4$.

24. Consider a Borel measurable function f on the measure space $(\Omega, \mathcal{F}, \mu)$,

$$f : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B}).$$

(a) Assume that μ is a finite measure. Prove that, if $f \in \mathrm{L}^s$, then $f \in \mathrm{L}^r$ for $0 \leq r \leq s < \infty$.

(b) Assume that μ is the Lebesgue measure on the σ-algebra $\mathcal{B}$ of the Borel sets of $\mathbb{R}$, i.e., $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{B}$, and

$$f(x) = \begin{cases} \dfrac{1}{x} & \text{for } x \geq 1 \\ \\ 0 & \text{for } x < 1. \end{cases}$$

Does f belong to the L^1 space? Does f belong to the L^2 space? Compare the answers to the previous two questions with the result of part (a) and make a comment.

25. (a) Find the distribution function $F(x, y)$ of the random vector $(\tilde{x}, \tilde{y})$ whose density function is

$$f(x, y) = \begin{cases} \dfrac{6}{5}(x + y^2) & \text{for } 0 < x < 1,\ 0 < y < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

and check that $f(x, y) = \dfrac{\partial^2 F(x, y)}{\partial x \partial y}$ a.e. with respect to Lebesgue measure.

(b) Find $\mathrm{Var}\,(\tilde{x} \,|\tilde{y} = 1/3)$.

26. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Assume that the sample space Ω is the closed interval of real numbers $[0, 1]$, $\mathcal{F}$ is the $\sigma-$algebra of Borel sets in the interval $[0, 1]$, and the probability P on $(\Omega, \mathcal{F})$ is the Lebesgue measure on $[0, 1]$. Consider the random variables $\tilde{x}$, $\tilde{y}$, and $\tilde{z}$ on $(\Omega, \mathcal{F}, P)$ defined as follows:

$$\tilde{x}(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[0, \dfrac{1}{2}\right] \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{y}(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\dfrac{1}{2}, 1\right] \\ \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\tilde{z}(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[0, \dfrac{1}{4}\right] \cup \left[\dfrac{1}{2}, \dfrac{3}{4}\right] \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the joint probability function of the random variables $\tilde{x}$, $\tilde{y}$, and $\tilde{z}$.

(b) Are $\tilde{x}$ and $\tilde{y}$ independent? Are $\tilde{x}$ and $\tilde{z}$ independent? Are $\tilde{y}$ and $\tilde{z}$ independent?

(c) Find $\mathrm{Cov}\left(\tilde{x},\tilde{y}\right)$, $\mathrm{Cov}\left(\tilde{x},\tilde{z}\right)$, and $\mathrm{Cov}\left(\tilde{y},\tilde{z}\right)$.

27. The distribution of the random vector $(\tilde{x},\tilde{y})$ has the following density function:

$$f_{(\tilde{x},\tilde{y})}(x,y)=\begin{cases} ky & \text{if } x<5,\ y>0,\ x-2y>3 \\ 0 & \text{otherwise.}\end{cases}$$

(a) Prove that $k=3$.

(b) Compute the expectations $\mathrm{E}(\tilde{x})$ and $\mathrm{E}(\tilde{y})$.

(c) Compute the variance $\mathrm{Var}(\tilde{x})$.

(d) Compute the covariance $\mathrm{Cov}(\tilde{x},\tilde{y})$ and the correlation coefficient ρ between $\tilde{x}$ and $\tilde{y}$. Are $\tilde{x}$ and $\tilde{y}$ independent random variables?

(e) Find the marginal density function of $\tilde{y}$.

(f) Find the conditional density function of $\tilde{x}$ given $\tilde{y}=3/4$, $f_{\tilde{x}|\tilde{y}}\left(x\,|3/4\right)$.

(g) Compute the conditional expectation of $\tilde{x}$ given $\tilde{y}=3/4$, $\mathrm{E}(\tilde{x}\,|\tilde{y}=3/4)$.

(h) Compute the conditional variance of $\tilde{x}$ given $\tilde{y}=3/4$, $\mathrm{Var}(\tilde{x}\,|\tilde{y}=3/4)$.

28. Consider the random experiment consisting of rolling a balanced dice. Let $\widetilde{x}:\left(\Omega,2^{\Omega}\right)\longrightarrow(\mathbb{R},\mathcal{B})$ be the random variable representing the prize in euros you get as a function of the number of dots $\omega\in\Omega$ obtained when rolling the dice. The random variable $\widetilde{x}$ is defined as follows:

$$\widetilde{x}(\omega)=\begin{cases} 0 & \text{if } \omega\in\{1,4\} \\ 1 & \text{if } \omega=3 \\ 2 & \text{if } \omega\in\{2,5\} \\ 3 & \text{if } \omega=6\end{cases}$$

(a) Find the σ-algebra $\mathcal{F}\left(\widetilde{x}\right)$ induced on Ω by the random variable $\widetilde{x}$. Recall that $\mathcal{F}\left(\widetilde{x}\right)$ is the smallest σ-algebra $\mathcal{F}$ on Ω for which the random variable $\widetilde{x}:(\Omega,\mathcal{F})\longrightarrow(\mathbb{R},\mathcal{B})$ is measurable.

Consider the following partition of the sample space Ω : $G=\{\{1\},\{2,3,4\},\{5,6\}\}$ and let $\mathcal{G}$ be the σ-algebra generated by the partition G.

(b) Compute the following three conditional expectations of $\widetilde{x}$ given a σ-algebra: $\mathrm{E}\left(\widetilde{x}\,|\mathcal{F}\left(\widetilde{x}\right)\right)$, $\mathrm{E}\left(\widetilde{x}\,|\mathcal{G}\right)$, and $\mathrm{E}\left(\widetilde{x}\,|\mathcal{H}\right)$, where $\mathcal{H}=\{\Omega,\emptyset\}$ is the coarsest σ-algebra on Ω. Recall that a conditional expectation given the σ-algebra $\mathcal{F}$ is a function defined on the sample space Ω that is $\mathcal{F}$-measurable, $\mathrm{E}\left(\widetilde{x}\,|\mathcal{F}\right):(\Omega,\mathcal{F})\longrightarrow(\mathbb{R},\mathcal{B})$.

29. Let $\tilde{y}$, $\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_m, \tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ be random variables on the probability space $(\Omega, \mathcal{F}, P)$ and b_j, $j = 1, ..., m$, c_i, $i = 1, ..., n$, are scalars.

(a) Prove that

$$\text{Cov}\left(\tilde{y}, \sum_{i=1}^{n} c_i \tilde{x}_i\right) = \sum_{i=1}^{n} c_i \text{Cov}\left(\tilde{y}, \tilde{x}_i\right).$$

(b) Prove that

$$\text{Cov}\left(\sum_{j=1}^{m} b_j \tilde{y}_j, \sum_{i=1}^{n} c_i \tilde{x}_i\right) = \sum_{j=1}^{m} \sum_{i=1}^{n} b_j c_i \text{Cov}\left(\tilde{y}_j, \tilde{x}_i\right).$$

30. Let $\tilde{x}$ and $\tilde{y}$ be real-valued random variables on the probability space $(\Omega, \mathcal{F}, P)$. Both random variables have a strictly positive, finite variance.

(a) One of the following equalities is always true. Which one? Justify your answer.

(i) $\text{Var}\left(\tilde{x}\right) = \text{Var}\left(\text{E}\left(\tilde{x} \,|\tilde{y}\right)\right)$

(ii) $\text{Var}\left(\tilde{x}\right) = \text{E}\left(\text{Var}\left(\tilde{x} \,|\tilde{y}\right)\right)$

(iii) $\text{Var}\left(\tilde{x}\right) = \text{E}\left(\text{Var}\left(\tilde{x} \,|\tilde{y}\right)\right) + \text{Var}\left(\text{E}\left(\tilde{x} \,|\tilde{y}\right)\right)$

(iv) $\text{Var}\left(\tilde{x}\right) = \text{E}\left(\text{Var}\left(\tilde{x}\right)\right) + \text{Var}\left(\text{E}\left(\tilde{x} \,|\tilde{y}\right)\right)$

(b) Assume now that $\tilde{x}$ and $\tilde{y}$ are independent. In this case, in addition to the one found in part (a), are there other equalities that are true? Which ones? Justify your answer.

(c) Let $h : (\mathbb{R}, \mathcal{B}) \longrightarrow (\mathbb{R}, \mathcal{B})$ be a Borel measurable function. Find the value of $\text{E}\left[\left(\tilde{x} - \text{E}\left(\tilde{x} \,|\tilde{y}\right)\right) \cdot h\left(\tilde{y}\right)\right]$ and the value of $\text{Cov}\left[\left(\tilde{x} - \text{E}\left(\tilde{x} \,|\tilde{y}\right)\right), h\left(\tilde{y}\right)\right]$.

Hint: In this exercise you should use the theorem of total expectation, according to which, if $\tilde{x}$ and $\tilde{y}$ are real-valued random variables on the probability space $(\Omega, \mathcal{F}, P)$, then

$$\text{E}\left(\tilde{x}\right) = \text{E}\left(\text{E}\left(\tilde{x} \,|\tilde{y}\right)\right).$$

31. Consider the random experiment consisting of rolling a balanced dice. Let $\tilde{x} : \left(\Omega, 2^{\Omega}\right) \longrightarrow (\mathbb{R}, \mathcal{B})$ be the random variable representing the prize in euros you get as a function of the number of dots $\omega \in \Omega$ obtained when rolling the dice. The random variable $\tilde{x}$ is defined as follows:

$$\tilde{x}(\omega) = \begin{cases} 1 & \text{if } \omega \in \{1, 2, 3\} \\ 2 & \text{if } \omega = \{4, 5\} \\ 3 & \text{if } \omega \in \{6\} \end{cases}$$

(a) Find the σ-algebra $\mathcal{F}\left(\tilde{x}\right)$ induced on Ω by the random variable $\tilde{x}$.

Consider the following partition of the sample space Ω :

$$G = \{\{1,2\},\{3,4\},\{5,6\}\}$$

and let $\mathcal{G}$ be the σ-algebra generated by the partition G.

(b) Compute the following three conditional expectations of $\tilde{x}$ given a σ-algebra: $\mathrm{E}(\tilde{x}\,|\mathcal{F}(\tilde{x}))$, $\mathrm{E}(\tilde{x}\,|\mathcal{G})$, and $\mathrm{E}(\tilde{x}\,|\mathcal{H})$, where $\mathcal{H} = \{\Omega, \emptyset\}$ is the coarsest σ-algebra on Ω.

32. Prove that the characteristic function $\varphi_{\tilde{x}}(t) = \mathrm{E}\left(e^{it\tilde{x}}\right)$ of the random variable $\tilde{x}$ is equal to the following complex valued function:

$$\mathrm{E}\left[\cos(t\tilde{x}) + i\sin(t\tilde{x})\right], \text{ which is equal to } \left(\mathrm{E}\left[\cos(t\tilde{x})\right], \mathrm{E}\left[\sin(t\tilde{x})\right]\right).$$

Hint: You should make the corresponding Taylor's expansions around zero.

33. In this exercise, we will use both Markov's and Jensen's inequalities to derive some lower bounds on the moment-generating function of a random variable.

(a) Use Markov's inequality (which is given in Exercise 4 of this list) to prove that the moment-generating function $M_{\tilde{y}}(t)$ of a random variable $\tilde{y}$ satisfies

$M_{\tilde{y}}(t) \geq e^{ct} \cdot P\{\tilde{y} \geq c\}$, for all $c \in \mathbb{R}$ and all $t \geq 0$ where $M_{\tilde{y}}(t)$ is well defined.

Similarly, prove that

$M_{\tilde{y}}(t) \geq e^{ct} \cdot P\{\tilde{y} \leq c\}$, for all $c \in \mathbb{R}$ and all $t \leq 0$ where $M_{\tilde{y}}(t)$ is well defined.

(b) Use Jensen's inequality to prove that the moment-generating function $M_{\tilde{x}}(t)$ of any random variable $\tilde{x}$ with finite mean μ is bounded below by $e^{\mu t}$ for all $t \in \mathbb{R}$ where $M_{\tilde{x}}(t)$ is well defined.

34. Consider the random experiment consisting of rolling a balanced dice. Let $\tilde{x} : \left(\Omega, 2^{\Omega}\right) \longrightarrow (\mathbb{R}, \mathcal{B})$ be the random variable representing the prize in euros you get as a function of the number of dots $\omega \in \Omega$ obtained when rolling the dice. The random variable $\tilde{x}$ is defined as follows:

$$\tilde{x}(\omega) = \begin{cases} 2 & \text{if } \omega \in \{1,2\} \\ \\ 3 & \text{if } \omega \in \{3,4,5,6\} \end{cases},$$

and $\mathcal{F}(\tilde{x})$ is the σ-algebra induced on Ω by the random variable $\tilde{x}$.

Consider the following partition of the sample space Ω :

$$G = \{\{1,2,3\},\{4,5,6\}\}$$

and $\mathcal{G}$ is the σ-algebra generated by the partition G.

Compute the following three conditional expectations of $\tilde{x}$ given a σ-algebra: $\mathrm{E}(\tilde{x}\,|\mathcal{F}(\tilde{x}))$, $\mathrm{E}(\tilde{x}\,|\mathcal{G})$, and $\mathrm{E}(\tilde{x}\,|\mathcal{H})$, where $\mathcal{H} = \{\Omega, \emptyset\}$ is the coarsest σ-algebra on Ω.

35. Prove that, for every non-negative random variable $\tilde{x}$, the following inequalities hold:

$$\mathrm{E}(\tilde{x}) \geq \mathrm{GE}(\tilde{x}) \geq \mathrm{HE}(\tilde{x}),$$

where $\mathrm{GE}(\tilde{x})$ and $\mathrm{HE}(\tilde{x})$ are the geometric and the harmonic mean of $\tilde{x}$, respectively

36. (a) Prove that, for every random variable $\tilde{x}$,

$$\left(\mathrm{E}\left[\tilde{x}^2\right]\right)^{1/2} \geq \mathrm{E}(\tilde{x}).$$

Note: the expresion in the left-hand side is called the quadratic expectation (or mean) of $\tilde{x}$.

(b) Let us define the generalized (or power) expectation (or mean) of a non-negative random variable $\tilde{x}$ as

$$\mathrm{E}_p(\tilde{x}) = \left(\mathrm{E}\left[\tilde{x}^p\right]\right)^{1/p}, \quad \text{with } p \in \mathbb{R}.$$

Note: $\mathrm{E}_1(\tilde{x})$ is the arithmetic mean of $\tilde{x}$ and $\mathrm{E}_2(\tilde{x})$ is the quadratic mean of $\tilde{x}$.

Prove the following five properties for a non-negative random variable $\tilde{x}$:

(i) $\mathrm{E}_0(\tilde{x}) \equiv \lim_{p\to 0} \mathrm{E}_p(\tilde{x}) = \mathrm{GE}(\tilde{x})$, where $\mathrm{GE}(\tilde{x})$ is the geometric mean of $\tilde{x}$.

(ii) $\mathrm{E}_{-1}(\tilde{x}) = \mathrm{HE}(\tilde{x})$, where $\mathrm{HE}(\tilde{x})$ is the harmonic mean of $\tilde{x}$.

(iii) If $q > p$, where $q \in \mathbb{R}$ and $p \in \mathbb{R}$, then $\mathrm{E}_q(\tilde{x}) \geq \mathrm{E}_p(\tilde{x})$.

Assume for the rest of the exercise that the support $supp(P_{\tilde{x}})$ of the distribution $P_{\tilde{x}}$ of $\tilde{x}$ is finite.

(iv) $\mathrm{E}_\infty(\tilde{x}) \equiv \lim_{p\to\infty} \mathrm{E}_p(\tilde{x}) = x_{\max}$, where $x_{\max} = \max\{supp(P_{\tilde{x}})\}$.

(v) $\mathrm{E}_{-\infty}(\tilde{x}) \equiv \lim_{p\to-\infty} \mathrm{E}_p(\tilde{x}) = x_{\min}$, where $x_{\min} = \min\{supp(P_{\tilde{x}})\}$.

37. Prove that the irrational number e satisfies the following:

(a)

$$e = \sum_{r=0}^{\infty} \frac{1}{r!} = \lim_{x\to\infty} \left(1 + \frac{1}{x}\right)^x.$$

Hint: Use Taylor's expansion for the first equality and Newton's binomial theorem for the second.

(b) $e \in (2,3)$. *Hint:* Compute the Lebesgue integral $\int_{(1,e]} (\ln x)^3 dx$.

4. Special Distributions

- $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^n, \mathcal{B})$
- A discrete distribution is fully characterized by its probability function (pmf), $f_{\tilde{x}} : \tilde{x}(\Omega) \longrightarrow [0, 1]$, since

$$P_{\tilde{x}}(B) = P\{\tilde{x} \in B\} = \sum_{x \in B} f_{\tilde{x}}(x), \text{ for all } B \in \mathcal{B}.$$

- An absolutely continuous distribution is fully characterized by its density function (pdf), $f_{\tilde{x}} : (\mathbb{R}, \mathcal{B}) \longrightarrow (\overline{\mathbb{R}}, \mathcal{B})$, since

$$P_{\tilde{x}}(B) = P\{\tilde{x} \in B\} = \int_B f_{\tilde{x}}(x)dx, \text{ for all } B \in \mathcal{B}.$$

- Let $\tilde{x}$ be an absolutely continuous random variable with density $f_{\tilde{x}}$. We assume, without loss of generality, that the Borel set $A = \{x \in \mathbb{R}^n \,|\, f_{\tilde{x}}(x) \neq 0\}$ is open in $\mathbb{R}^n$.
- Thus, under this convention, the closure of A is the support $supp\,(P_{\tilde{x}})$ of the distribution of the absolutely continuous random variable $\tilde{x}$.

4.1. The discrete uniform distribution and the Dirac distribution

- The random variable/vector $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^n, \mathcal{B})$ has the **discrete uniform** distribution if its probability function is

$$f(x) = \frac{1}{k}, \text{ for } x = \underbrace{x_1, \ldots, x_k}_{\tilde{x}(\Omega)}, \text{ with } x_i \neq x_j \text{ for } i \neq j.$$

- If $k = 1$, then $\tilde{x}$ is a constant a.s.
- **Definition:** An extended random variable/vector $\tilde{x} : (\Omega, \mathcal{F}, P) \to \left(\overline{\mathbb{R}}^n, \mathcal{B}\right)$ is called "degenerate" when there is a value $x_0 \in \overline{\mathbb{R}}^n$ such that $P_{\tilde{x}}(x_0) = 1$. The distribution associated with this random variable is also called "degenerate".

- **Example:** The discrete random vector $(\tilde{x}_1, \tilde{x}_2)$, where $\tilde{x}_1$ is the number of points when rolling a dice and $\tilde{x}_2$ is the number of heads when tossing a coin has a probability function $f(x_1, x_2)$ summarized in the following table:

$x_2 \backslash x_1$	1	2	3	4	5	6
0	1/12	1/12	1/12	1/12	1/12	1/12
1	1/12	1/12	1/12	1/12	1/12	1/12

- Therefore,

$$f(x_1, x_2) = \frac{1}{12}, \text{ for } (x_1, x_2) \in \{1, 2, 3, 4, 5, 6\} \times \{0, 1\}.$$

- The **Dirac** distribution $P_{\tilde{x}}$ of the random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ satisfies

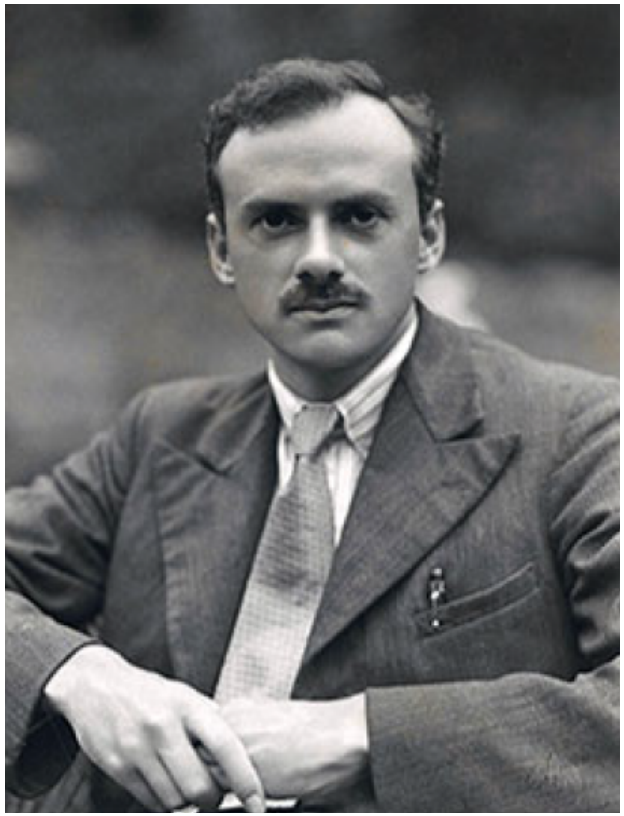
$$P_{\tilde{x}}(A) = \begin{cases} 1 & \text{if } a \in A \\ \\ 0 & \text{if } a \notin A, \end{cases}$$

  for all $A \in \mathcal{B}$.

- Note that $P_{\tilde{x}}\{a\} = 1$, i.e., the value a is taken almost surely by the random variable $\tilde{x}$.

- Some people define the density function of the previous Dirac distribution as a function (called the delta function) such that $f_{\tilde{x}}(x) = \infty$ if $x = a$, $f_{\tilde{x}}(x) = 0$ if $x \neq a$, and $\int_{\mathbb{R}} f_{\tilde{x}}(x)\, dx = 1$.

- Obviously, there is no density function with those properties, even though it can be viewed as the limit of a sequence of strictly positive density functions converging to zero for all $x \neq a$.



Paul Dirac (1902 - 1984)

4.2. The Bernoulli, binomial, Pascal, geometric, and hypergeometric distributions

Jacob Bernoulli (1654 - 1705)

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ has the **Bernoulli** distribution if its probability function (pmf) is

$$f(x;\theta) = \theta^x (1-\theta)^{1-x}, \quad \text{for } x = 0, 1.$$

- or

$$f(x;\theta) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \\ \theta & \text{for } x = 1. \end{cases}$$

- Mean and variance:

$$\mu = \theta$$

and

$$\sigma^2 = \theta(1-\theta).$$

- **The binomial distribution.**
- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has a binomial distribution (or is binomial) if its probability function is

$$b(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad \text{for } x = 0, 1, \ldots, n.$$

- *Motivation:* θ is the probability of a success in each trial. Then, $b(x; n, \theta)$ gives the probability of x successes in n independent trials.
- *Note:*

$$b(x; 1, \theta) = \theta^x (1-\theta)^{1-x}, \text{ for } x = 0, 1. \leftarrow \text{Bernoulli pmf}$$

- Obviously, if $\tilde{x}_1, ..., \tilde{x}_n$ are independently distributed random variables having a Bernoulli distribution with parameter θ, then its sum $\tilde{S} = \sum_{i=1}^{n} \tilde{x}_i$ has a binomial distribution with parameters n and θ.
- Therefore, if $\tilde{x}_1, ..., \tilde{x}_m$ are independently distributed random variables having a binomial distribution with parameters n_i and θ, for $i = 1, ..., m$, then its sum $\tilde{y} = \sum_{i=1}^{m} \tilde{x}_i$ has a binomial distribution with parameters $\sum_{i=1}^{m} n_i$ and θ.

- **Example:** Let $\tilde{x}$ be the number of heads when tossing 4 coins.

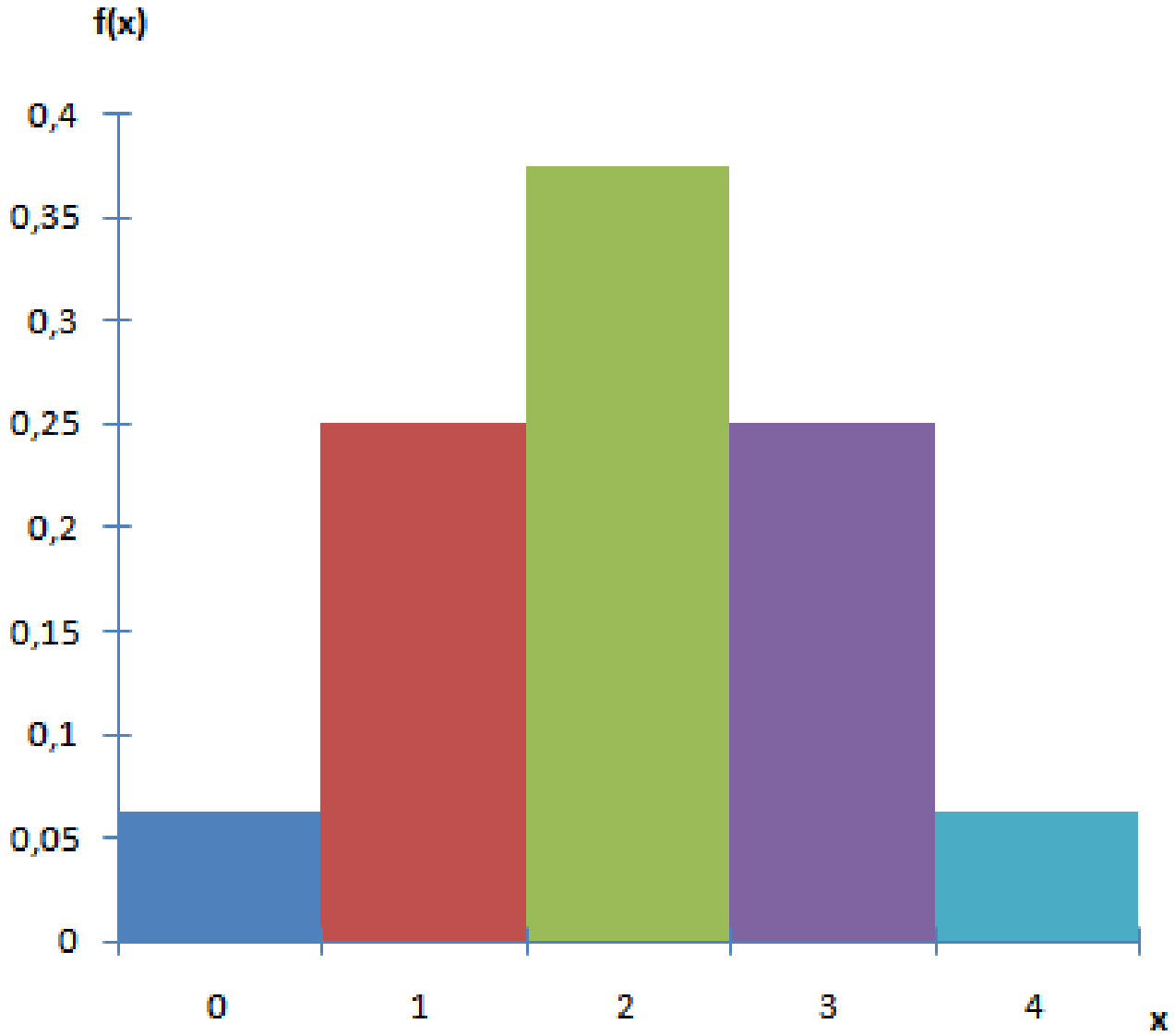
$$P\{\tilde{x}=x\}=f_{\tilde{x}}(x)=\underbrace{\binom{4}{x}\left(\frac{1}{2}\right)^{x}\left(\frac{1}{2}\right)^{4-x}}_{b\left(x;4,\frac{1}{2}\right)}=\binom{4}{x}\left(\frac{1}{2}\right)^{4}=\frac{1}{16}\binom{4}{x},$$

$$\text{for } x=\underbrace{0,1,2,3,4}_{\tilde{x}(\Omega)}$$

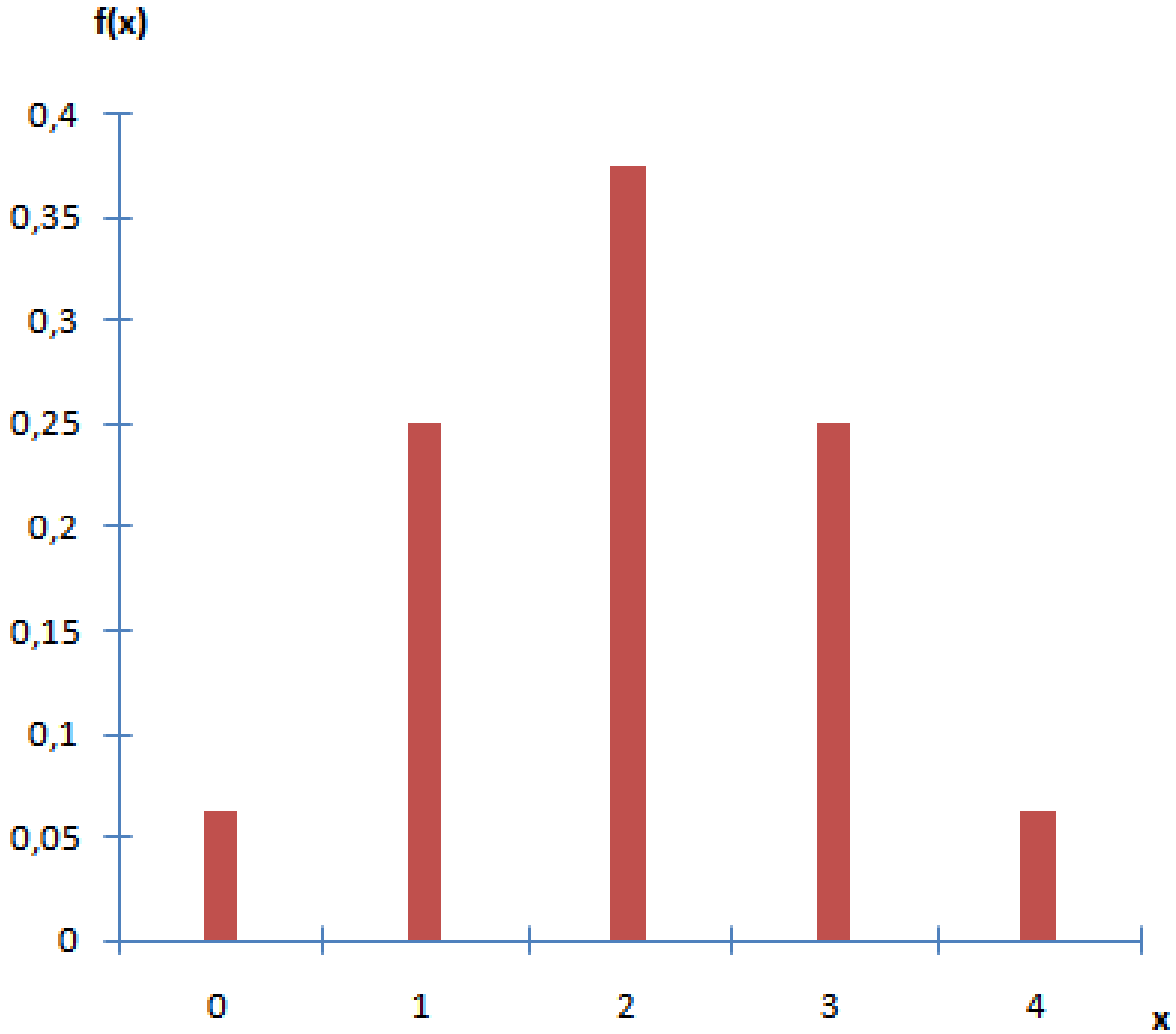
or

$$f_{\tilde{x}}(x)=b\left(x;4,\frac{1}{2}\right)=\begin{cases}1/16=0.0625 & \text{for } x=0\\ 4/16=0.25 & \text{for } x=1\\ 6/16=0.375 & \text{for } x=2\\ 4/16=0.25 & \text{for } x=3\\ 1/16=0.0625 & \text{for } x=4.\end{cases}$$

- Probability Histogram (which is symmetric iff $\theta = 1/2$):



- Probability Bar Chart (which is symmetric iff $\theta = 1/2$):



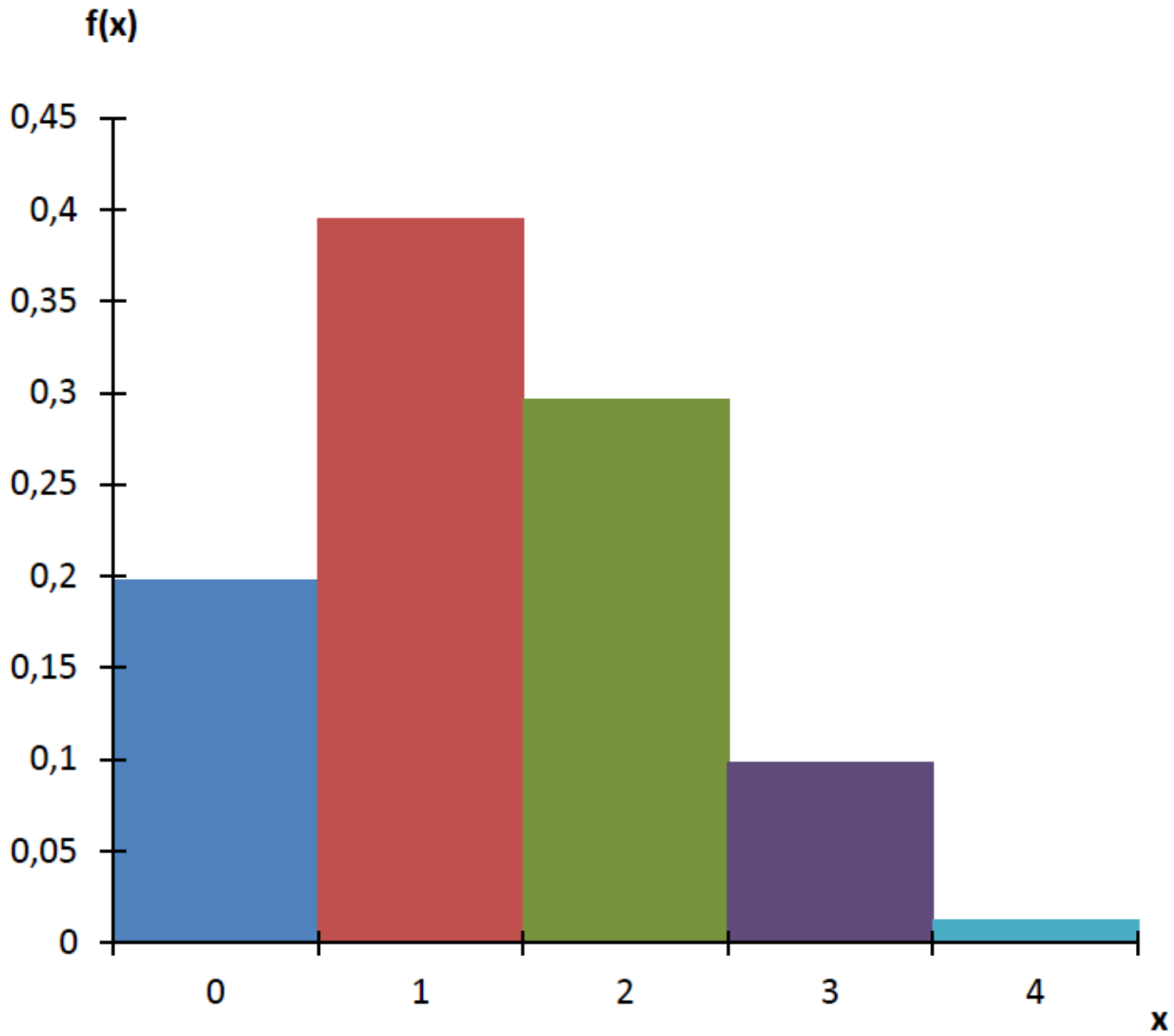
- **Example:** Let $\tilde{x}$ be the number of heads when tossing 4 coins. The coin is unbalanced and the probability of head is $\theta = 1/3$.

$$P\left\{\tilde{x} = x\right\} = f_{\tilde{x}}(x) = \underbrace{\binom{4}{x}\left(\frac{1}{3}\right)^{x}\left(\frac{2}{3}\right)^{4-x}}_{b\left(x;4,\frac{1}{3}\right)} \quad \text{for } x = \underbrace{0,1,2,3,4}_{\tilde{x}(\Omega)}$$

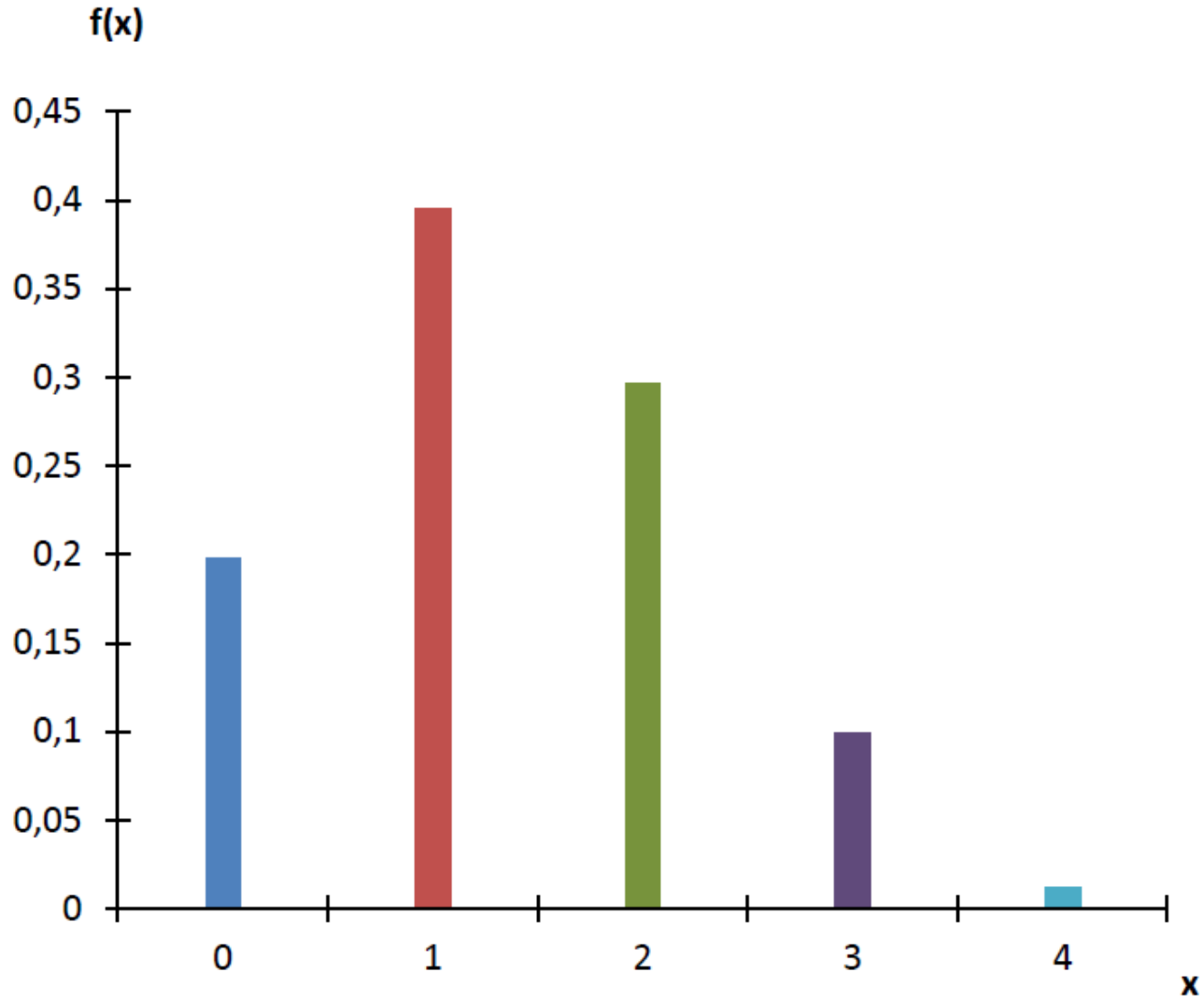
or

$$f_{\tilde{x}}(x) = b\left(x;4,\frac{1}{3}\right) = \begin{cases} 0.1975 & \text{for } x = 0 \\ 0.3951 & \text{for } x = 1 \\ 0.2963 & \text{for } x = 2 \\ 0.0988 & \text{for } x = 3 \\ 0.0123 & \text{for } x = 4. \end{cases}$$

- Probability Histogram (which is not symmetric since $\theta = 1/3$):



- Probability Bar Chart (which is not symmetric since $\theta = 1/3$):



- Moment-generating function of a binomial random variable:

$$M_{\tilde{x}}(t) = [1 + \theta(e^t - 1)]^n.$$

- **Proof:**

$$M_{\tilde{x}}(t) = \mathrm{E}\left(e^{t\tilde{x}}\right) = \sum_{x=0}^{n} e^{tx} \underbrace{\binom{n}{x} \theta^x (1-\theta)^{n-x}}_{b(x;n,\theta)}$$

$$= \sum_{x=0}^{n} \binom{n}{x} \left(\theta e^t\right)^x (1-\theta)^{n-x} = [\theta e^t + (1-\theta)]^n = [1 + \theta(e^t - 1)]^n.$$

- Then,

$$\mu = M'_{\tilde{x}}(0) = n[1 + \theta(e^t - 1)]^{n-1} \theta e^t \Big|_{t=0} = n\theta$$

and

$$\sigma^2 = \mathrm{E}\left(\tilde{x}^2\right) - \left[\mathrm{E}\left(\tilde{x}\right)\right]^2 = M''_{\tilde{x}}(0) - \mu^2 =$$
$$\left\{n(n-1)[1 + \theta(e^t - 1)]^{n-2} \theta^2 e^{2t} + n[1 + \theta(e^t - 1)]^{n-1} \theta e^t\right\}\Big|_{t=0}$$
$$- n^2\theta^2 = n(n-1)\theta^2 + n\theta - n^2\theta^2 = -n\theta^2 + n\theta = n\theta(1-\theta).$$

- **The Pascal (negative binomial or binomial waiting-time) distribution.**

Blaise Pascal (1623 - 1662)

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has a Pascal distribution if its probability function is

$$b^*(x;\, k, \theta) = \binom{x-1}{k-1} \theta^k (1-\theta)^{x-k}, \quad \text{for } x = k, k+1, \ldots$$

- *Motivation:* θ is the probability of success in each trial. Then, $b^*(x;\, k, \theta)$ gives the probability that the k^{th} success will occur on the x^{th} independent trial.
- *Note:* $b^*(x; k, \theta) = \theta b(k-1; x-1, \theta)$.
- Mean, variance, and moment-generating function:

$$\mu = \frac{k}{\theta}, \quad \sigma^2 = \frac{k(1-\theta)}{\theta^2},$$

$$M_{\tilde{x}}(t) = \left[\frac{\theta e^t}{1-(1-\theta)e^t}\right]^k \quad \text{for } t < \underbrace{-\ln(1-\theta)}_{+}.$$

- **The geometric distribution.**
- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has a geometric distribution if its probability function is

$$g(x;\theta) = \theta(1-\theta)^{x-1}, \quad \text{for } x = 1, 2, \ldots$$

- *Motivation:* θ is the probability of success in each trial. Then, $g(x;\theta)$ gives the probability that the first success will occur on the x^{th} independent trial.
- *Note:* $g(x;\theta) = b^*(x;1,\theta)$.
- Thus, the moment-generating function is

$$M_{\tilde{x}}(t) = \frac{\theta e^t}{1-(1-\theta)e^t} \quad \text{for } t < \underbrace{-\ln(1-\theta)}_{+}.$$

- Mean and variance:

$$\mu = M'_{\tilde{x}}(0) = \frac{1}{\theta} \quad \text{and} \quad \sigma^2 = M''_{\tilde{x}}(0) - \mu^2 = \frac{1-\theta}{\theta^2}.$$

- **The hypergeometric distribution.**
- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has a hypergeometric distribution if its probability function is

$$h(x; n, N, a) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}},$$

for $x = 0, 1 \ldots, n,\ x \leq a,\ n - x \leq N - a$.

- *Motivation:* Consider a set of N elements of which a are successes and $N-a$ are failures. We choose, without replacement, n of the N elements contained in the set. Then, $h(x;\, n, N, a)$ gives the probability of getting x successes.

- Mean and variance:

$$\mu = \frac{na}{N}$$

and

$$\sigma^2 = \frac{na(N-a)(N-n)}{N^2(N-1)}.$$

- *Note:*

$$\lim_{\substack{N \to \infty \\ a/N = \theta}} h(x; n, N, a) = b(x; n, \theta)$$

4.3. The multinomial and multivariate hypergeometric distributions.

- **The multinomial distribution.**
- The random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k) : (\Omega, \mathcal{F}, P) \to \left(\mathbb{R}^k, \mathcal{B}\right)$ has the multinomial distribution if its probability function is

$$m(x_1, x_2, \ldots, x_k; n, \theta_1, \theta_2, \ldots, \theta_k) = \binom{n}{x_1, x_2, \ldots, x_k} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k},$$

  for $x_i = 0, 1, \ldots, n$, $\sum_{i=1}^k x_i = n$ and $\sum_{i=1}^k \theta_i = 1$.
- Recall that

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1! \cdot x_2! \cdot \ldots \cdot x_k!}.$$

- *Motivation:* There are n independent trials permitting k exclusive outcomes, whose respective probabilities are $\theta_1, \theta_2, \ldots, \theta_k$ (with $\sum_{i=1}^{k} \theta_i = 1$). We shall refer to the outcomes as being of the first type, the second type, ... and the k^{th} type. Then, $m(x_1, x_2, \ldots, x_k; n, \theta_1, \theta_2, \ldots, \theta_k)$ gives the probability of getting x_1 outcomes of the first type, x_2 outcomes of the second type, ... and x_k outcomes of the k^{th} type (with $\sum_{i=1}^{k} x_i = n$).

- *Note:* $m(x, n-x; n, \theta, 1-\theta) = b(x;\, n, \theta)$.

- **Example:** Consider a very large population. 50% of the individuals have brown eyes, 30% have black eyes, and 20% have blue eyes. We pick randomly 8 individuals. The probability of picking 5 individuals with brown eyes, 2 with black eyes, and 1 with blue eyes is

$$m(5, 2, 1; 8, 0.5, 0.3, 0.2) = \frac{8!}{5!2!1!} \cdot 0.5^5 \cdot 0.3^2 \cdot 0.2^1 = 0.0945.$$

- **The multivariate hypergeometric distribution.**
- The random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_k) : (\Omega, \mathcal{F}, P) \to \left(\mathbb{R}^k, \mathcal{B}\right)$ has the multivariate hypergeometric distribution if its probability function is

$$f\left(x_1, x_2, \ldots, x_k; n, N, a_1, a_2, \ldots, a_k\right) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_k}{x_k}}{\binom{N}{n}},$$

  for $x_i = 0, 1, \ldots, n$, and $x_i \leq a_i$, where $\sum_{i=1}^{k} x_i = n$, $\sum_{i=1}^{k} a_i = N$.
- *Motivation:* There is a set of N elements, of which a_1 are elements of the first type, a_2 are elements of the second type, ..., a_k are elements of the k^{th} type (with $\sum_{i=1}^{k} a_i = N$). We choose, without replacement, n of the N elements in the set. Then
  $f\left(x_1, x_2, \ldots, x_k; n, N, a_1, a_2, \ldots, a_k\right)$ gives the probability of getting x_1 outcomes of the first type, x_2 outcomes of the second type, ... and x_k outcomes of the k^{th} type (with $\sum_{i=1}^{k} x_i = n$).

- *Note:* $f(x, n-x;\ n, N, a, N-a) = h(x;\ n, N, a)$.
- *Note:*

$$\begin{array}{c} \lim \\ N \to \infty \\ a_1/N = \theta_1 \\ a_2/N = \theta_2 \\ \dots \\ a_k/N = \theta_k \end{array} f\left(x_1, x_2, \dots, x_k; n, N, a_1, a_2, \dots, a_k\right)$$

$$= m\left(x_1, x_2, \dots, x_k; n, \theta_1, \theta_2, \dots, \theta_k\right).$$

- Example of the multinomial with $N = 2000$:

$$f\left(5, 2, 1; 8, 2000, 1000, 600, 400\right) = \frac{\binom{1000}{5}\binom{600}{2}\binom{400}{1}}{\binom{2000}{8}} = 0.0947$$

$$\approx m\left(5, 2, 1; 8, 0.5, 0.3, 0.2\right) = 0.0945.$$

4.4. Integration by parts for Lebesgue-Stieltjes integrals.

Recall that, if μ is the Lebesgue-Stieltjes measure associated with F,

$$\mu(a,b] = \int_{(a,b]} d\mu(x) \equiv \int_{(a,b]} dF(x) = F(b) - F(a),$$

$$\mu[a,b] = \mu\{a\} + \int_{(a,b]} dF(x)$$

$$= F(a) - F(a^-) + \int_{(a,b]} dF(x) = F(b) - F(a^-),$$

$$\mu(a,b) = -\mu\{b\} + \int_{(a,b]} dF(x)$$

$$= -\left[F(b) - F(b^-)\right] + \int_{(a,b]} dF(x) = F(b^-) - F(a),$$

$$\mu[a,b) = \mu\{a\} - \mu\{b\} + \int_{(a,b]} dF(x)$$

$$= F(a) - F(a^-) - \left[F(b) - F(b^-)\right] + \int_{(a,b]} dF(x) = F(b^-) - F(a^-).$$

- **Integration by parts.**
- From now on, whenever we write the integral of a function w.r.t. a measure it should be understood that the function is integrable w.r.t. that measure.
- Assume that $F : M \longrightarrow \mathbb{R}$ and $G : M \longrightarrow \mathbb{R}$ are continuously differentiable functions, where M is an open subset of $\mathbb{R}$, and $[a, b] \subset M$. Then,

$$\int_{[a,b]} G(x)F'(x)dx = \underbrace{F(b)G(b) - F(a)G(a)}_{F(x)G(x)|_a^b \equiv [F(x)G(x)]_a^b} - \int_{[a,b]} G'(x)F(x)dx.$$

- *Note:* The interval of integration can be replaced by (a, b) , $(a, b]$, or $[a, b)$.
- The previous Lebesgue integrals are equal to their Riemann counterparts as the functions $G \cdot F'$ and $G' \cdot F$ are continuous on $[a, b]$.

- Remember that, if F is a distribution function, then

$$\int_{(a,b]} G(x)dF(x) \equiv \int_{(a,b]} G(x)d\mu(x),$$

where μ is the Lebesgue-Stieltjes measure associated with F.

- Assume that $F : \mathbb{R} \longrightarrow \mathbb{R}$ is a distribution function, $G : M \longrightarrow \mathbb{R}$ is a continuously differentiable function, where M is an open subset of $\mathbb{R}$, and $[a, b] \subset M$. Then,

$$\int_{(a,b]} G(x)dF(x) = \underbrace{F(b)G(b) - F(a)G(a)}_{F(x)G(x)|_a^b \equiv [F(x)G(x)]_a^b} - \int_{(a,b]} F(x)G'(x)dx. \quad (\star)$$

- Note that $F(a^+) = F(a)$, because of the right-continuity of F.

- Since G is continuous on $[a,b]$, then

$$\underbrace{\int_{(a,b]} G(x)dF(x)}_{\text{Lebesgue-Stieltjes integral}} = \underbrace{\int_a^b G(x)dF(x)}_{\text{Riemann-Stieltjes integral}}.$$

- The last (Lebesgue) integral in $(\star)$ obviously satisfies

$$\int_{(a,b]} F(x)G'(x)dx = \int_{(a,b)} F(x)G'(x)dx$$
$$= \int_{[a,b]} F(x)G'(x)dx = \int_{[a,b)} F(x)G'(x)dx.$$

-

$$\int_{[a,b]} G(x)dF(x) = \mu\{a\}\, G(a) + \int_{(a,b]} G(x)dF(x) =$$
$$\left[F(a) - F(a^-)\right] G(a) + F(b)G(b) - F(a)G(a) - \int_{[a,b]} F(x)G'(x)dx$$
$$= F(b)G(b) - F(a^-)G(a) - \int_{[a,b]} F(x)G'(x)dx.$$

- $$\int_{(a,b)} G(x)dF(x) = -\mu\{b\} G(b) + \int_{(a,b]} G(x)dF(x) =$$

$$-\left[F(b)-F(b^-)\right] G(b)+F(b)G(b)-F(a)G(a) - \int_{(a,b]} F(x)G'(x)dx$$

$$= F(b^-)G(b) - F(a)G(a) - \int_{(a,b)} F(x)G'(x)dx.$$

- $$\int_{[a,b)} G(x)dF(x) = \mu\{a\} G(a) + \int_{(a,b)} G(x)dF(x) =$$

$$\left[F(a) - F(a^-)\right] G(a) + F(b^-)G(b) - F(a)G(a) - \int_{(a,b)} F(x)G'(x)dx$$

$$= F(b^-)G(b) - F(a^-)G(a) - \int_{[a,b)} F(x)G'(x)dx.$$

4.5. Lebesgue integration by change of variable: polar coordinates.

- Assume that (i) $g : M \longrightarrow \mathbb{R}$ is a continuously differentiable function, where M is an open subset of $\mathbb{R}$, and (ii) the function g restricted to $g^{-1}([a,b])$, $g : g^{-1}([a,b]) \longrightarrow [a,b]$, where $[a,b] \subset g(M)$ (or $g^{-1}([a,b]) \subset M$), is a bijective function (or one-to-one correspondence). Let $f : ([a,b], \mathcal{B}([a,b])) \longrightarrow (\mathbb{R}, \mathcal{B})$ be a Lebesgue integrable function. Then,

$$\int_{[a,b]} f(x)dx = \int_{g^{-1}([a,b])} f\underbrace{(g(y))}_{x} \left|g'(y)\right| dy.$$

- See the handout for an informal proof.
- Note that $g^{-1}([a,b])$ is the interval $\left[g^{-1}(a), g^{-1}(b)\right]$ if g is increasing or is the interval $\left[g^{-1}(b), g^{-1}(a)\right]$ if g is decreasing.

- Assume that (i) $g : M \longrightarrow \mathbb{R}^n$ is a continuously differentiable function, where M is an open subset of $\mathbb{R}^n$, and (ii) the function g restricted to $g^{-1}(B)$, $g : g^{-1}(B) \longrightarrow B$, where B is a Borel set in $\mathbb{R}^n$ such that $B \subset g(M)$ (or $g^{-1}(B) \subset M$), is a bijective function (or one-to-one correspondence).
- Let $J_g(y_1, y_2, ..., y_n)$ be the Jacobian matrix of the function

$$g : \underbrace{(y_1, ..., y_n)}_{y \in \mathbb{R}^n} \longmapsto \underbrace{(g_1(y_1, y_2, ..., y_n), ..., g_n(y_1, y_2, ..., y_n))}_{g(y) \in \mathbb{R}^n},$$

which is given by

$$J_g(y_1, y_2, ..., y_n) = \begin{pmatrix} \frac{\partial g_1(y)}{\partial y_1} & \frac{\partial g_1(y)}{\partial y_2} & \dots & \dots & \frac{\partial g_1(y)}{\partial y_n} \\ \frac{\partial g_2(y)}{\partial y_1} & \frac{\partial g_2(y)}{\partial y_2} & \dots & \dots & \frac{\partial g_2(y)}{\partial y_n} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{\partial g_n(y)}{\partial y_1} & \frac{\partial g_n(y)}{\partial y_2} & \dots & \dots & \frac{\partial g_n(y)}{\partial y_n} \end{pmatrix},$$

- Let $f : (B, \mathcal{B}(B)) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be a Lebesgue integrable function. Then,

$$\int_B f(x_1, x_2, ..., x_n) d(x_1, ..., x_n) =$$

$$\int_{g^{-1}(B)} f(\underbrace{g(y_1, y_2, ..., y_n)}_{(x_1, x_2, ..., x_n)}) \left| J_g(y_1, y_2, ..., y_n) \right| d(y_1, ..., y_n),$$

where $\left| J_g(y_1, y_2, ..., y_n) \right|$ is the absolute value of the determinant of $J_g(y_1, y_2, ..., y_n)$.

- Alternatively, we can assume that (i) $g : M \longrightarrow \mathbb{R}^n$ is a continuously differentiable function, where M is an open subset of $\mathbb{R}^n$, and (ii) the function g restricted to A, $g : A \longrightarrow g(A)$, where A is a Borel set in $\mathbb{R}^n$ such that $A \subset M$, is a bijective function (or one-to-one correspondence). Let $f : (g(A), \mathcal{B}(g(A))) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be a Lebesgue integrable function. Then,

$$\int_{g(A)} f(x_1, x_2, ..., x_n) d(x_1, ..., x_n) =$$

$$\int_A f(\underbrace{g(y_1, y_2, ..., y_n)}_{(x_1, x_2, ..., x_n)}) \left| J_g(y_1, y_2, ..., y_n) \right| d(y_1, ..., y_n).$$

- A very useful change of variable (or substitution) when we integrate over a circular region is the change to **polar coordinates** in $\mathbb{R}^2$.
- Consider the function

$$g : \mathbb{R}_{++} \times [0, 2\pi) \longrightarrow \mathbb{R}^2,$$

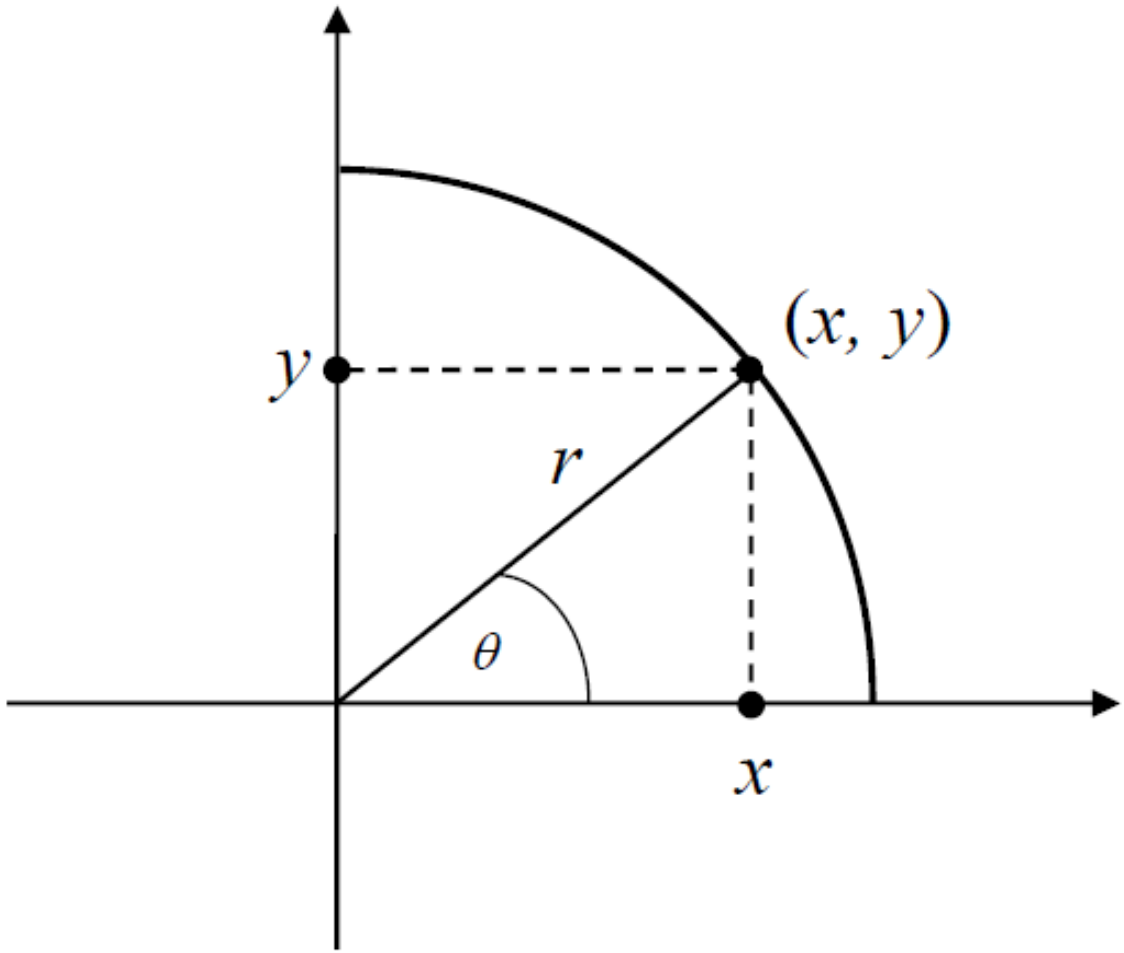
given by

$$(x, y) = g(r, \theta), \text{ with } \begin{cases} x = r \cdot \cos\theta \\ \\ y = r \cdot \sin\theta. \end{cases}$$

- Note that

$$x^2 + y^2 = r^2 \cos^2 \theta + r^2 \sin^2 \theta = r^2,$$

which is the equation of a circumference with the radius r.



- Clearly, $g\left(\mathbb{R}_{++} \times [0, 2\pi)\right) = \mathbb{R}^2 \backslash (0, 0)$, where $\mathbb{R}^2 \backslash (0, 0)$ is the real plane without the point $(0, 0)$, which has zero Lebesgue measure.
- Moreover, the function g restricted to $g^{-1}\left(\mathbb{R}^2 \backslash (0, 0)\right) = \mathbb{R}_{++} \times [0, 2\pi)$, $g : \mathbb{R}_{++} \times [0, 2\pi) \longrightarrow \mathbb{R}^2 \backslash (0, 0)$, is bijective.
- C is a circular region with center at $(0, 0)$ in $\mathbb{R}^2 \backslash (0, 0)$ if $g^{-1}(C)$ is a measurable rectangle on $\mathbb{R}_{++} \times [0, 2\pi)$, that is, if $g^{-1}(C)$ is the Cartesian product of two Borel sets, $g^{-1}(C) = D_1 \times D_2$, with $D_1 \in \mathcal{B}\left(\mathbb{R}_{++}\right)$ and $D_2 \in \mathcal{B}\left([0, 2\pi)\right)$.
- If the point (x, y) belongs to a circular region with center at (c_1, c_2), then the point (s, t), where $s = x - c_1$ and $t = y - c_2$, belongs to a circular region with center at $(0, 0)$.
- Thus, if we are integrating w.r.t. (x, y) over a circular region with center at (c_1, c_2), we should first make the change of variable $(x, y) = h(s, t) = (s + c_1, t + c_2)$, where the Jacobian matrix $J_h(s, t)$ is the identity matrix and, thus, $|J_h(s, t)| = 1$, and then integrate w.r.t. (s, t) over a circular region with center at $(0, 0)$.

- Obviously, we can define the inverse function g^{-1} defined on $\mathbb{R}^2 \setminus (0,0)$ as

$$(r,\theta) = g^{-1}(x,y), \text{ with } \begin{cases} r = \left(x^2+y^2\right)^{1/2} \\ \\ \theta = \arctan\left(\dfrac{y}{x}\right). \end{cases}$$

- Moreover,

$$J_g(r,\theta) = \begin{bmatrix} \cos\theta & -r\cdot\sin\theta \\ \\ \sin\theta & r\cdot\cos\theta \end{bmatrix},$$

so that

$$|J_g(r,\theta)| = \left|r\cos^2\theta + r\sin^2\theta\right| = |r| = r.$$

- Therefore, by making the change to polar coordinates, we can compute the following integral over a circular region C with center at $(0,0)$:

$$\int_C f(x,y)\, d(x,y) = \int_{g^{-1}(C)} f(r\cos\theta, r\sin\theta) r\; d(r,\theta)$$
$$= \int_{D_1}\int_{D_2} f(r\cos\theta, r\sin\theta) r d\theta dr,$$

where we use Fubini's theorem in the last equality since $g^{-1}(C)$ is a measurable rectangle on $\mathbb{R}_{++} \times [0, 2\pi)$, i.e., it is the Cartesian product of two Borel sets, $g^{-1}(C) = D_1 \times D_2$, with $D_1 \in \mathcal{B}(\mathbb{R}_{++})$ and $D_2 \in \mathcal{B}([0,2\pi))$.

- Moreover, we can easily compute the following integral over a circular region C with center at $(0,0)$:

$$\int_C h\left(x^2+y^2\right) d(x,y) = \int_{g^{-1}(C)} h(r^2) r\; d(r,\theta)$$
$$= \int_{D_1}\int_{D_2} h(r^2) r d\theta dr = \left(\int_{D_1} h(r^2) r dr\right)\left(\int_{D_2} d\theta\right).$$

- **Example:**

$$\int_0^\infty \int_0^\infty e^{-\left(x^2+y^2\right)} dxdy = \int_{\mathbb{R}_+ \times \mathbb{R}_+} e^{-\left(x^2+y^2\right)} d(x,y) =$$

$$= \left(\int_{\mathbb{R}_{++}} e^{-r^2} r dr \right) \left(\int_{[0,\pi/2)} d\theta \right) = \left[\frac{-e^{-r^2}}{2} \right]_0^\infty \cdot \left[\theta \right]_0^{\pi/2} = \frac{1}{2} \cdot \frac{\pi}{2} = \frac{\pi}{4}.$$

- Note that

$$\int_0^\infty \int_0^\infty e^{-\left(x^2+y^2\right)} dxdy = \left(\int_0^\infty e^{-x^2} dx \right) \left(\int_0^\infty e^{-y^2} dy \right) = M^2,$$

where $M = \int_0^\infty e^{-x^2} dx = \int_0^\infty e^{-y^2} dy$.

- Therefore, given the symmetry of e^{-x^2} with respect to zero,

$$M = \int_0^\infty e^{-x^2} dx = \left(\frac{\pi}{4} \right)^{1/2} = \frac{\sqrt{\pi}}{2} \Longrightarrow \int_{-\infty}^\infty e^{-x^2} dx = \sqrt{\pi}.$$

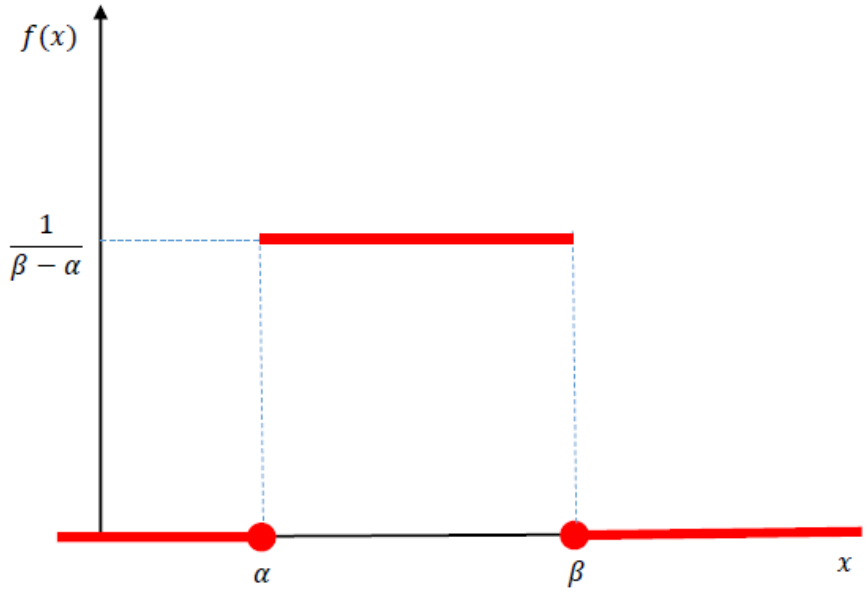
4.6. The uniform density

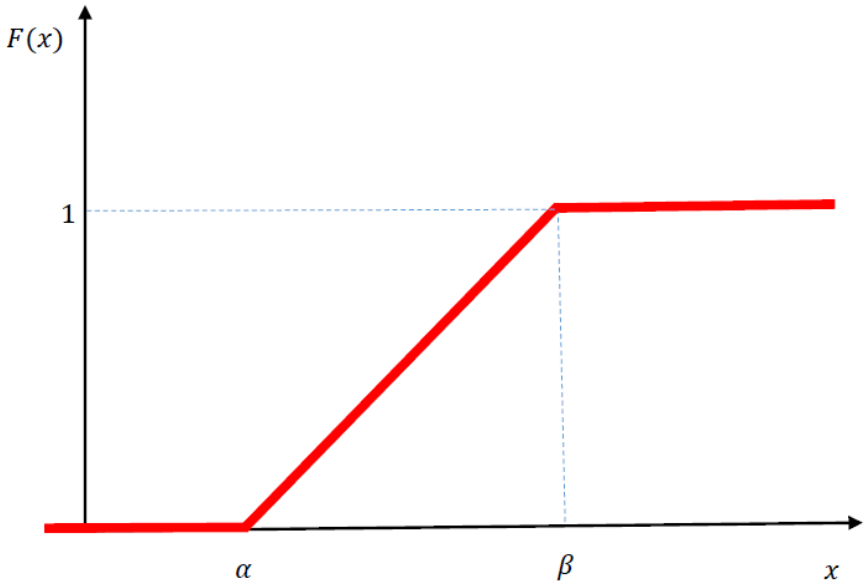
- The absolutely continuous random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ is uniform (or has a uniform distribution) on (α, β) if its density function is

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{for } x \in (\alpha, \beta) \\ 0 & \text{otherwise.} \end{cases}$$

- The distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x \leq \alpha \\ \dfrac{x - \alpha}{\beta - \alpha} & \text{for } x \in (\alpha, \beta) \\ 1 & \text{for } x \geq \beta. \end{cases}$$


$f(x)$
$\frac{1}{\beta - \alpha}$
α
β
x


$F(x)$
1
α
β
x

- Quantile function:

$$Q(p) = \alpha + (\beta - \alpha)p \quad \text{for } p \in (0,1).$$

- If $\alpha = 0$ and $\beta = 1$, then $\tilde{x}$ has the uniform distribution on $(0,1)$ (also called the standard uniform distribution). In this case, $Q(p) = p = F(p)$ for $p \in (0,1)$.
- **Proposition.** Let F be a distribution function and Q is its associated quantile function (i.e., $Q(p) = \inf\{x \,|\, F(x) \geq p\}$ for $p \in (0,1)$). If the random variable $\tilde{p}$ has the standard uniform distribution, then the random variable $\tilde{x} = Q(\tilde{p})$ has the distribution function F.
- **Proof.** Recall that the quantile function satisfies the following:

$$Q(p) \leq x \quad \text{if and only if} \quad F(x) \geq p.$$

Since the random variable $\tilde{p}$ has the standard uniform distribution, it holds that $F_{\tilde{p}}(p) = P\{\tilde{p} \leq p\} = p$ for $p \in (0,1)$. Therefore,

$$P\{\tilde{x} \leq x\} = P\{Q(\tilde{p}) \leq x\} = P\{\tilde{p} \leq F(x)\} = F(x). \qquad Q.E.D.$$

- The previous proposition allows us to generate random values from a distribution having an arbitrarily given distribution function F.
- To do so, generate random values p_i, $i = 1, 2, ..., N$, from the standard uniform distribution and apply to these values the quantile function, $x_i = Q(p_i)$. Then, the values x_i, $i = 1, 2, ..., N$, are random values from a distribution having the distribution function F.

- Let $\alpha \leq c < d \leq \beta$, then,

$$P\left\{c \underset{(<)}{\leq} \widetilde{x} \underset{(<)}{\leq} d\right\} = P_{\tilde{x}}[c,d] = \frac{d-c}{\beta-\alpha}.$$

- Mean:

$$\mu = \mathrm{E}(\tilde{x}) = \int_{-\infty}^{\infty} xf(x)dx = \int_{\alpha}^{\beta} x\left(\frac{1}{\beta-\alpha}\right)dx = \frac{1}{\beta-\alpha}\left[\frac{x^2}{2}\right]_{\alpha}^{\beta}$$

$$= \frac{1}{\beta-\alpha}\left[\frac{\beta^2-\alpha^2}{2}\right] = \frac{\beta+\alpha}{2},$$

where the last inequality follows since $(\beta+\alpha)(\beta-\alpha) = \beta^2-\alpha^2$.

- Moreover,

$$\mathrm{E}\left(\tilde{x}^2\right) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_{\alpha}^{\beta} x^2 \left(\frac{1}{\beta - \alpha}\right) dx$$

$$= \frac{1}{\beta - \alpha}\left[\frac{x^3}{3}\right]_{\alpha}^{\beta} = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)},$$

  so that

$$\mathrm{Var}\left(\tilde{x}\right) = \mathrm{E}\left(\tilde{x}^2\right) - \left[\mathrm{E}\left(\tilde{x}\right)\right]^2 = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} - \left[\frac{\beta + \alpha}{2}\right]^2 = \frac{(\beta - \alpha)^2}{12},$$

  where the last equality is obtained after some tedious algebra.

- Moment-generating function:

$$M_{\tilde{x}}(t) = \begin{cases} \dfrac{e^{t\beta} - e^{t\alpha}}{t\left(\beta - \alpha\right)} & \text{for } t \neq 0 \\ \\ 1 & \text{for } t = 0. \end{cases}$$

4.7. The gamma, exponential, and chi-square distributions.

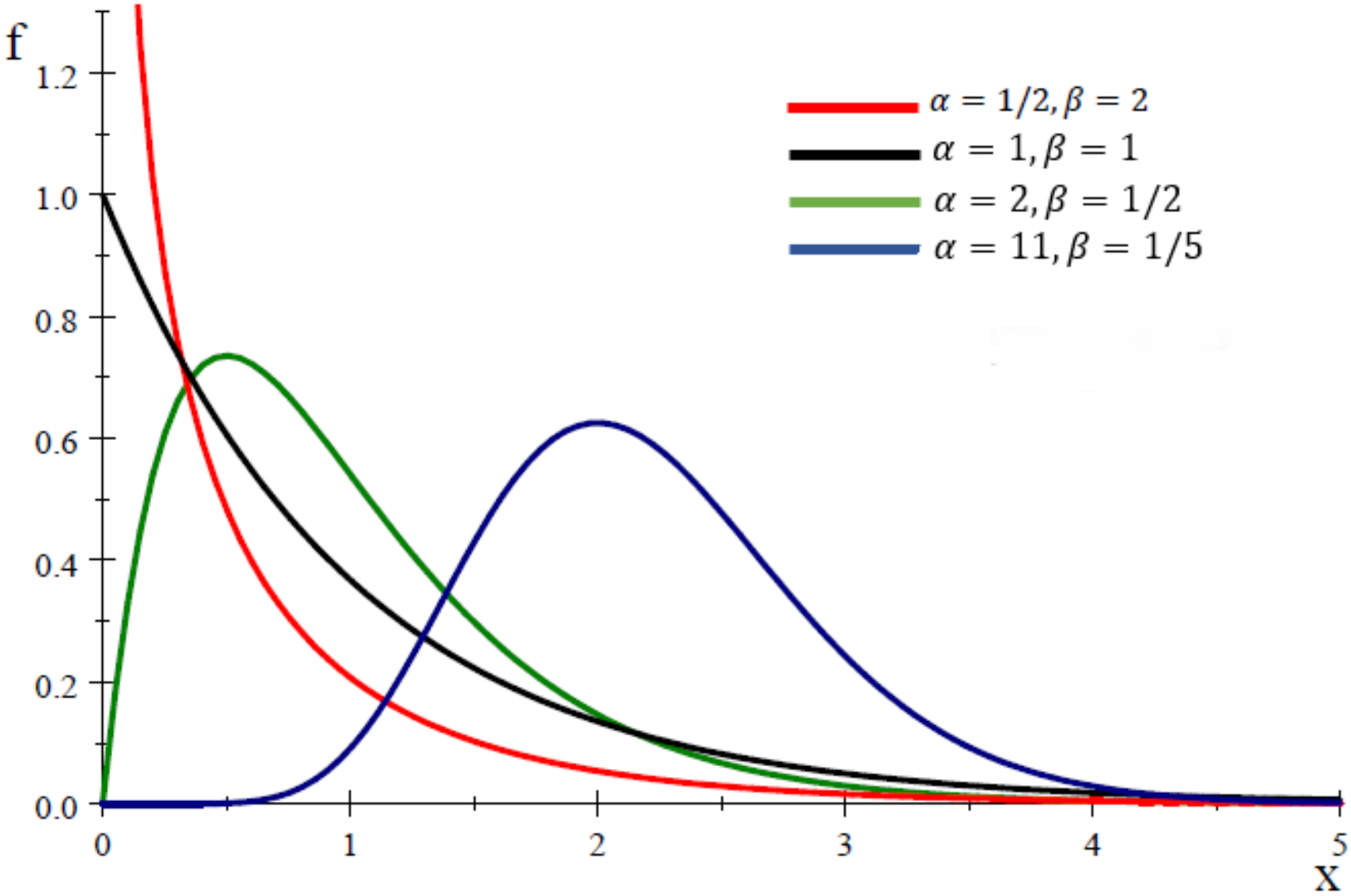
- **The gamma distribution.**
- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has the gamma distribution if its density is

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

with $\alpha > 0$, $\beta > 0$ and where

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy, \text{ for } \alpha > 0. \qquad \text{(the gamma function)}$$

Densities of the gamma distribution:



- Making the change of variable

$$x = g(y) = \beta y \Leftrightarrow y = g^{-1}(x) = \frac{x}{\beta}, \text{ so that } \frac{dx}{dy} = g'(y) = \beta > 0,$$

we can check that

$$\int_0^\infty f(x;\alpha,\beta)dx = \int_0^\infty kx^{\alpha-1}e^{-x/\beta}dx = k\int_0^\infty (\beta y)^{\alpha-1}e^{-y}\beta dy$$

$$= k\beta^\alpha \underbrace{\int_0^\infty y^{\alpha-1}e^{-y}dy}_{\Gamma(\alpha)} = 1.$$

Therefore,

$$k = \frac{1}{\beta^\alpha \Gamma(\alpha)}.$$

Properties of the gamma function $\Gamma(\alpha)$.

- **(a)** $\Gamma(1) = \int_0^\infty e^{-y}\,dy = [-e^{-y}]_0^\infty = \lim_{y\to\infty}(-e^{-y}) + e^0 = 1.$
- **(b)** $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for $\alpha > 1$.
- **Proof.** Integrating by parts:

$$\Gamma(\alpha) = \int_0^\infty \underbrace{y^{\alpha-1}}_{F(y)}\underbrace{e^{-y}}_{G'(y)}dy = \left[-y^{\alpha-1}e^{-y}\right]_0^\infty - \int_0^\infty \underbrace{(\alpha-1)y^{\alpha-2}}_{F'(y)}\underbrace{\left(-e^{-y}\right)}_{G(y)}dy$$

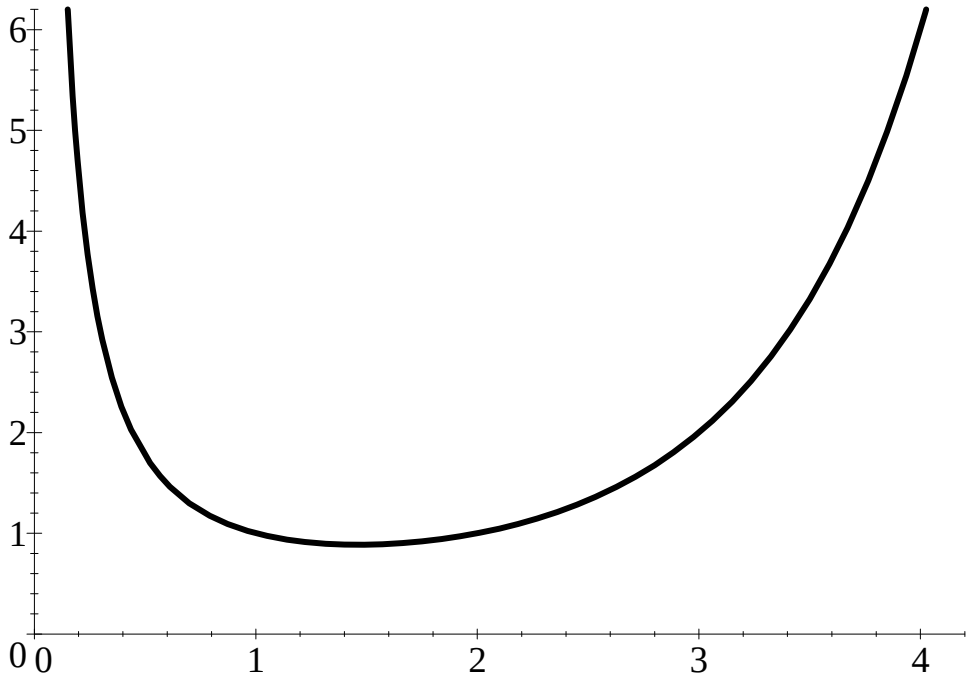
$$= 0 + (\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}\,dy = (\alpha-1)\Gamma(\alpha-1).$$

- Note that (a), (b), and the continuity of the gamma function imply that

$$1 = \lim_{\alpha\to 1^+}\Gamma(\alpha) = \lim_{x\equiv\alpha-1\to 0^+}[x\cdot\Gamma(x)] \Longrightarrow \lim_{x\to 0^+}\Gamma(x) = \infty.$$

- **(c)** $\Gamma(\alpha) = (\alpha-1)!$ when α is a strictly positive integer.

- **(d)** $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
  - **Proof:** See handout 1.A.
- **Corollary.** $\int_0^\infty e^{-\frac{1}{2}z^2}\,dz = \sqrt{\frac{\pi}{2}}$.
  - **Proof:** See handout 1.B.



The gamma function $\Gamma(\alpha)$

- Let $\tilde{x}$ be a random variable that has a gamma distribution with parameters α and β. Then,
- (a) $\mu'_r = \dfrac{\beta^r \Gamma(\alpha + r)}{\Gamma(\alpha)}$.
  - **Proof.** See handout 1.C.
- (b) $M_{\tilde{x}}(t) = (1 - \beta t)^{-\alpha}$ for $t < 1/\beta$.
  - **Proof.** See handout 1.D.
- **Corollary.**

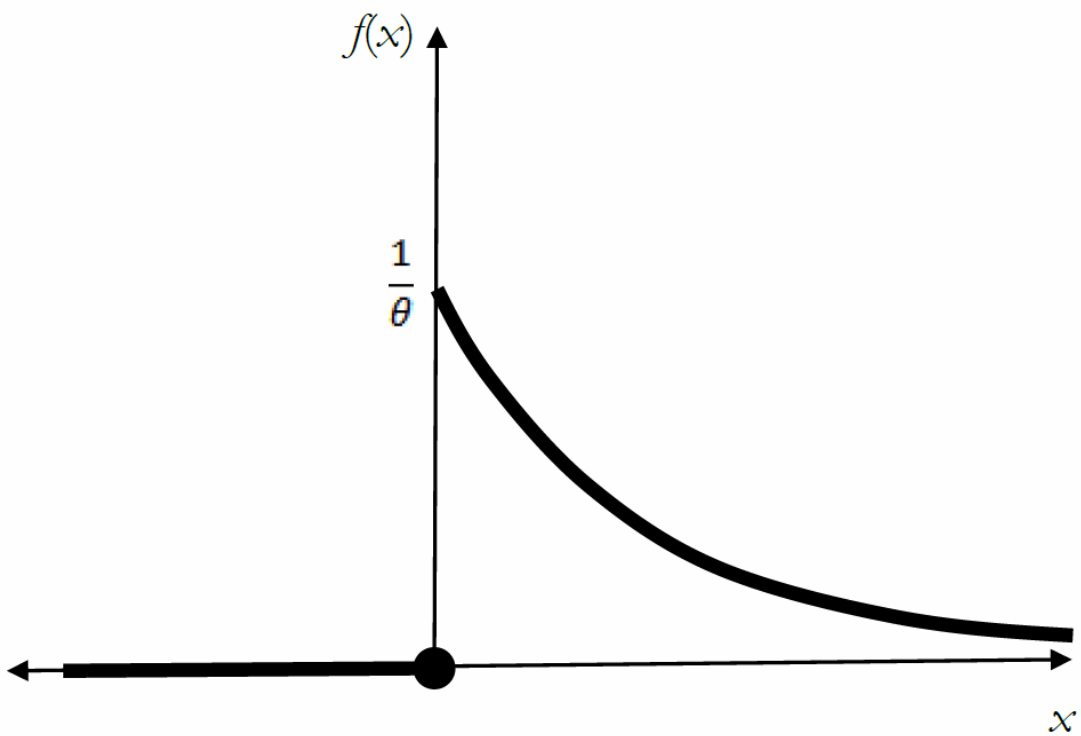
  (i) $\mu = \mu'_1 = \alpha\beta$,

  (ii) $\mu'_2 = \alpha(\alpha + 1)\beta^2$,

  (iii) $\sigma^2 = \mu'_2 - \mu^2 = \alpha\beta^2$.

- **The exponential distribution.**
- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has the exponential distribution if its density is the gamma density with $\alpha = 1$ and $\beta = \theta$,

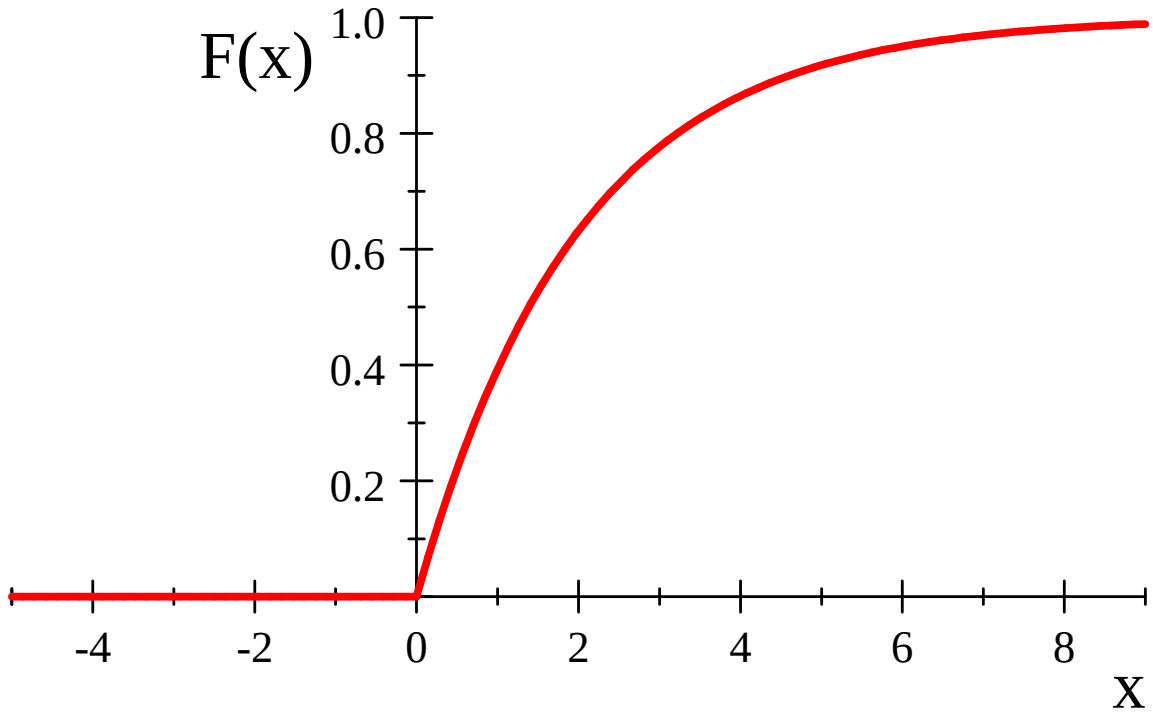
$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$



- This distribution is used to model waiting time.
- The distribution function is

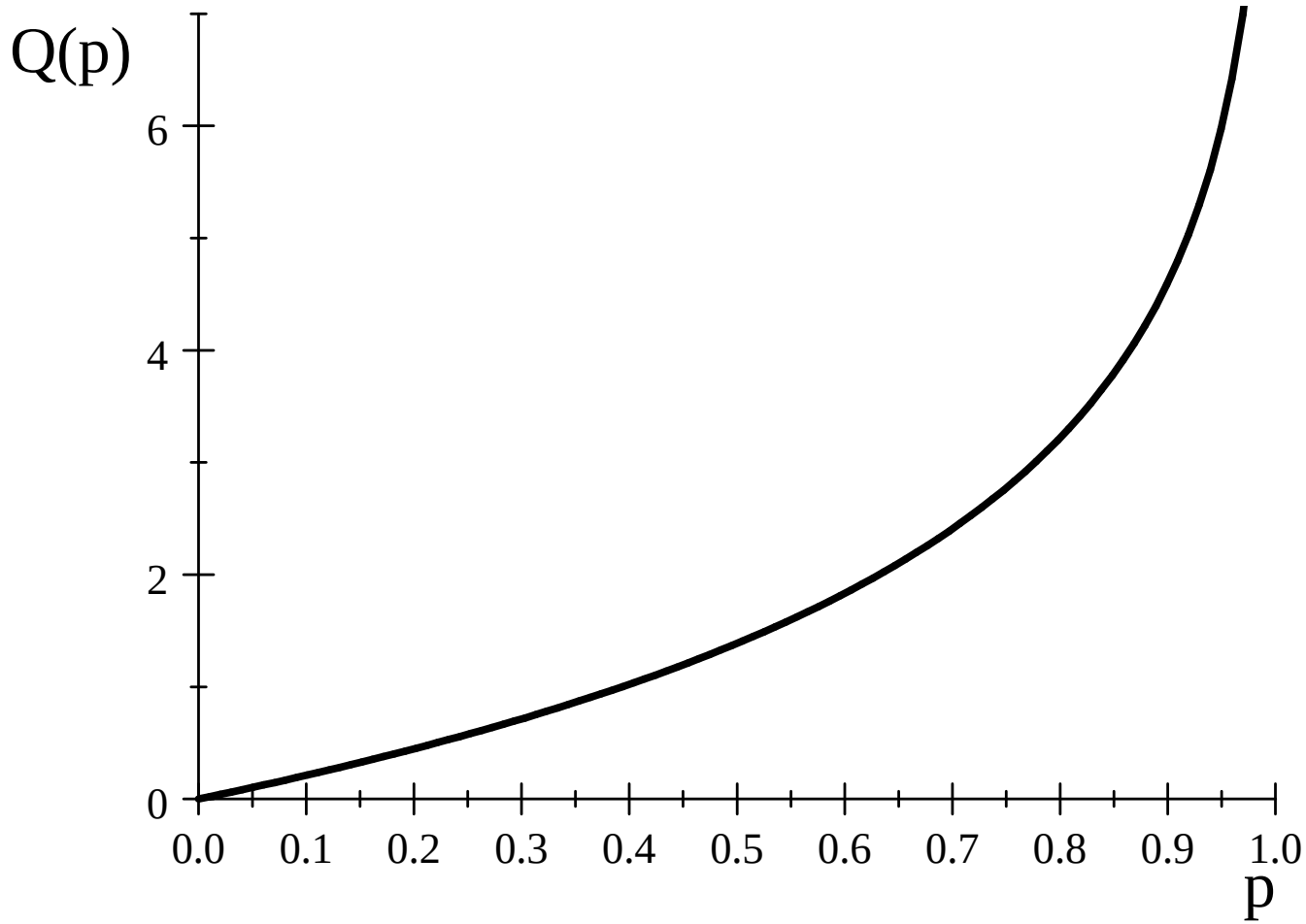
$$P\{\tilde{x} \leq x\} = F(x) = \begin{cases} 1 - e^{-x/\theta} & \text{for } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

which gives the probability of waiting less than x units of time.



- Quantile function:

$$Q(p) = -\theta \ln(1-p) \quad \text{for } p \in (0,1).$$



- Mean and variance:

$$\mu = \alpha\beta = \theta,$$
$$\sigma^2 = \alpha\beta^2 = \theta^2.$$

- Moment-generating function:

$$M_{\tilde{x}}(t) = (1 - \beta t)^{-\alpha} = \frac{1}{1 - \theta t} \quad \text{for } t < 1/\theta.$$

- **The chi-square $\left(\chi^2\right)$ distribution.**

Karl Pearson (1857 – 1936)

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ has the χ^2 (chi-square) distribution if its density is the gamma density with $\alpha = \nu/2$ and $\beta = 2$,

$$f(x;\nu) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu-2}{2}} e^{-x/2} & \text{for } x > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

- Mean and variance:

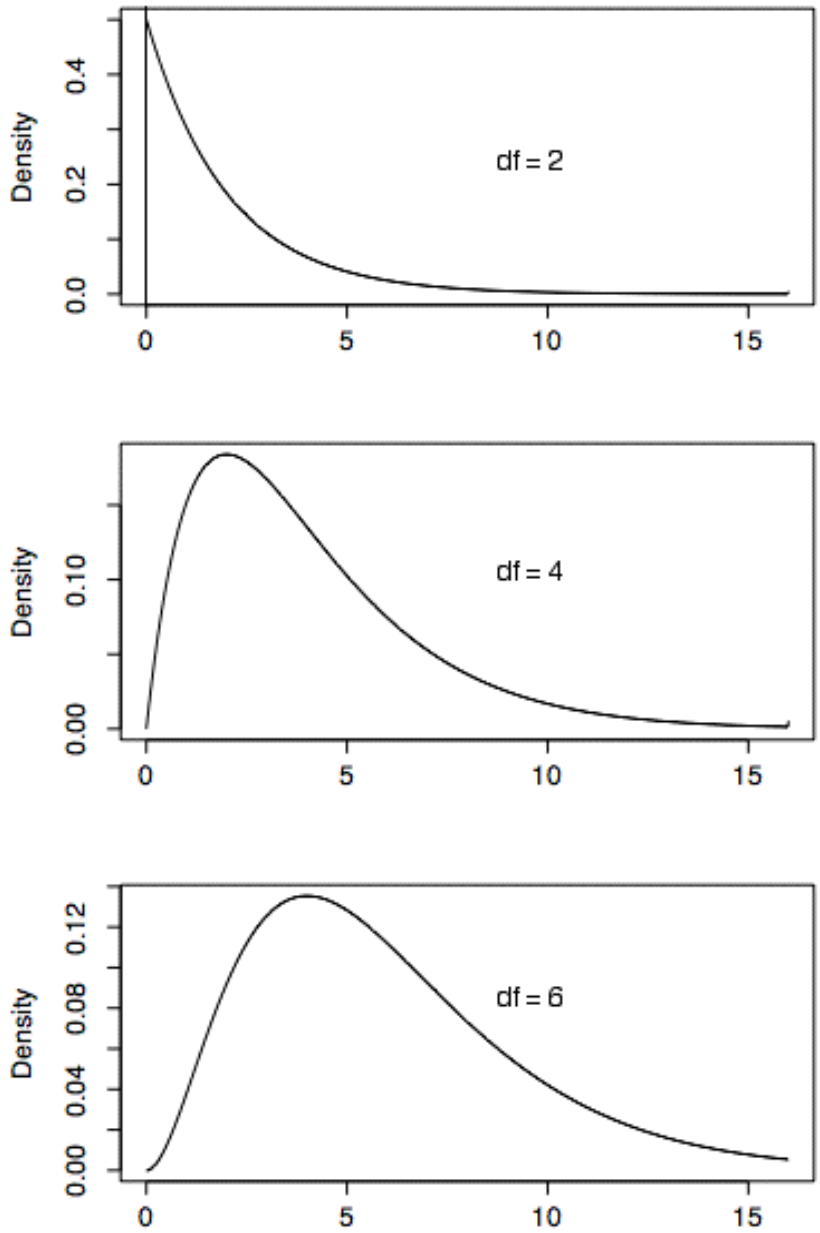
$$\mu = \alpha\beta = \nu,$$
$$\sigma^2 = \alpha\beta^2 = 2\nu.$$

- Moment-generating function:

$$M_{\tilde{x}}(t) = (1 - \beta t)^{-\alpha} = (1 - 2t)^{-\nu/2} \quad \text{for } t < 1/2.$$

- *Notation:*
  - $\tilde{x} \sim \mathrm{B}(n, \theta)$ $\longleftarrow$ $\tilde{x}$ has the binomial distribution
  - $\tilde{x} \sim \mathrm{U}(\alpha, \beta)$ $\longleftarrow$ $\tilde{x}$ has the uniform distribution on (α, β)
  - $\tilde{x} \sim \Gamma(\alpha, \beta)$ $\longleftarrow$ $\tilde{x}$ has the gamma distribution
  - $\tilde{x} \sim \chi^2_\nu$ $\longleftarrow$ $\tilde{x}$ has the chi-square distribution with ν degrees of freedom.

Densities of the chi-square distributions with 2, 4, and 6 degrees of freedom:



4.8. The beta distribution.

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ has the beta distribution if its density is

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } x \in (0,1) \\ \\ 0 & \text{otherwise,} \end{cases}$$

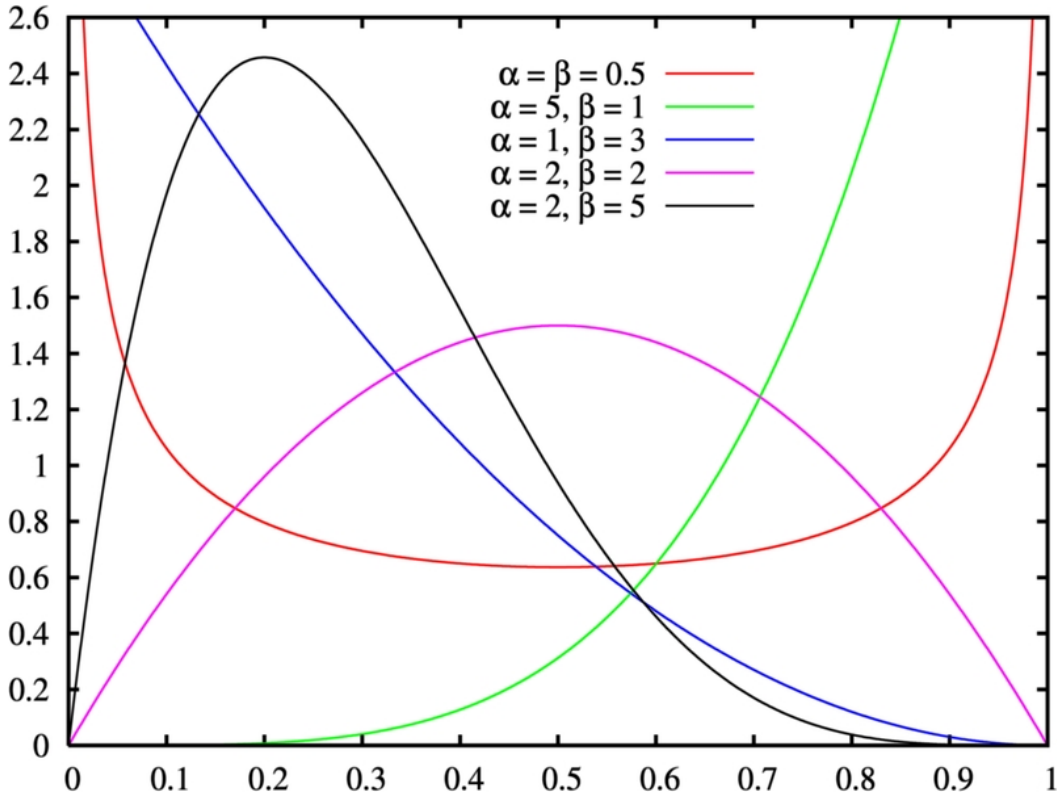
  with $\alpha > 0$, $\beta > 0$.

- If $\alpha = 1$ and $\beta = 1$, then the beta distribution becomes the uniform distribution on $(0,1)$.

- Obviously, the following holds:

$$\int_0^1 f(x; \alpha, \beta) dx = 1 \qquad \text{(see the handout)}$$

- Densities of the beta distribution:



- Mean and variance:

$$\mu = \frac{\alpha}{\alpha + \beta},$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}.$$

4.9. The normal distribution

Carl Friedrich Gauss (1777 – 1855)

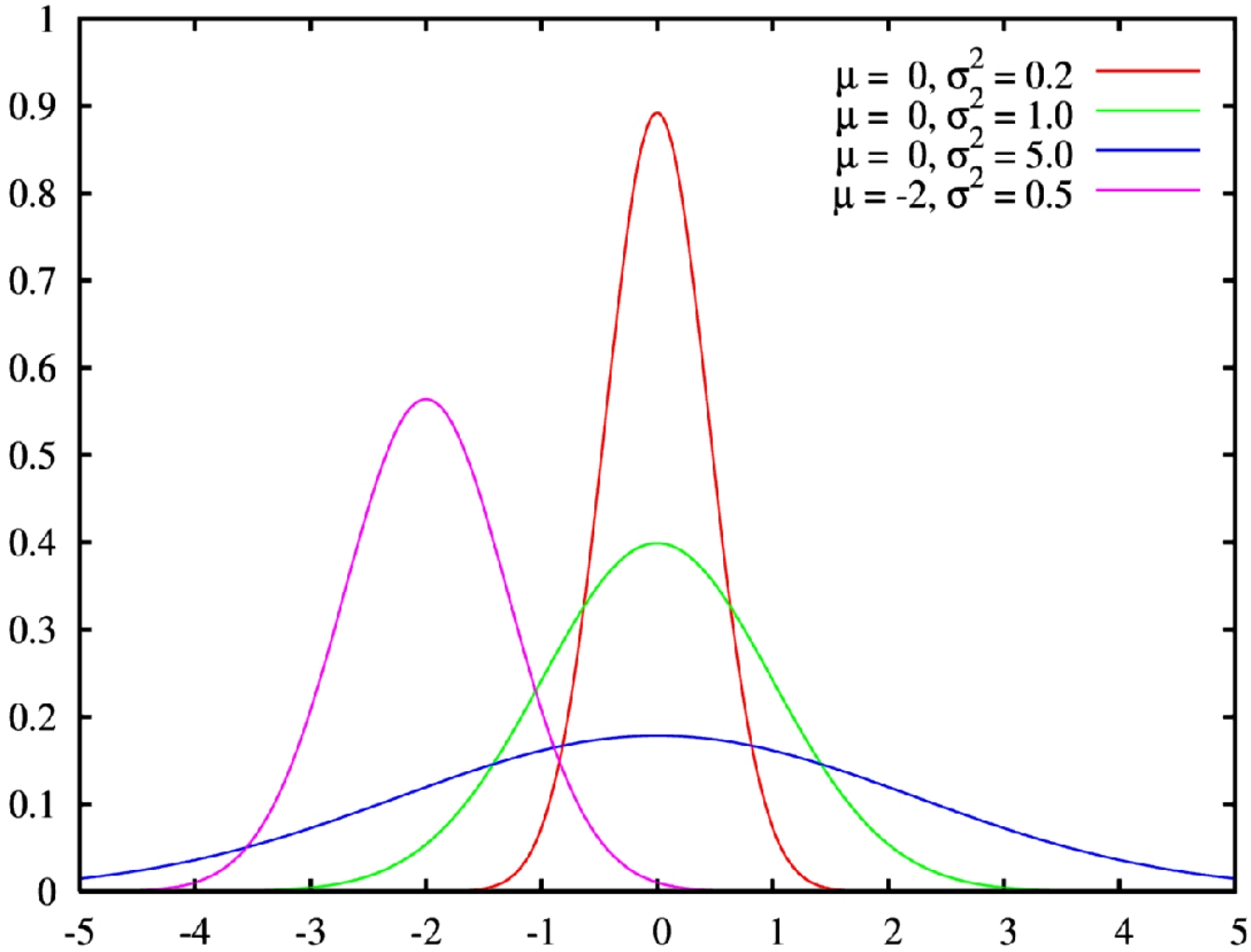
- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has the normal distribution (or is normal) if its density is

$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \text{with } \sigma > 0, \quad \text{for all } x \in \mathbb{R}.$$

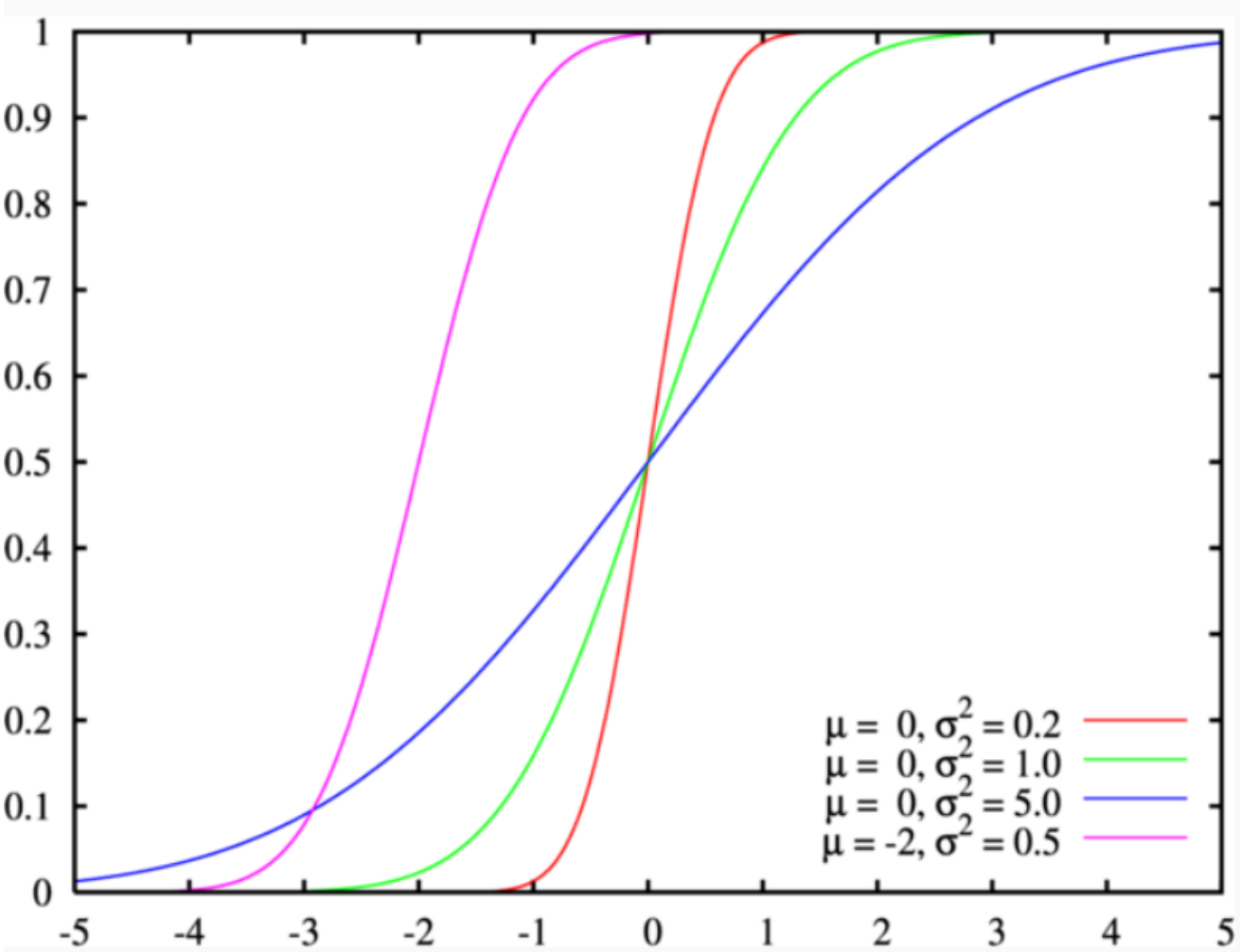
- *Notation:* $\tilde{x} \sim \mathrm{N}(\mu, \sigma^2)$.
- A random variable $\tilde{z}$ has the standard normal distribution (or is standard normal) if $\tilde{z} \sim \mathrm{N}(0, 1)$. Thus, the density of a standard normal random variable is

$$n(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad \text{for all } z \in \mathbb{R}.$$

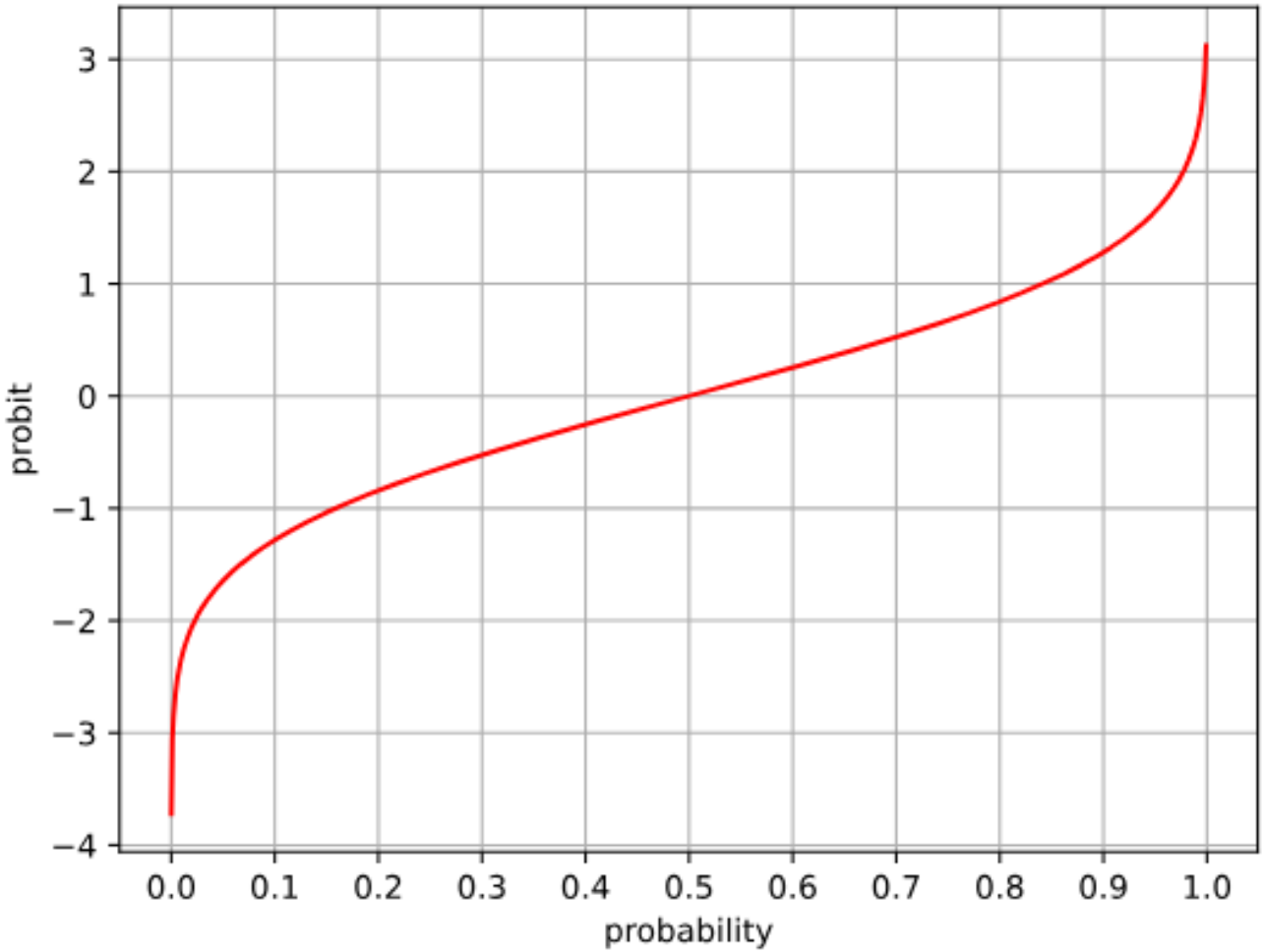
- Normal densities (i.e., densities of normal distributions) $n(x; \mu, \sigma)$ are symmetric around μ. The graph of the normal density is the Gauss bell.



Distribution functions $N(x; \mu, \sigma^2)$ do not have a closed-form expression:



- The quantile function of the standard normal distribution is the **probit** function used in probit regressions/models:



Properties of the normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$.

- **(1)** $\int_{-\infty}^{\infty} n(x; \mu, \sigma) = 1$.
  - **Proof:** See handout 3.A.
- **(2)** $M_{\tilde{x}}(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. In particular, if $\tilde{z} \sim \mathrm{N}(0,1)$, then $M_{\tilde{z}}(t) = e^{t^2/2}$.
  - **Proof:** See handout 3.B.
- **(3)**

$$M'_{\tilde{x}}(t) = (\mu + \sigma^2 t) M_{\tilde{x}}(t) \Rightarrow M'_{\tilde{x}}(0) = \mu = \mathrm{E}(\tilde{x}).$$

$$M''_{\tilde{x}}(t) = \sigma^2 M_{\tilde{x}}(t) + (\mu + \sigma^2 t)^2 M_{\tilde{x}}(t) \Rightarrow M''_{\tilde{x}}(0) = \sigma^2 + \mu^2 = \mathrm{E}(\tilde{x}^2).$$

$$\mathrm{Var}(\tilde{x}) = \mathrm{E}(\tilde{x}^2) - [\mathrm{E}(\tilde{x})]^2 = \sigma^2.$$

- *Note:* If $\mu_{\tilde{x}}$ and $\sigma_{\tilde{x}} > 0$ are the mean and the standard deviation, respectively, of the random variable $\tilde{x}$, then the "standardized" random variable $\tilde{z} = \dfrac{\tilde{x} - \mu_{\tilde{x}}}{\sigma_{\tilde{x}}}$ has $\mu_{\tilde{z}} = 0$ and $\sigma^2_{\tilde{z}} = \sigma_{\tilde{z}} = 1$.

- **(4)** If $\tilde{x} \sim \mathrm{N}(\mu, \sigma^2)$ and $\tilde{z} = \dfrac{\tilde{x} - \mu}{\sigma}$, then $\tilde{z} \sim \mathrm{N}(0, 1)$.

- **Proof:** Let $x = g(z) = \mu + \sigma z$ so that $z = g^{-1}(x) = \dfrac{x - \mu}{\sigma}$ and $g'(z) = \sigma > 0$.
  For all $A \in \mathcal{B}$, we have

$$P_{\tilde{z}}(A) = P\left\{\tilde{z} \in A\right\} = P\left\{g^{-1}(\tilde{x}) \in A\right\} = P\left\{\tilde{x} \in g\left(A\right)\right\}$$

$$= P_{\tilde{x}}(g\left(A\right)) = \int_{g(A)} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \int_A \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \sigma dz$$

$$= \int_A \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \int_A n(z; 0, 1) dz.$$

  Therefore, $\tilde{z} \sim \mathrm{N}(0, 1)$. *Q.E.D.*

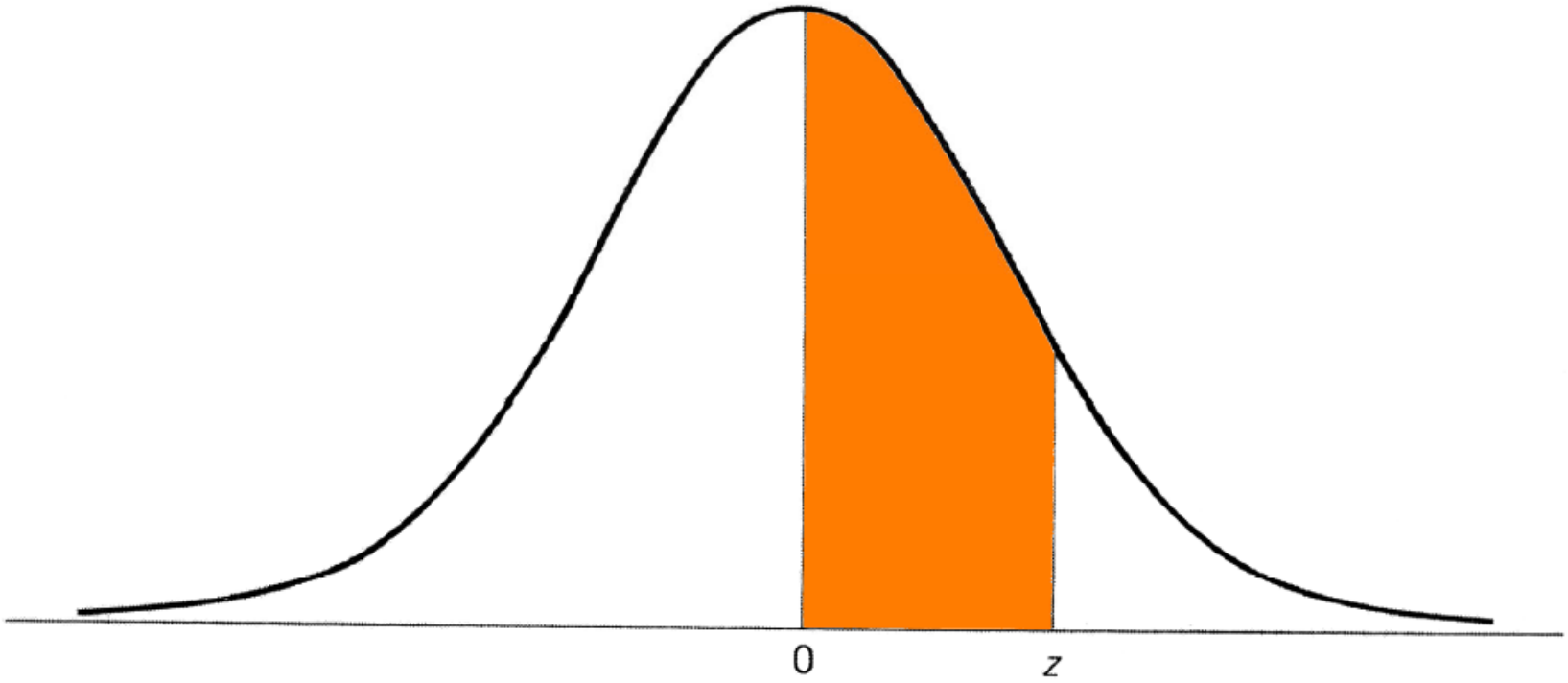
The distribution function of the standard normal distribution is tabulated:

The Standard Normal Distribution Function N(0,1)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Also, for $z = 4.0$, 5.0, and 6.0 the probabilities are 0.49997, 0.4999997, 0.499999999.

The previous table gives the area of the shaded region.



$$\int_0^z n(z;0,1)\,dz$$

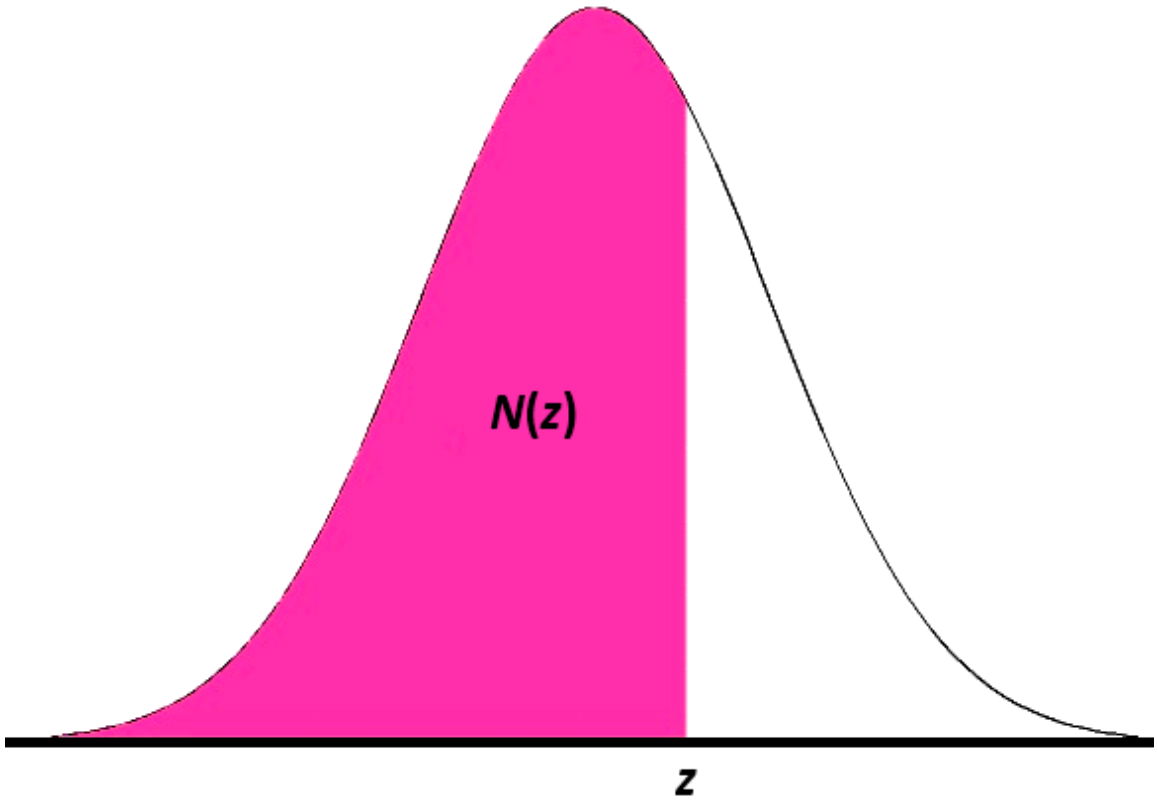
or

The Standard Normal Distribution Function N(0,1)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0000	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1000	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2000	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3000	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4000	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5000	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6000	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7000	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.8000	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9000	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0000	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1000	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2000	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3000	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4000	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5000	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6000	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7000	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8000	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9000	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0000	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1000	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2000	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3000	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4000	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5000	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6000	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7000	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8000	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9000	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0000	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1000	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2000	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3000	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4000	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5000	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6000	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

For $z = 4.0$, 5.0 and 6.0 the probabilities are 0.99997, 0.9999997 and 0.999999999.

The previous table gives the area of the shaded region.



$$N(z) = \int_{-\infty}^{z} n(z; 0, 1) dz$$

- If $a < 0$, then $\mathrm{N}(a) = 1 - \mathrm{N}(-a)$.
- If $a < b < 0$, then

$$\mathrm{N}(b) - \mathrm{N}(a) = 1 - \mathrm{N}(-b) - [1 - \mathrm{N}(-a)] = \mathrm{N}(-a) - \mathrm{N}(-b).$$

- if $a < 0$ and $b > 0$, then

$$\mathrm{N}(b) - \mathrm{N}(a) = \mathrm{N}(b) - [1 - \mathrm{N}(-a)] = \mathrm{N}(b) + \mathrm{N}(-a) - 1.$$

- Then, if $\tilde{x} \sim \mathrm{N}(\mu, \sigma^2)$,

$$P\{a \leq \tilde{x} \leq b\} = P\left\{\frac{a-\mu}{\sigma} \leq \frac{\tilde{x}-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right\} =$$

$$P\left\{\frac{a-\mu}{\sigma} \leq \tilde{z} \leq \frac{b-\mu}{\sigma}\right\} = \mathrm{N}\left(\frac{b-\mu}{\sigma}\right) - \mathrm{N}\left(\frac{a-\mu}{\sigma}\right),$$

where $\tilde{z} \sim \mathrm{N}(0,1)$ and $\mathrm{N}(\cdot)$ is the distribution function of the standard normal distribution.

- **Example.** If $\tilde{x} \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu = 4$ and $\sigma^2 = 49$, then

$$P\{-2 \leq \tilde{x} \leq 5\} = P\left\{\frac{-2-4}{7} \leq \frac{\tilde{x}-\mu}{\sigma} \leq \frac{5-4}{7}\right\}$$

$$= P\{-0.8571 \leq \tilde{z} \leq 0.1429\} = \mathrm{N}(0.1429) - \mathrm{N}(-0.8571)$$

$$= \mathrm{N}(0.1429) + \mathrm{N}(0.8571) - 1 = 0.5568 + 0.8043 - 1 = 0.3611.$$

4.10. The multivariate normal distribution and its properties

The random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^\intercal$: $(\Omega, \mathcal{F}, P) \to (\mathbb{R}^n, \mathcal{B})$ has the multivariate normal distribution ($\tilde{x} \sim \mathrm{MN}(\mu, \Sigma)$) if its density function is

$$f(x;\, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^\intercal \Sigma^{-1}(x-\mu)\right], \ \forall\, x \in \mathbb{R}^n,$$

where

$$x = (x_1, x_2, \ldots, x_n)^\intercal,$$

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \mu_n \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix},$$

Σ is a symmetric positive definite matrix, and $|\Sigma| > 0$ is the (absolute value of the) determinant of Σ.

- If $n = 1$, then $f(x;\, \mu, \Sigma) = n(x;\, \mu, \sigma)$, with $\sigma = \sqrt{\Sigma}$ as $\Sigma \in \mathbb{R}$ in this case.

- **Properties of the multivariate normal distribution.**

- **(1)** The marginal distribution of any sub-vector of the multivariate normally distributed random vector $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)^\intercal$ will be multivariate normally distributed and the corresponding sub-vector of μ and the corresponding sub-matrix of Σ will be the mean vector and the variance-covariance matrix of that random sub-vector. In particular,

$$\tilde{x}_i \sim \mathrm{N}(\mu_i, \sigma_i^2), \text{ where } \sigma_i^2 \equiv \sigma_{ii},$$

and

$$\mathrm{Cov}(\tilde{x}_i, \tilde{x}_j) = \sigma_{ij}.$$

- However, it is not true that the joint distribution of (multivariate) normal random vectors/variables is multivariate normal.

- **(2)** Moment-generating function of the multivariate normal distribution:

$$M_{\tilde{x}}\underbrace{(t_1, t_2, ..., t_n)}_{t\in\mathbb{R}^n} = e^{t^\intercal \mu + \frac{1}{2} t^\intercal \Sigma t}.$$

- **(3)** The random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^\intercal$ is multivariate normally distributed, $\tilde{x} \sim \mathrm{MN}(\mu, \Sigma)$, with Σ diagonal (i.e., $\sigma_{ij} = 0$, for all $i \neq j$) if and only if the random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$ are normally distributed and independent.
- Observe that, for Σ diagonal,

$$f(x_1, x_2, \ldots, x_n;\ \mu, \Sigma) = \prod_{i=1}^{n} n(x_i; \mu_i, \sigma_i), \quad \text{where } \sigma_i \equiv (\sigma_{ii})^{1/2}.$$

- Under multivariate normality, zero covariance implies independency! However, it is not true that the joint distribution of uncorrelated normally distributed random variables is multivariate normal.
- *Note:* When Σ is diagonal and σ_i^2 is the same for all i, we say that $\tilde{x}$ has a multivariate circular or spherical normal distribution.

- **(4)** Consider the following partitioned vectors/matrices of x, μ, Σ and $\tilde{x}$:

$$x = \begin{pmatrix} x_1 \\ \cdots \\ x_2 \end{pmatrix}, \; x_1 \in \mathbb{R}^{n_1}, \;\; x_2 \in \mathbb{R}^{n_2}, \;\; n_1 + n_2 = n,$$

$$\mu = \begin{pmatrix} \mu_1 \\ \cdots \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \vdots & \Sigma_{12} \\ \cdots & \cdots & \cdots \\ \Sigma_{21} & \vdots & \Sigma_{22} \end{pmatrix}, \quad \Sigma_{12}^{\intercal} = \Sigma_{21},$$

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \cdots \\ \tilde{x}_2 \end{pmatrix}, \quad \text{where } \tilde{x}_1 \sim \mathrm{MN}(\mu_1, \Sigma_{11}) \text{ and } \tilde{x}_2 \sim \mathrm{MN}(\mu_2, \Sigma_{22}),$$

and $\Sigma_{ij} = \mathrm{Cov}(\tilde{x}_i, \tilde{x}_j) = \mathrm{E}\left[(\tilde{x}_i - \mu_i)(\tilde{x}_j - \mu_j)^{\intercal}\right]$ is a $n_i \times n_j$ matrix, for $i = 1, 2, \; j = 1, 2$.

- Then, the conditional density of the random vector $\tilde{x}_1$ given $\tilde{x}_2 = x_2$,

$$f_{\tilde{x}_1|\tilde{x}_2}\left(x_1|x_2; \mu_{\tilde{x}_1|\tilde{x}_2=x_2}, \Sigma_{\tilde{x}_1|\tilde{x}_2=x_2}\right) = \frac{f(x;\, \mu, \Sigma)}{f_{\tilde{x}_2}(x_2;\, \mu_2, \Sigma_{22})}, \quad \forall x_1 \in \mathbb{R}^{n_1},$$

  is the density of a multivariate normal random vector with mean $\mu_{\tilde{x}_1|\tilde{x}_2=x_2}$ and covariance matrix $\Sigma_{\tilde{x}_1|\tilde{x}_2=x_2}$.

- Moreover, the conditional mean of the random vector $\tilde{x}_1$ given $\tilde{x}_2 = x_2$ is the following:

$$\mu_{\tilde{x}_1|\tilde{x}_2=x_2} \equiv \mathrm{E}(\tilde{x}_1|\tilde{x}_2 = x_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \in \mathbb{R}^{n_1}.$$

- *Note:* If $n_1 = n_2 = 1$, then

$$\mathrm{E}(\tilde{x}_1|\tilde{x}_2 = x_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2).$$

- Finally, let

$$\Sigma_{\tilde{x}_1|\tilde{x}_2=x_2} \equiv \text{Var}(\tilde{x}_1|\tilde{x}_2 = x_2) =$$

$$\text{E}([\tilde{x}_1 - \text{E}(\tilde{x}_1|\tilde{x}_2 = x_2)][\tilde{x}_1 - \text{E}(\tilde{x}_1|\tilde{x}_2 = x_2)]^\intercal|\tilde{x}_2 = x_2)$$

be the $n_1 \times n_1$ conditional covariance matrix of the random vector $\tilde{x}_1$ given $\tilde{x}_2 = x_2$. Then,

$$\Sigma_{\tilde{x}_1|\tilde{x}_2=x_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

which does not depend on the value x_2 taken by the random vector $\tilde{x}_2$.

- *Note:* If $n_1 = n_2 = 1$, then

$$\text{Var}(\tilde{x}_1|\tilde{x}_2 = x_2) = \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2} = \sigma_1^2(1 - \rho^2),$$

so that the random variable $\text{Var}(\tilde{x}_1|\tilde{x}_2) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a constant.

- **(5)** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^\intercal \sim \mathrm{MN}(\mu, \Sigma)$ and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^\intercal$ is a non-zero vector of scalars, then

$$\alpha^\intercal \tilde{x} = \sum_{i=1}^{n} \alpha_i \tilde{x}_i \sim \mathrm{N}(\alpha^\intercal \mu, \alpha^\intercal \Sigma \alpha).$$

- Note that in the previous result the vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^\intercal$ has to be multivariate normal. It is not enough that each component of that vector be normal.

- **General Proposition.** Let $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^\intercal \sim \mathrm{MN}(\mu, \Sigma)$ and $\tilde{y} = c + A\tilde{x}$ be an affine transformation of $\tilde{x}$, where $c \in \mathbb{R}^m$ is a column vector and A is $m \times n$ matrix with $\mathrm{rank}(A) = m \leq n$. Then, $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m)^\intercal \sim \mathrm{MN}(c + A\mu, A\Sigma A^\intercal)$.

- The previous Proposition implies Property 5 (when $m = 1$, $A = \alpha^{\intercal}$, and $c = \underline{0}$) and Property 1. For the later, consider the following example: to extract the sub-vector $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_4)^{\intercal}$ from the random vector $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^{\intercal}$ use the vector $c = \underline{0}$ and the $3 \times n$ matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix},$$

which extracts the desired sub-vector directly.

4.11. Multivariate normality and linear models

- **From linearity to multivariate normality:**
- Consider the random variable $\tilde{y}$,

$$\tilde{y} = \alpha + \underbrace{\beta^{\intercal}\tilde{x}}_{=\tilde{x}^{\intercal}\beta} + \tilde{\varepsilon} = \alpha + \sum_{i=1}^{n} \beta_i \tilde{x}_i + \tilde{\varepsilon},$$

  where $\alpha \in \mathbb{R}$, $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^{\intercal} \in \mathbb{R}^n$ is a column vector, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^{\intercal} \sim \mathrm{MN}(\mu_x, \Sigma_x)$, $\tilde{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$, and $\tilde{\varepsilon}$ and $\tilde{x}$ are independent.
- We know from Property 3 above that the vector $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n, \tilde{\varepsilon})^{\intercal}$ is multivariate normal. Thus, from the General Proposition above, the random variable $\tilde{y}$ is normal since it is an affine transformation of the random variables appearing in the multivariate normal random vector $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n, \tilde{\varepsilon})^{\intercal}$. Finally, also from the General Proposition above, the random vector $(\tilde{y}, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^{\intercal}$ is multivariate normal.

- The mean μ_y and the variance σ_y^2 of $\tilde{y}$ are computed as follows:
- Mean:

$$\mu_y = \mathrm{E}(\tilde{y}) = \mathrm{E}\left(\alpha + \beta^\intercal \tilde{x} + \tilde{\varepsilon}\right) = \alpha + \beta^\intercal \mu_x + \mathrm{E}\left(\tilde{\varepsilon}\right).$$

$$= \alpha + \beta^\intercal \mu_x + 0 = \alpha + \underbrace{\beta^\intercal \mu_x}_{=\mu_x^\intercal \beta}. \tag{1}$$

- Variance:

$$\sigma_y^2 = \mathrm{Var}(\tilde{y}) = \mathrm{Var}(\alpha + \beta^\intercal \tilde{x} + \tilde{\varepsilon}) = \beta^\intercal \Sigma_x \beta + \sigma_\varepsilon^2. \tag{2}$$

- Therefore, $\tilde{y} \sim \mathrm{N}(\underbrace{\alpha + \beta^\intercal \mu_x}_{\mu_y}, \underbrace{\beta^\intercal \Sigma_x \beta + \sigma_\varepsilon^2}_{\sigma_y^2})$.

- Let us compute the conditional expectation of $\tilde{y}$ given $\tilde{x} = x$,

$$\begin{aligned}
\mathrm{E}(\tilde{y}|\tilde{x} = x) &= \mathrm{E}\left(\alpha + \beta^{\intercal}\tilde{x} + \tilde{\varepsilon}|\, \tilde{x} = x\right) \\
&= \alpha + \mathrm{E}\left(\beta^{\intercal}\tilde{x}|\, \tilde{x} = x\right) + \mathrm{E}\left(\tilde{\varepsilon}|\tilde{x} = x\right) \\
&= \alpha + \mathrm{E}\left(\beta^{\intercal}x\right) + \mathrm{E}\left(\tilde{\varepsilon}\right) = \alpha + \underbrace{\beta^{\intercal}x}_{=x^{\intercal}\beta},
\end{aligned}$$

  where the third equality holds because $\tilde{\varepsilon}$ and $\tilde{x}$ are independent.

- Therefore, the conditional expectation $\mathrm{E}(\tilde{y}|\tilde{x} = x)$ is an affine transformation of x. This agrees with our previous Property 4, according to which,

$$\begin{aligned}
\mathrm{E}(\tilde{y}|\tilde{x} = x) &= \mu_y + \Sigma_{y,x}\Sigma_x^{-1}(x - \mu_x) \\
&= \left(\mu_y - \Sigma_{y,x}\Sigma_{\tilde{x}}^{-1}\mu_x\right) + \Sigma_{y,x}\Sigma_x^{-1}x.
\end{aligned}$$

- Thus, the following equalities should be true:

$$\alpha = \mu_y - \Sigma_{y,x}\Sigma_x^{-1}\mu_x \tag{3}$$

and

$$\beta^\intercal = \Sigma_{y,x}\Sigma_x^{-1} \quad \text{or} \quad \beta = \Sigma_x^{-1}\Sigma_{y,x}^\intercal = \Sigma_x^{-1}\Sigma_{x,y}. \tag{4}$$

- We can check that (4) holds since the $1 \times n$ matrix $\Sigma_{y,x}$ satisfies

$$\Sigma_{y,x} = \text{Cov}\left(\tilde{y}, \tilde{x}\right) = \text{Cov}\left(\alpha + \beta^\intercal \tilde{x} + \tilde{\varepsilon}, \tilde{x}\right) = \beta^\intercal \Sigma_x$$

so that

$$\Sigma_{y,x}\Sigma_{\tilde{x}}^{-1} = \beta^\intercal \Sigma_x \Sigma_x^{-1} = \beta^\intercal.$$

- To check that (3) holds note that, from (1) and (4), we get

$$\alpha = \mu_y - \beta^\intercal \mu_x = \mu_y - \Sigma_{y,x}\Sigma_x^{-1}\mu_x.$$

- We can compute now the conditional variance of $\tilde{y}$ given $\tilde{x} = x$,

$$\mathrm{Var}(\tilde{y}|\tilde{x} = x) = \mathrm{Var}\left(\ \alpha + \beta^\intercal \tilde{x} + \tilde{\varepsilon}|\ \tilde{x} = x\right)$$

$$= \mathrm{Var}\left(\alpha + \beta^\intercal x + \tilde{\varepsilon}\right) = \mathrm{Var}\left(\tilde{\varepsilon}\right) = \sigma_\varepsilon^2,$$

  which agrees with our previous Property 4, according to which,

$$\mathrm{Var}(\tilde{y}|\tilde{x} = x) = \sigma_y^2 - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}$$

  and, thus, $\mathrm{Var}(\tilde{y}|\tilde{x}= x)$ does not depend on the value x taken by the random vector $\tilde{x}$, i.e., the random variable $\mathrm{Var}(\tilde{y}|\tilde{x})$ is a constant.

- Therefore, we should have

$$\sigma_{\varepsilon}^2 = \sigma^2{}_y - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}.$$

- We can check that the previous equality holds indeed since, from (2) and (4), we get

$$\sigma_{\varepsilon}^2 = \sigma_y^2 - \beta^{\intercal}\Sigma_x\beta = \sigma_y^2 - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_x\Sigma_x^{-1}\Sigma_{x,y}$$

$$= \sigma_y^2 - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}.$$

- **From multivariate normality to linearity:**
- Assume that the random vector $(\tilde{y}, \underbrace{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n}_{\tilde{x}^\intercal})^\intercal$ is $\mathrm{MN}(\mu, \Sigma)$, with

$$\mu = (\mu_y, \underbrace{\mu_1, \mu_2, ..., \mu_n}_{\mu_x^\intercal})^\intercal$$

and

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \Sigma_{y,x} \\ \Sigma_{x,y} & \Sigma_x \end{pmatrix}.$$

- Then, we know from Property 4 above that

$$\mathrm{E}(\tilde{y}|\tilde{x}=x)=\mu_y+\Sigma_{y,x}\Sigma_x^{-1}(x-\mu_x)=\left(\mu_y-\Sigma_{y,x}\Sigma_{\tilde{x}}^{-1}\mu_x\right)+\Sigma_{y,x}\Sigma_x^{-1}x.$$

or equivalently,

$$\mathrm{E}(\tilde{y}|\tilde{x}) = \left(\mu_y - \Sigma_{y,x}\Sigma_{\tilde{x}}^{-1}\mu_x\right) + \Sigma_{y,x}\Sigma_x^{-1}\tilde{x},$$

so that $\mathrm{E}(\tilde{y}|\tilde{x})$ is an affine transformation of $\tilde{x}$. Thus, since $\tilde{x} \sim \mathrm{MN}(\mu_x, \Sigma_x)$, the random variable $\mathrm{E}(\tilde{y}|\tilde{x})$ is normal as dictated by the General Proposition above.

- Define the random variable

$$\tilde{\varepsilon} = \tilde{y} - \mathrm{E}(\tilde{y}|\tilde{x}) = \tilde{y} - \left(\mu_y - \Sigma_{y,x}\Sigma_{\tilde{x}}^{-1}\mu_x\right) - \Sigma_{y,x}\Sigma_x^{-1}\tilde{x}.$$

- According to the General Proposition above, the random variable $\tilde{\varepsilon}$ is normal since it is an affine transformation of the random variables appearing in the multivariate normal random vector $(\tilde{y}, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)^\intercal$. Moreover, also from the General Proposition above, the random vector $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n, \tilde{\varepsilon})^\intercal$ is multivariate normal.

- Thanks to the theorem of total expectation, we can find the expectation for $\tilde{\varepsilon}$:

$$\mathrm{E}(\tilde{\varepsilon}) = \mathrm{E}\left[\tilde{y} - \mathrm{E}(\tilde{y}|\tilde{x})\right] = \mathrm{E}\left(\tilde{y}\right) - \mathrm{E}\left[\mathrm{E}(\tilde{y}|\tilde{x})\right] = \mathrm{E}\left(\tilde{y}\right) - \mathrm{E}\left(\tilde{y}\right) = 0.$$

- Moreover, the variance of $\tilde{\varepsilon}$ is

$$\sigma_\varepsilon^2 = \text{Var}(\tilde{\varepsilon}) = \text{Var}\left[\tilde{y} - \text{E}(\tilde{y}|\tilde{x})\right]$$

$$= \text{Var}\left[\tilde{y} - \left(\mu_y - \Sigma_{y,x}\Sigma_{\tilde{x}}^{-1}\mu_x\right) - \Sigma_{y,x}\Sigma_x^{-1}\tilde{x}\right]$$

$$= \text{Var}\left(\tilde{y}\right) + \text{Var}\left(\Sigma_{y,x}\Sigma_x^{-1}\tilde{x}\right) - 2\text{Cov}\left(\tilde{y}, \Sigma_{y,x}\Sigma_x^{-1}\tilde{x}\right)$$

$$= \sigma_y^2 + \Sigma_{y,x}\Sigma_x^{-1}\Sigma_x\Sigma_x^{-1}\Sigma_{x,y} - 2\Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}$$

$$= \sigma_y^2 + \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y} - 2\Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y} = \sigma_y^2 - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}.$$

- Therefore $\tilde{\varepsilon} \sim \text{N}(0, \underbrace{\sigma_y^2 - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}}_{\sigma_\varepsilon^2})$.

- Moreover, from generalizing Exercise 30, part (c), of List 3, we know that $\tilde{\varepsilon} = \tilde{y} - \text{E}(\tilde{y}|\tilde{x})$ has zero covariance with $\tilde{x}$, $\text{Cov}\left(\tilde{\varepsilon}, \tilde{x}\right) = \underline{0}^\intercal \in \mathbb{R}^n$ and, thus, from Property 3 above, $\tilde{\varepsilon}$ and $\tilde{x}$ are independent.

- Note that, from the definition of $\tilde{\varepsilon}$, we have

$$\tilde{y} = \mathrm{E}(\tilde{y}|\tilde{x}) + \tilde{\varepsilon} = \left(\mu_y - \Sigma_{y,x}\Sigma_{\tilde{x}}^{-1}\mu_x\right) + \Sigma_{y,x}\Sigma_x^{-1}\tilde{x} + \tilde{\varepsilon}.$$

- Then, we can define the scalar $\alpha = \mu_y - \Sigma_{y,x}\Sigma_{\tilde{x}}^{-1}\mu_x$ and the column vector $\beta = \Sigma_x^{-1}\Sigma_{x,y} \in \mathbb{R}^n$ so that the previous equation becomes

$$\tilde{y} = \alpha + \underbrace{\beta^{\intercal}\tilde{x}}_{=\tilde{x}^{\intercal}\beta} + \tilde{\varepsilon}.$$

where the random vector $\tilde{x}$ and the random variable $\tilde{\varepsilon}$ are independent, and $\tilde{\varepsilon} \sim \mathrm{N}(0, \sigma_y^2 - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y})$.

- We can summarize all this section with the following proposition, which is related with regression analysis:
- **Proposition.** The random vector $(\tilde{y}, \underbrace{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n}_{\tilde{x}^\intercal})^\intercal$ has a multivariate normal distribution $\mathrm{MN}(\mu, \Sigma)$ with

$$\mu = (\mu_y, \underbrace{\mu_1, \mu_2, ..., \mu_n}_{\mu_x^\intercal})^\intercal \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \Sigma_{y,x} \\ \Sigma_{x,y} & \Sigma_x \end{pmatrix}$$

if and only if there exist a scalar α, a vector of scalars $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^\intercal$, and a random variable $\tilde{\varepsilon} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ such that

$$\tilde{y} = \alpha + \beta^\intercal \tilde{x} + \tilde{\varepsilon} \quad (\text{or} \quad \tilde{y} = \alpha + \tilde{x}^\intercal \beta + \tilde{\varepsilon}),$$

where the random vector $\tilde{x} \sim \mathrm{MN}(\mu_x, \Sigma_x)$ and the random variable $\tilde{\varepsilon}$ are independent.
Moreover, the following equalities hold:

$$\alpha = \mu_y - \Sigma_{y,x}\Sigma_x^{-1}\mu_x\,, \quad \beta = \Sigma_x^{-1}\Sigma_{x,y}\,, \quad \text{and} \quad \sigma_\varepsilon^2 = \sigma_y^2 - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}\,;$$

$$\mu_y = \alpha + \beta^\intercal \mu_x\,, \quad \sigma_y^2 = \beta^\intercal \Sigma_x \beta + \sigma_\varepsilon^2\,, \quad \text{and} \quad \Sigma_{y,x} = \beta^\intercal \Sigma_x\,.$$

- It is straightforward to generalize the previous proposition when $\tilde{y} = (\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_m)^\intercal$ is a random vector as follows:
- **Proposition.** The random vector $(\underbrace{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m}_{\tilde{y}^\intercal}, \underbrace{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n}_{\tilde{x}^\intercal})^\intercal$ has a multivariate normal distribution $\mathrm{MN}(\mu, \Sigma)$ with

$$\mu = (\underbrace{\mu_{y_1}, \mu_{y_2}, ..., \mu_{y_m}}_{\mu_y^\intercal}, \underbrace{\mu_{x_1}, \mu_{x_2}, ..., \mu_{x_n}}_{\mu_x^\intercal})^\intercal \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_y & \Sigma_{y,x} \\ \Sigma_{x,y} & \Sigma_x \end{pmatrix}$$

if and only if there exist a vector $\alpha \in \mathbb{R}^m$, a $n \times m$ matrix B, and a random vector $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, ..., \tilde{\varepsilon}_m)^\intercal \sim \mathrm{MN}(\underline{0}, \Sigma_\varepsilon)$ such that

$$\tilde{y} = \alpha + \mathrm{B}^\intercal \tilde{x} + \tilde{\varepsilon} \quad (\text{or} \quad \tilde{y}^\intercal = \alpha^\intercal + \tilde{x}^\intercal \mathrm{B} + \tilde{\varepsilon}^\intercal),$$

where the random vectors $\tilde{x} \sim \mathrm{MN}(\mu_x, \Sigma_x)$ and $\tilde{\varepsilon}$ are independent. Moreover, the following equalities hold:

$$\alpha = \mu_y - \Sigma_{y,x}\Sigma_x^{-1}\mu_x\,, \quad \mathrm{B} = \Sigma_x^{-1}\Sigma_{x,y}\,, \quad \text{and} \quad \Sigma_\varepsilon = \Sigma_y - \Sigma_{y,x}\Sigma_x^{-1}\Sigma_{x,y}\,;$$

$$\mu_y = \alpha + \mathrm{B}^\intercal \mu_x\,, \quad \Sigma_y = \mathrm{B}^\intercal \Sigma_x \mathrm{B} + \sigma_\varepsilon^2\,, \quad \text{and} \quad \Sigma_{y,x} = \mathrm{B}^\intercal \Sigma_x\,.$$

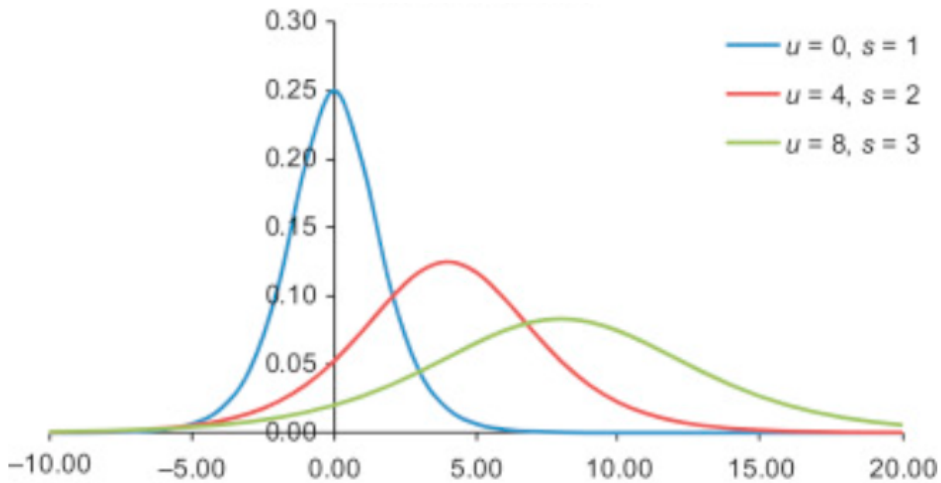
4.11. The logistic distribution

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has the logistic distribution (or is logistic) if its density is

$$l(x; m, s) = \frac{e^{-\left(\frac{x-m}{s}\right)}}{s\left[1 + e^{-\left(\frac{x-m}{s}\right)}\right]^2} = \frac{e^{\frac{x-m}{s}}}{s\left[1 + e^{\frac{x-m}{s}}\right]^2}, \quad \text{for all } x \in \mathbb{R},$$

  where m is a location parameter (which is equal to the mean μ) and $s > 0$ is the scale parameter.

- Density:



- The logistic density is symmetric around m (Exercise).
- $$\mu = \mathrm{E}(\tilde{x}) = m \quad \text{and} \quad \mathrm{Var}(\tilde{x}) = \frac{s^2\pi^2}{3}.$$
- Distribution function:

$$L(x; m, s) = \frac{1}{1 + e^{-\left(\frac{x-m}{s}\right)}} = \frac{e^{\frac{x-m}{s}}}{1 + e^{\frac{x-m}{s}}},$$

as $L'(x; m, s) = l(x; m, s)$, $\lim_{x \to -\infty} L(x; m, s) = 0$, and $\lim_{x \to \infty} L(x; m, s) = 1$.

- *Notation:* $\tilde{x} \sim L(m, s)$.
- Quantile function:

$$Q(p; m, s) = m + s \ln\left(\frac{p}{1-p}\right).$$

- Standard logistic distribution when $m = 0$ and $s = 1$,

$$l(x; 0, 1) = \frac{e^{-x}}{\left(1 + e^{-x}\right)^2} = \frac{e^{x}}{\left(1 + e^{x}\right)^2}, \quad \text{for all } x \in \mathbb{R}.$$

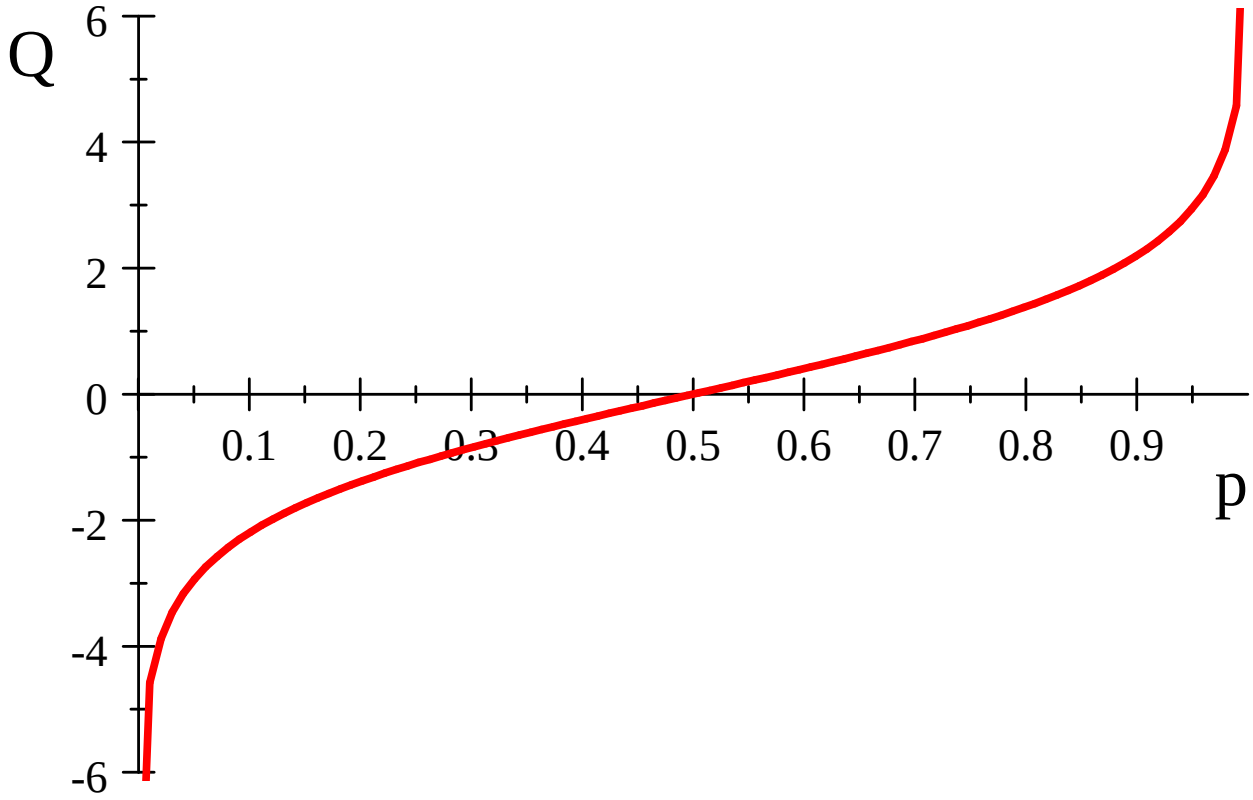
- Distribution function:

$$L(x; 0, 1) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{1 + e^{x}}.$$

- Quantile function:

$$Q(p; 0, 1) = \ln\left(\frac{p}{1 - p}\right).$$

- The quantile function of the standard logistic distribution is the **logit** function used in logit regressions/models:



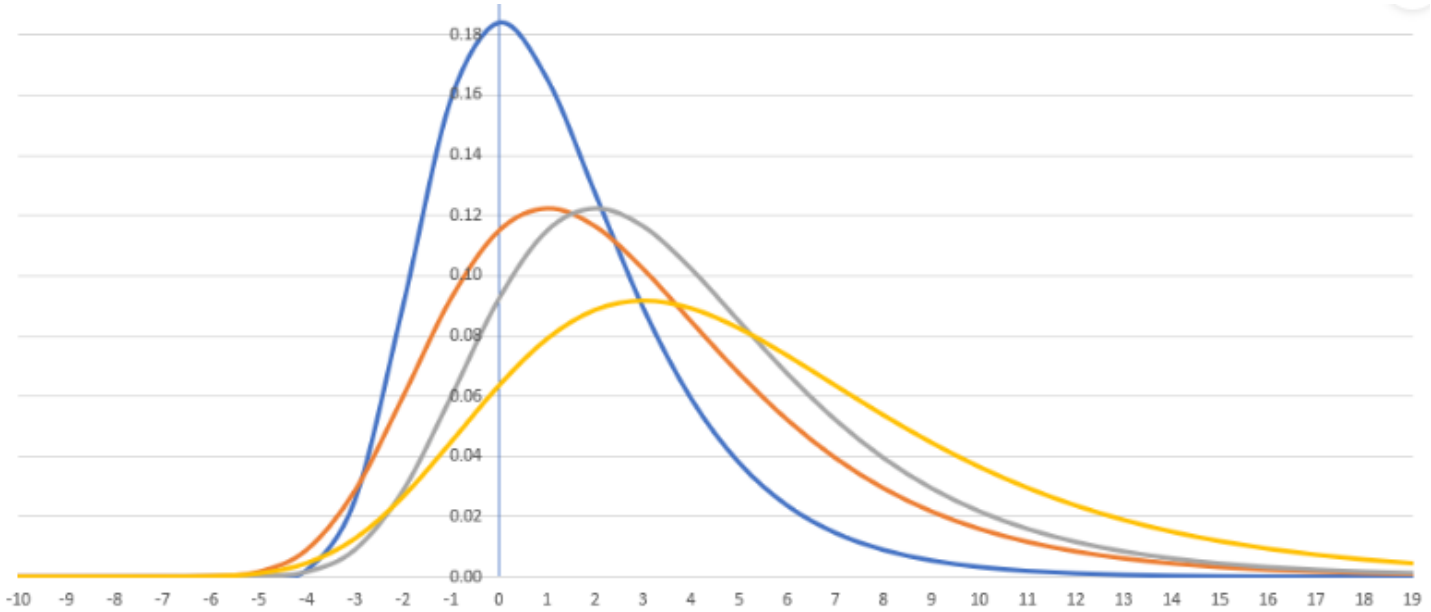
4.11. Generalized extreme value distributions

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has the generalized **extreme value type I** (or Gumbel) distribution if its density is

$$g(x; m, s) = \frac{1}{s} e^{-\left(\frac{x-m}{s}\right)} e^{-e^{-\left(\frac{x-m}{s}\right)}}, \quad \text{for all } x \in \mathbb{R},$$

  where m is the location parameter and $s > 0$ is the scale parameter.

- Density:



- $$\mathrm{E}(\tilde{x}) = m + \gamma s \quad \text{and} \quad \mathrm{Var}(\tilde{x}) = \frac{s^2 \pi^2}{6},$$

  where $\gamma = \lim_{n \to \infty} \left(-\ln n + \sum_{k=1}^{n} \frac{1}{k} \right) \approx 0.57722$ is the Euler(-Mascheroni) constant.

- Standard extreme value type I distribution when $m = 0$ and $s = 1$,

$$g(x; 0, 1) = e^{-x} e^{-e^{-x}}, \quad \text{for all } x \in \mathbb{R}.$$

- Distribution function:

$$G(x; m, s) = e^{-e^{-\left(\frac{x-m}{s}\right)}}, \quad \text{for all } x \in \mathbb{R},$$

$$G(x; 0, 1) = e^{-e^{-x}}, \quad \text{for all } x \in \mathbb{R},$$

  as $G'(x; m, s) = g(x; m, s)$, $\lim_{x \to -\infty} G(x; m, s) = 0$, and $\lim_{x \to \infty} G(x; m, s) = 1$.

- *Notation:* $\tilde{x} \sim G(m, s)$.

- Quantile function:

$$Q(p; m, s) = m - s \ln(-\ln p),$$

$$Q(p; 0, 1) = -\ln(-\ln p).$$

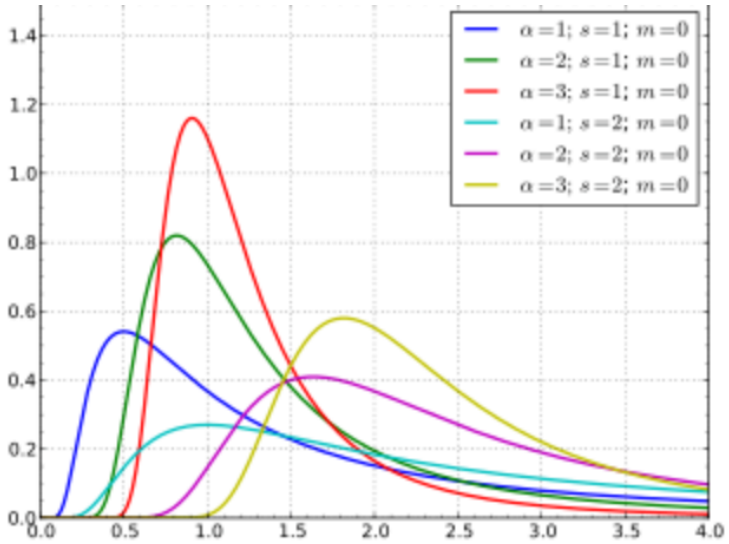
- The following proposition is very important for (logit) multinomial discrete choice models:
- **Proposition.** If $\tilde{x}_1 \sim G(m_1, s)$ and $\tilde{x}_2 \sim G(m_2, s)$ are independent Gumbel-distributed random variables, then the random variable $\tilde{x}_1 - \tilde{x}_2$ has the logistic distribution $L(m_1 - m_2, s)$.
- **Proof.** Exercise 26 in List 5.
- *Note* 1: If $\tilde{x}_1 \sim G(m_1, s)$ and $\tilde{x}_2 \sim G(m_2, s)$ are independent, then $\mathrm{E}(\tilde{x}_1 - \tilde{x}_2) = m_1 + \gamma s - (m_2 + \gamma s) = m_1 - m_2$ and $\mathrm{Var}(\tilde{x}_1 - \tilde{x}_2) = \mathrm{Var}(\tilde{x}_1) + \mathrm{Var}(\tilde{x}_2) = \frac{s^2\pi^2}{6} + \frac{s^2\pi^2}{6} = \frac{s^2\pi^2}{3}$, which is the variance of a logistic distribution with the scale parameter s.
- *Note* 2: If $\tilde{x}_1 \sim G(m_1, s)$ and $\tilde{x}_2 \sim G(m_2, s)$ are independent, then $\tilde{x}_1 + \tilde{x}_2$ does not have the logistic distribution $L(m_1 + m_2, s)$ since $\tilde{x}_1 + \tilde{x}_2$ has an asymmetric density and, moreover, $\mathrm{E}(\tilde{x}_1 + \tilde{x}_2) = m_1 + m_2 + 2\gamma s \neq m_1 + m_2$.

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \rightarrow (\mathbb{R}, \mathcal{B})$ has the generalized **extreme value type II** (or Fréchet) distribution if its density is

$$f(x; m, s, \alpha) = \begin{cases} \frac{\alpha}{s}\left(\frac{x-m}{s}\right)^{-1-\alpha} e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} & \text{for } x > m \\ 0 & \text{otherwise,} \end{cases}$$

  where m is the location parameter, $s > 0$ is the scale parameter, and $\alpha > 0$ is the shape parameter.

- Density:



- Standard extreme value type II distribution when $m = 0$, $s = 1$, and $\alpha > 0$,

$$f(x; 0, 1) = \alpha x^{-1-\alpha} e^{-x^{-\alpha}}, \quad \text{for all } x \in \mathbb{R}.$$

- Distribution function:

$$F(x; m, s) = \begin{cases} 0 & \text{for } x < m \\ \\ e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} & \text{for } x \geq m, \end{cases}$$

$$F(x; 0, 1) = \begin{cases} 0 & \text{for } x < 0 \\ \\ e^{-x^{-\alpha}} & \text{for } x \geq 0. \end{cases}$$

- $$\mathrm{E}(\tilde{x}) = m + s\Gamma\left(1 - \frac{1}{\alpha}\right) \quad \text{if } \alpha > 1,$$

$$\mathrm{Var}(\tilde{x}) = s^2\left[\Gamma\left(1 - \frac{2}{\alpha}\right) + \left(\Gamma\left(1 - \frac{1}{\alpha}\right)\right)^2\right] \quad \text{if } \alpha > 2.$$

- Moments of order k are finite if $\alpha > k$.
- Quantile function:

$$Q(p; m, s, \alpha) = m + s\left(-\ln p\right)^{-1/\alpha},$$

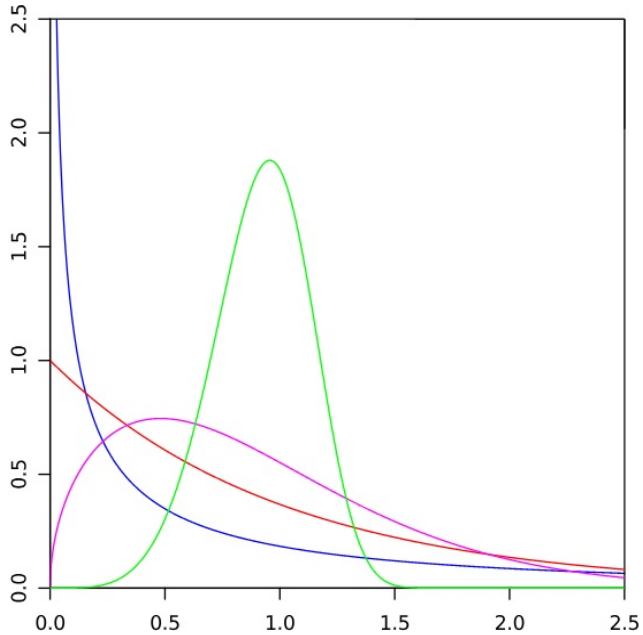
$$Q(p; 0, 1, \alpha) = \left(-\ln p\right)^{-1/\alpha}.$$

- The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$ has the generalized **extreme value type III** (or Weibull) distribution if its density is

$$f(x; m, s, \alpha) = \begin{cases} \frac{\alpha}{s} \left(\frac{x-m}{s} \right)^{\alpha-1} e^{-\left(\frac{x-m}{s}\right)^{\alpha}} & \text{for } x > m \\ 0 & \text{otherwise,} \end{cases}$$

  where m is the location parameter, $s > 0$ is the scale parameter, and $\alpha > 0$ is the shape parameter.

- Density (with $m = 0$):



- Standard extreme value type III distribution when $m = 0$, $s = 1$, and $\alpha > 0$,

$$f(x; 0, 1, \alpha) = \begin{cases} \alpha x^{\alpha-1} e^{-x^{\alpha}} & \text{for } x > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

- For $m = 0$ and $\alpha = 1$, the Weibull distribution becomes exponential with the parameter s,

$$f(x; 0, s, 1) = \begin{cases} \frac{1}{s} e^{-x/s} & \text{for } x > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

- Distribution function:

$$F(x; m, s, \alpha) = \begin{cases} 0 & \text{for } x < m \\ \\ 1 - e^{-\left(\frac{x-m}{s}\right)^{\alpha}} & \text{for } x \geq m, \end{cases}$$

$$F(x; 0, 1, \alpha) = \begin{cases} 0 & \text{for } x < 0 \\ \\ 1 - e^{-x^{\alpha}} & \text{for } x \geq 0. \end{cases}$$

-

$$\mathrm{E}(\tilde{x}) = m + s\Gamma\left(1 + \frac{1}{\alpha}\right),$$

$$\mathrm{Var}(\tilde{x}) = s^2\left[\Gamma\left(1 + \frac{2}{\alpha}\right) + \left(\Gamma\left(1 + \frac{1}{\alpha}\right)\right)^2\right].$$

- Quantile function:

$$Q(p; m, s, \alpha) = m + s\left[-\ln(1-p)\right]^{1/\alpha},$$

$$Q(p; 0, 1, \alpha) = \left[-\ln(1-p)\right]^{1/\alpha}.$$

- Other famous absolutely continuous univariate distributions appearing in the exercises are the **log-normal**, the **Cauchy**, the **Pareto**, the **Rayleigh**, the **Laplace**, the **triangular**, and the **trapezoidal** distributions. In Chaper 7, we will also see the **Wishart**, the **Student's t** and the **F** distributions.

Informal proof of the formula of integration by change of variable

We want to compute the following Lebesgue integral:

$$\int_{[a,b]} f(x)dx.$$

Note that when we consider an interval $[a,b]$, we must have $a < b$. From now on, whenever we write the integral of a function w.r.t. the Lebesgue measure it should be understood that the function is not only integrable w.r.t. that measure, but also that it is Riemann integrable so that

$$\int_{[a,b]} f(x)dx = \int_a^b f(x)dx. \tag{1}$$

Let $x = g(y)$ and assume that g is differentiable on $g^{-1}([a,b])$. This requirement is fulfilled if we assume that the function $g : M \longrightarrow \mathbb{R}$ is differentiable and M is an open subset of $\mathbb{R}$ with $g^{-1}([a,b]) \subset M$ (or equivalently with $[a,b] \subset g(M)$).

Consider the inverse function $g^{-1} : [a,b] \longrightarrow \mathbb{R}$ so that $y = g^{-1}(x)$. This inverse function g^{-1} exists if and only if the function g restricted to $g^{-1}([a,b])$, i.e., $g : g^{-1}([a,b]) \longrightarrow [a,b]$, is bijective (or a one-to-one correspondence). That is, g must be strictly increasing ($g' > 0$ a.e.) or strictly decreasing ($g' < 0$ a.e.) on $g^{-1}([a,b])$. Therefore, if $x = a$ then $y = g^{-1}(a)$, whereas if $x = b$ then $y = g^{-1}(b)$.

From the theory of Riemann integration, recall that

$$\int_a^b f(x)dx = F(b) - F(a), \text{ where } F' = f \text{ on } [a,b].$$

Therefore,

$$\int_b^a f(x)dx = F(a) - F(b) = -\int_a^b f(x)dx. \tag{2}$$

Note that the primitive (or antiderivative) of $f(g(y)) \cdot g'(y)$ is $F(g(y))$ as follows from the chain rule,

$$\frac{dF(g(y))}{dy} = F'(g(y)) \cdot g'(y) = f(g(y)) \cdot g'(y).$$

Therefore,

$$\begin{aligned} \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(y))g'(y)dy &= F(g(y))\Big|_{g^{-1}(a)}^{g^{-1}(b)} = F(g(g^{-1}(b))) - F(g(g^{-1}(a))) \\ &= F(b) - F(a) = \int_a^b f(x)dx. \end{aligned} \tag{3}$$

On the one hand, if $g' > 0$ a.e. (g is strictly increasing) then $g^{-1}(a) < g^{-1}(b)$ and, thus, $g^{-1}([a,b])$ is the interval $\left[g^{-1}(a), g^{-1}(b)\right]$. On the other hand, if $g' < 0$

a.e. (g is strictly decreasing) then $g^{-1}(a) > g^{-1}(b)$ and, thus, $g^{-1}([a,b])$ is the interval $\left[g^{-1}(b), g^{-1}(a)\right]$.

If $g' > 0$ a.e. then

$$\int_{g^{-1}(a)}^{g^{-1}(b)} f(g(y))g'(y)dy = \int_{[g^{-1}(a),g^{-1}(b)]} f(g(y))g'(y)dy = \int_{g^{-1}([a,b])} f(g(y))\underbrace{g'(y)}_{>0}dy. \quad (4)$$

However, if $g' < 0$ a.e. then

$$\int_{g^{-1}(a)}^{g^{-1}(b)} f(g(y))g'(y)dy = -\int_{g^{-1}(b)}^{g^{-1}(a)} f(g(y))g'(y)dy = \int_{g^{-1}(b)}^{g^{-1}(a)} f(g(y))\left[-g'(y)\right]dy$$

$$= \int_{[g^{-1}(b),g^{-1}(a)]} f(g(y))\left[-g'(y)\right]dy = \int_{g^{-1}([a,b])} f(g(y))\underbrace{\left[-g'(y)\right]}_{>0}dy, \quad (5)$$

where the first equality comes from (2).

Combining (1), (3), (4), and (5), we get

$$\int_{[a,b]} f(x)dx = \int_a^b f(x)dx = \int_{g^{-1}(a)}^{g^{-1}(b)} f(g(y))g'(y)dy = \int_{g^{-1}([a,b])} f(g(y))\left|g'(y)\right|dy.$$

The previous formula holds both for g strictly increasing and for g strictly decreasing.

1. The Gamma Distribution

1.A. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Proof.

1st step: Let $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}\,dy$. We make the following change of variable:

$$y = g(z) = \frac{1}{2}z^2 \Rightarrow \frac{dy}{dz} = g'(z) = z, \quad \text{for } y > 0,\ z > 0.$$

$$\Rightarrow \Gamma(\alpha) = \int_0^\infty \frac{1}{2^{\alpha-1}} z^{2\alpha-2} e^{-\frac{1}{2}z^2} z\,dz = 2^{1-\alpha}\int_0^\infty z^{2\alpha-1}e^{-\frac{1}{2}z^2}dz.$$

$$\Longrightarrow \Gamma\left(\frac{1}{2}\right) = \sqrt{2}\int_0^\infty e^{-\frac{1}{2}z^2}dz.$$

2nd step:

$$\left[\Gamma\left(\frac{1}{2}\right)\right]^2 = 2\left[\int_0^\infty e^{-\frac{1}{2}z^2}dz\right]\left[\int_0^\infty e^{-\frac{1}{2}x^2}dx\right] = 2\int_0^\infty\int_0^\infty e^{-\frac{1}{2}(z^2+x^2)}dzdx.$$

Let us make the change to polar coordinates:

$$\Rightarrow \int_0^\infty\int_0^\infty e^{-\frac{1}{2}(z^2+x^2)}dzdx = \int_{\mathbb{R}^2_+} e^{-\frac{1}{2}(z^2+x^2)}d(z,x) = \int_0^\infty\int_0^{\pi/2} e^{-\frac{1}{2}r^2} r\,d\theta dr$$

$$= \left(\int_0^\infty e^{-\frac{1}{2}r^2}r dr\right)\left(\int_0^{\pi/2} d\theta\right) = \left[-e^{-\frac{1}{2}r^2}\right]_0^\infty \cdot [\theta]_0^{\pi/2} = \frac{\pi}{2}.$$

$$\Longrightarrow \left[\Gamma\left(\frac{1}{2}\right)\right]^2 = 2\left(\frac{\pi}{2}\right) = \pi \Longrightarrow \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

– – – – – – – – – – – – – – – –

1.B. $\int_0^\infty e^{-\frac{1}{2}z^2}dz = \sqrt{\frac{\pi}{2}}$.

Proof. Observe that,

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{2}\int_0^\infty e^{-\frac{1}{2}z^2}dz = \sqrt{\pi},$$

where the first equality comes from step 1 in 1.A. and the second one comes from step 2 in 1.A. Then,

$$\int_0^\infty e^{-\frac{1}{2}z^2}dz = \sqrt{\frac{\pi}{2}}.$$

– – – – – – – – – – – – – – – –

1.C. $\mu'_r = \dfrac{\beta^r\Gamma(\alpha+r)}{\Gamma(\alpha)}$.

Proof.

$$\mu'_r = \int_0^\infty x^r \frac{1}{\beta^\alpha\Gamma(\alpha)}\, x^{\alpha-1}e^{-x/\beta}\, dx.$$

Making the change of variable

$$x = g(y) = \beta y \Leftrightarrow y = g^{-1}(x) = \frac{x}{\beta}, \text{ so that } \frac{dx}{dy} = g'(y) = \beta > 0,$$

$$\begin{aligned}\mu'_r &= \int_0^\infty \beta^r y^r \frac{1}{\beta^\alpha\Gamma(\alpha)}\beta^{\alpha-1}y^{\alpha-1}e^{-y}\beta\, dy \\ &= \frac{\beta^r}{\Gamma(\alpha)}\underbrace{\int_0^\infty y^{\alpha+r-1}e^{-y}dy}_{\Gamma(\alpha+r)} = \frac{\beta^r\Gamma(\alpha+r)}{\Gamma(\alpha)}.\end{aligned}$$

– – – – – – – – – – – – – – – –

1.D. $M_{\tilde{x}}(t) = (1-\beta t)^{-\alpha}$ if $t < 1/\beta$.

Proof.

$$M_{\tilde{x}}(t) = \int_0^\infty e^{tx} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}\, dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x\left(\frac{1}{\beta}-t\right)} dx.$$

Change of variable:

$$x = g(y) = \frac{y}{\frac{1}{\beta}-t} \Leftrightarrow y = g^{-1}(x) = x\left(\frac{1}{\beta}-t\right),$$

so that

$$\frac{dx}{dy} = g'(y) = \frac{1}{\frac{1}{\beta}-t} > 0, \quad \text{if } t < 1/\beta.$$

$$M_{\tilde{x}}(t) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \left(\frac{y}{\frac{1}{\beta}-t}\right)^{\alpha-1} e^{-y} \left(\frac{1}{\frac{1}{\beta}-t}\right) dy$$

$$= \frac{1}{\beta^\alpha \Gamma(\alpha) \left(\frac{1}{\beta}-t\right)^\alpha} \underbrace{\int_0^\infty y^{\alpha-1} e^{-y} dy}_{\Gamma(\alpha)} = \frac{1}{\beta^\alpha \left(\frac{1}{\beta}-t\right)^\alpha}$$

$$= \frac{1}{(1-\beta t)^\alpha} = (1-\beta t)^{-\alpha} \quad \text{if } t < 1/\beta.$$

– – – – – – – – – – – – – – – – –

2. The Beta Distribution

$$\int_0^1 f(x;\alpha,\beta)dx = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}dx = 1, \quad \text{with } \alpha>0,\ \beta>0.$$

Proof. We just have to prove that

$$\Gamma(\alpha)\Gamma(\beta) = \Gamma(\alpha+\beta)\cdot\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx. \tag{1}$$

Note that

$$\Gamma(\alpha)\Gamma(\beta) = \left[\int_0^\infty y^{\alpha-1}e^{-y}dy\right]\cdot\left[\int_0^\infty z^{\beta-1}e^{-z}dz\right]$$

$$= \int_0^\infty\int_0^\infty y^{\alpha-1}z^{\beta-1}e^{-(y+z)}dydz = \int_{\mathbb{R}^2_{++}} y^{\alpha-1}z^{\beta-1}e^{-(y+z)}d(y,z).$$

Let us make the following change of variable $(y,z)=g(s,x)$, where $y=sx\in(0,\infty)$ and $z=s(1-x)\in(0,\infty)$. Note that the inverse function $(s,x)=g^{-1}(y,z)$ is given by $s=y+z\in(0,\infty)$ and $x=\dfrac{y}{y+z}\in(0,1)$. Thus,

$$\left|J_{g(s,x)}\right| = \left|\det\begin{bmatrix} x & s \\ 1-x & -s\end{bmatrix}\right| = |-s| = s,$$

so that

$$\Gamma(\alpha)\Gamma(\beta) = \int_{\mathbb{R}^2_{++}} y^{\alpha-1}z^{\beta-1}e^{-(y+z)}d(y,z) = \int_{(0,\infty)\times(0,1)} (sx)^{\alpha-1}\left(s(1-x)\right)^{\beta-1}e^{-s}sd(s,x) =$$

$$\int_0^\infty\int_0^1 (sx)^{\alpha-1}\left(s(1-x)\right)^{\beta-1}e^{-s}sdxds = \underbrace{\left[\int_0^\infty s^{\alpha-1}s^{\beta-1}e^{-s}sds\right]}_{\underbrace{\int_0^\infty s^{\alpha+\beta-1}e^{-s}ds}_{\Gamma(\alpha+\beta)}}\cdot\left[\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx\right].$$

Therefore, we have proved (1).

Note: The function

$$\Xi(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

is called the beta function. Thus, the density of the beta distribution can be written as

$$f(x;\alpha,\beta) = \frac{1}{\Xi(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}.$$

– – – – – – – – – – – – – – – –

3. The Normal Distribution

3.A. $\int_{-\infty}^{\infty} n(x;\mu,\sigma) = 1.$

Proof. Let

$$x = g(z) = \mu + \sigma z \iff z = g^{-1}(x) = \frac{x-\mu}{\sigma} \Rightarrow \frac{dx}{dz} = g'(z) = \sigma > 0.$$

Then,

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sigma} e^{-\frac{1}{2}z^2} \sigma \, dz \\ &= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{1}{2}z^2} \, dz = \sqrt{\frac{2}{\pi}} \cdot \sqrt{\frac{\pi}{2}} = 1,\end{aligned}$$

where the second equality comes from the symmetry of the function $e^{-\frac{1}{2}z^2}$ with respect to zero and the third equality comes from 1.B.

– – – – – – – – – – – – – – – – –

3.B. $M_{\tilde{x}}(t) = e^{\mu t + \frac{1}{2}t^2\sigma^2}$.

Proof. We will use in this proof a technique called "completing the square".

$$M_{\tilde{x}}(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[-2xt\sigma^2 + (x-\mu)^2]} dx$$

Observe that

$$-2xt\sigma^2 + (x-\mu)^2 = [x - (\mu + t\sigma^2)]^2 - 2\mu t\sigma^2 - t^2\sigma^4.$$

Then, the last integral equals

$$e^{\frac{1}{2\sigma^2}[2\mu t\sigma^2 + t^2\sigma^4]} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[x-(\mu+t\sigma^2)]^2} dx}_{1} = e^{\mu t + \frac{1}{2}t^2\sigma^2},$$

since $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[x-(\mu+t\sigma^2)]^2} dx$ is the integral of the normal density function with parameters $\mu + t\sigma^2$ and σ, $n(x;\mu + t\sigma^2, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[x-(\mu+t\sigma^2)]^2}$.

The Standard Normal Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Also, for z = 4.0, 5.0, and 6.0 the probabilities are 0.49997, 0.4999997, 0.499999999.

Exercises. Probability and Statistics. IDEA.
4. Special Distributions

1. If $\tilde{x}$ has the discrete uniform distribution with $f(x) = 1/k$ for $x = 1, 2, ..., k$ show that

   (a) its mean is $\mu = \dfrac{k+1}{2}$;

   (b) its variance is $\sigma^2 = \dfrac{k^2-1}{12}$.

2. If $\tilde{x}$ has the discrete uniform distribution $f(x) = 1/k$ for $x = 1, 2, ..., k$, show that its moment-generating function is given by

$$M_{\tilde{x}}(t) = \frac{e^t\left(1 - e^{kt}\right)}{k\left(1 - e^t\right)}, \quad \text{for } t \neq 0.$$

   Also find the mean of this distribution by evaluating $\lim_{t\to 0} M'_{\tilde{x}}(t)$, and compare the result obtained in part (a) of Exercise 1.

3. A multiple-choice test consists of eight questions and three answers to each question (of which only one is correct). If a student answers each question by rolling a balanced dice and checking the first answer if he gets a 1 or 2, the second answer if he gets a 3 or 4, and the third answer if he gets a 5 or 6, what is the probability that he will get exactly four correct answers?

4. A quality control engineer wants to check whether (in accordance with specifications) 95 percent of the electronic components shipped by her company are without flaws. To this end, she randomly selects 20 components from each large lot ready to be shipped and passes the lot if they are all without flaws; otherwise each component in the lot is checked. Assuming that the lots are so large that we can use the binomial distribution as an approximation, find the probabilities that she will commit the error of

   (a) holding a lot for a complete check even though 95 percent of the components are without flaws;

   (b) passing a lot even though only 90 percent of the components are without flaws;

   (c) passing a lot even though only 80 percent of the components are without flaws;

   (d) passing a lot even though only 70 percent of the components are without flaws.

5. Show that the moment-generating function of the geometric distribution is given by

$$M_{\tilde{x}}(t) = \frac{\theta e^t}{1 - e^t(1-\theta)}$$

   in a neighborhood of $t = 0$ and use it to verify that $\mu = \dfrac{1}{\theta}$ and $\sigma^2 = \dfrac{1-\theta}{\theta^2}$.

6. According to the Mendelian theory of heredity, if plants with round yellow seeds are crossbred with plants with wrinkled green seeds, the probabilities of getting a plant that produces round yellow seeds, wrinkled yellow seeds, round green seeds, or wrinkled green seeds are, respectively, 9/16, 3/16, 3/16, and 1/16. What is the probability that among nine plants thus obtained there will be four that produce round yellow seeds, two that produce wrinkled yellow seeds, three that produce round green seeds, and none that produce wrinkled green seeds?

7. Show that if a random variable has a uniform density on the interval (α, β), the probability that it will take on a value less than $\alpha + p(\beta - \alpha)$ is equal to p.

8. A random variable is said to have a **Cauchy distribution** if its density is given by

$$f(x) = \frac{\beta/\pi}{(x-\alpha)^2 + \beta^2} \quad \text{for } -\infty < x < \infty$$

Show that for this distribution the moment μ_1' does not exist and the moment μ_2' is not finite. Note that this means that a random variable having the Cauchy distribution does not belong to the L^1 space. For the purpose of this example set $\alpha = 0$ and $\beta = 1$.

Hint: If $f(x) = \arctan(x)$, where the function $\arctan(\cdot) : \mathbb{R} \longrightarrow (-\pi/2, \pi/2)$ is the inverse function of $\tan(\cdot) : (-\pi/2, \pi/2) \longrightarrow \mathbb{R}$, then the derivative of f is $f'(x) = \dfrac{1}{1+x^2}$. Recall also that $\tan\left(\dfrac{\pi}{4}\right) = 1$, $\tan\left(-\dfrac{\pi}{4}\right) = -1$, $\lim\limits_{\theta \to \pi/2} \tan(\theta) = \infty$, $\lim\limits_{\theta \to -\pi/2} \tan(\theta) = -\infty$ and $\tan(0) = 0$.

9. Show that if a random variable has an exponential density with parameter θ, the probability that it will take on a value less than $-\theta \cdot \ln(1-p)$ is equal to p for $0 \leq p < 1$.

10. A random variable $\tilde{x}$ has a **Pareto distribution** if its density is given by

$$f(x) = \begin{cases} \dfrac{\alpha k^\alpha}{x^{\alpha+1}} & \text{for } x > k \\ \\ 0 & \text{elsewhere,} \end{cases}$$

with $\alpha > 0$ and $k > 0$. Show that the moment μ_r' is finite only if $r < \alpha$. Note that this means that $\tilde{x} \in L^r$ only for $r < \alpha$.

11. Karl Pearson, one of the founders of modern statistics, showed that the differential equation

$$\frac{f'(x)}{f(x)} = \frac{d-x}{a+bx+cx^2}$$

yields (for appropriate values of the constants a, b, c, and d) the densities of the most important distributions of statistics. Verify that the differential equation gives the density f of

(a) the gamma distribution when $a = c = 0$, $b > 0$, and $d > -b$;

(b) the exponential distribution when $a = c = d = 0$ and $b > 0$;

(c) the beta distribution when $a = 0$, $b = -c$, $\frac{d-1}{b} < 1$, and $\frac{d}{b} > -1$;

(d) the normal distribution when $b = c = 0$ and $a > 0$.

12. A point X is chosen on the segment AB, whose midpoint is C and whose length is a. If $\tilde{x}$, the distance from X to A, is a random variable having a uniform density on the interval $(0, a)$, what is the probability that AX, BX, and AC will form a triangle?

13. The mileage (in thousands of miles) which car owners get with a certain kind of tire is a random variable having an exponential distribution with $\theta = 40$. Find the probabilities that one of this tires will last

(a) at least 20 000 miles;

(b) at most 30 000 miles.

14. Twice more differentiating the moment-generating function of the normal distribution, verify that $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$.

15. If $\tilde{x}$ is a random variable having a normal distribution, what are the probabilities of getting a value

(a) within one standard deviation of the mean;

(b) within two standard deviations of the mean;

(c) within three standard deviations of the mean;

(d) within four standard deviations of the mean?

16. Suppose that during periods of transcendental meditation the reduction of a person's oxygen consumption is a random variable having a normal distribution with $\mu = 37.6$ cc per minute and $\sigma = 4.6$ cc per minute. Find the probabilities that during a period of transcendental meditation a person's oxygen consumption will be reduced by

(a) at least 44.5 cc per minute;

(b) at most 35.0 cc per minute;

(c) anywhere from 30.0 to 40.0 cc per minute.

17. If $\tilde{x}$ and $\tilde{y}$ have the circular normal distribution with $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 12$, find

(a) the probability of getting a point (x, y) inside the circle $x^2 + y^2 \leq 36$;

(b) the value of c for which the probability of getting a point (x, y) inside the circle $x^2 + y^2 \leq c^2$ is 0.80.

18. The joint distribution of the random variables $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$ is $\mathrm{MN}(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 9 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

(a) Compute $P\{\tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 \leq 3\}$.

(b) Find the conditional density of $(\tilde{x}_1, \tilde{x}_2)^\intercal$ given $\tilde{x}_3 = 0$.

(c) Find a so that $\tilde{z} = a\tilde{x}_1 + \tilde{x}_2$ is independent of $\tilde{x}_3$.

19. Find the moment-generating function of the Pascal (or negative binomial) distribution having the probability function $b^*(x; k, \theta)$ for $x = k, k+1, ...$ and use this moment-generating function to prove that the mean of the Pascal distribution is k/θ. *Hint:* You should use the results of Exercise 6 of List 1.

20. The probability of giving birth to a male is 0.51. Abstracting from multiple pregnancies, find

(a) the probability that a couple had to wait until the fourth child in order to get exactly one male (i.e., the first male is obtained in the fourth trial);

(b) the probability that a couple had to wait until the fourth child in order to get exactly two males (i.e., the second male is obtained in the fourth trial);

(c) the expected number of children that a couple should have in order to get exactly two males.

Hint: For (c) you should use the result of the previous Exercise 19 of this list.

21. The random vector $(\tilde{x}, \tilde{y})$ is distributed according to the following density function:

$$f(x,y) = \begin{cases} \dfrac{1}{\pi} e^{-\frac{1}{2}(x^2+y^2)} & \text{if } x > 0 \text{ and } y \leq 0, \text{ or if } x \leq 0 \text{ and } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that the marginal distributions of the random variables $\tilde{x}$ and $\tilde{y}$ are standard normal.

(b) Are the random variables $\tilde{x}$ and $\tilde{y}$ independent?

(c) Is the random vector $(\tilde{x}, \tilde{y})$ multivariate normally distributed? Any comment about these results?

22. The distribution of a random vector $(\tilde{x}, \tilde{y})$ is characterized by the following density function:

$$f(x,y) = \begin{cases} k\,(x^2+y^2)^{1/2} & \text{when } (x,y) \in A \\ 0 & \text{when } (x,y) \notin A, \end{cases}$$

Find the value of the constant k

(a) when $A = \{(x, y) \in \mathbb{R}^2 \mid 4 < x^2 + y^2 < 9,\ 0 < y < x\}$;

(b) when $A = \{(x, y) \in \mathbb{R}^2 \mid 9 < x^2 + y^2 < 16,\ y > 0\}$.

23. Find the area of a circle with radius r by means of the following two methods:

(a) Compute the integral with respect to the Lebesgue measure on $\mathbb{R}^2$,

$$\int_C d(x, y), \quad \text{where } C = \left\{(x, y) \in \mathbb{R}^2 \,\middle|\, x^2 + y^2 \leq r^2\right\},$$

through a change to polar coordinates. Why this integral is equal to the area of a circle?

(b) Compute an integral with respect to the Lebesgue measure on $\mathbb{R}$ using the change of variable $x = r \sin \alpha$.

Hint: Remember that

$$\cos^2 \alpha = \frac{1 + \cos(2\alpha)}{2} \text{ since } \cos(\alpha + \beta) = \cos \alpha \cdot \cos \beta - \sin \alpha \cdot \sin \beta.$$

24. Find the following integral using the change to polar coordinates:

$$\int_C \frac{e^{-\left[(x+4)^2 + (y-3)^2\right]}}{\pi} d(x, y),$$

where C is the region of $\mathbb{R}^2$ such that $(x + 4)^2 + (y - 3)^2 \leq 36$, that is, the region C is a circle with its radius length equal to 6 and its center located at the point $(-4, 3)$.

25. The four blood groups are distributed in a given (very large) population according to the following percentages:

$$0 = 45\% \qquad A = 43\% \qquad B = 8\% \qquad AB = 4\%$$

If we randomly and independently choose people, compute the probability:

(a) That when we pick 10 individuals, 4 of them will be of the A group, 2 of the B group, and 1 of the AB group;

(b) That when we pick three individuals, one of them will be able to donate blood to the other two (the 0 group is the universal donor and the AB group is the universal receiver);

(c) That we will have to wait until the fifth volunteer to find someone of the AB group, if volunteers are called for an emergency.

26. We have 7 white balls, 3 red balls, and 2 blue balls in an urn.

    (a) What is the probability that, if we pick 5 balls without replacement, we will get 3 white balls, 2 red balls and none blue?

    (b) What is the probability that, if we pick 5 balls with replacement (that is, if we put back in the urn the ball we have picked before the next extraction), we will also get 3 white balls, 2 red balls and none blue?

    (c) Which distribution is used to solve for part (a)? The multinomial or the multivariate hypergeometric? And for part (b)? Why?

27. (a) Prove that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}.$$

    (b) Find $\int_0^{\infty} e^{-x^2} dx$.

28. Suppose that the random variable $\tilde{z}$ has a standard normal distribution, $\tilde{z} \sim \mathrm{N}(0,1)$. Let $\tilde{x}$ be a discrete random variable with the following probability function:

$$f(x) = \begin{cases} \dfrac{1}{2} & \text{for } x = 1 \\ \\ \dfrac{1}{2} & \text{for } x = -1. \end{cases}$$

    Assume that $\tilde{z}$ and $\tilde{x}$ are independent. Define the random variable $\tilde{y} = \tilde{x} \cdot \tilde{z}$.

    Prove that

    (a) the random variables $\tilde{z}$ and $\tilde{y}$ are not multivariate normally distributed. *Hint:* In order to prove it you could compute the probability $P\{\tilde{z} + \tilde{y} = 0\}$ and extract the appropriate conclusion;

    (b) the random variable $\tilde{y}$ has a standard normal distribution. *Hint:* you should use the theorem of total probability to compute the probability $P\{\tilde{y} \leq y\}$ and check that this probability is equal to the distribution function $\mathrm{N}(y)$ of a standard normal random variable;

    (c) $\mathrm{Cov}(\tilde{z}, \tilde{y}) = 0$, i.e., the random variables $\tilde{z}$ and $\tilde{y}$ are uncorrelated. *Hint:* you should use the fact that $\mathrm{E}(\tilde{z} \cdot \tilde{y}) = \mathrm{E}(\mathrm{E}(\tilde{z} \cdot \tilde{y} \,|\tilde{x}))$;

    (d) the random variables $\tilde{z}$ and $\tilde{y}$ are not independent. *Hint:* In order to prove it you could compute $P(\{\tilde{y} > 1\} \cap \{\tilde{z} \in [0, 1/2)\})$ and extract the appropriate conclusion.

    Note: this exercise provides an example of uncorrelated (but not independent) normal random variables ($\tilde{z}$ and $\tilde{y}$) that are not multivariate normal.

29. Prove that the mean of the beta distribution with parameters α and β is $\dfrac{\alpha}{\alpha + \beta}$.

30. The distribution of a random vector $(\tilde{x}, \tilde{y})$ is characterized by the following density function:

$$f(x,y) = \begin{cases} \dfrac{k}{(x^2+y^2)^3} & \text{when } (x,y) \in D \\ \\ 0 & \text{when } (x,y) \notin D, \end{cases}$$

where $D = \{(x,y) \in \mathbb{R}^2 \,|\, 1 < x^2 + y^2 < 3,\ x < 0,\ y < 0\}$.

Find the value of the constant k.

31. We randomly and independently pick two points A and B on a circumference with radius r. Find the expected value of the distance between A and B.

32. Let $\tilde{z}$ be a standard normal random variable. Prove that

$$P\{\tilde{z} \geq b\} = P\{\tilde{z} \leq -b\} \leq e^{-b^2/2}, \quad \text{for all } b \geq 0.$$

Hint: for this exercise use the result in part (a) of Exercise 33 of List 3.

33. Assume that there are 6 balls in a box and 3 of them are white, 2 are black, and 1 is green.

(a) We extract 4 balls with no replacement from the box. What is the probability of extracting 2 white balls and 2 black balls.

(b) We extract 4 balls with replacement from the box. What is the probability of extracting 2 white balls and 2 black balls.

34. Prove that the logistic density function is symmetric around the location parameter m. That is, $l(m-x, m, s) = l(m+x, m, s)$ for all $x \in \mathbb{R}$.

35. The joint distribution of the random variables $\tilde{x}_1,\ \tilde{x}_2,\ \tilde{x}_3$ is multivariate normal, $\text{MN}(\mu, \Sigma)$, where the mean vector and the covariance matrix are

$$\mu = \begin{pmatrix} 1 \\ \\ 2 \\ \\ 1 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 5 & -1 & 1 \\ \\ -1 & 3 & 1 \\ \\ 1 & 1 & 2 \end{pmatrix}.$$

(a) Find the distribution of the random variable $\tilde{w} = \tilde{x}_1 + 2\tilde{x}_2 + 3\tilde{x}_3$.

(b) Compute $P\{\tilde{x}_1 + 2\tilde{x}_2 + 3\tilde{x}_3 \leq 6\}$.

(c) Find the value of the constant c that makes the random variable $\tilde{y} = \tilde{x}_1 + 2\tilde{x}_2 + c\tilde{x}_3$ independent of $\tilde{x}_2$.

5. Functions of Random Variables

5.1. The distribution of a function of a random object

- Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object with distribution $P_{\tilde{x}}$ and $g : (\Omega', \mathcal{F}') \longrightarrow (\Omega'', \mathcal{F}'')$ be a measurable function. Then, the distribution of the random object $\tilde{y} = g(\tilde{x}) : (\Omega, \mathcal{F}) \longrightarrow (\Omega'', \mathcal{F}'')$ is given by

$$P_{\tilde{y}}(B) = P_{g(\tilde{x})}(B) = P\{g(\tilde{x}) \in B\}$$

$$= P\left\{\tilde{x} \in g^{-1}(B)\right\} = P_{\tilde{x}}\left(g^{-1}(B)\right), \quad \text{for all } B \in \mathcal{F}''.$$

5.2. The distribution function of a vector-valued function of a random object

- Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object with distribution $P_{\tilde{x}}$ and $g : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be a vector-valued Borel measurable function. Then, the distribution function $F_{\tilde{y}} : \mathbb{R}^n \longrightarrow \mathbb{R}$ of the random vector $\tilde{y} = g(\tilde{x}) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ is given by

$$\begin{aligned} F_{\tilde{y}}(y) &= P\{\tilde{y} \leqq y\} = P\{g(\tilde{x}) \leqq y\} = P\{g(\tilde{x}) \in (-\infty, y]\} \\ &= P\left\{\tilde{x} \in g^{-1}(-\infty, y]\right\} = P_{\tilde{x}}\left(g^{-1}(-\infty, y]\right) \\ &= P_{\tilde{x}}\left\{x \in \Omega' \,|g(x) \leqq y\right\}. \end{aligned}$$

- Recall that $y = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$ and

$$\begin{aligned} (-\infty, y] &= \{z \in \mathbb{R}^n \,|z \leqq y\} \\ &= \{z \in \mathbb{R}^n \,|z_1 \leq y_1, z_2 \leq y_2, ..., z_n \leq y_n\}. \end{aligned}$$

- If $F_{\tilde{y}}(y)$ is absolutely continuous, then the density function will be

$$f_{\tilde{y}}\left(y_1, y_2, ..., y_n\right) = \frac{\partial^n F_{\tilde{y}}\left(y_1, y_2, ..., y_n\right)}{\partial y_1 \partial y_2 ... \partial y_n}.$$

  when this nth crossed partial derivative of $F_{\tilde{y}}$ exists. Moreover, this derivative exists a.e. w.r.t. Lebesgue measure on $(\mathbb{R}^n, \mathcal{B})$.

- **Example:** Let $\tilde{z}$ be a random variable and $\tilde{x} = |\tilde{z}| \geq 0$. Then

$$F_{\tilde{x}}(x) = P\{\tilde{x} \leq x\} = P\{|\tilde{z}| \leq x\} = P\{-x \leq \tilde{z} \leq x\}$$

$$= P_{\tilde{z}}[-x, x] = F_{\tilde{z}}(x) - \lim_{s \to -x^{-}} F_{\tilde{z}}(s), \text{ for } x \geq 0.$$

and $F_{\tilde{x}}(x) = 0$ for $x < 0$.

- Assume now that $\tilde{z} \sim \mathrm{N}(0,1)$. Then, $F_{\tilde{z}}$ is absolutely continuous (and, hence, continuous), so that the distribution function of $\tilde{x} = |\tilde{z}|$ is

$$F_{\tilde{x}}(x) = \begin{cases} F_{\tilde{z}}(x) - F_{\tilde{z}}(-x) & \text{for } x \geq 0 \\ \\ 0 & \text{for } x < 0, \end{cases}$$

and its density is

$$f_{\tilde{x}}(x) = \begin{cases} f_{\tilde{z}}(x) + f_{\tilde{z}}(-x) = n(x;0,1) + n(-x;0,1) & \text{for } x > 0 \\ \\ 0 & \text{for } x \leq 0. \end{cases}$$

- Given the symmetry of the density function $n(\cdot; 0, 1)$ with respect to 0, we get

$$f_{\tilde{x}}(x) = \begin{cases} 2n(x; 0, 1) = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}x^2} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0. \end{cases}$$

5.3. The probability function of a discrete vector-valued function of a random object

- Let $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$ be a random object with distribution $P_{\tilde{x}}$, $g : (\Omega', \mathcal{F}') \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be a vector-valued Borel measurable function, and the random vector $\tilde{y} = g(\tilde{x}) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ be discrete (i.e., $\tilde{y}(\Omega)$ is discrete or countable). Then, the probability function $f_{\tilde{y}} : \tilde{y}(\Omega) \longrightarrow [0, 1]$ of the random vector $\tilde{y}$ is given by

$$f_{\tilde{y}}(y) = P_{\tilde{y}}\{y\} = P_{\tilde{x}}\left(g^{-1}(y)\right), \quad \text{for all } y \in g(\tilde{x}(\Omega)) \equiv \tilde{y}(\Omega).$$

- If either $\tilde{x}$ is discrete (i.e., $\tilde{x}(\Omega)$ is discrete or countable) or g is discrete (i.e., $g(\Omega')$ is discrete or countable), then the random vector $\tilde{y} = g(\tilde{x})$ is discrete (i.e., $\tilde{y}(\Omega) \equiv g(\tilde{x}(\Omega))$ is discrete or countable).

- If $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}^m, \mathcal{B})$ is a discrete random vector with probability function $f_{\tilde{x}} : \tilde{x}(\Omega) \longrightarrow [0,1]$ and $g : (\mathbb{R}^m, \mathcal{B}) \longrightarrow (\mathbb{R}^n, \mathcal{B})$, then the probability function $f_{\tilde{y}} : \tilde{y}(\Omega) \longrightarrow [0,1]$ of the discrete random vector $\tilde{y} = g(\tilde{x}) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B})$ is given by

$$f_{\tilde{y}}(y) = P_{\tilde{y}}\{y\} = P_{\tilde{x}}\left(g^{-1}(y)\right)$$

$$= \sum_{x \in g^{-1}(y)} f_{\tilde{x}}(x), \quad \text{for all } y \in g(\tilde{x}(\Omega)) \equiv \tilde{y}(\Omega).$$

- If the restriction of the function g to the range of $\tilde{x}$, $g : \tilde{x}(\Omega) \longrightarrow g(\tilde{x}(\Omega))$, is a one-to-one correspondence, then

$$f_{\tilde{y}}(y) = P_{\tilde{y}}\{y\} = P_{\tilde{x}}\left(g^{-1}(y)\right)$$

$$= f_{\tilde{x}}\left(g^{-1}(y)\right), \quad \text{for all } y \in g(\tilde{x}(\Omega)) \equiv \tilde{y}(\Omega).$$

- **Example:** Let $\tilde{x}$ be a binomial random variable with probability function

$$b_{\tilde{x}}(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, ..., n$$

and let $\tilde{y} = g(\tilde{x}) = \frac{\tilde{x}}{n}$ be the percentage of successes in n independent trials. Note that $x = g^{-1}(y) = ny$. Then, $g(\tilde{x}(\Omega)) = \{0, 1/n, 2/n, ..., 1\}$ and

$$f_{\tilde{y}}(y; n, \theta) = b_{\tilde{x}}(g^{-1}(y); n, \theta)$$

$$= \binom{n}{ny} \theta^{ny} (1-\theta)^{\overbrace{n-ny}^{n(1-y)}}, \quad y = 0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1.$$

5.4. Convolutions

- **Definition.** The convolution of two independent random variables on the same probability space, $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ and $\tilde{y} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$, is the distribution of their sum, $\tilde{x} + \tilde{y}$, i.e., the distribution of the random variable $\tilde{z} = \tilde{x} + \tilde{y}$, with $\tilde{z} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$.

- **Convolution distribution:** Consider two random variables $\tilde{x}$ and $\tilde{y}$ on the same probability space $(\Omega, \mathcal{F}, P)$, then the distribution of the random variable $\tilde{z} = \tilde{x} + \tilde{y}$ is

$$P_{\tilde{z}}(B) = P_{\tilde{x}+\tilde{y}}(B) = P\left\{\tilde{x} + \tilde{y} \in B\right\}$$

$$= P\left\{\omega \in \Omega \mid \tilde{x}(\omega) + \tilde{y}(\omega) \in B\right\}, \quad \text{for all } B \in \mathcal{B}.$$

- Independency is not required for the previous expression.

- **Convolution distribution function:** If the two random variables $\tilde{x}$ and $\tilde{y}$ on the same probability space $(\Omega, \mathcal{F}, P)$ are independent, then the distribution function of the random variable $\tilde{z} = \tilde{x} + \tilde{y}$ is

$$F_{\tilde{z}}(z) = F_{\tilde{x}+\tilde{y}}(z) = P\{\tilde{x} + \tilde{y} \leq z\}$$

$$= \int_{\mathbb{R}} P\{\tilde{x} + \tilde{y} \leq z \,|\tilde{x} = x\}\, dF_{\tilde{x}}(x) = \int_{\mathbb{R}} P\{x + \tilde{y} \leq z\}\, dF_{\tilde{x}}(x)$$

$$= \int_{\mathbb{R}} P\{\tilde{y} \leq z - x\}\, dF_{\tilde{x}}(x) = \int_{\mathbb{R}} F_{\tilde{y}}(z - x) dF_{\tilde{x}}(x),$$

where the third equality comes from the theorem of total probability and the fourth from the independence between $\tilde{x}$ and $\tilde{y}$. Note that, if these two variables where not independent, then we will have that $P\{\tilde{x} + \tilde{y} \leq z \,|\tilde{x} = x\} = P\{x + \tilde{y} \leq z \,|\tilde{x} = x\}$, which is not necessarily equal to $P\{x + \tilde{y} \leq z\}$.

- **Proposition.** Assume that the two random variables $\tilde{x}$ and $\tilde{y}$ on the same probability space $(\Omega, \mathcal{F}, P)$ are independent.

  (a) If $\tilde{x}$ and $\tilde{y}$ are discrete with the probability functions $f_{\tilde{x}} : \tilde{x}(\Omega) \longrightarrow [0,1]$ and $f_{\tilde{y}} : \tilde{y}(\Omega) \longrightarrow [0,1]$, respectively, then the probability function of their sum $\tilde{z} = \tilde{x} + \tilde{y}$ (i.e., the convolution pmf) is

$$f_{\tilde{x}+\tilde{y}}(z) = f_{\tilde{z}}(z) = \sum_{x \in \tilde{x}(\Omega)} f_{\tilde{y}}(z-x) f_{\tilde{x}}(x), \quad \text{for } z \in \tilde{z}(\Omega),$$

  with $z - x \in \tilde{y}(\Omega)$.

  (b) If $\tilde{x}$ and $\tilde{y}$ are absolutely continuous with the densities $f_{\tilde{x}} : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$ and $f_{\tilde{y}} : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$, respectively, then the density of their sum $\tilde{z} = \tilde{x} + \tilde{y}$ (i.e., the convolution pdf) is

$$f_{\tilde{x}+\tilde{y}}(z) = f_{\tilde{z}}(z) = \int_{\mathbb{R}} f_{\tilde{y}}(z-x) f_{\tilde{x}}(x) dx.$$

- **Proof. (a)** Define the discrete random variable $\tilde{z} = \tilde{x} + \tilde{y}$. Then, the probability function of $\tilde{z}$, $f_{\tilde{z}} : \tilde{z}(\Omega) \longrightarrow [0,1]$, satisfies

$$f_{\tilde{z}}(z) = f_{\tilde{x}+\tilde{y}}(z) = P\{\tilde{x} + \tilde{y} = z\}$$

$$= \sum_{x \in \tilde{x}(\Omega)} P\{\tilde{x} + \tilde{y} = z \,|\tilde{x} = x\} \cdot P\{\tilde{x} = x\}$$

$$= \sum_{x \in \tilde{x}(\Omega)} P\{x + \tilde{y} = z\} \cdot P\{\tilde{x} = x\} = \sum_{x \in \tilde{x}(\Omega)} P\{\tilde{y} = z - x\} \cdot P\{\tilde{x} = x\}$$

$$= \sum_{x \in \tilde{x}(\Omega)} f_{\tilde{y}}(z - x) f_{\tilde{x}}(x), \quad \text{for } z \in \tilde{z}(\Omega), \text{ with } z - x \in \tilde{y}(\Omega),$$

where the third equality comes from the theorem of total probability and the fourth from the independence between $\tilde{x}$ and $\tilde{y}$. Note that, if these two variables where not independent, then we will have that $P\{\tilde{x} + \tilde{y} = z \,|\tilde{x} = x\} = P\{x + \tilde{y} = z \,|\tilde{x} = x\}$, which is not necessarily equal to $P\{x + \tilde{y} = z\}$.

(b) We know that the distribution function $F_{\tilde{z}}$ of the absolutely continuous random variable $\tilde{z} = \tilde{x} + \tilde{y}$ is

$$F_{\tilde{z}}(z) = \int_{\mathbb{R}} F_{\tilde{y}}(z - x) dF_{\tilde{x}}(x) = \int_{\mathbb{R}} F_{\tilde{y}}(z - x) f_{\tilde{x}}(x) dx.$$

Differentiating the last expression with respect to z, and exchanging the order of differentiation and integration, we obtain the density of the random variable $\tilde{z} = \tilde{x} + \tilde{y}$,

$$f_{\tilde{x}+\tilde{y}}(z) = f_{\tilde{z}}(z) = \frac{dF_{\tilde{z}}(z)}{dz} = \int_{\mathbb{R}} \frac{dF_{\tilde{y}}(z - x)}{dz} f_{\tilde{x}}(x) dx$$

$$= \int_{\mathbb{R}} f_{\tilde{y}}(z - x) f_{\tilde{x}}(x) dx, \quad \text{for } z \in \mathbb{R}. \qquad Q.E.D.$$

- See the handout for a non-trivial application of the previous proposition to the convolution of two independent absolutely continuous random variables.

5.5. The density of a vector-valued function of an absolutely continuous random vector

- Let $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ be a random vector with density $f_{\tilde{x}} : \mathbb{R}^m \longrightarrow \overline{\mathbb{R}}$ and $g : (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m)) \longrightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ be a vector-valued Borel measurable function. Then the distribution $P_{\tilde{y}}$ of the random vector $\tilde{y} = g(\tilde{x}) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is given by

$$P_{\tilde{y}}(B) = P_{\tilde{x}}\left(g^{-1}(B)\right) = \int_{g^{-1}(B)} dP_{\tilde{x}} = \int_{g^{-1}(B)} f_{\tilde{x}}(x)dx,$$

where $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^m$, for all $B \in \mathcal{B}(\mathbb{R}^n)$.

• Let $m = n$ and assume now that (i) the function $g : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)) \longrightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ restricted to the (open) set A of values of its domain satisfying $f_{\tilde{x}}(x) \neq 0$, $g : A \longrightarrow g(A)$, is a one-to-one correspondence (or bijective function), and (ii) the inverse function $g^{-1} : g(A) \longrightarrow A$ is continuously differentiable (which implies that $g(A)$ is open as A is open). Then, the density function $f_{\tilde{y}}$ of the absolutely continuous random vector $\tilde{y} = g(\tilde{x}) : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is given by

$$f_{\tilde{y}}(y) = \begin{cases} f_{\tilde{x}}(g^{-1}(y)) \left| J_{g^{-1}}(y) \right| & \text{for } y \in g(A) \\ \\ 0 & \text{otherwise.} \end{cases}$$

• **Proof:** For every Borel set $B \subset g(A) \equiv \left\{ y \in \mathbb{R}^n \left| f_{\tilde{x}}\left(g^{-1}(y)\right) \neq 0 \right. \right\}$,

$$P_{\tilde{y}}(B) = P_{\tilde{x}}\left(g^{-1}(B)\right) = \int_{g^{-1}(B)} f_{\tilde{x}}(x)dx = \int_B \underbrace{f_{\tilde{x}}(g^{-1}(y)) \left| J_{g^{-1}}(y) \right|}_{f_{\tilde{y}}(y)} dy,$$

where $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ and $y = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$. $Q.E.D.$

- **Example:** Let $\tilde{z} \sim \mathrm{N}(0,1)$ and $\tilde{y} = h(\tilde{z}) = \tilde{z}^2$. Find the density of $\tilde{y}$.

- Note that $h : \mathbb{R} \longrightarrow \mathbb{R}$ is not a one-to-one correspondence. However, we can make $\tilde{y} = g(\tilde{x}) = \tilde{x}^2$, where $\tilde{x} = |\tilde{z}|$ and, then, $g : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is a one-to-one correspondence, whose inverse is $\tilde{x} = g^{-1}(\tilde{y}) = \tilde{y}^{1/2}$ and

$$\frac{dx}{dy} = \frac{dg^{-1}(y)}{dy} = \frac{1}{2} y^{-1/2} > 0.$$

- Recall that the density of $\tilde{x}$ is

$$f_{\tilde{x}}(x) = \begin{cases} 2n(x;0,1) & \text{for } x > 0 \\ \\ 0 & \text{for } x \leq 0. \end{cases}$$

- Then,

$$f_{\tilde{y}}(y) = \begin{cases} 2n(y^{1/2};0,1)\cdot\left(\frac{1}{2}y^{-1/2}\right) = n(y^{1/2};0,1)\cdot y^{-1/2} & \text{for } y > 0 \\ 0 & \text{for } y \leq 0. \end{cases}$$

- Note that

$$n(y^{1/2};0,1)\cdot y^{-1/2} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(y^{1/2}\right)^2}\cdot y^{-1/2} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y}\cdot y^{-1/2}.$$

- Therefore,

$$f_{\tilde{y}}(y) = \begin{cases} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y}y^{-1/2} & \text{for } y > 0 \\ 0 & \text{for } y \leq 0. \end{cases}$$

- Recall also that a random variable $\tilde{y}:(\Omega,\mathcal{F})\rightarrow(\mathbb{R},\mathcal{B})$ has the χ^2 (chi-square) distribution with ν degrees of freedom $\left(\tilde{y}\sim\chi^2_\nu\right)$ if its density is

$$f_{\tilde{y}}(y)=\begin{cases}\dfrac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}y^{\frac{\nu-2}{2}}e^{-y/2} & \text{for } y>0\\ \\ 0 & \text{otherwise.}\end{cases}$$

- Then, $\tilde{y}\sim\chi^2_1$ if

$$f_{\tilde{y}}(y)=\begin{cases}\dfrac{1}{2^{1/2}\Gamma\left(\frac{1}{2}\right)}y^{-1/2}e^{-y/2}=\dfrac{1}{\sqrt{2\pi}}e^{-y/2}y^{-1/2} & \text{for } y>0\\ \\ 0 & \text{otherwise.}\end{cases}$$

- Therefore, the random variable $\tilde{y}=\tilde{z}^2$, with $\tilde{z}\sim\mathrm{N}(0,1)$, is chi-square with 1 degree of freedom $\left(\tilde{y}\sim\chi^2_1\right)$.

- Even if $m > n$ we can apply this method by introducing additional variables to enlarge the dimension of the image space and, thus, to make the function g a one-to-one correspondence for all the values $x \in \mathbb{R}^m$ of the domain of g for which $f_{\tilde{x}}(x) \neq 0$. Then, we will need to compute the marginal density to get rid of the additional variables we have introduced. See the handout.

- Note that, in order to find the density of $\tilde{y} = g(\tilde{x})$ we could find first the distribution function of $\tilde{y}$ from the density of $\tilde{x}$ and, then, by performing the corresponding derivative, we obtain the density of $\tilde{y}$ (**distribution function method**). This method does not require that g be a bijection for all the values $x \in \mathbb{R}^n$ of the domain of g for which $f_{\tilde{x}}(x) \neq 0$ (see Section 5.2).

- In this Section 5.5 we find the density of $\tilde{y} = g(\tilde{x})$ directly from the density function of $\tilde{x}$ (**density function method**). This method requires that g be a bijection for all the values $x \in \mathbb{R}^n$ of the domain of g for which $f_{\tilde{x}}(x) \neq 0$.

5.6. Characteristic function, moment-generating function, and Laplace transform of a function of a random variable

- If $\tilde{y} = g(\tilde{x})$, then we can compute the characteristic function $\varphi_{\tilde{y}}$ of $\tilde{y}$ (or its moment-generating function $M_{\tilde{y}}$ or its Laplace transform $\Lambda_{\tilde{y}}$ if they are well-defined in a neighborhood of $t = 0$), using the distribution $P_{\tilde{x}}$ of $\tilde{x}$,

$$\varphi_{\tilde{y}}(t) = \mathrm{E}(e^{it\tilde{y}}) = \mathrm{E}(e^{itg(\tilde{x})}) = \int_{\mathbb{R}} e^{itg(x)} dP_{\tilde{x}}(x),$$

$$= (\mathrm{E}[\cos(tg(\tilde{x}))], \mathrm{E}[\sin(tg(\tilde{x}))])$$

$$= \left(\int_{\mathbb{R}} \cos(tg(x)) dP_{\tilde{x}}(x), \int_{\mathbb{R}} \sin(tg(x)) dP_{\tilde{x}}(x) \right),$$

$$M_{\tilde{y}}(t) = \mathrm{E}(e^{t\tilde{y}}) = \mathrm{E}(e^{tg(\tilde{x})}) = \int_{\mathbb{R}} e^{tg(x)} dP_{\tilde{x}}(x),$$

$$\Lambda_{\tilde{y}}(t) = \mathrm{E}(e^{-t\tilde{y}}) = \mathrm{E}(e^{-tg(\tilde{x})}) = \int_{\mathbb{R}} e^{-tg(x)} dP_{\tilde{x}}(x),$$

and try to identify the distribution $P_{\tilde{y}}$ associated with $\varphi_{\tilde{y}}$ (or with $M_{\tilde{y}}$ or with $\Lambda_{\tilde{y}}$).

- This is called **the characteristic function (or moment-generating function or Laplace transform) method.**

5.7. Mixture distributions

- Consider a discrete set of random variables $\{\tilde{x}_j\}_{j=1}^{N}$, with $\tilde{x}_j : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ for $j = 1, ..., N$, where N can be either a finite number or infinity.
- **Definition.** The random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ is a (discrete) mixture of the random variables $\{\tilde{x}_j\}_{j=1}^{N}$, if $\tilde{x} = \tilde{x}_j$ with probability p_j.
- Obviously, $p_j \in [0, 1]$ and $\sum_j p_j = 1$.
- The distribution of $\tilde{x}$ is called the **mixture distribution** and the random variables $\{\tilde{x}_j\}_{j=1}^{N}$ in the mixture are called the **components** of the mixture.
- **Notation for a mixture:**

$$\tilde{x} = (p_1 \odot \tilde{x}_1) \oplus (p_2 \odot \tilde{x}_2) \oplus (p_3 \odot \tilde{x}_3) \oplus ...$$

- **Proposition.** Assume that the component $\tilde{x}_j$ of the mixture $\tilde{x}$ has the distribution $P_j : \mathcal{B} \longrightarrow [0,1]$ for $j = 1, ..., N$. Then, the mixture distribution $P_{\tilde{x}}$ is

$$P_{\tilde{x}}(B) = \sum_j p_j P_j(B), \quad \text{for all } B \in \mathcal{B}.$$

Proof.

$$P_{\tilde{x}}(B) = \sum_j p_j P\left\{\tilde{x} \in B \,|\tilde{x} = \tilde{x}_j\right\}$$

$$= \sum_j p_j P\left\{\tilde{x}_j \in B\right\} = \sum_j p_j P_j(B), \quad \text{for all } B \in \mathcal{B},$$

where the first equality comes from the theorem of total probability. $Q.E.D.$

- **Corollary.** Assume that the component $\tilde{x}_j$ of the mixture $\tilde{x}$ has the distribution function $F_j : \mathbb{R} \longrightarrow [0,1]$ for $j = 1, ..., N$. Then, the distribution function $F_{\tilde{x}}$ of the mixture distribution is

$$F_{\tilde{x}}(x) = \sum_j p_j F_j(x), \quad \text{for } x \in \mathbb{R}.$$

- **Proof.** Using the previous proposition we get

$$F_{\tilde{x}}(x) = P\{\tilde{x} \leq x\} = P_{\tilde{x}}(-\infty, x] = \sum_j p_j P_j(-\infty, x]$$

$$= \sum_j p_j P\{\tilde{x}_j \leq x\} = \sum_j p_j F_j(x), \quad \text{for } x \in \mathbb{R}. \qquad Q.E.D.$$

- **Corollary.** Assume that the component $\tilde{x}_j$ of the mixture $\tilde{x}$ is discrete with the probability function $f_j : \tilde{x}_j(\Omega) \longrightarrow [0,1]$ for $j = 1, ..., N$. Then, the mixture distribution is discrete with the probability function

$$f_{\tilde{x}}(x) = \sum_j p_j f_j(x), \quad \text{for } x \in \tilde{x}(\Omega) \equiv \bigcup_j \tilde{x}_j(\Omega),$$

where we make $f_j(x) = 0$ if $x \notin \tilde{x}_j(\Omega)$.

- **Proof.** Using again the proposition we get

$$f_{\tilde{x}}(x) = P\{\tilde{x} = x\} = \sum_j p_j P\{\tilde{x}_j = x\} = \sum_j p_j f_j(x), \quad \text{for } x \in \tilde{x}(\Omega).$$

$Q.E.D.$

- **Proposition.** Assume that the component $\tilde{x}_j$ of the mixture $\tilde{x}$ is absolutely continuous with the density $f_j : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$ for $j = 1, ..., N$. Then, the mixture distribution is absolutely continuous with the density

$$f_{\tilde{x}}(x) = \sum_j p_j f_j(x), \quad \text{for } x \in \mathbb{R}.$$

- **Proof.** Since the distribution function of the mixture satisfies $F_{\tilde{x}}(x) = \sum_j p_j F_j(x)$, we can take the derivative with respect to x to obtain the density of the mixture $\tilde{x}$,

$$f_{\tilde{x}}(x) = \frac{dF_{\tilde{x}}(x)}{dx} = \frac{d\left[\sum_j p_j F_j(x)\right]}{dx}$$

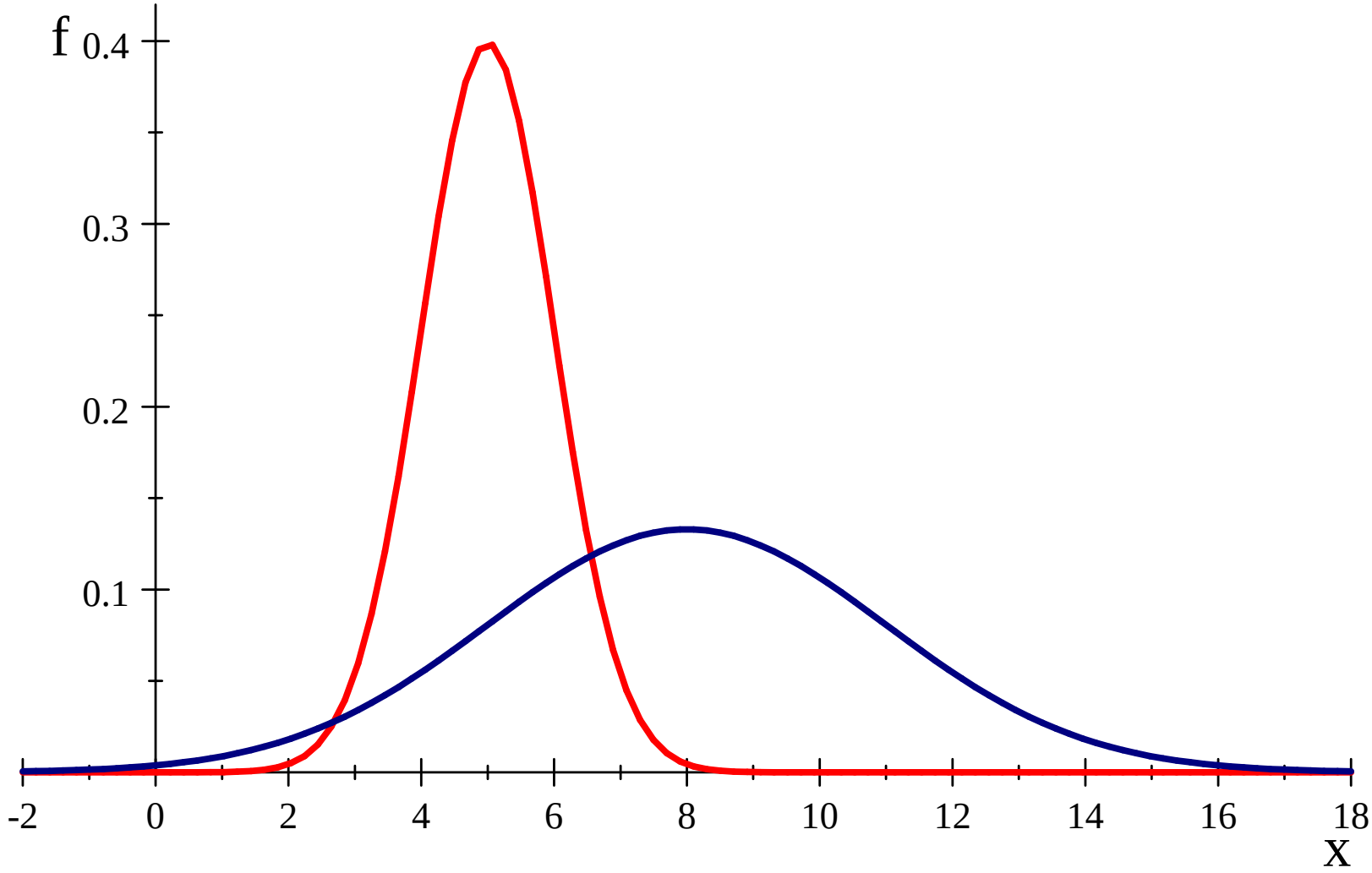
$$= \sum_j p_j \frac{dF_j(x)}{dx} = \sum_j p_j f_j(x), \quad \text{for } x \in \mathbb{R}. \qquad Q.E.D.$$

- Note that the distribution of a mixture of random variables is not the same as the distribution of their weighted sum.

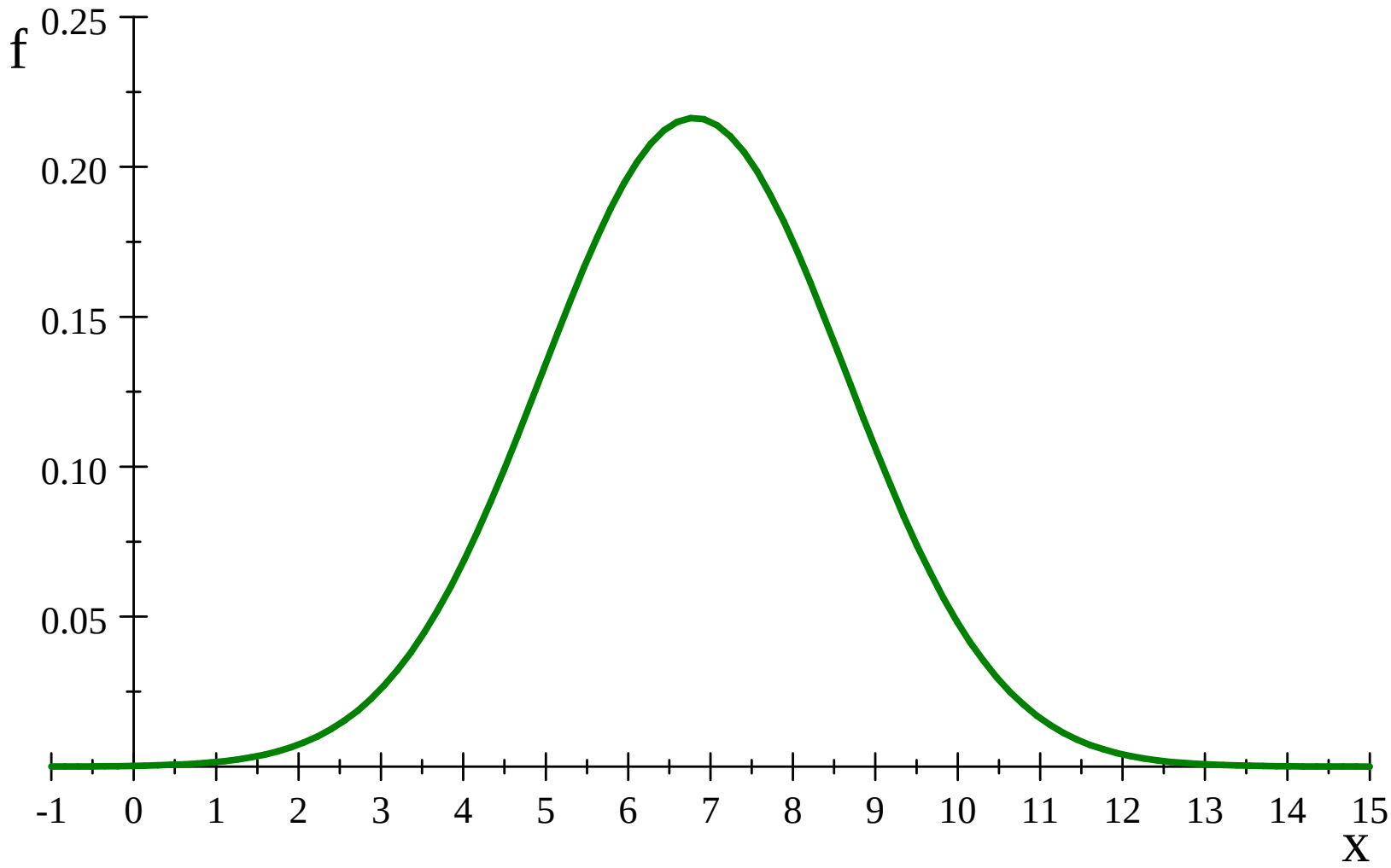
- **Example:** Consider the two independent random variables $\tilde{x}_1 \sim \mathrm{N}\left(\mu_1, \sigma_1^2\right)$ and $\tilde{x}_2 \sim \mathrm{N}\left(\mu_2, \sigma_2^2\right)$. Then, $p_1\tilde{x}_1 + p_2\tilde{x}_2 \sim \mathrm{N}\left(p_1\mu_1 + p_2\mu_2, p_1^2\sigma_1^2 + p_2^2\sigma_2^2\right)$, while $\tilde{x} = (p_1 \odot \tilde{x}_1) \oplus (p_2 \odot \tilde{x}_2)$ is not normal.

- $\mu_1 = 5$, $\sigma_1 = 1$, $\mu_2 = 8$, $\sigma_2 = 3$, $p_1 = 0.4$, $p_2 = 0.6$.

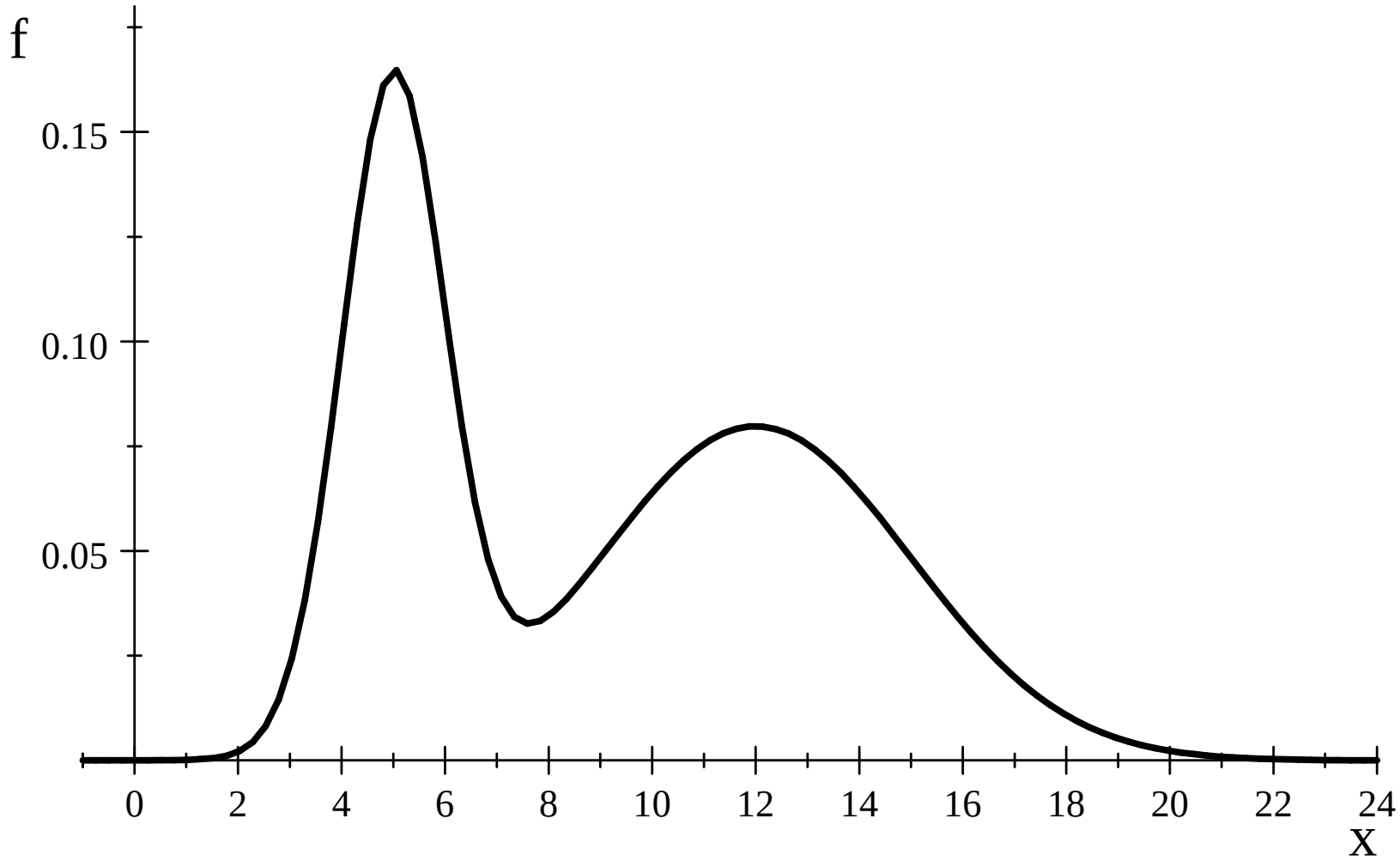
- Densities of $\tilde{x}_1$ and $\tilde{x}_2$, $n(x; 5, 1)$ and $n(x; 8, 3)$, respectively:



- Density of the weighted sum $p_1\tilde{x}_1 + p_2\tilde{x}_2$:
  $n\left(x; p_1\mu_1 + p_2\mu_2, \sqrt{p_1^2\sigma_1^2 + p_2^2\sigma_2^2}\right) = n(x; 6.8, 1.8439)$:



- Density of the mixture $\tilde{x} = (p_1 \odot \tilde{x}_1) \oplus (p_2 \odot \tilde{x}_2)$:
  $0.6n(x; 5, 1) + 0.4n(x; 8, 3)$:



- **Proposition.** Assume that the component $\tilde{x}_j$ of the mixture $\tilde{x}$ has the finite mean μ_j (i.e., $\tilde{x}_j \in \mathrm{L}^1$) for $j = 1, ..., N$. Then, the mean μ of the mixture distribution is

$$\mu = \sum_j p_j \mu_j.$$

- **Proof.**

$$\mu = \mathrm{E}(\tilde{x}) = \sum_j p_j \mathrm{E}(\tilde{x} \,|\tilde{x} = \tilde{x}_j) = \sum_j p_j \mathrm{E}(\tilde{x}_j) = \sum_j p_j \mu_j,$$

where the second equality comes from the theorem of total expectation. $Q.E.D.$

- **Proposition.** Assume that the component $\tilde{x}_j$ of the mixture $\tilde{x}$ has the finite mean μ_j and the finite variance σ_j^2 (i.e., $\tilde{x}_j \in \mathrm{L}^2$) for $j = 1, ..., N$, and let $\mu = \sum_j p_j \mu_j$ be the mean of the mixture $\tilde{x}$. Then, the variance σ^2 of the mixture distribution is

$$\sigma^2 = \sum_j p_j \left(\sigma_j^2 + \mu_j^2 - \mu^2 \right) = \sum_j p_j \left(\sigma_j^2 + \mu_j^2 \right) - \mu^2. \qquad \text{(Exercise)}$$

- **Proposition.** Assume that the component $\tilde{x}_j$ of the mixture $\tilde{x}$ has a finite kth central moment (i.e., $\tilde{x}_j \in \mathrm{L}^k$) and the finite mean μ_j for $j = 1, ..., N$, and let $\mu = \sum_j p_j \mu_j$ be the mean of the mixture $\tilde{x}$. Then, the kth central moment of the mixture is given by

$$\mathrm{E}\left[(\tilde{x} - \mu)^k \right] =$$

$$\sum_j p_j \left(\sum_{n=0}^{k} \binom{k}{n} \left(\mu_j - \mu \right)^{k-n} \mathrm{E}\left[\left(\tilde{x}_j - \mu_j \right)^n \right] \right). \qquad \text{(Exercise)}$$

- **General mixtures (not necessarily discrete)**

- Consider a set of random variables. A generic element of this set is $\tilde{x}_\theta$, where $\theta \in \Theta \subset \mathbb{R}$. The index set Θ can be either discrete or continuous. The distribution of the random variable $\tilde{x}_\theta$ is $P_\theta : \mathcal{B} \longrightarrow [0,1]$, for $\theta \in \Theta$.

- Consider also a probability P on $(\mathbb{R}, \mathcal{B})$, where $\Theta \in \mathcal{B}$ with $P(\Theta) = 1$.

- In the mixture $\tilde{x}$, the random variables $\tilde{x}_\theta$, with $\theta \in \Theta$, are "mixed" according to the probability P.

- Thus, the **mixture distribution** $P_{\tilde{x}}$ is

$$P_{\tilde{x}}(B) = \int_\Theta P_\theta(B) dP(\theta) = \int_{\mathbb{R}} P_\theta(B) dP(\theta), \quad \text{for all } B \in \mathcal{B}.$$

- The previous two integrals are equal since the probability P is concentrated on Θ.

- If the random variables $\tilde{x}_\theta$, for $\theta \in \Theta$, are discrete (absolutely continuous) with probability function (density) f_θ, then the probability function (density) of the mixture distribution is

$$f_{\tilde{x}}(x) = \int_\Theta f_\theta(x) dP(\theta) = \int_{\mathbb{R}} f_\theta(x) dP(\theta).$$

- Moreover, if $\mu_\theta = \mathrm{E}(\tilde{x}_\theta)$ and $\sigma_\theta^2 = \mathrm{Var}(\tilde{x}_\theta)$, for $\theta \in \Theta$, then the mean of the mixture $\tilde{x}$ is

$$\mu = \mathrm{E}(\tilde{x}) = \int_\Theta \mu_\theta dP(\theta)$$

and the variance of the mixture $\tilde{x}$ is

$$\sigma^2 = \mathrm{Var}(\tilde{x}) = \int_\Theta \left(\sigma_\theta^2 + \mu_\theta^2 - \mu^2\right) dP(\theta) = \int_\Theta \left(\sigma_\theta^2 + \mu_\theta^2\right) dP(\theta) - \mu^2.$$

- **Particular case.** Assume now that the probability P on $(\mathbb{R}, \mathcal{B})$ is absolutely continuous with respect to Lebesgue measure with the density $f : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$ so that

$$P(C) = \int_C f(\theta) d\theta, \quad \text{for all } C \in \mathcal{B}.$$

- Assume also that each component $\tilde{x}_\theta$ of the mixture $\tilde{x}$ is also absolutely continuous with the density $f_\theta : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$, for $\theta \in \Theta \subset \mathbb{R}$.
- Then, the mixture distribution is absolutely continuous with the density

$$f_{\tilde{x}}(x) = \int_\Theta f_\theta(x) dP(\theta) = \int_{\mathbb{R}} f_\theta(x) f(\theta) d\theta, \quad \text{for } x \in \mathbb{R},$$

with the mean

$$\mu = \mathrm{E}(\tilde{x}) = \int_{\mathbb{R}} \mu_\theta f(\theta) d\theta$$

and the variance

$$\sigma^2 = \mathrm{Var}(\tilde{x}) = \int_{\mathbb{R}} \left(\sigma_\theta^2 + \mu_\theta^2 - \mu^2\right) f(\theta) d\theta = \int_{\mathbb{R}} \left(\sigma_\theta^2 + \mu_\theta^2\right) f(\theta) d\theta - \mu^2.$$

Example of Convolution

Consider two independent, absolutely continuous random variables $\tilde{x}_1$ and $\tilde{x}_2$ having both the uniform density on $(0, \theta)$,

$$\tilde{x}_i \sim f_i(x_i) = f(x_i) = \begin{cases} \dfrac{1}{\theta} & \text{for} \quad x_i \in (0, \theta) \\ \\ 0 & \text{elsewhere,} \end{cases}$$

for $i = 1, 2$. Let us find the density $f_{\tilde{x}_1 + \tilde{x}_2}$ of their sum, $\tilde{x}_1 + \tilde{x}_2$. This density will characterize the distribution of $\tilde{x}_1 + \tilde{x}_2$ (i.e., the convolution of $\tilde{x}_1$ and $\tilde{x}_2$).

Let us define the random variable $\tilde{y} = \tilde{x}_1 + \tilde{x}_2$. The density $f_{\tilde{y}}$ of $\tilde{y}$ is the following:

$$f_{\tilde{y}}(y) = f_{\tilde{x}_1 + \tilde{x}_2}(y) = \int_{\mathbb{R}} f_{\tilde{x}_2}(y - x_1) f_{\tilde{x}_1}(x_1) dx_1 = \int_{\mathbb{R}} f(y - x_1) f(x_1) dx_1.$$

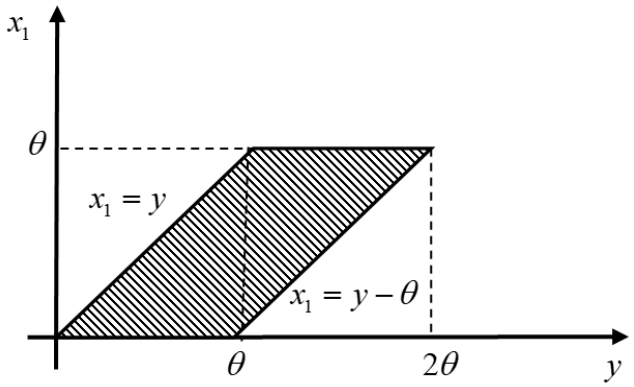
Note that $f(y - x_1)$ is equal to $1/\theta$ if $y - x_1 \in (0, \theta)$ and is equal to zero otherwise. Moreover, $f(x_1)$ is equal to $1/\theta$ if $x_1 \in (0, \theta)$ and is equal to zero otherwise. Therefore, $f(y - x_1) f(x_1)$ is equal to $(1/\theta)^2$ if $y - x_1 \in (0, \theta)$ and $x_1 \in (0, \theta)$ while it is equal to zero otherwise. Note that the set $C \subset \mathbb{R}^2$ where $f(y - x_1) f(x_1) = (1/\theta)^2 \neq 0$ is given by the points (y, x_1) satisfying the following two inequalities:

$$x_1 < y < x_1 + \theta \Longleftrightarrow y - \theta < x_1 < y$$

and

$$0 < x_1 < \theta$$

This set C is given by the interior of the shaded region in the following figure:



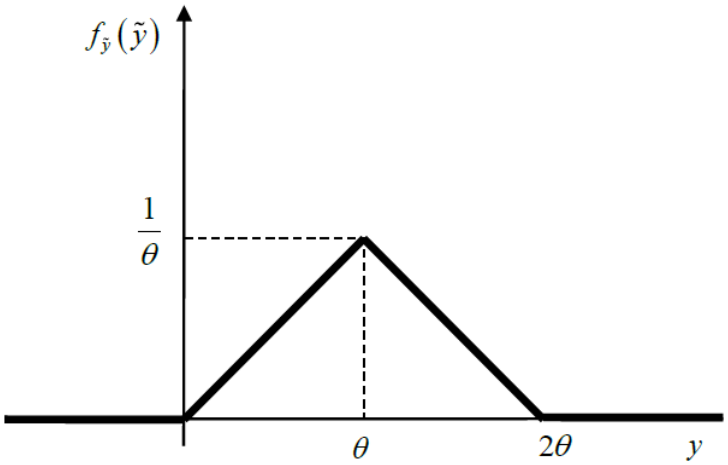
Therefore, the density of the convolution for $y \in (0, 2\theta)$ is

$$f_{\tilde{x}_1 + \tilde{x}_2}(y) = f_{\tilde{y}}(y) = \int_{\mathbb{R}} \mathbb{I}_C(y, x_1) f(y - x_1) f(x_1) dx_1 = \int_{C(y)} f(y - x_1) f(x_1) dx_1,$$

where $C(y)$ is the section of y, $C(y) = \{x_1 \in \mathbb{R} \,|(x_1, y) \in C\}$. Thus,

$$f_{\tilde{x}_1+\tilde{x}_2}(y) = f_{\tilde{y}}(y) = \begin{cases} \int_0^y \frac{1}{\theta^2} dx_1 = \frac{1}{\theta^2} y, & \text{for} \quad 0 < y \leq \theta \\ \int_{y-\theta}^{\theta} \frac{1}{\theta^2} dx_1 = -\frac{1}{\theta^2} y + \frac{2}{\theta}, & \text{for} \quad \theta < y < 2\theta, \\ 0 & \text{elsewhere,} \end{cases}$$

whose graph is



This type of densities (and their corresponding distributions) are called "triangular" for obvious reasons.

Examples of densities of functions of random vectors

Example 1.

Assume that the joint density function of the random variables $\tilde{x}_1$ and $\tilde{x}_2$ is

$$f_{\tilde{x}_1,\tilde{x}_2}(x_1,x_2) = \begin{cases} e^{-(x_1+x_2)} & \text{for } x_1 > 0,\ x_2 > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the density function of the random variable $\tilde{y} = \dfrac{\tilde{x}_1}{\tilde{x}_1 + \tilde{x}_2}$.

Solution: For each given value of x_1, $\dfrac{x_1}{x_1+x_2}$ is strictly decreasing in x_2. Therefore, we can define the bijective function $g : \mathbb{R}_{++} \times \mathbb{R}_{++} \longrightarrow (0,1) \times \mathbb{R}_{++}$ as

$$(y,z) = g(x_1,x_2) : \begin{cases} y = \dfrac{x_1}{x_1+x_2} \in (0,1) \\ \\ z = x_1 > 0, \end{cases}$$

and thus $g^{-1} : (0,1) \times \mathbb{R}_{++} \longrightarrow \mathbb{R}_{++} \times \mathbb{R}_{++}$ is given by

$$(x_1,x_2) = g^{-1}(y,z) : \begin{cases} x_1 = z > 0 \\ \\ x_2 = \dfrac{z(1-y)}{y} > 0. \end{cases}$$

The joint density of $\tilde{y}$ and $\tilde{z}$ is given by

$$f_{\tilde{y},\tilde{z}}(y,z) = \begin{cases} f_{\tilde{x}_1,\tilde{x}_2}\left(g^{-1}(y,z)\right) \left|J_{g^{-1}}(y,z)\right| & \text{for } y \in (0,1),\ z > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where,

$$|J_{g^{-1}}| = \left| \det \begin{pmatrix} \dfrac{\partial x_1}{\partial y} & \dfrac{\partial x_1}{\partial z} \\ \\ \dfrac{\partial x_2}{\partial y} & \dfrac{\partial x_2}{\partial z} \end{pmatrix} \right| = \left| \det \begin{pmatrix} 0 & 1 \\ \\ -\dfrac{z}{y^2} & \dfrac{1-y}{y} \end{pmatrix} \right| = \left| \frac{z}{y^2} \right| = \frac{z}{y^2} > 0.$$

Therefore, since $x_1 + x_2 = z + \dfrac{z(1-y)}{y} = \dfrac{z}{y}$, we get

$$f_{\tilde{y},\tilde{z}}(y,z) = \begin{cases} e^{-z/y}\dfrac{z}{y^2} & \text{for } y \in (0,1) \text{ and } z > 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

The set A where $f_{\tilde{x}_1,\tilde{x}_2}(x_1,x_2) \neq 0$ is defined by the points $(x_1,x_2) \in \mathbb{R}$ satisfying the following two inequalities:

$$x_1 > 0 \text{ and } x_2 > 0.$$

Thus, the set $g(A)$ where $f_{\tilde{y},\tilde{z}}(y,z) \neq 0$ is defined by the points $(y,z) \in \mathbb{R}$ satisfying the following two inequalities:

$$0 < y < 1 \text{ and } z > 0.$$

The marginal density of $\tilde{y}$ is

$$f_{\tilde{y}}(y) = \int_{-\infty}^{\infty} f_{\tilde{y},\tilde{z}}(y,z)dz = \int_0^{\infty} e^{-z/y}\,\frac{z}{y^2}\,dz, \quad \text{if } y \in (0,1)$$

and $f_{\tilde{y}}(y) = 0$, otherwise.

Let us make a change of variable: $u = \dfrac{z}{y} \Leftrightarrow z = yu \Rightarrow \dfrac{dz}{du} = y \in (0,1)$.

$$\Longrightarrow f_{\tilde{y}}(y) = \int_0^{\infty} e^{-u}\,\frac{u}{y}y\,du = \int_0^{\infty} e^{-u}\,u\,du = \Gamma(2) = (2-1)! = 1 \text{ if } y \in (0,1).$$

Alternatively, note that we can integrate by parts,

$$\begin{aligned}\int_0^{\infty} e^{-u}\,u\,du &= \left[-e^{-u}u\right]_0^{\infty} - \int_0^{\infty} -e^{-u}\,du = 0 + \int_0^{\infty} e^{-u}\,du \\ &= \left[-e^{-u}\right]_0^{\infty} = 0 - (-1) = 1.\end{aligned}$$

$$\Longrightarrow f_{\tilde{y}}(y) = \begin{cases} 1 & \text{for } y \in (0,1) \\ \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\tilde{y}$ is uniform on $(0,1)$.

Note that, in this example, we can name the new variable $\tilde{z}$ with the old name $\tilde{x}_1$ so that the function g would be

$$(y,x_1) = g(x_1,x_2) : \begin{cases} y = \dfrac{x_1}{x_1+x_2} \in (0,1) \\ \\ x_1 = x_1 > 0. \end{cases}$$

Example 2.

Assume that the two independent random variables $\tilde{x}_1$ and $\tilde{x}_2$ have both the uniform density on $(0,\theta)$,

$$\tilde{x}_i \sim f_i(x_i) = f(x_i) = \begin{cases} \dfrac{1}{\theta} & \text{for} \quad x_i \in (0,\theta) \\ \\ 0 & \text{elsewhere,} \end{cases}$$

for $i = 1, 2$. Find the density function of the random variable $\tilde{y} = \tilde{x}_1 + \tilde{x}_2$. Note that this is the same example as in the previous handout on convolutions.

Solution: From independence, we obtain the joint density of $\tilde{x}_1$ and $\tilde{x}_2$,

$$f_{\tilde{x}_1,\tilde{x}_2}(x_1, x_2) = f(x_1) \cdot f(x_2) = \begin{cases} \dfrac{1}{\theta^2} & \text{for} \quad x_1 \in (0,\theta) \text{ and } x_2 \in (0,\theta) \\ \\ 0 & \text{elsewhere,} \end{cases}$$

We have

$$g : \begin{cases} y = x_1 + x_2 \in (x_1, x_1 + \theta) \\ \\ z = x_1 \in (0,\theta) \end{cases} \Longrightarrow g^{-1} : \begin{cases} x_1 = z \in (0,\theta) \\ \\ x_2 = y - z \in (0,\theta) \end{cases}$$

Note that $y - z \in (0,\theta) \Longleftrightarrow y \in (z, z+\theta) \Longleftrightarrow z \in (y-\theta, y)$.

$$|J_{g^{-1}}(y,z)| = \left| \det \begin{pmatrix} \dfrac{\partial x_1}{\partial z} & \dfrac{\partial x_1}{\partial y} \\ \\ \dfrac{\partial x_2}{\partial z} & \dfrac{\partial x_2}{\partial y} \end{pmatrix} \right| = \left| \det \begin{pmatrix} 1 & 0 \\ \\ -1 & 1 \end{pmatrix} \right| = 1.$$

Then,

$$f_{\tilde{y},\tilde{z}}(y,z) = \frac{1}{\theta^2} \cdot 1 = \frac{1}{\theta^2}, \text{ for } 0 < z < \theta \text{ and } y - \theta < z < y.$$

Therefore,

$$f_{\tilde{y},\tilde{z}}(y,z) = \begin{cases} \dfrac{1}{\theta^2} & \text{for} \quad 0 < z < \theta \text{ and } y - \theta < z < y \\ \\ 0 & \text{elsewhere} \end{cases}$$

The set A where $f_{\tilde{x}_1,\tilde{x}_2} \neq 0$ is defined by the points $(x_1, x_2) \in \mathbb{R}$ satisfying the following two inequalities:

$$0 < x_1 < \theta \quad \text{and} \quad 0 < x_2 < \theta.$$

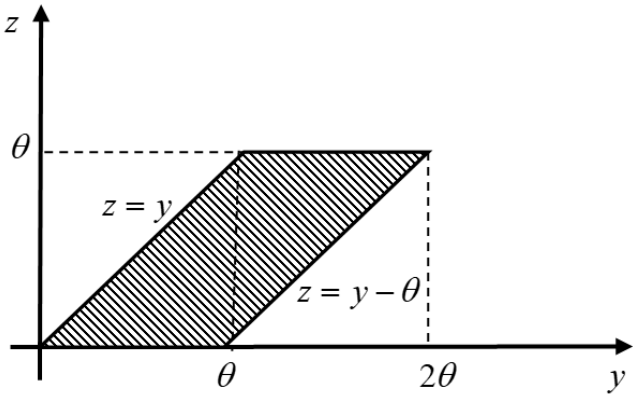
Thus, the set $g(A)$ where $f_{\tilde{y},\tilde{z}} \neq 0$ is defined by the points $(y, z) \in \mathbb{R}$ satisfying the following two inequalities:

$$0 < z < \theta \quad \text{and} \quad z < y < z + \theta$$

or, equivalently, the following two inequalities:

$$0 < z < \theta \quad \text{and} \quad y - \theta < z < y.$$

Therefore, $g(A) = C$, where the set C is given by the interior of the shaded region in the following figure:



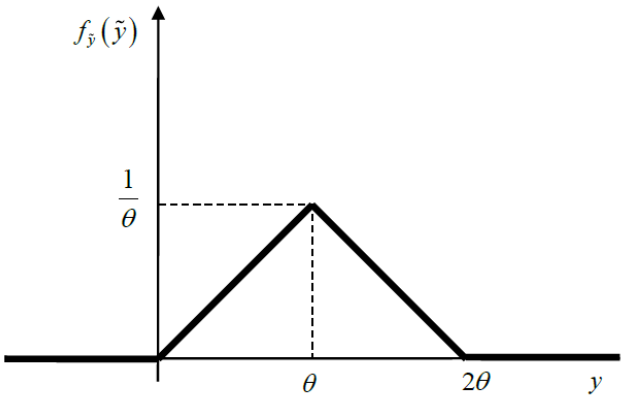
Then, the density of y for $y \in (0, 2\theta)$ is

$$f_{\tilde{y}}(y) = \int_{-\infty}^{\infty} f_{\tilde{y},\tilde{z}}(y,z)\,dz = \int_{\mathbb{R}} \mathbb{I}_C(y,z)\,\underbrace{f_{\tilde{y},\tilde{z}}(y,z)}_{1/\theta^2} dz = \int_{C(y)} f_{\tilde{y},\tilde{z}}(y,z)\,dz,$$

where $C(y)$ is the section of y, $C(y) = \{z \in \mathbb{R} \,|(y,z) \in C\}$. Therefore

$$f_{\tilde{x}_1+\tilde{x}_2}(y)\,(y) = \begin{cases} \int_0^y \frac{1}{\theta^2} dz = \frac{1}{\theta^2} y, & \text{for} \quad 0 < y \leq \theta \\ \int_{y-\theta}^{\theta} \frac{1}{\theta^2} dz = -\frac{1}{\theta^2} y + \frac{2}{\theta}, & \text{for} \quad \theta < y < 2\theta, \\ 0 & \text{elsewhere,} \end{cases}$$

so that the density is "triangular",



Again, in this example, we can name the new variable $\tilde{z}$ with the old name $\tilde{x}_1$ so that the function g would be

$$(y, x_1) = g(x_1, x_2) : \begin{cases} y = x_1 + x_2 \in (x_1, x_1 + \theta) \\ x_1 = x_1 \in (0, \theta). \end{cases}$$

Exercises. Probability and Statistics. IDEA.
5. Functions of Random Variables

1. A random variable has the **log-normal distribution** if its logarithm has a normal distribution.

   Let $\tilde{x} = \ln \tilde{y}$ have a normal distribution with the mean μ and the variance σ^2.

   (a) find the density of the strictly positive log-normal random variable $\tilde{y}$**.**

   (b) find the mean m, the variance s^2, and the coefficient of variation $\mathrm{CV}_{\tilde{y}}$ of $\tilde{y}$ as functions of μ and σ^2.

   Let $\tilde{y}$ be a log-normally distributed random variable with the mean m and the variance s^2.

   (c) find the mean μ and the variance σ^2 of the normal random variable $\tilde{x} = \ln \tilde{y}$ as functions of m and s^2.

2. If $\tilde{x}$ has a uniform density on the interval $(0, 1)$, show that the random variable $\tilde{y} = -2 \cdot \ln \tilde{x}$ has a gamma distribution. What are its parameters?

3. If the joint probability function of $\tilde{x}_1$ and $\tilde{x}_2$ is given by $f(x_1, x_2) = \dfrac{x_1 x_2}{36}$ for $x_1 = 1, 2, 3$, and $x_2 = 1, 2, 3$, find

   (a) the probability function of $\tilde{y} = \tilde{x}_1 \tilde{x}_2$;

   (b) the probability function of $\tilde{z} = \dfrac{\tilde{x}_1}{\tilde{x}_2}$.

4. If n independent random variables have the same gamma distribution with the parameters α and β, find the moment-generating function of their sum, and, if possible, identify its distribution. What is the distribution of the sum of n independent random variables having the same exponential distribution with the parameter θ?

5. If $\tilde{\theta}$ is the angle that a "random spinner" makes with the positive side of the horizontal axis, what is the density function of $\tan \tilde{\theta}$? In other words, what is the density of the "random tangent"?

   Hint: If $\psi(x) = \arctan(x)$, where the function $\arctan(\cdot) : \mathbb{R} \longrightarrow (-\pi/2, \pi/2)$ is the inverse function of $\tan(\cdot) : (-\pi/2, \pi/2) \longrightarrow \mathbb{R}$, then the derivative of ψ is $\psi'(x) = \dfrac{1}{1 + x^2}$. Recall also that $\tan\left(\dfrac{\pi}{4}\right) = 1$, $\tan\left(-\dfrac{\pi}{4}\right) = -1$, $\lim\limits_{\theta \to \pi/2} \tan(\theta) = \infty$, $\lim\limits_{\theta \to -\pi/2} \tan(\theta) = -\infty$ and $\tan(0) = 0$.

6. If the joint density of the random variables $\tilde{x}_1$ and $\tilde{x}_2$ is

$$f(x_1, x_2) = \begin{cases} 1 & \text{for } 0 < x_1 < 1 \text{ and } 0 < x_2 < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

find the joint density of the random variables $\tilde{u}_1$ and $\tilde{u}_2$, where

$$\tilde{u}_1 = (-2\ln \tilde{x}_1)^{\frac{1}{2}} \sin(2\pi \cdot \tilde{x}_2)$$

$$\tilde{u}_2 = (-2\ln \tilde{x}_1)^{\frac{1}{2}} \cos(2\pi \cdot \tilde{x}_2).$$

Note: You should use the hint of the previous exercise.

7. If the joint density of $\tilde{x}_1$ and $\tilde{x}_2$ is given by

$$f(x_1, x_2) = \begin{cases} 1 & \text{for } 0 < x_1 < 1 \text{ and } 0 < x_2 < 1 \\ \\ 0 & \text{elsewhere} \end{cases}$$

find the density $h(y)$ of $\tilde{y} = \tilde{x}_1 + \tilde{x}_2$. Sketch the density $h(y)$.

8. Consider the two-dimensional vector $(\tilde{x}, \tilde{y})$, which has coordinates that are independent and bivariate normally distributed with zero mean and variance σ^2.

(a) Prove that the variable "modulus" $\tilde{r} = (\tilde{x}^2 + \tilde{y}^2)^{1/2}$ has the **Rayleigh** distribution with the scale parameter σ, whose density is given by

$$f_{\tilde{r}}(r) = \begin{cases} \dfrac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} & \text{for } r \in (0, \infty) \\ \\ 0 & \text{otherwise,} \end{cases}$$

and find the density function of the random variable "angle" $\tilde{\theta} = \arctan(\tilde{y}/\tilde{x})$ with $0 \leq \tilde{\theta} < 2\pi$.

(b) Are $\tilde{r}$ and $\tilde{\theta}$ independent?

(c) Prove that the square of the modulus $\tilde{r}^2 = \tilde{x}^2 + \tilde{y}^2$ has an exponential distribution with parameter $2\sigma^2$ and, if $\sigma^2 = 1$, then $\tilde{r}^2$ has the chi-square distribution with 2 degrees of freedom χ^2_2.

(d) Find the expected value and the variance of the modulus $\tilde{r}$.

(e) Find the expected value and the variance of the angle $\tilde{\theta}$.

9. The joint density function of the two random variables $\tilde{x}$ and $\tilde{y}$ is

$$f(x, y) = \begin{cases} \dfrac{1}{12}(x + 2y) & \text{if } \; x \in (0, 2), \; y \in (0, 2) \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the joint distribution function of $\tilde{x}$ and $\tilde{y}$, $F(x, y)$.

(b) Find the marginal density functions of $\tilde{x}$ and $\tilde{y}$, $f_{\tilde{x}}(x)$ and $f_{\tilde{y}}(y)$.

(c) Find the conditional density function of $\tilde{y}$ given $\tilde{x} = 1/2$, $f_{\tilde{y}|\tilde{x}}(y|\, 1/2)$.

(d) Find the expectation of $\tilde{y}$, $\mathrm{E}(\tilde{y})$, and the conditional expectation of $\tilde{y}$ given $\tilde{x} = 1/2$, $\mathrm{E}(\tilde{y}|\,\tilde{x} = 1/2)$.

(e) Find $P\left(\{0 < \tilde{x} < 1\} \cap \{1 < \tilde{y} < 3/2\}\right)$.

(f) Could you find a value x for which the conditional density of $\tilde{y}$ given $\tilde{x} = x$ is equal to the marginal density of $\tilde{y}$? Are $\tilde{x}$ and $\tilde{y}$ independent?

(g) Find the density function of the random variable $\tilde{w} = 2 - \tilde{y}^{1/3}$ and the expectation of $\tilde{w}$, $\mathrm{E}(\tilde{w})$.

(h) Find the density function of the random variable $\tilde{z} = \tilde{x} \cdot \tilde{y}$.

10. The density function of an absolutely continuous random variable $\tilde{x}$ is

$$f_{\tilde{x}}(x) = \frac{1}{2}e^{-|x|}\ , \quad -\infty < x < \infty$$

Find the distribution function of $\tilde{x}$ and the density function of

$$\tilde{y} = (\tilde{x})^3 \qquad \text{and} \qquad \tilde{z} = \ln\left(|\tilde{x}| + 1\right).$$

11. Let $\tilde{x}$ be an absolutely continuous random variable with the density function

$$f(x) = \begin{cases} \dfrac{1}{3} & \text{if } x \in (0,3) \\ \\ 0 & \text{otherwise.} \end{cases}$$

Consider the transformation $\tilde{y} = 9 - (\tilde{x})^2$. Find and draw the distribution and the density functions of $\tilde{y}$.

12. Let $F_{\tilde{x}}(x)$ be the distribution function of the univariate random variable $\tilde{x}$.

(a) Assume that $\tilde{x}$ is a mixed random variable (that is, $\tilde{x}$ is neither discrete nor absolutely continuous). Find the distribution function of the absolute value of $\tilde{x}$, that is, find the distribution function $F_{\tilde{y}}(y)$ of the random variable $\tilde{y} = |\tilde{x}|$.

(b) Answer the same question of part (a) but assuming now that the distribution function $F_{\tilde{x}}(x)$ is continuous for all $x \in (-\infty, \infty)$.

(c) Assume now that the random variable $\tilde{x}$ is absolutely continuous with the density function $f_{\tilde{x}}(x)$. Moreover, assume that the distribution function $F_{\tilde{x}}(x)$ of the variable $\tilde{x}$ is differentiable for all $x \in (-\infty, \infty)$. Using the density function $f_{\tilde{x}}(x)$ of the variable $\tilde{x}$, find the density function $f_{\tilde{y}}(y)$ of the random variable $\tilde{y} = |\tilde{x}|$.

(d) Answer the same question of part (c) but assuming now that the density function $f_{\tilde{x}}(x)$ is also symmetric with respect to $x = 0$, that is, $f_{\tilde{x}}(x) = f_{\tilde{x}}(-x)$ for all $x \in (-\infty, \infty)$.

13. The density function of the random variable $\tilde{x}$ is

$$f_{\tilde{x}}(x) = \begin{cases} \dfrac{x}{8} & \text{if } x \in (0,4) \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the density function $f_{\tilde{y}}(y)$ and the distribution function $F_{\tilde{y}}(y)$ of the random variable $\tilde{y} = \tilde{x}^2 - 8$.

(b) Compute $P\{1 < \tilde{y} < 3\}$ and $P\{\tilde{y} \geq -2\}$.

14. The absolutely continuous random variable $\tilde{x}$ has the following density function:

$$f_{\tilde{x}}(x) = \begin{cases} 6x(1-x) & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the mean and the variance of $\tilde{y} = (\tilde{x})^3$.

(b) Find the density function and the distribution function of $\tilde{y}$.

15. The density function of the random vector $(\tilde{x}, \tilde{y})$ is

$$f(x,y) = \begin{cases} k\,(x+y) & \text{for } x > 0,\ x^2 + y^2 < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of the constant k.

(b) Find the density function of the random variable $\tilde{z} = (\tilde{x}^2 + \tilde{y}^2)^{1/2}$.

16. The density function of the random vector $(\tilde{x}_1, \tilde{x}_2)$ is

$$f_{\tilde{x}_1,\tilde{x}_2}(x_1,x_2) = \begin{cases} 6e^{-3x_1-2x_2} & \text{for } x_1 > 0,\ x_2 > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the conditional expectation of $\tilde{x}_2$ given $\tilde{x}_1 = 5$, $\mathrm{E}(\tilde{x}_2 \mid \tilde{x}_1 = 5)$. Are the random variables $\tilde{x}_1$ and $\tilde{x}_2$ independent?

(b) Find the moment-generating function for the random variable $\tilde{x}_1$ and use it to find the mean, the variance, and the coefficient of variation of $\tilde{x}_1$. Do the same for the random variable $\tilde{x}_2$.

For the rest of this exercise, consider the random variable $\tilde{y} = 4\tilde{x}_1 + \tilde{x}_2$.

(c) Without finding first the density of the random variable $\tilde{y}$, find directly the distribution function $F_{\tilde{y}}$ of the random variable $\tilde{y}$.

(d) Find the joint density of the random variables $\tilde{y}$ and $\tilde{x}_2$, $f_{\tilde{y},\tilde{x}_2}(y,x_2)$.

(e) Use the joint density obtained in (d) to find the density $f_{\tilde{y}}$ of the random variable $\tilde{y}$.

(f) Check that the results you have obtained in parts (c) and (e) are mutually consistent, i.e., check that $F_{\tilde{y}}$ is the distribution function associated with the density $f_{\tilde{y}}$, and that $f_{\tilde{y}}$ is the density associated with the distribution function $F_{\tilde{y}}$.

(g) Find the conditional density of the random variable $\tilde{y}$ given $\tilde{x}_2 = 3$, $f_{\tilde{y}\,|\,\tilde{x}_2}\,(y\,|\,3)$.

(h) Find both the (unconditional) expectation of $\tilde{y}$, $\mathrm{E}\,(\tilde{y})$, and the conditional expectation of $\tilde{y}$ given $\tilde{x}_2 = 3$, $\mathrm{E}\,(\tilde{y}\,|\,\tilde{x}_2 = 3)$.

(i) Find the covariance between $\tilde{y}$ and $\tilde{x}_2$, $\mathrm{Cov}\,(\tilde{y}, \tilde{x}_2)$.

17. The density function of the random variable $\tilde{x}$ is

$$f_{\tilde{x}}\,(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Consider the random variable $\tilde{y} = \tilde{x}^{1/2}$, i.e., $\tilde{y}$ is the positive square root of $\tilde{x}$.

(a) Without finding first the density of the random variable $\tilde{y}$, find directly the distribution function $F_{\tilde{y}}$ of the random variable $\tilde{y}$.

(b) Without making use of the distribution function you have found in part (a), find directly the density $f_{\tilde{y}}$ of the random variable $\tilde{y}$.

(c) Check that the results you have obtained in parts (a) and (b) are mutually consistent, i.e., check that $F_{\tilde{y}}$ is the distribution function associated with the density $f_{\tilde{y}}$, and that $f_{\tilde{y}}$ is the density associated with the distribution function $F_{\tilde{y}}$.

(d) Is the distribution function $F_{\tilde{x}}(x)$ of the random variable $\tilde{x}$ differentiable for all $x \in \mathbb{R}$? Is the distribution function $F_{\tilde{y}}(y)$ of the random variable $\tilde{y}$ differentiable for all $y \in \mathbb{R}$? Draw the functions $f_{\tilde{x}}\,(x)$, $F_{\tilde{x}}(x)$, $f_{\tilde{y}}(y)$, and $F_{\tilde{y}}(y)$.

18. The density function of the random vector $(\tilde{x}, \tilde{y})$ is

$$f(x, y) = \begin{cases} e^{-(x+y)} & \text{if } \;\; x > 0, \;\; y > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find

(a) The distribution function of the random vector $(\tilde{x}, \tilde{y})$.

(b) The marginal density functions of $\tilde{x}$ and $\tilde{y}$. Are the random variables $\tilde{x}$ and $\tilde{y}$ independent?

(c) The conditional density function of $\tilde{y}$ given $\tilde{x} = 3$.

(d) $P\,(\{0 < \tilde{x} < 1\} \cap \{\tilde{x} < \tilde{y} < 1\})$.

Consider the random variable $\tilde{z} = \tilde{x} - 3\tilde{y}$.

(e) Without finding first the density of the random variable $\tilde{z}$, find the distribution function $F_{\tilde{z}}$ of the random variable $\tilde{z}$. That is, use the distribution function method.

(f) Without making use of the distribution function you have found in part (a), find the density $f_{\tilde{z}}$ of the random variable $\tilde{z}$. That is, use the density function method.

(g) Check that the results you have obtained in parts (e) and (f) are mutually consistent, i.e., check that $F_{\tilde{z}}$ is the distribution function associated with the density $f_{\tilde{z}}$, and that $f_{\tilde{z}}$ is the density associated with the distribution function $F_{\tilde{z}}$.

(h) Draw the functions $F_{\tilde{z}}(z)$ and $f_{\tilde{z}}(z)$.

19. The density of the random vector $(\tilde{x}_1, \tilde{x}_2)$ is given by

$$f(x_1, x_2) = \begin{cases} \dfrac{1}{2} & \text{for } 0 < x_1 < 2 \text{ and } 0 < x_2 < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Find the density $f_{\tilde{y}}(y)$ of the random variable $\tilde{y} = \tilde{x}_1 + \tilde{x}_2$. To do so, find first the density of the vector $(\tilde{y}, \tilde{x}_2)$. Draw the density $f_{\tilde{y}}(y)$ and check that the area below this density is equal to one.

(b) Find the distribution function of the random variable $\tilde{y}$, $F_{\tilde{y}}(y)$.

(c) Find the conditional density of $\tilde{y}$ given $\tilde{x}_2 = x_2$, $f_{\tilde{y}|\tilde{x}_2}(y\,|x_2)$, and the conditional density of $\tilde{x}_2$ given $\tilde{y} = y$, $f_{\tilde{x}_2|\tilde{y}}(x_2\,|y)$. Compute the following conditional expectations: $\mathrm{E}(\tilde{y}\,|\tilde{x}_2 = 1/4)$, $\mathrm{E}(\tilde{x}_2\,|\tilde{y} = 5/2)$ and $\mathrm{E}(\tilde{x}_2\,|\tilde{y} = 3/2)$.

20. The density function of the random vector $(\tilde{x}, \tilde{y})$ is

$$f_{\tilde{x},\tilde{y}}(x, y) = \begin{cases} kx(y+1) & \text{for } 0 < x < 1,\ 0 < y < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Find the value of the constant k.

(b) Compute the product moment $\mathrm{E}(\tilde{x} \cdot \tilde{y})$ and the covariance $\mathrm{Cov}(\tilde{x}, \tilde{y})$ between $\tilde{x}$ and $\tilde{y}$. Are $\tilde{x}$ and $\tilde{y}$ independent?

(c) Compute the conditional expectation of $\tilde{y}$ given $\tilde{x} = 1/3$, $\mathrm{E}(\tilde{y}\,|\tilde{x} = 1/3)$.

(d) Use the marginal density of the random variable $\tilde{x}$ to find the moment-generating function $M_{\tilde{x}}(t)$ of $\tilde{x}$ and use this moment-generating function to find the expectation of $\tilde{x}$, $\mathrm{E}(\tilde{x})$.

(e) Find the (cumulative) distribution function $F_{\tilde{x},\tilde{y}}(x, y)$ of the random vector $(\tilde{x}, \tilde{y})$ and check that the cross partial derivative of $F_{\tilde{x},\tilde{y}}(x, y)$ is the density $f_{\tilde{x},\tilde{y}}(x, y)$ a.e. with respect to Lebesgue measure in $\mathbb{R}^2$.

(f) Use the marginal density of the random variable $\tilde{x}$ to compute the expectation of $\ln \tilde{x}$, $\mathrm{E}(\ln \tilde{x})$

(g) Use the marginal density of the random variable $\tilde{x}$ to find the density of the random variable $\tilde{w} = \ln \tilde{x}$. Draw this density.

(h) Find the moment-generating function of the random variable $\tilde{w}$ and use this moment-generating function to compute both the expectation and the variance of $\tilde{w}$, $\mathrm{E}(\tilde{w})$ and $\mathrm{Var}(\tilde{w})$.

(i) Use the density function method to find the density $f_{\tilde{w},\tilde{z}}(w, z)$ of the random vector $(\tilde{w}, \tilde{z})$, where

$$\tilde{w} = \ln \tilde{x} \quad \text{and} \quad \tilde{z} = \ln \left(\frac{\tilde{x}}{\tilde{y}} \right).$$

(j) Use the density $f_{\tilde{w},\tilde{z}}(w, z)$ of the random vector $(\tilde{w}, \tilde{z})$ to find the density $f_{\tilde{w}}(w)$ of the random variable $\tilde{w}$ and the density $f_{\tilde{z}}(z)$ of the random variable $\tilde{z}$. Draw the density of $f_{\tilde{z}}(z)$.

21. Assume that the density function of the random variable $\tilde{x}$ is

$$f_{\tilde{x}}(x) = \begin{cases} \dfrac{2 - x}{4} & \text{for } x \in (0, 2) \\ \dfrac{x + 2}{4} & \text{for } x \in (-2, 0] \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the distribution function $F_{\tilde{x}}(x)$ of the random variable $\tilde{x}$. Draw both $f_{\tilde{x}}(x)$ and $F_{\tilde{x}}(x)$.

(b) Find and draw the distribution function $F_{\tilde{y}}(y)$ and the density function $f_{\tilde{y}}(y)$ of the random variable $\tilde{y} = \tilde{x}^2$.

(c) Consider the random vector $(\tilde{x}, \tilde{y})$, where $\tilde{x}$ has the density defined above and $\tilde{y}$ is defined in part (b). Is the random vector $(\tilde{x}, \tilde{y})$ absolutely continuous, i.e., does its distribution have a density function?

(d) Compute the covariance between $\tilde{x}$ and $\tilde{y}$, $\mathrm{Cov}(\tilde{x}, \tilde{y})$. Are $\tilde{x}$ and $\tilde{y}$ independent? Answer the latter question using the definition of independence of two random variables.

22. Assume that the three random variables $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{x}_3$ on the same probability space $(\Omega, \mathcal{F}, P)$ are independent with the probability/density functions f_1, f_2 and f_3, respectively. Find the probability function/density of their convolution (i.e., of the distribution of their sum).

23. Consider a mixture $\tilde{x}$ of a discrete set $\{\tilde{x}_j\}_{j=1}^{N}$ of random variables, where N can be either a finite number or infinity and $\tilde{x} = \tilde{x}_j$ with probability p_j, for $j = 1, ..., N$. Let μ_j and σ_j^2 be the finite mean and the finite variance,

respectively, of each component $\tilde{x}_j$ of the mixture, for $j = 1, ..., N$, and $\mu = \sum_j p_j \mu_j$ is thus the mean of the mixture $\tilde{x}$.

(a) Prove that the kth central moment of the mixture $\tilde{x}$ is

$$\mathrm{E}\left[(\tilde{x} - \mu)^k\right] = \sum_j p_j \left(\sum_{n=0}^{k} \binom{k}{n} (\mu_j - \mu)^{k-n} \mathrm{E}\left[(\tilde{x}_j - \mu_j)^n\right] \right).$$

(b) Prove that the variance σ^2 of the mixture $\tilde{x}$ is

$$\sigma^2 = \sum_j p_j \left(\sigma_j^2 + \mu_j^2 - \mu^2\right) = \sum_j p_j \left(\sigma_j^2 + \mu_j^2\right) - \mu^2.$$

24. The density function of the random vector $(\tilde{x}, \tilde{y})$ is

$$f(x,y) = \begin{cases} \dfrac{2x+y}{12} & \text{if } x \in (0,2) \text{ and } y \in (0,2) \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute the covariance between $\tilde{x}$ and $\tilde{y}$. Are $\tilde{x}$ and $\tilde{y}$ independent?

(b) Compute the conditional expectation of $\tilde{x}$ given $\tilde{y} = 1/2$, $\mathrm{E}(\tilde{x} \,|\tilde{y} = 1/2)$.

(c) Use the density function method to find the density $f_{\tilde{z}}(z)$ of the random variable

$$\tilde{z} = 3\tilde{y} - \frac{\tilde{x}}{2}$$

and draw the density $f_{\tilde{z}}(z)$ of $\tilde{z}$.

25. Prove that, if $\tilde{x}$ has the logistic or the generalized extreme value type I, type II or type III distribution with location parameter m and scale parameter $s > 0$, then $\tilde{z} = c\tilde{x} + b$, with $c > 0$, has the the logistic or the generalized extreme value type I, type II or type III distribution, respectively, with location parameter $cm + b$ and scale parameter $cs > 0$. In particular, if $\tilde{x}$ is standard logistic or standard extreme value type I, type II or type III then $\tilde{z}$ is logistic or generalized extreme value type I, type II or type III, respectively, with location parameter b and scale parameter $c > 0$. Conversely, if $c = 1/s$ and $b = -m/s$, (i.e., $\tilde{z} = (\tilde{x} - m)/s$) then $\tilde{z}$ is standard logistic or standard extreme value type I, type II or type III, respectively. Moreover, if $\tilde{x}$ has the generalized extreme value type II or type III distribution, then the respective shape parameter α remains the same after applying that affine transformation.

26. In this exercise, you are asked to prove that the difference between two independent Gumbel-distributed random variables with the same scale parameter has a logistic distribution. To do so, follow the four steps below:

(a) Prove that $\tilde{w}$ has the exponential distribution with parameter equal to 1 (i.e., it has unitary mean) if and only if the random variable $\tilde{x} = m - s \ln \tilde{w}$

has the generalized extreme value type I (or Gumbel) distribution with location parameter m and scale parameter $s > 0$. Note that $\tilde{w} = e^{-(\tilde{x}-m)/s}$.

(b) Assume that $\tilde{w}_1$ and $\tilde{w}_2$ are independent random variables having the exponential distribution with parameter equal to 1. Find the distribution function and the density function of the random variable $\tilde{z} = \tilde{w}_1/\tilde{w}_2$.

(c) Assume that $\tilde{x}_1$ and $\tilde{x}_2$ are independent and have generalized extreme value type I (or Gumbel) distributions with location parameters m_1 and m_2, respectively, with the same scale parameter $s > 0$. Use the results of parts (a) and (b) to prove that the random variable

$$\tilde{q} = \left(\frac{\tilde{x}_1 - m_1}{s}\right) - \left(\frac{\tilde{x}_2 - m_2}{s}\right) = \ln\left(\frac{e^{-\left(\frac{\tilde{x}_2 - m_2}{s}\right)}}{e^{-\left(\frac{\tilde{x}_1 - m_1}{s}\right)}}\right)$$

has the standard logistic distribution, $\tilde{q} \sim L(0,1)$.

(d) Use the result in part (c) to prove that, if $\tilde{x}_1$ and $\tilde{x}_2$ are independent and have generalized extreme value type I (or Gumbel) distributions with location parameters m_1 and m_2, respectively, with the same scale parameter $s > 0$, then the random variable $\tilde{y} = \tilde{x}_1 - \tilde{x}_2$ has the logistic distribution with location parameter (or mean) $m_1 - m_2$ and scale parameter s, $\tilde{y} \sim L(m_1 - m_2, s)$. In particular, if $\tilde{x}_1$ and $\tilde{x}_2$ are independent and have standard extreme value type I distributions, then the random variable $\tilde{y} = \tilde{x}_1 - \tilde{x}_2$ has the standard logistic distribution, $\tilde{y} \sim L(0,1)$.

27. The density of the random vector $(\tilde{x}_1, \tilde{x}_2)$ is given by

$$f(x_1, x_2) = \begin{cases} 1 & \text{for } 1 < x_1 < 2 \text{ and } 3 < x_2 < 4 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Consider the random variable $\tilde{y} = \tilde{x}_1 + \tilde{x}_2$. Find the joint density of the vector $(\tilde{y}, \tilde{x}_2)$, $f_{\tilde{y},\tilde{x}_2}(y, x_2)$.

(b) Use the density you have found in part (a) of this exercise to find the density $f_{\tilde{y}}(y)$ of the random variable $\tilde{y} = \tilde{x}_1 + \tilde{x}_2$. Moreover, draw the density $f_{\tilde{y}}(y)$ of the random variable $\tilde{y}$ and check that the area below this density is equal to one.

(c) Find the distribution function $F_{\tilde{y}}(y)$ of the random variable $\tilde{y}$.

(d) Find the conditional density of $\tilde{y}$ given $\tilde{x}_2 = x_2$, $f_{\tilde{y}|\tilde{x}_2}(y\,|x_2)$, and the conditional density of $\tilde{x}_2$ given $\tilde{y} = y$, $f_{\tilde{x}_2|\tilde{y}}(x_2\,|y)$.

(e) Compute the following three conditional expectations:

$$\mathrm{E}(\tilde{y}\,|\tilde{x}_2 = 13/4), \quad \mathrm{E}(\tilde{x}_2\,|\tilde{y} = 11/2) \quad \text{and} \quad \mathrm{E}(\tilde{x}_2\,|\tilde{y} = 17/4).$$

(f) Prove that the random variables $\tilde{x}_1$ and $\tilde{x}_2$ are independent and compute the covariance between $\tilde{x}_2$ and $\tilde{y}$, $\mathrm{Cov}(\tilde{x}_2, \tilde{y})$.

6. Stochastic Processes and Limiting Distributions

6.1. Stochastic processes

- **Definition.** A stochastic process is a sequence (a vector) of random variables $\{\tilde{x}_t\}_{t=0}^{T}$ defined on the probability space $(\Omega, \mathcal{F}, P)$.
- *Note:* More generally, we can define a stochastic process of random objects $\{\tilde{x}_t\}_{t=0}^{T}$, where $\tilde{x}_t : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$.
- *Note:* We can have $T = \infty$ in all the previous and forthcoming expressions.
- **Definition.** A stochastic process $\{\tilde{x}_t\}_{t=0}^{\infty}$ defined on the probability space $(\Omega, \mathcal{F}, P)$ is strongly (or strictly) stationary if the distributions

$$P_{\tilde{x}_t, \tilde{x}_{t+1}, \tilde{x}_{t+2}, ..., \tilde{x}_{t+n}}(C) = P\left\{(\tilde{x}_t, \tilde{x}_{t+1}, \tilde{x}_{t+2}, ..., \tilde{x}_{t+n}) \in C\right\}$$

do not depend on t, for all non-negative integers $n = 0, 1, 2, ...$, and all $C \in \mathcal{B}\left(\mathbb{R}^{n+1}\right)$.

- **Definition.** A stochastic process $\{\tilde{x}_t\}_{t=0}^{\infty}$ on the probability space $(\Omega, \mathcal{F}, P)$ with $\tilde{x}_t \in \mathrm{L}^2$ for all t is weakly stationary if

$$\mathrm{E}(\tilde{x}_t) = \mu \text{ for } t = 0, 1, ...$$

and $\mathrm{Cov}(\tilde{x}_t, \tilde{x}_{t+n})$ depends only on n, for all t, i.e.,

$$\mathrm{Cov}(\tilde{x}_t, \tilde{x}_{t+n}) = \hat{\gamma}(n), \quad \text{for } t = 0, 1, ... \text{ and } n = 0, 1, ...$$

- The covariance $\mathrm{Cov}(\tilde{x}_t, \tilde{x}_{t+n}) \equiv \gamma(t, n)$ is called autocovariance since it is the covariance between the members of a single stochastic process.
- Note that $\gamma(t, n) = \hat{\gamma}(n)$ for all t when the stochastic process is weakly stationary. Moreover, in this case $\mathrm{Var}(\tilde{x}_t) = \hat{\gamma}(0)$ for all t.

- **Definition.** The random variables $\{\tilde{x}_t\}_{t=0}^{\infty}$ on the probability space $(\Omega, \mathcal{F}, P)$ are independent if the random variables of any finite subset of $\{\tilde{x}_t\}_{t=0}^{\infty}$ are independent.
- **Definition.** The random variables $\{\tilde{x}_t\}_{t=0}^{T}$ on the probability space $(\Omega, \mathcal{F}, P)$ are identically distributed if they have the same (marginal) distribution.
- We say that the random variables $\{\tilde{x}_t\}_{t=0}^{T}$ on the probability space $(\Omega, \mathcal{F}, P)$ are i.i.d. (identically and independently distributed) or that the stochastic process is i.i.d. when they simultaneously satisfy the previous two definitions.
- *Note:* i.i.d. processes are strongly stationary.

- **Definition.** The stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ on the probability space $(\Omega, \mathcal{F}, P)$ with $\tilde{x}_t \in \mathrm{L}^1$ for all t is white noise if it is i.i.d. and $\mathrm{E}(\tilde{x}_t) = 0$ for all t.

- **Definition.** The stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ on the probability space $(\Omega, \mathcal{F}, P)$ is a random walk if $\tilde{x}_{t+1} = \tilde{x}_t + \tilde{y}_{t+1}$ for $t = 0, 1, ...,$ where $\{\tilde{y}_t\}_{t=1}^{T}$ is i.i.d. and the value x_0 taken by the random variable $\tilde{x}_0$ is exogenously given.

- If the stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ on the probability space $(\Omega, \mathcal{F}, P)$ is a random walk with $\tilde{x}_{t+1} = \tilde{x}_t + \tilde{y}_{t+1}$ and $\mathrm{E}(\tilde{y}_t) = \mu \neq 0$ for all t, then $\{\tilde{x}_t\}_{t=0}^{T}$ is called a random walk with drift μ. If $\mathrm{E}(\tilde{y}_t) = 0$ for all t, then $\{\tilde{x}_t\}_{t=0}^{T}$ is called a (symmetric) random walk.

- Thus, a random walk $\{\tilde{x}_t\}_{t=0}^{T}$ with drift μ can be written as $\tilde{x}_{t+1} = \mu + \tilde{x}_t + \tilde{\varepsilon}_{t+1}$, where $\{\tilde{\varepsilon}_t\}_{t=1}^{T}$ is white noise since $\tilde{\varepsilon}_t \equiv \tilde{y}_t - \mu$ for all t.

6.2. Filtrations and martingales

- **Definition.** A filtration $\mathbb{F}$ on $(\Omega, \mathcal{F})$ is a sequence of σ-algebras, $\mathbb{F} = \{\mathcal{F}_t\}_{t=0}^{T}$ such that $\mathcal{F}_t \subset \mathcal{F}_{t+1} \subset \mathcal{F}$ for all t.
- **Definition.** If $\mathbb{F} = \{\mathcal{F}_t\}_{t=0}^{T}$ is a filtration on $(\Omega, \mathcal{F})$ and $\{\tilde{x}_t\}_{t=0}^{T}$ is a stochastic process on $(\Omega, \mathcal{F}, P)$ such that $\tilde{x}_t$ is $\mathcal{F}_t$-measurable for $t = 0, 1, ...T$, then we say that the stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ is adapted to $\mathbb{F}$.
- The **natural filtration** associated to the stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ on $(\Omega, \mathcal{F}, P)$ is the simplest filtration which the stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ is adapted to. All information concerning the process, and only that information, is available in the natural filtration.

- Formally, the natural filtration $\mathbb{F}^X = \left\{\mathcal{F}_t^X\right\}_{t=0}^T$ associated to the stochastic process $X = \left\{\tilde{x}_t\right\}_{t=0}^T$ on $(\Omega, \mathcal{F}, P)$ is given by

$$\mathcal{F}_t^X = \sigma\left\{\tilde{x}_j^{-1}(B)\middle|\ j \leq t,\ B \in \mathcal{B}\right\},$$

  i.e., $\mathcal{F}_t^X$ is the smallest σ-algebra on Ω that contains all pre-images of the Borel sets of $\mathbb{R}$ for all j up to t.

- Any stochastic process is adapted to its natural filtration.

- **Definition.** Let $\left\{\tilde{x}_t\right\}_{t=0}^T$ be a stochastic process on $(\Omega, \mathcal{F}, P)$ adapted to the filtration $\mathbb{F} = \left\{\mathcal{F}_t\right\}_{t=0}^T$ on $(\Omega, \mathcal{F})$ and $\tilde{x}_t \in \mathrm{L}^1$ for all t. Then, we say that the stochastic process $\left\{\tilde{x}_t\right\}_{t=0}^T$ is a martingale (submartingale) (supermartingale) relative to $\mathbb{F}$ if, $\mathrm{E}\left(\tilde{x}_{t+1}\middle|\mathcal{F}_t\right) = (\geq)\,(\leq)\,\tilde{x}_t$ a.s. $[P]$, for $t = 0, 1, \ldots T-1$.

- When we say that $X = \left\{\tilde{x}_t\right\}_{t=0}^T$ is a martingale (submartingale) (supermartingale) without mentioning the filtration, we mean that $X = \left\{\tilde{x}_t\right\}_{t=0}^T$ is a martingale (submartingale) (supermartingale) relative to its associated natural filtration $\mathbb{F}^X$.

- **Proposition.** The stochastic process $X = \{\tilde{x}_t\}_{t=0}^{T}$ on $(\Omega, \mathcal{F}, P)$ is a martingale (submartingale) (supermartingale) relative to its associated natural filtration $\mathbb{F}^X = \{\mathcal{F}_t^X\}_{t=0}^{T}$ if and only if

$$\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_1, \tilde{x}_0 \right.\right) = (\geq)(\leq)\tilde{x}_t \ ,$$

$$\text{a.s.}\,[P], \text{ for } t = 0, 1, ..., T-1, \qquad (*)$$

- **Proof.** Just note that

$$\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_0 \right.\right) = \mathrm{E}\left(\tilde{x}_{t+1} \left| \mathcal{F}_t^X \right.\right), \ \text{a.s.}\,[P]. \qquad \textit{Q.E.D.}$$

- Note that the condition $(*)$ in the previous proposition is equivalent to the following:

$$\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t = x_t, \tilde{x}_{t-1} = x_{t-1}, ..., \tilde{x}_0 = x_0 \right.\right) = (\geq)\,(\leq)\,x_t,$$

$$\text{a.s.} \ \left[P_{\tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_0}\right], \ \text{for } t = 0, 1, ..., T-1.$$

- A random walk with no drift (with positive drift) (with negative drift) is a martingale (submartingale) (supermartingale) relative to its associated natural filtration. Recall that every stochastic process is adapted to its natural filtration.
- **Proposition.** If the stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ on $(\Omega, \mathcal{F}, P)$ adapted to the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t=0}^{T}$ on $(\Omega, \mathcal{F})$ is a martingale (submartingale) (supermartingale) relative to $\mathbb{F} = \{\mathcal{F}_t\}_{t=0}^{T}$ then

(a)

$$\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_1, \tilde{x}_0 \right.\right) = (\geq)(\leq)\tilde{x}_t \text{ , a.s. } [P], \text{ for } t = 0, 1, ..., T-1,$$

i.e., $\{\tilde{x}_t\}_{t=0}^{T}$ is a martingale relative to its natural filtration $\mathbb{F}^X$; and

(b)

$$\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t = x_t, \tilde{x}_{t-1} = x_{t-1}, ..., \tilde{x}_0 = x_0 \right.\right) = (\geq)(\leq) x_t,$$

$$\text{a.s. } \left[P_{\tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_0}\right], \text{ for } t = 0, 1, ..., T-1.$$

- **Proof. (a)** Recall that $\mathcal{F}(\tilde{x}_t)$ is the $\sigma-$algebra on Ω induced by the random variable $\tilde{x}_t$, i.e., the coarsest $\sigma-$algebra that makes $\tilde{x}_t$ measurable. Observe that $\mathcal{F}(\tilde{x}_t) \subset \mathcal{F}_t^X \subset \mathcal{F}_t$ so that $\tilde{x}_t$ is $\mathcal{F}_t^X-$measurable. Therefore,

$$\mathrm{E}(\tilde{x}_{t+1} | \tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_0) = \mathrm{E}\left(\tilde{x}_{t+1} \middle| \mathcal{F}_t^X\right) = \mathrm{E}\left(\mathrm{E}(\tilde{x}_{t+1} | \mathcal{F}_t) \middle| \mathcal{F}_t^X\right)$$

$$= (\geq)(\leq)\mathrm{E}\left(\tilde{x}_t \middle| \mathcal{F}_t^X\right) = \tilde{x}_t, \text{ a.s. } [P],$$

where the first equality follows from the definition of $\mathcal{F}_t^X$, the second equality follows from the law of iterated expectations, the third equality (inequality) follows from the martingale (submartingale) (supermartingale) assumption, $\mathrm{E}(\tilde{x}_{t+1} | \mathcal{F}_t) = (\geq)(\leq)\tilde{x}_t$, and the fourth equality follows from the $\mathcal{F}_t^X-$measurability of $\tilde{x}_t$.
(b) Obvious since (a) and (b) are equivalent. *Q.E.D.*

- **Corollary 1.** If the stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ on $(\Omega, \mathcal{F}, P)$ is a martingale (submartingale) (supermartingale) relative to the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t=0}^{T}$, then

$$\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t \right.\right) = (\geq)\,(\leq)\, \tilde{x}_t, \text{ a.s. } [P], \text{ for } t = 0, 1, ..., T-1.$$

and

$$\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t = x_t \right.\right) = (\geq)\,(\leq)\, x_t, \text{ a.s. } [P_{\tilde{x}_t}], \text{ for } t = 0, 1, ..., T-1.$$

- **Proof.** Note that $\mathcal{F}(\tilde{x}_t) \subset \mathcal{F}_t$, where $\mathcal{F}(\tilde{x}_t)$ is the $\sigma-$algebra on Ω induced by the random variable $\tilde{x}_t$, and, hence,

$$\begin{aligned}\mathrm{E}\left(\tilde{x}_{t+1} \left| \tilde{x}_t \right.\right) &= \mathrm{E}\left(\tilde{x}_{t+1} \left| \mathcal{F}(\tilde{x}_t) \right.\right) = \mathrm{E}\left(\mathrm{E}\left(\tilde{x}_{t+1} \left| \mathcal{F}_t \right.\right) \left| \mathcal{F}(\tilde{x}_t) \right.\right) \\ &= (\geq)\,(\leq)\, \mathrm{E}\left(\tilde{x}_t \left| \mathcal{F}(\tilde{x}_t) \right.\right) = \tilde{x}_t, \text{ a.s. } [P],\end{aligned}$$

where the second equality follows from the law of iterated expectations and the third equality (inequality) follows from the martingale (submartingale) (supermartingale) assumption, $\mathrm{E}\left(\tilde{x}_{t+1} \left| \mathcal{F}_t \right.\right) = (\geq)\,(\leq)\, \tilde{x}_t$. *Q.E.D.*

- **Corollary 2.** If the stochastic process $\{\tilde{x}_t\}_{t=0}^{T}$ on $(\Omega, \mathcal{F}, P)$ is a martingale relative to the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t=0}^{T}$, then

$$\mathrm{E}\left(\tilde{x}_{t+n} \left| \mathcal{F}_t \right.\right) = \tilde{x}_t, \quad \text{a.s.} \quad [P], \text{ for } t = 0, 1, ...T-1 \text{ and } n = 1, 2, ...$$

and

$$\mathrm{E}\left(\tilde{x}_{t+n} \left| \tilde{x}_t = x_t \right.\right) = x_t, \quad \text{a.s.} \quad [P_{\tilde{x}_t}], \text{ for } t = 0, 1, ...T-1 \text{ and } n = 1, 2, ...$$

- **Proof.** Since $\mathcal{F}_t \subset \mathcal{F}_{t+1} \subset ... \subset \mathcal{F}_{t+n-1}$, we apply the law of iterated expectations and the martingale property,

$$\mathrm{E}\left(\tilde{x}_{t+n} \left| \mathcal{F}_t \right.\right) = \mathrm{E}\left(\mathrm{E}\left(\tilde{x}_{t+n} \left| \mathcal{F}_{t+n-1} \right.\right) \left| \mathcal{F}_t \right.\right) = \mathrm{E}\left(\tilde{x}_{t+n-1} \left| \mathcal{F}_t \right.\right),$$

and we get the desired result by induction, i.e., by repeating the previous procedure n times until we get

$$\mathrm{E}\left(\tilde{x}_{t+n} \left| \mathcal{F}_t \right.\right) = \mathrm{E}\left(\tilde{x}_t \left| \mathcal{F}_t \right.\right) = \tilde{x}_t. \qquad \textit{Q.E.D.}$$

- *Notation:* In Dynamic Economics we use the notation $\mathrm{E}_t\left(\tilde{x}_j\right)$ to mean $\mathrm{E}\left(\tilde{x}_j \left| \mathcal{F}_t \right.\right)$.
- If the stochastic process $\left\{\tilde{x}_t\right\}_{t=0}^{\infty}$ is "independently distributed" then $\mathrm{E}_t\left(\tilde{x}_{t+j}\right) = \mathrm{E}\left(\tilde{x}_{t+j}\right)$ for all t and $j > 0$.
- If the stochastic process $\left\{\tilde{x}_t\right\}_{t=0}^{\infty}$ is "identically distributed" then $\mathrm{E}\left(\tilde{x}_t\right) = \mathrm{E}\left(\tilde{x}\right)$ for all t, where $\tilde{x}$ is a random variable having the same distribution as the random variable $\tilde{x}_t$ for all t.
- Thus, if the stochastic process $\left\{\tilde{x}_t\right\}_{t=0}^{\infty}$ is "identically and independently distributed (i.i.d.)" then $\mathrm{E}_t\left(\tilde{x}_{t+j}\right) = \mathrm{E}\left(\tilde{x}\right)$ for all t and $j > 0$, where $\tilde{x}$ is a random variable having the same distribution as the random variable $\tilde{x}_s$ for all s.
- Obviously, $\mathrm{E}_t\left(\tilde{x}_{t-j}\right) = \tilde{x}_{t-j}$, for all t and $j \geq 0$.

6.3. Convergence in probability, in mean square, in distribution, and almost sure convergence

- Let $\{\tilde{x}_n\}_{n=1}^{\infty}$ be a stochastic process or sequence of random variables on $(\Omega, \mathcal{F}, P)$. Note that we are making $n = t + 1$.

• **Convergence in probability.** The sequence of random variables $\{\tilde{x}_n\}_{n=1}^{\infty}$ is said to converge to the random variable $\tilde{x}$ in probability if

$$\lim_{n\to\infty} P\left\{|\tilde{x}_n - \tilde{x}| \geq \varepsilon\right\} \equiv \lim_{n\to\infty} P\left\{\omega \in \Omega \,||\tilde{x}_n(\omega) - \tilde{x}(\omega)| \geq \varepsilon\right\} = 0, \text{ for all } \varepsilon > 0,$$

or

$$\lim_{n\to\infty} P\left\{|\tilde{x}_n - \tilde{x}| < \varepsilon\right\} \equiv \lim_{n\to\infty} P\left\{\omega \in \Omega \,||\tilde{x}_n(\omega) - \tilde{x}(\omega)| < \varepsilon\right\} = 1, \text{ for all } \varepsilon > 0.$$

- This concept is the same as that of "convergence in measure".
- We use the notation $\tilde{x}_n \xrightarrow{p} \tilde{x}$ or $\underset{n\to\infty}{\text{plim}}\, \tilde{x}_n = \tilde{x}$.

- **Convergence in mean square.** The sequence of random variables $\{\tilde{x}_n\}_{n=1}^{\infty}$ belonging to L^2 is said to converge to the random variable $\tilde{x} \in \mathrm{L}^2$ in mean square if

$$\lim_{n \to \infty} \mathrm{E}\left[(\tilde{x}_n - \tilde{x})^2\right] = 0.$$

- We write $\tilde{x}_n \xrightarrow{m} \tilde{x}$.

- **Proposition.** If $\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right) = c$, where c is a constant, and $\lim_{n\to\infty} \mathrm{Var}\left(\tilde{x}_n\right) = 0$, then $\tilde{x}_n \xrightarrow{m} c$.

- **Proof.**

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Var}\left(\tilde{x}_n\right) &= \lim_{n\to\infty} \mathrm{Var}\left(\tilde{x}_n - c\right) \\
&= \lim_{n\to\infty} \mathrm{E}\left[\left(\tilde{x}_n - c\right)^2\right] - \lim_{n\to\infty} \left[\mathrm{E}\left(\tilde{x}_n - c\right)\right]^2 \\
&= \lim_{n\to\infty} \mathrm{E}\left[\left(\tilde{x}_n - c\right)^2\right] - \underbrace{\left[\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right) - c\right]^2}_{=0} \\
&= \lim_{n\to\infty} \mathrm{E}\left[\left(\tilde{x}_n - c\right)^2\right] = 0. \qquad \textit{Q.E.D.}
\end{aligned}$$

- **Corollary.** If $\mathrm{E}\left(\tilde{x}_n\right) = c$ for all n, where c is a constant, and $\lim_{n\to\infty} \mathrm{Var}\left(\tilde{x}_n\right) = 0$, then $\tilde{x}_n \xrightarrow{m} c$.

- **Almost sure convergence.** The sequence of random variables $\{\tilde{x}_n\}_{n=1}^{\infty}$ is said to converge to the random variable $\tilde{x}$ almost surely (or strongly) if

$$P\left\{\lim_{n\to\infty}\tilde{x}_n=\tilde{x}\right\}\equiv P\left\{\omega\in\Omega\,\middle|\,\lim_{n\to\infty}\tilde{x}_n(\omega)=\tilde{x}(\omega)\right\}=1,$$

  or

$$P\left\{\lim_{n\to\infty}\tilde{x}_n\neq\tilde{x}\right\}\equiv P\left\{\omega\in\Omega\,\middle|\,\lim_{n\to\infty}\tilde{x}_n(\omega)\neq\tilde{x}(\omega)\right\}=0.$$

- This concept is the same as that of "convergence almost everywhere".
- We write $\tilde{x}_n \xrightarrow{a.s} \tilde{x}$.

- **Sure convergence.** The sequence of random variables $\{\tilde{x}_n\}_{n=1}^{\infty}$ is said to converge to the random variable $\tilde{x}$ surely if

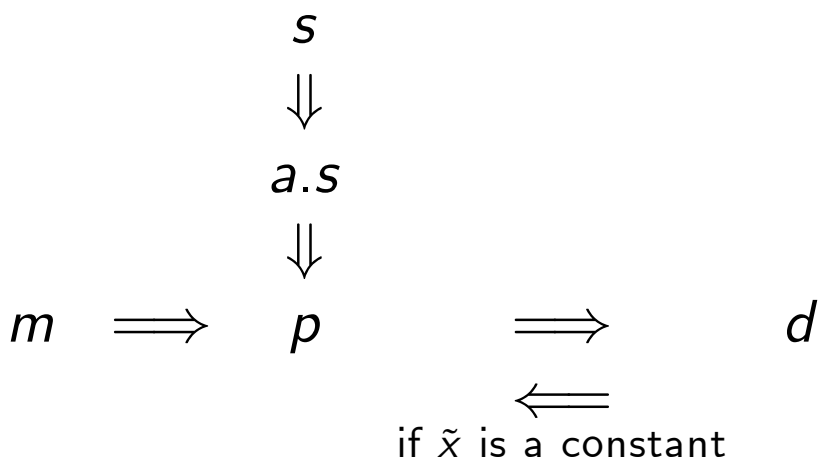
$$\lim_{n\to\infty} \tilde{x}_n = \tilde{x}, \text{ that is, } \lim_{n\to\infty} \tilde{x}_n(\omega) = \tilde{x}(\omega) \text{ for all } \omega \in \Omega.$$

- This concept is the same as that of "pointwise convergence" and is rarely used in Statistics.
- We write $\tilde{x}_n \xrightarrow{s} \tilde{x}$.

- **Convergence in distribution.** The sequence of random variables $\{\tilde{x}_n\}_{n=1}^{\infty}$ is said to converge to the random variable $\tilde{x}$ in distribution (or weakly) if the distribution function F_n of $\tilde{x}_n$ converges pointwise to the distribution F of $\tilde{x}$ at every continuity point of F.

- We write $\tilde{x}_n \xrightarrow{d} \tilde{x}$, $\tilde{x}_n \longrightarrow F$, $\tilde{x}_n \longrightarrow \widehat{P}$, $\tilde{x}_n \overset{a}{\sim} \tilde{x}$ ($\tilde{x}_n$ is asymptotically distributed as $\tilde{x}$), $\tilde{x}_n \overset{a}{\sim} F$ or $\tilde{x}_n \overset{a}{\sim} \widehat{P}$, where F and $\widehat{P}$ are called the asymptotic (or limiting) distribution function and the asymptotic distribution of $\{\tilde{x}_n\}_{n=1}^{\infty}$, respectively. Obviously, $\widehat{P}$ is the distribution associated with the distribution function F.

- All the previous convergence concepts can be applied to sequences of random vectors, $\left\{\tilde{X}_n\right\}_{n=1}^{\infty}$ with $\tilde{X}_n : (\Omega, \mathcal{F}, P) \longrightarrow \left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right)$. For convergence in probability we just need to replace the absolute value $|\tilde{x}_n(\omega) - \tilde{x}(\omega)|$ by the Euclidean distance $\left\|\tilde{X}_n(\omega) - \tilde{X}(\omega)\right\|$, where $\tilde{X}$ is the limiting random vector. For convergence in mean square, we just need to replace $\lim_{n\to\infty} \mathrm{E}\left[(\tilde{x}_n - \tilde{x})^2\right] = 0$ by $\lim_{n\to\infty} \mathrm{E}\left[\left\|\tilde{X}_n - \tilde{X}\right\|^2\right] = 0$.


s
a.s
m
p
d
if $\tilde{x}$ is a constant

- **Proof of** $s \Longrightarrow a.s$ **(trivial).**
- **Proof of** $a.s \Longrightarrow p$ **(trivial):**

$$P\left\{\lim_{n\to\infty} \tilde{x}_n = \tilde{x}\right\} = 1 \Longrightarrow \lim_{n\to\infty} P\left\{\left|\tilde{x}_n - \tilde{x}\right| < \varepsilon\right\} = 1.$$

- **Proof of** $p \Longrightarrow d$ (see the handout).
- **Proof of** $d \Longrightarrow p$ **when** $\tilde{x}$ **is a constant** (see the handout).

- **Proof of $m \Longrightarrow p$ (easy):**
- Recall the Markov inequality: If $\tilde{y}$ is a non-negative random variable, then

$$P\left\{\tilde{y} \geq a\right\} \leq \frac{\mathrm{E}\left(\tilde{y}\right)}{a} \text{ for all } a > 0.$$

Make $a = \varepsilon^2 > 0$ and $\tilde{y} = \left(\tilde{x}_n - \tilde{x}\right)^2 \geq 0$ so that

$$P\left\{\left(\tilde{x}_n - \tilde{x}\right)^2 \geq \varepsilon^2\right\} \leq \frac{\mathrm{E}\left[\left(\tilde{x}_n - \tilde{x}\right)^2\right]}{\varepsilon^2} \text{ for all } \varepsilon \neq 0.$$

Therefore,

$$\lim_{n\to\infty} P\left\{\left(\tilde{x}_n - \tilde{x}\right)^2 \geq \varepsilon^2\right\} \leq \lim_{n\to\infty} \frac{\mathrm{E}\left[\left(\tilde{x}_n - \tilde{x}\right)^2\right]}{\varepsilon^2} \text{ for all } \varepsilon \neq 0.$$

If $\tilde{x}_n \xrightarrow{m} \tilde{x}$, then

$$\lim_{n\to\infty} \mathrm{E}\left[(\tilde{x}_n - \tilde{x})^2\right] = 0$$

so that

$$0 \leq \lim_{n\to\infty} P\left\{(\tilde{x}_n - \tilde{x})^2 \geq \varepsilon^2\right\} \leq \lim_{n\to\infty} \frac{\mathrm{E}\left[(\tilde{x}_n - \tilde{x})^2\right]}{\varepsilon^2} = 0 \text{ for all } \varepsilon \neq 0.$$

Thus,

$$\lim_{n\to\infty} P\left\{|\tilde{x}_n - \tilde{x}| \geq \varepsilon\right\} = 0 \text{ for all } \varepsilon > 0,$$

and, hence, $\tilde{x}_n \xrightarrow{p} \tilde{x}$. *Q.E.D.*

6.4. Weak convergence of distribution functions and of probability measures

- **Definition.** The sequence of distribution functions $\{F_n\}_{n=1}^{\infty}$ on $\mathbb{R}^k$ converges weakly to the distribution function F if F_n converges pointwise to F at every continuity point of F, i.e., $\lim_{n\to\infty} F_n(x) = F(x)$ if F is continuous at $x \in \mathbb{R}^k$.
- We write $F_n \xrightarrow{w} F$ for weak convergence, while $F_n \longrightarrow F$ denotes pointwise convergence of F, i.e., $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}^k$.
- Therefore, if $\tilde{x}_n \sim F_n(x)$ and $\tilde{x} \sim F(x)$, then $\tilde{x}_n \xrightarrow{d} \tilde{x} \iff F_n \xrightarrow{w} F$.

- **Example:**

$$\tilde{x}_n \sim F_n(x) = \begin{cases} 0 & \text{if } x < \frac{1}{n} \\ \\ 1 & \text{if } x \geq \frac{1}{n} \end{cases}$$

and

$$\tilde{x} \sim F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \\ 1 & \text{if } x \geq 0. \end{cases}$$

- Then,

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ for all } x \neq 0,$$

and

$$\lim_{n\to\infty} F_n(0) = 0 \neq F(0) = 1.$$

Therefore,

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \\ 1 & \text{if } x > 0. \end{cases}$$

However, F is discontinuous at 0. Hence, $F_n \xrightarrow{w} F \iff \tilde{x}_n \xrightarrow{d} \tilde{x}$.

- **Definition.** The sequence of measures $\{\mu_n\}_{n=1}^{\infty}$ on $\left(\mathbb{R}^k, \mathcal{B}\right)$ converges weakly to the measure μ if $\lim_{n\to\infty} \mu_n(B) = \mu(B)$ for all Borel sets $B \in \mathcal{B}\left(\mathbb{R}^k\right)$ such that $\mu(\partial B) = 0$, where ∂B is the boundary of B.
- We write $\mu_n \xrightarrow{w} \mu$.
- **Proposition.** $F_n \xrightarrow{w} F$ if and only if $P_n \xrightarrow{w} \widehat{P}$, where P_n and $\widehat{P}$ are the distributions (probability measures on $\left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right)$) associated with F_n and F, respectively.
- *Note:* To say that $\lim_{n\to\infty} F_n(x) = F(x)$ when F is continuous at $x \in \mathbb{R}$ is equivalent to say that

$$\lim_{n\to\infty} P_n(-\infty, x] = \widehat{P}(-\infty, x]$$

  at every continuity point of F, i.e., at every point where $\widehat{P}\{x\} = 0$.
- Therefore, $\tilde{x}_n \xrightarrow{d} \tilde{x} \iff F_n \xrightarrow{w} F \iff P_n \xrightarrow{w} \widehat{P}$.

- **Proposition.** The sequence of measures $\{\mu_n\}_{n=1}^{\infty}$ on $\left(\mathbb{R}^k, \mathcal{B}\right)$ converges weakly to the measure μ, $\mu_n \xrightarrow{w} \mu$, if and only if

$$\lim_{n\to\infty} \int_{\mathbb{R}^k} g(x) d\mu_n = \int_{\mathbb{R}^k} g(x) d\mu.$$

  for all bounded and continuous functions $g : \left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right) \longrightarrow (\mathbb{R}, \mathcal{B})$.

- **Corollary (Portmanteau theorem).** The sequence of random vectors $\left\{\tilde{X}_n\right\}_{n=1}^{\infty}$, $\tilde{X}_n : (\Omega, \mathcal{F}, P) \longrightarrow \left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right)$, converges in distribution to the random vector $\tilde{X}$ if and only if

$$\lim_{n\to\infty} \mathrm{E}\left[g\left(\tilde{X}_n\right)\right] = \mathrm{E}\left[g\left(\tilde{X}\right)\right]$$

  for all bounded and continuous functions $g : \left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right) \longrightarrow (\mathbb{R}, \mathcal{B})$.

- Recall that $\varphi_{\tilde{y}}(t) = \mathrm{E}\left(e^{it\tilde{y}}\right)$ is the characteristic function, $M_{\tilde{y}}(t) = \mathrm{E}\left(e^{t\tilde{y}}\right)$ is the moment-generating function, and $\Lambda_{\tilde{y}}(t) = \mathrm{E}\left(e^{-t\tilde{y}}\right)$ is the Laplace transform of the random variable $\tilde{y}$ (or of its distribution).

- **Theorem (Lévy's convergence theorem).** $\tilde{x}_n \xrightarrow{d} \tilde{x}$ if and only if $\lim_{n\to\infty} \varphi_{\tilde{x}_n}(t) = \varphi_{\tilde{x}}(t)$ for all $t \in \mathbb{R}$.

  or

  $\tilde{x}_n \xrightarrow{d} \tilde{x}$ if and only if $\lim_{n\to\infty} M_{\tilde{x}_n}(t) = M_{\tilde{x}}(t)$ for all t in the domain of $M_{\tilde{x}}$, provided $M_{\tilde{x}}$ is well-defined (i.e., finite) in a neighborhood of $t = 0$.

  or

  $\tilde{x}_n \xrightarrow{d} \tilde{x}$ if and only if $\lim_{n\to\infty} \Lambda_{\tilde{x}_n}(t) = \Lambda_{\tilde{x}}(t)$ for all t in the domain of $\Lambda_{\tilde{x}}$, provided $\Lambda_{\tilde{x}}$ is well-defined (i.e., finite) in a neighborhood of $t = 0$.

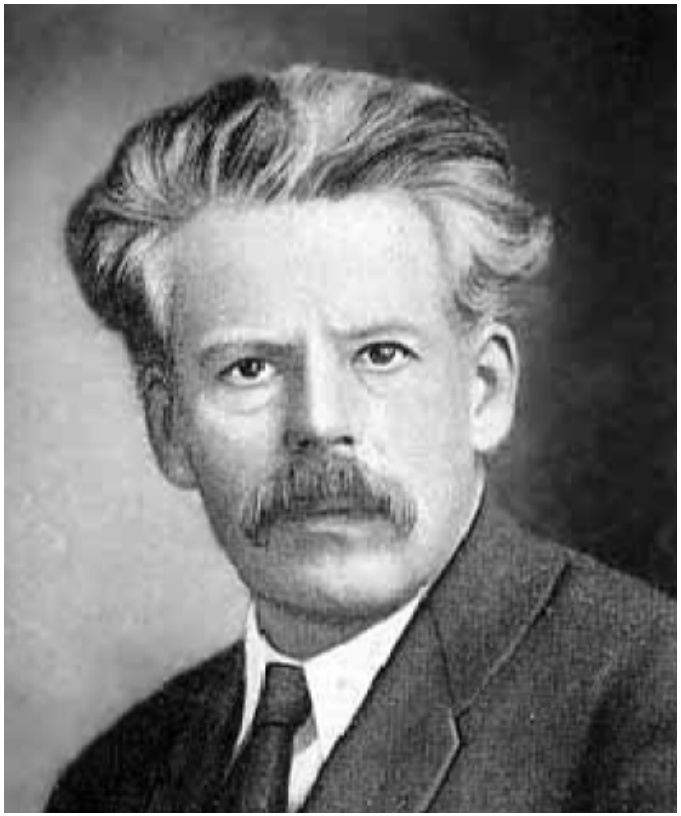
Paul Lévy (1886 - 1971)

- **Continuous mapping theorem.** If a sequence of random vectors $\left\{\tilde{X}_n\right\}_{n=1}^{\infty}$, $\tilde{X}_n : (\Omega, \mathcal{F}, P) \longrightarrow \left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right)$, converges in distribution (in probability) (almost surely) to the random vector $\tilde{X}$ and the function $g : \left(\mathbb{R}^k, \mathcal{B}\left(\mathbb{R}^k\right)\right) \longrightarrow \left(\mathbb{R}^m, \mathcal{B}\left(\mathbb{R}^m\right)\right)$ is continuous, then $g(\tilde{X}_n)$ converges in distribution (in probability) (almost surely) to $g\left(\tilde{X}\right)$.

- **Slutsky's theorem.** If the sequence of random variables $\left\{\tilde{x}_n\right\}_{n=1}^{\infty}$ converges in distribution to the random variable $\tilde{x}$ and the sequence of random variables $\left\{\tilde{y}_n\right\}_{n=1}^{\infty}$ converges in distribution (or in probability) to the constant c, then

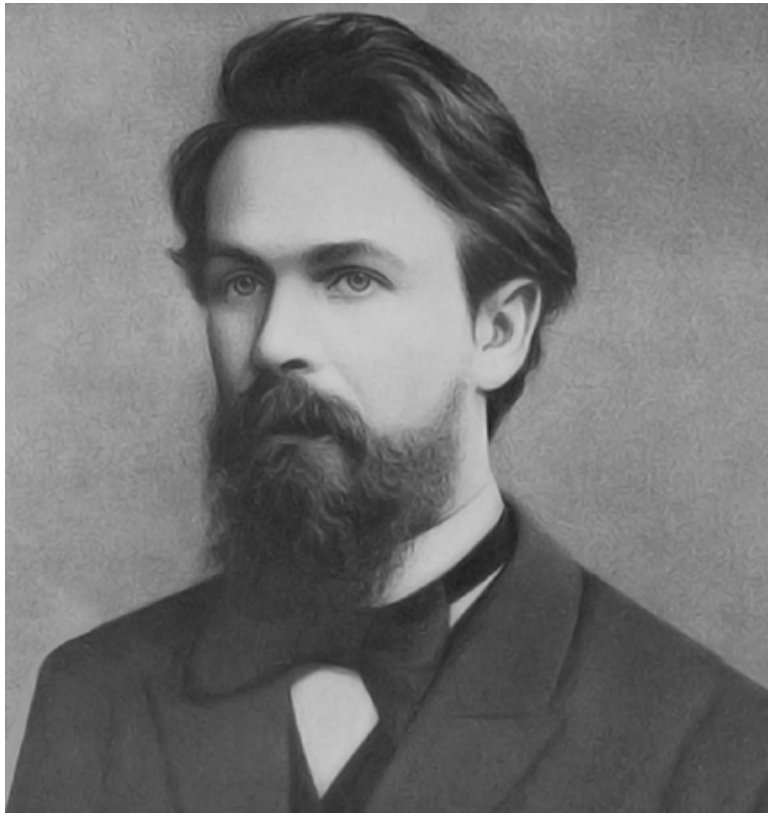
  1) $\tilde{x}_n + \tilde{y}_n \xrightarrow{d} \tilde{x} + c$,
  2) $\tilde{x}_n \cdot \tilde{y}_n \xrightarrow{d} c\tilde{x}$,
  3) $\tilde{x}_n / \tilde{y}_n \xrightarrow{d} \tilde{x}/c$ for $c \neq 0$.

- All the previous discussion about convergence can be extended to a framework where the random variable $\tilde{x}_r$ is characterized by a vector $r \in \mathbb{R}^k$ and r converges to $a \in \overline{\mathbb{R}}^k$.



Evgeny Slutsky (1880 – 1948)

6.5. Markov processes



Andrei Markov (1856 - 1922)

- **Discrete Markov Chains.**

- A discrete Markov process (or Markov chain) is a sequence of discrete random objects $\{\tilde{x}_t\}_{t=0}^{\infty}$ defined on the probability space $(\Omega, \mathcal{F}, P)$ taking values on the same countable set S (the state space), $\tilde{x}_t : (\Omega, \mathcal{F}, P) \longrightarrow \left(S, 2^S\right)$, where the conditional distribution of $\tilde{x}_{t+1}$ given $(\tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_0) = (x_t, x_{t-1}, ..., x_0)$ depends only on the value x_t taken by $\tilde{x}_t$ for $t = 0, 1, ...$, i.e.,

$$P\left\{\tilde{x}_{t+1} = x_{t+1} \middle| \tilde{x}_t = x_t, \tilde{x}_{t-1} = x_{t-1}, ..., \tilde{x}_0 = x_0\right\}$$

$$= P\left\{\tilde{x}_{t+1} = x_{t+1} \middle| \tilde{x}_t = x_t\right\},$$

for all $(x_0, x_1, ..., x_t, x_{t+1}) \in S^{t+2}$ and $t = 0, 1, ...$

- The elements of S are called states, $S = \{s_1, s_2, ...\}$.

- A Markov chain is said to be finite if the state space S is finite.

- Usually, the subindex t refers to "time" in Markov processes.
- The Markovian assumption is not so restrictive. To see this, define $\widetilde{X}_t = (\tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_{t-j})$.
- If $\left\{\widetilde{X}_t\right\}_{t=0}^{\infty}$ is a Markov chain then

$$P\left\{\widetilde{X}_{t+1} = X_{t+1} | \widetilde{X}_t = X_t\right\} =$$

$$P\left\{(\tilde{x}_{t+1}, ..., \tilde{x}_{t-j+1}) = (x_{t+1}, ..., x_{t-j+1}) \,|\, (\tilde{x}_t, ..., \tilde{x}_{t-j}) = (x_t, ..., x_{t-j})\right\}$$

- Therefore, the conditional distribution of $\tilde{x}_{t+1}$ depends on the values x_{t-s} taken by $\tilde{x}_{t-s}$ for $s = 0, 1, ..., j$.

- Consider from now on a finite Markov chain with $\#S = N < \infty$.
- A transition matrix Π is a matrix whose element π_{ij} at the (i,j)-entry is the probability of $\tilde{x}_{t+1} = s_j$ given $\tilde{x}_t = s_i$,

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1N} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2N} \\ \cdots & \cdots & \ddots & \cdots \\ \pi_{N1} & \pi_{N2} & \cdots & \pi_{NN} \end{pmatrix},$$

  where

$$\pi_{ij} = P\left\{\tilde{x}_{t+1} = s_j \middle| \tilde{x}_t = s_i\right\}.$$

- π_{ij} is called the transition probability from state s_i to state s_j.

- Although one could consider Markov chains whose transition matrices vary with t, Markov chains are usually understood to have time-invariant transition matrices unless otherwise specified.
- Any row i of a transition matrix consists of nonnegative elements which must add up to unity, $\sum_j \pi_{ij} = 1$. A matrix with this property is called a **stochastic matrix** and its rows are called **probabilistic (or probability) vectors .**
- The product of transition matrices is also a stochastic matrix that gives the conditional probabilities after a sequence of transitions.

- To see this, observe that, according to the theorem of total (conditional) probability, we have the following:

$$P\left\{\tilde{x}_2 = s_j \middle| \tilde{x}_0 = s_i\right\} = \sum_{m=1}^{N} P\left\{\tilde{x}_2 = s_j, \tilde{x}_1 = s_m \middle| \tilde{x}_0 = s_i\right\}$$

$$= \sum_{m=1}^{N} P\left\{\tilde{x}_1 = s_m \middle| \tilde{x}_0 = s_i\right\} \cdot P\left\{\tilde{x}_2 = s_j \middle| \tilde{x}_1 = s_m, \tilde{x}_0 = s_i\right\}$$

$$= \sum_{m=1}^{N} P\left\{\tilde{x}_1 = s_m \middle| \tilde{x}_0 = s_i\right\} \cdot P\left\{\tilde{x}_2 = s_j \middle| \tilde{x}_1 = s_m\right\} = \sum_{m=1}^{N} \pi_{im} \cdot \pi_{mj},$$

where the first equality follows as $\left\{\tilde{x}_2 = s_j\right\} = \bigcup_{m=1}^{N} \left\{\tilde{x}_2 = s_j, \tilde{x}_1 = s_m\right\}$ and the events appearing in the previous union are disjoint, the second equality is obvious from the definition of conditional probability, and the third equality follows from the fact that $P\left\{\tilde{x}_2 = s_j \middle| \tilde{x}_1 = s_m, \tilde{x}_0 = s_i\right\} = P\left\{\tilde{x}_2 = s_j \middle| \tilde{x}_1 = s_m\right\}$ as dictated by the assumption that $\left\{\tilde{x}_t\right\}_{t=0}^{\infty}$ is a Markov chain.

- Note that $\pi_{ij,2} \equiv \sum_{m=1}^{N} \pi_{im} \cdot \pi_{mj}$ is the value appearing in the (i,j)-entry of the square of the time-invariant transition matrix, Π^2.
- Therefore, in general, the element $\pi_{ij,t}$ in the (i,j)-entry of the matrix Π^t is

$$\pi_{ij,t} = P\left\{\tilde{x}_t = s_j | \tilde{x}_0 = s_i\right\} = P\left\{\tilde{x}_{n+t} = s_j | \tilde{x}_n = s_i\right\},$$

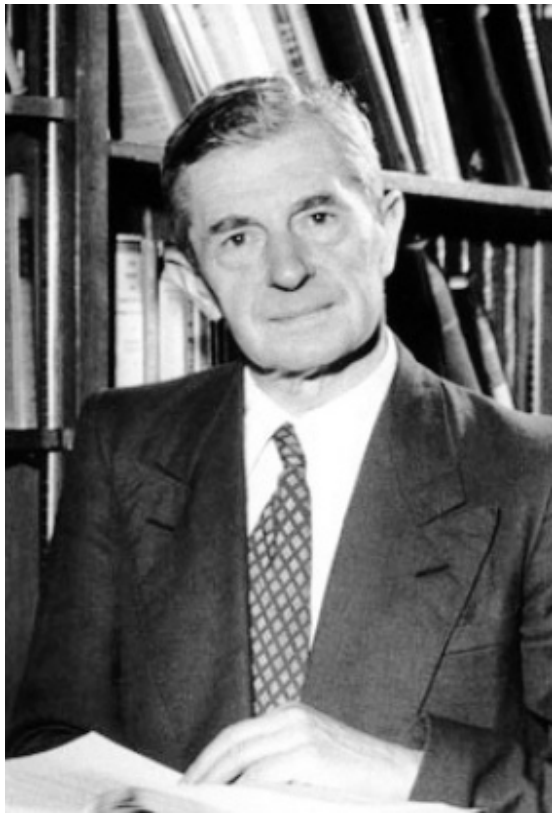
  for all $n = 0, 1, ...$
- Chapman-Kolmogorov equation:

$$\pi_{ij,t+s} \equiv \sum_{m=1}^{N} \pi_{im,t} \cdot \pi_{mj,s} \Longleftrightarrow \Pi^{t+s} = \Pi^t \Pi^s.$$

- Let the probabilistic row vector $p_0 = (p_{01}, p_{02}, ..., p_{0N})$, where $p_{0i} = P\left\{\tilde{x}_0 = s_i\right\}$ with $s_i \in S$, correspond to the initial distribution of states. A Markov chain with the transition matrix Π satisfies

$$P\left\{\tilde{x}_0 = s_{i_0}, \tilde{x}_1 = s_{i_1}, \tilde{x}_2 = s_{i_2}, ..., \tilde{x}_T = s_{i_T}\right\} = p_{0i_0} \cdot \pi_{i_0 i_1} \cdot \pi_{i_1 i_2} \cdot ... \cdot \pi_{i_{T-1} i_T},$$

  for $T = 1, 2, ...$ and all $(s_{i_0}, s_{i_1}, ..., s_{i_T}) \in S^{T+1}$.



Sydney Chapman (1888 - 1970)

- After t transitions, the unconditional probability to be in the state s_j is the jth coordinate of a row vector p_t, where

$$p_t = p_0 \Pi^t.$$

- Clearly, p_t evolves according to

$$p_{t+1} = p_t \Pi.$$

- The stationary distribution of a Markov chain is characterized by a probabilistic row vector p^* such that

$$p^* = p^* \Pi,$$

and, thus,

$$p^* = p^* \Pi^t, \text{ for all } t.$$

- p^* is called the **stationary probabilistic vector** and its components are called the **stationary probabilities.**

- **Definition.** Given a square matrix A with n rows/columns, a non-zero column vector $x \in \mathbb{R}^n$ is a (right-)**eigenvector** of A if it satisfies the equation $Ax = \lambda x$ for some scalar $\lambda \in \mathbb{R}$. In this case, the scalar λ is the **eigenvalue** of A associated with the (right-)eigenvector x, or the vector x is a (right-)eigenvector associated with the eigenvalue λ.

- Note that the equation $Ax = \lambda x$ can be written as $Ax = \lambda \mathrm{I} x$ or as $(A - \lambda \mathrm{I})x = \underline{0}$, where $\underline{0}$ is a column vector of zeroes and I is the identity matrix.

- Therefore, since x is a non-zero vector, an eigenvalue is a solution (real or complex) to the equation

$$\det(A - \lambda \mathrm{I}) = 0,$$

  and, hence, the set of vectors x satisfying the equation $(A - \lambda \mathrm{I})x = \underline{0}$, where λ is an eigenvalue of A, constitutes a vector subspace of $\mathbb{R}^n$.

- **Definition.** The (right-)**eigenspace** associated with a particular eigenvalue of a matrix is the set of the (right-)eigenvectors associated with this eigenvalue, together with the zero vector (which has no direction). That is, the (right-)eigenspace associated with the eigenvalue λ of the matrix A is the set of vectors x satisfying the equation $(A - \lambda \mathrm{I})x = \underline{0}$.

- *Note:* If x is a (right-)eigenvector associated with the eigenvalue λ, then the vector αx, where $\alpha \in \mathbb{R}$ is an arbitrary scalar, belongs to the (right-)eigenspace associated with the eigenvalue λ. Moreover, any linear combination of (right-)eigenvectors associated with the eigenvalue λ belongs to the (right-)eigenspace associated with the eigenvalue λ.

- **Definition.** Given a square matrix A with n rows/columns, a non-zero row vector $x \in \mathbb{R}^n$ is a **left-eigenvector** of A if it satisfies the equation $xA = \lambda x$ for some scalar $\lambda \in \mathbb{R}$. In this case, the vector x is the left-eigenvector associated with the "left-eigenvalue" λ.

- *Note:* A scalar λ satisfying the equation $xA = \lambda x$ (or $xA = \lambda x\mathrm{I}$ or $x(A - \lambda\mathrm{I}) = \underline{0}^\intercal$) for some non-zero row vector x must be also a solution to the equation $\det(A - \lambda\mathrm{I}) = 0$. That is, the set of "left-eigenvalues" coincides with the set of eigenvalues. However, the sets of (transposed) left-eigenvectors and of (right-)eigenvectors are not equal.

- **Definition.** The **left-eigenspace** associated with a particular eigenvalue of a matrix is the set of the left-eigenvectors associated with this eigenvalue, together with the zero vector (which has no direction). That is, the left-eigenspace associated with the eigenvalue λ of the matrix A is the set of row vectors x satisfying the equation $x(A - \lambda\mathrm{I}) = \underline{0}^\intercal$.

- A stochastic matrix has always an eigenvalue equal to 1 and the column vector $\underline{1} \equiv (1, ..., 1)^\intercal$ is one of the (right-)eigenvectors associated with the eigenvalue 1. Obviously, since

$$\sum_j \pi_{ij} = 1, \text{ for all } i \iff \Pi \underline{1} = \underline{1} \iff (\Pi - \mathrm{I})\,\underline{1} = \underline{0},$$

  we must have

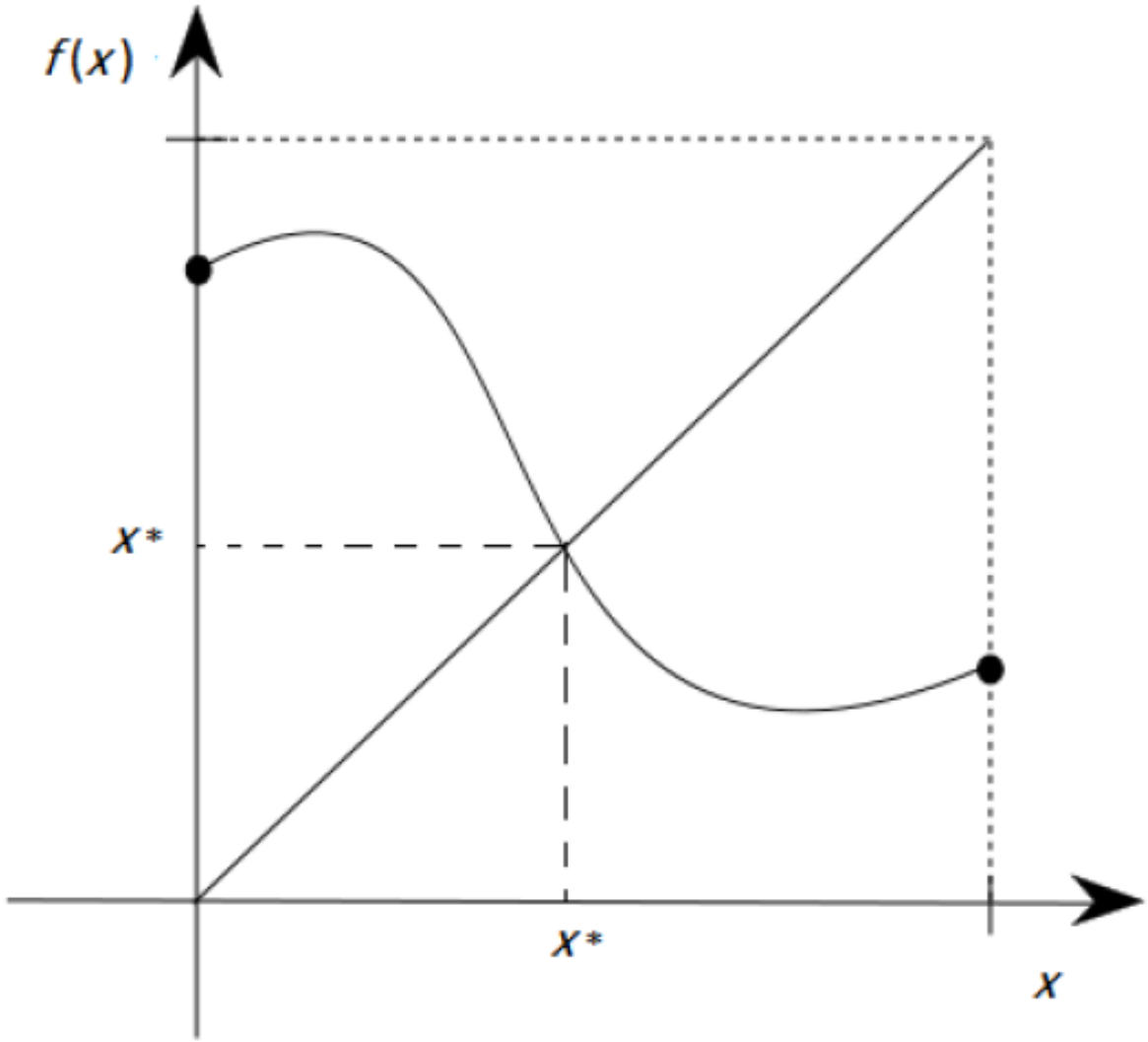
$$\det(\Pi - \mathrm{I}) = 0.$$

- Therefore, any column vector with identical values in its entries, $\underline{\alpha} = \alpha \underline{1} = (\alpha, \alpha, ..., \alpha)^\intercal$, belongs to the (right-)eigenspace associated with the eigenvalue 1 of the stochastic matrix Π.
- The stationary probabilistic vector p^* of the Markov chain is a left-eigenvector associated with the eigenvalue 1 since

$$p^* = p^* \Pi \iff p^* (\Pi - \mathrm{I}) = \underline{0}^\intercal,$$

  where $\underline{0}^\intercal$ is a row vector of zeroes ($\underline{0}$ is a column vector of zeroes). Moreover, the elements of the vector p^* must be non-negative and add up to unity.

- **Brouwer's fixed point theorem:** Every continuous function f from a convex compact subset D of $\mathbb{R}^n$ to itself has a fixed point x^*, $f\left(x^*\right) = x^*$.



- Consider the unitary simplex Δ of $\mathbb{R}^n$,

$$\Delta = \left\{ (x_1, x_2, ..., x_n) \in \mathbb{R}^n \;\middle|\; x_i \in [0,1], \text{ for } i = 1, 2, ..., n, \text{ and } \sum_{i=1}^{n} x_i = 1 \right\}.$$

  - *Note:* The dimension of a set in $\mathbb{R}^n$ is the minimum number of coordinates needed to specify any point within it.
  - Hence, the unitary simplex has dimension $n-1$.
  - The set of probabilistic vectors is the unitary simplex Δ.
  - The unitary simplex Δ is convex and compact and the function $f : \Delta \longrightarrow \Delta$ defined by $f(p) = p\Pi$, where Π is a stochastic matrix, is continuous. Therefore, according to Brouwer's fixed point theorem, f has a fixed point p^*, $f(p^*) = p^*$, that is, there exists a probabilistic vector p^* such that $p^*\Pi = p^*$.

- Therefore, the stationary probabilistic vector p^* always exists but may fail to be unique, since the eigenvalue 1 could exhibit a multiplicity.
- The dimension of the (non-empty) set of stationary probabilistic vectors is equal to the multiplicity of the eigenvalue 1 of the matrix Π minus 1.
- If there exists a matrix $\widehat{\Pi}$ such that $\lim_{t\to\infty} \Pi^t = \widehat{\Pi}$, then the row i of the limiting matrix $\widehat{\Pi}$ gives the conditional long-run (or ergodic) probabilities of the different states when the initial state is s_i. These conditional ergodic probabilities define the "conditional ergodic distribution" of the chain given $\tilde{x}_0 = s_i$.
- The element $\widehat{\pi}_{ij}$ at row i and column j of the matrix $\widehat{\Pi}$ is thus

$$\widehat{\pi}_{ij} = \lim_{t\to\infty} P\left\{\tilde{x}_t = s_j \middle| \tilde{x}_0 = s_i\right\}.$$

- Moreover, if all the rows of $\widehat{\Pi}$ are identical and equal to the row vector $\widehat{p}$, then this probabilistic vector $\widehat{p}$ gives the "ergodic distribution" of the Markov chain, which is the long-run distribution of the Markov chain for all initial probabilistic vectors p_0. The row vector $\widehat{p}$ is called the **ergodic probabilistic vector** of the Markov chain and its components are called the **ergodic probabilities.** Clearly, in this case,

$$\lim_{t\to\infty} p_0\Pi^t = p_0\widehat{\Pi} = \widehat{p}, \text{ for all } p_0.$$

- Moreover, as

$$p^* = \lim_{t\to\infty} p^*\Pi^t = p^*\widehat{\Pi} = \widehat{p},$$

the ergodic distribution, if it exists, coincides with the stationary distribution, which is unique in this case.

- However, $\lim_{t\to\infty} \Pi^t$ may fail to exist. Moreover, even if this limit exists, it may occur that the limiting matrix $\widehat{\Pi}$ has non-identical rows. In these cases, there is no ergodic probabilistic vector and, hence, the ergodic distribution does not exist.

- Recall that a stochastic matrix has always an eigenvalue equal to 1 and that the column vector $\underline{1} \equiv (1, ..., 1)^\intercal$ is one of the (right-)eigenvectors associated with the eigenvalue 1. Therefore, any column vector with identical values in its entries, $\underline{\alpha} = \alpha\underline{1} = (\alpha, \alpha, ..., \alpha)^\intercal$, belongs to the (right-)eigenspace associated with the eigenvalue 1.

- While stochastic matrices have at least one eigenvalue equal to 1, the other (real or complex) eigenvalues have modulus smaller or equal than 1.

- A sufficient condition for the existence of the ergodic distribution is that the transition matrix Π be **regular** (i.e., there exists a natural number t for which the matrix Π^t has only strictly positive elements). In this case, $\widehat{p} = p^* \gg 0$.

- **Doeblin's Theorem:** A Markov chain has an ergodic distribution if and only if there is a natural number t for which Π^t has a column j whose elements are all strictly positive. In this case, $\widehat{p}_j = p_j^* > 0$.
- Note that, if the column j has all its elements strictly positive, then the state s_j has always a strictly positive probability of being reached for all the states of the chain in the previous period. The state s_j is going to visited again and again and, after a while, the chain's initial distribution is going to get "forgotten."
- Please, read the Introduction and Sections 1 and 2 of Bru, B. and Yor, M. (2002). "Comments on the life and mathematical legacy of Wolfgang Doeblin." *Finance and Stochastics* 6: 3-47. https://link.springer.com/content/pdf/10.1007%2Fs780-002-8399-0.pdf
- An equivalent necessary and sufficient condition for the existence of the ergodic distribution is that the matrix Π has a unique eigenvalue equal to 1 and all the other eigenvalues have modulus strictly smaller than 1.

Wolfgang Doeblin (1915 – 1940)

- If the limiting matrix $\widehat{\Pi}$ exists, then

$$\Pi \cdot \widehat{\Pi} = \Pi \cdot \left(\lim_{t \to \infty} \Pi^t \right) = \lim_{t \to \infty} \Pi^{t+1} = \widehat{\Pi}.$$

- Therefore, if the column vector $\overline{\pi}_j$ is the column j of $\widehat{\Pi}$, then $\Pi \overline{\pi}_j = \overline{\pi}_j$ (or $(\Pi - \mathrm{I})\,\overline{\pi}_j = \underline{0}$) so that $\overline{\pi}_j$ belongs to the (right-)eigenspace associated with the eigenvalue 1 of the matrix Π.

- In fact, the column vectors of $\widehat{\Pi}$ span the (right-)eigenspace associated with the eigenvalue 1 of the matrix Π.

- Thus, if the matrix Π has only one eigenvalue equal to 1 and all the other eigenvalues have modulus strictly smaller than 1, then the limiting matrix $\widehat{\Pi}$ exists and all the columns of $\widehat{\Pi}$ will be proportional to a column vector $\underline{\alpha} = (\alpha, \alpha, ..., \alpha)^{\intercal}$, with identical values in its entries, which implies in turn that all the rows of $\widehat{\Pi}$ are identical and, hence, the ergodic distribution exists. Moreover, in this case the stationary probabilistic vector is unique. **(Case I)**

- If the matrix Π has several eigenvalues equal to 1 and all the other eigenvalues have modulus strictly smaller than one, then the limiting matrix $\widehat{\Pi}$ exists but its rows are non-identical. Therefore, the ergodic distribution does not exist. Moreover, in this case the set of stationary probabilistic vectors has dimension equal to the multiplicity of the eigenvalue 1 of the matrix Π minus 1. **(Case II)**

- If the matrix Π has some eigenvalues different from 1 but with modulus equal to 1, then the limiting matrix $\widehat{\Pi}$ fails to exist. Therefore, the ergodic distribution does not exist. Moreover, in this case the set of stationary probabilistic vectors has also dimension equal to the multiplicity of the eigenvalue 1 of the matrix Π minus 1. **(Case III)**

- Moreover, if the limiting matrix $\widehat{\Pi}$ exists, then

$$\widehat{\Pi} \cdot \Pi = \left(\lim_{t \to \infty} \Pi^t \right) \cdot \Pi = \lim_{t \to \infty} \Pi^{t+1} = \widehat{\Pi}.$$

- Therefore, if the row vector $\overline{\pi}_i$ is the row i of $\widehat{\Pi}$, then $\overline{\pi}_i \Pi = \overline{\pi}_i$ (or $\overline{\pi}_i (\Pi - \mathrm{I}) = \underline{0}^\intercal$) so that $\overline{\pi}_i$ belongs to the left-eigenspace associated with the eigenvalue 1 of the matrix Π.
- In fact, the row vectors of $\widehat{\Pi}$ span the left-eigenspace associated with the eigenvalue 1 of the matrix Π.
- Recall also that the row i of the limiting matrix $\widehat{\Pi}$ gives the conditional long-run (or ergodic) probabilities of the different states when the initial state is s_i. These conditional ergodic probabilities define the "conditional ergodic distribution" of the Markov chain given $\tilde{x}_0 = s_i$.

- **Definition.** A state $s_i \in S$ is absorbing if

$$\pi_{ii} = P\left\{\tilde{x}_{t+1} = s_i \middle| \tilde{x}_t = s_i\right\} = 1.$$

- If the state s_i is absorbing then a probabilistic vector that assigns probability 1 to the state s_i (and zero probability to the other states) is a stationary probabilistic vector.
- Assume that the limiting matrix $\widehat{\Pi}$ exists. If the state s_j is absorbing then the (i,j)-entry $\widehat{\pi}_{ij}$ of the matrix $\widehat{\Pi}$ gives the probability that a chain starting in state s_i ends up being absorbed by state s_j.
- **Definition.** A set E is ergodic if

  (a) E is a subset of the state space S.

  (b) $P\left\{\tilde{x}_{t+1} \in E \middle| \tilde{x}_t \in E\right\} = 1$.

  (c) No proper (or strict) subset of E satisfies property (b).
- *Note:* D is a proper (or strict) subset of E if $D \subset E$ and $D \neq E$.
- *Note:* Absorbing states are ergodic sets with $\#E = 1$.

- **Definition.** A state $s_i \in S$ is transient if, given $\tilde{x}_0 = s_i$, there is a non-zero probability that we will never return to s_i, i.e., if

$$P\left(\bigcap_{t=1}^{\infty}\{\tilde{x}_t \neq s_i\}\middle|\,\tilde{x}_0 = s_i\right) > 0.$$

- **Definition.** A state $s_i \in S$ is recurrent (or persistent) if it is not transient, i.e., if

$$P\left(\bigcap_{t=1}^{\infty}\{\tilde{x}_t \neq s_i\}\middle|\,\tilde{x}_0 = s_i\right) = 0$$

or

$$P\left(\bigcup_{t=1}^{\infty}\{\tilde{x}_t = s_i\}\middle|\,\tilde{x}_0 = s_i\right) = 1.$$

- **Definition.** A state $s_i \in S$ has period k if any return to state s_i must occur in multiples of k transitions (or steps or time periods).
- Formally, the period of a state s_i is defined as the greatest common divisor of $\{t \mid P\{\tilde{x}_t = s_i | \tilde{x}_0 = s_i\} > 0\}$, where t is a strictly positive natural number.
- Note that even though a state has period k, it may not be possible to reach the state in k steps. For example, suppose it is possible to return to the state in $\{6, 9, 12, 15, ...\}$ steps; k would be 3, even though 3 does not appear in this list.
- If $k = 1$, then the state s_i is said to be aperiodic: returns to state s_i can occur at irregular times. Otherwise $(k > 1)$, the state s_i is said to be periodic with period k. Thus, absorbing states are aperiodic.

- **Examples.** The state space is $S = \{s_1, s_2, s_3\}$.
- **Case I.**

$$\lim_{t\to\infty} \underbrace{\begin{pmatrix} 3/10 & 0 & 7/10 \\ 4/5 & 1/5 & 0 \\ 0 & 3/5 & 2/5 \end{pmatrix}}_{\Pi}^{t} = \begin{pmatrix} 24/73 & 21/73 & 28/73 \\ 24/73 & 21/73 & 28/73 \\ 24/73 & 21/73 & 28/73 \end{pmatrix} = \widehat{\Pi}$$

- None of the columns of the transition matrix Π has all its elements strictly positive. However, the eigenvalues of the matrix Π are 1 and $-0.05 \pm 0.59791i$ (which have modulus smaller than 1) so that the ergodic distribution exists. In fact, the matrix Π is regular. In this case,

$$p^* = \widehat{p} = \left(\frac{24}{73}, \frac{21}{73}, \frac{28}{73}\right)$$

and the single ergodic set is the state space $S = \{s_1, s_2, s_3\}$. All the states are recurrent (or persistent) and aperiodic.

- **Case II.**

$$\lim_{t\to\infty} \underbrace{\begin{pmatrix} 2/5 & 1/5 & 2/5 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^t}_{\Pi} = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \widehat{\Pi}$$

- The matrix $\widehat{\Pi}$ exists but has non-identical rows. The eigenvalues of the transition matrix Π are $1, 1$ and 0.4. Hence, there is no ergodic probabilistic vector and the ergodic distribution fails to exist. Moreover, all the columns of $\widehat{\Pi}$ belong to (and they span) the (right-)eigenspace associated with the eigenvalue 1 of Π.
- The Markov chain has the following stationary probabilistic vectors:

$$p^* = (0,\ \beta,\ 1-\beta) \text{ for all } \beta \in [0,1],$$

which form a subset of $\mathbb{R}^3$ of dimension $1\ (= 2-1)$. Clearly, $(0, \beta, 1-\beta)\cdot\Pi = (0, \beta, 1-\beta)$.

- The states s_2 and s_3 are absorbing and there are two ergodic sets: $\{s_2\}$ and $\{s_3\}$. The state s_1 is transient whereas the states s_2 and s_3 are obviously recurrent (or persistent). All three states are aperiodic.

- **Case III.**

$$\lim_{t\to\infty} \underbrace{\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 3/10 & 2/10 & 1/2 \end{pmatrix}}_{\Pi}^{t}$$

does not exists since

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 3/10 & 2/10 & 1/2 \end{pmatrix}^{t} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ a_t & b_t & c_t \end{pmatrix}$$

if t is odd, whereas

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 3/10 & 2/10 & 1/2 \end{pmatrix}^{t} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ d_t & e_t & f_t \end{pmatrix}$$

if t is even. Thus, there is no ergodic distribution.

- The eigenvalues of the transition matrix Π are $1, -1$ and 0.5. Thus, since there is an eigenvalue different from 1 with modulus equal to 1, $\lim_{t \to \infty} \Pi^t$ does not exist and there is no ergodic probabilistic vector.
- Since there is only one eigenvalue equal to 1, the Markov chain has the following unique stationary probabilistic vector:

$$p^* = \left(\frac{1}{2}, \ \frac{1}{2}, \ 0 \right).$$

- The single ergodic set is $\{s_1, s_2\}$. The state s_3 is transient and aperiodic. The states s_1 and s_2 are recurrent (or persistent) and periodic with period 2.

- **A 4×4 matrix example (Case II) - Exercise**
- The matrix

$$\Pi' = \begin{pmatrix} 1/5 & 0 & 4/5 & 0 \\ 0 & 3/5 & 0 & 2/5 \\ 7/10 & 0 & 3/10 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$

could be rewritten, without loss of generality, as

$$\Pi = \begin{pmatrix} 1/5 & 4/5 & 0 & 0 \\ 7/10 & 3/10 & 0 & 0 \\ 0 & 0 & 3/5 & 2/5 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

by permuting or relabeling the states s_2 and s_3. In this way, the two ergodic sets become more apparent.

$$\lim_{t\to\infty} \underbrace{\begin{pmatrix} 1/5 & 4/5 & 0 & 0 \\ 7/10 & 3/10 & 0 & 0 \\ 0 & 0 & 3/10 & 4/10 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}}_{\Pi}^{t} = \underbrace{\begin{pmatrix} 7/15 & 8/15 & 0 & 0 \\ 7/15 & 8/15 & 0 & 0 \\ 0 & 0 & 5/9 & 4/9 \\ 0 & 0 & 5/9 & 4/9 \end{pmatrix}}_{\widehat{\Pi}}$$

- The matrix $\widehat{\Pi}$ exists but has non-identical rows. The eigenvalues of the transition matrix Π are $1, 1, 0.1$, and -0.5. Hence, there is no ergodic probabilistic vector and the ergodic distribution fails to exist. Moreover, all the columns of $\widehat{\Pi}$ belong to (and they span) the (right-)eigenspace associated with the eigenvalue 1 of Π.

- The Markov chain has the following stationary probabilistic vectors:

$$p^* = \left(\alpha,\ \frac{8\alpha}{7},\ \frac{5}{9} - \frac{25\alpha}{21},\ \frac{4}{9} - \frac{20\alpha}{21}\right) \quad \text{for all } \alpha \in [0, 7/15],$$

which form a subset of $\mathbb{R}^4$ of dimension 1 $(= 2 - 1)$. Clearly,

$$\left(\alpha,\ \frac{8\alpha}{7},\ \frac{5}{9} - \frac{25\alpha}{21},\ \frac{4}{9} - \frac{20\alpha}{21}\right) \cdot \Pi = \left(\alpha,\ \frac{8\alpha}{7},\ \frac{5}{9} - \frac{25\alpha}{21},\ \frac{4}{9} - \frac{20\alpha}{21}\right).$$

- If $\alpha = 7/15$, then we remain all the time in the (relabeled) ergodic set $\{s_1, s_2\}$. If $\alpha = 0$, then we remain all the time in the (relabeled) ergodic set $\{s_3, s_4\}$.
- There are no absorbing states and there are two (relabeled) ergodic sets: $\{s_1, s_2\}$ and $\{s_3, s_4\}$. All the states are recurrent and aperiodic.

- **Absorbing Markov Chains:**
- **Definition.** A Markov chain is **absorbing** if it has at least one absorbing state and if from every state it is possible to go to an absorbing state (not necessarily in one step).
- **Questions:**
  1. What is the probability that the chain will eventually reach an absorbing state?
  2. On the average, how many times will the chain be in each transient state?
  3. On the average, how long will it take for the chain to be absorbed?
  4. What is the probability that the chain will end up in a given absorbing state?
- See the handout for the answers to these four questions.

- **Irreducible Markov Chains:**
- **Definition.** A Markov chain is called **irreducible** (also called **ergodic**) if it is possible to go from every state to every state (not necessarily in one step).
- **Questions:**
  1. On the average, how long will it take for the chain to reach state s_j for the first time if it started in state s_i?
  2. If the chain is started in state s_i, how long will it take on average to return to s_i for the first time?
- See the handout for the answers to these two questions.

- **General Markov processes:**
- **Definition.** A Markov process (or chain) is a sequence of random objects $\{\tilde{x}_t\}_{t=0}^{\infty}$, $\tilde{x}_t : (\Omega, \mathcal{F}, P) \longrightarrow (\Omega', \mathcal{F}')$, where the conditional distribution of $\tilde{x}_{t+1}$ given $(\tilde{x}_t, \tilde{x}_{t-1}, ..., \tilde{x}_0) = (x_t, x_{t-1}, ..., x_0)$ depends only on the value x_t taken by $\tilde{x}_t$ for $t = 0, 1, ...,$

$$P\left\{\tilde{x}_{t+1} \in B \middle| \tilde{x}_t = x_t, \tilde{x}_{t-1} = x_{t-1}, ..., \tilde{x}_0 = x_0\right\}$$

$$= P\left\{\tilde{x}_{t+1} \in B \middle| \tilde{x}_t = x_t\right\},$$

for all $B \in \mathcal{F}'$, all $(x_0, x_1, ..., x_t) \in \left(\Omega'\right)^{t+1}$, and $t = 0, 1, ...$

- A Markov process $\{\tilde{x}_t\}_{t=0}^{\infty}$ has time-invariant transition if there exists a time-invariant function called the **transition function** $Q : \Omega' \times \mathcal{F}' \longrightarrow [0,1]$, which is given by

$$Q(x, B) = P\{\tilde{x}_{t+1} \in B \,|\tilde{x}_t = x\}, \text{ for all } x \in \Omega', B \in \mathcal{F}', t = 0, 1, ...$$

so that the distribution P_{t+1} of the random object $\tilde{x}_{t+1}$ is given by the following recursive formula (which arises from the theorem of total probability):

$$P_{t+1}(B) = \int_{\Omega'} Q(x, B) dP_t(x), \text{ for all } B \in \mathcal{F}', t = 0, 1, ...$$

- Thus, the stationary distribution P^* of the Markov process satisfies

$$P^*(B) = \int_{\Omega'} Q(x, B) dP^*(x), \text{ for all } B \in \mathcal{F}'.$$

- Assume now that the random objects $\{\tilde{x}_t\}_{t=0}^{\infty}$ are random variables, $\tilde{x}_t : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$.
- The Markov process of random variables $\{\tilde{x}_t\}_{t=0}^{\infty}$ has an ergodic distribution $\widehat{P}$ if $P_t \xrightarrow{w} \widehat{P}$ for all P_0. In this case, $\widehat{P} = P^*$ and the stationary distribution P^* is unique.
- Thus, a Markov process of random variables $\{\tilde{x}_t\}_{t=0}^{\infty}$ having the ergodic distribution $\widehat{P}$ converges in distribution to a random variable $\tilde{x}$ having the distribution $\widehat{P}$, $\tilde{x}_t \xrightarrow{d} \tilde{x}$ or $\tilde{x}_t \longrightarrow \widehat{P}$.

- If the random objects $\{\tilde{x}_t\}_{t=0}^{\infty}$ are absolutely continuous random variables, $\tilde{x}_t : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$, with the density f_t, and there exists a time-invariant conditional density $f_{\tilde{x}'|\tilde{x}}(\cdot\,|x) : \mathbb{R} \longrightarrow \overline{\mathbb{R}}$ such that

$$P\{\tilde{x}_{t+1} \in B \,|\tilde{x}_t = x\} = \int_B f_{\tilde{x}'|\tilde{x}}(x'\,|x)dx',$$

$$\text{for all } x \in \mathbb{R},\ B \in \mathcal{B},\ t = 0, 1, ...,$$

then

$$P_{t+1}(B) = \int_B f_{t+1}(x')dx' = \int_B \int_{\mathbb{R}} f_{\tilde{x}_{t+1},\tilde{x}_t}(x', x)dxdx'$$

$$= \int_B \underbrace{\int_{\mathbb{R}} f_{\tilde{x}'|\tilde{x}}(x'\,|x)f_t(x)dx}_{f_{t+1}(x')}dx', \quad \text{for all } B \in \mathcal{B},\ t = 0, 1, ...$$

- Therefore, the densities evolve according to

$$f_{t+1}(x') = \int_{\mathbb{R}} f_{\tilde{x}'|\tilde{x}}(x'|x) f_t(x) dx, \quad \text{for all } x' \in \mathbb{R}.$$

- Thus, the stationary density f^* of the Markov process satisfies

$$f^*(x') = \int_{\mathbb{R}} f_{\tilde{x}'|\tilde{x}}(x'|x) f^*(x) dx, \quad \text{for all } x' \in \mathbb{R}.$$

6.6. The Poisson distribution as the limit of binomial distributions

Siméon Denis Poisson (1781 - 1840)

- See the handout for the proof of the following statements:
- Let

$$b(x; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, ..., n.$$

be the probability function of a binomial random variable with the parameters n and θ. Then

$$\lim_{\substack{n\to\infty \\ n\theta=\lambda\in(0,\infty)}} b(x; n, \theta) = \lim_{\substack{\theta\to 0 \\ n\theta=\lambda\in(0,\infty)}} b(x; n, \theta) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{for } x = 0,\ 1, \ldots$$

- The random variable $\tilde{x}$ has a Poisson distribution if its probability function is

$$p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{with } \lambda > 0, \quad \text{for } x = 0,\ 1, \ldots$$

- We write $\tilde{x} \sim \mathrm{P}(\lambda)$.
- Thus, if $\tilde{x}_n \sim \mathrm{B}(n, \theta)$, then $\tilde{x}_n \longrightarrow \mathrm{P}(\lambda)$ when $n \longrightarrow \infty$ and $n\theta = \lambda \in (0, \infty)$ (or when $\theta \longrightarrow 0$ and $n\theta = \lambda \in (0, \infty)$).
- Mean and variance of the Poisson distribution:

$$\mu = n\theta = \lambda,$$

and

$$\sigma^2 = n\theta(1 - \theta) = \lambda,$$

since $n\theta = \lambda$ and $\theta \longrightarrow 0$.

- Moment generating function of the Poisson distribution, $M_{\tilde{x}}(t)$:

$$M_{\tilde{x}}(t) = e^{\lambda(e^t - 1)}.$$

- The Poisson approximation to the binomial distribution is good when $n \geq 30$ and $\theta \leq 0.1$.
- **Example.** Let $n = 500$ and $\theta = 0.01$. Thus, $\lambda = n\theta = 5$. Find $P\{\tilde{x} = 7\}$.
- $$P\{\tilde{x} = 7\} \approx p(7;5) = \frac{5^7 e^{-5}}{7!} = 0.10444,$$
- whereas

$$b(7;500,0.01) = \binom{500}{7} (0.01)^7 (0.99)^{493} = 0.10476.$$

Thus, the approximation error is 0.00032.

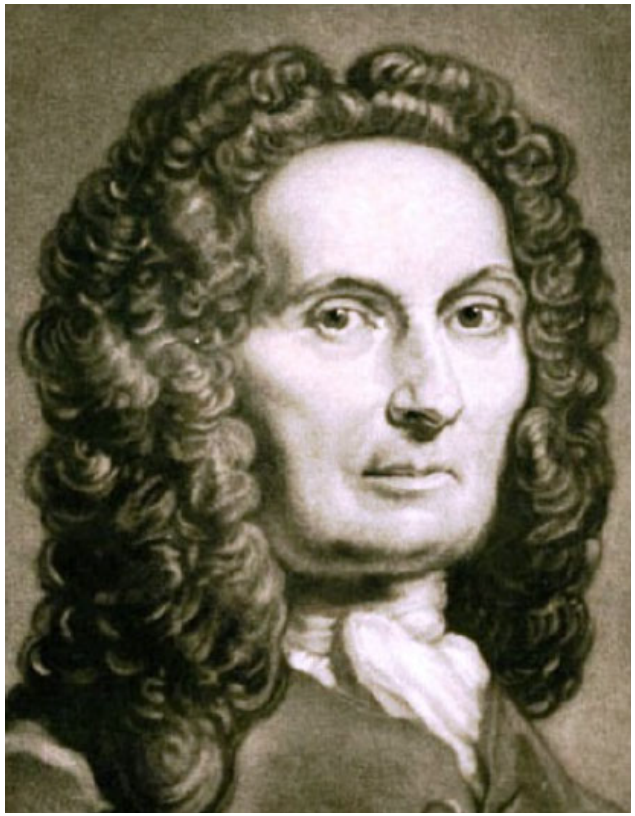
6.7. The standard normal distribution as the limit of standardized binomial distributions

- **Theorem (De Moivre-Laplace).** If $\tilde{x}_n$ is a random variable having a binomial distribution with parameters n and $\theta \in (0,1)$, i.e., its probability function is $b(x; n, \theta)$, then the moment generating function of the standardization of $\tilde{x}_n$,

$$\tilde{z}_n = \frac{\tilde{x}_n - \mathrm{E}(\tilde{x}_n)}{\sqrt{\mathrm{Var}(\tilde{x}_n)}} = \frac{\tilde{x}_n - n\theta}{\sqrt{n\theta(1-\theta)}},$$

  tends to that of the standard normal distribution as $n \longrightarrow \infty$. Hence, $\tilde{z}_n \longrightarrow \mathrm{N}(0,1)$.

- **Proof.** See the handout.



Abraham de Moivre (1667 - 1754)

- Let $\tilde{x} \sim \mathrm{B}(n, \theta)$ and a and b are natural numbers smaller or equal than n, with $a \leq b$. Then, if n is large,

$$P\{a \leq \tilde{x} \leq b\} = P\left\{a - \frac{1}{2} \leq \tilde{x} \leq b + \frac{1}{2}\right\}$$

$$= P\left\{\frac{a - \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}} \leq \frac{\tilde{x} - n\theta}{\sqrt{n\theta(1-\theta)}} \leq \frac{b + \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right\}$$

$$= P\left\{\frac{a - \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}} \leq \tilde{z} \leq \frac{b + \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right\}$$

$$\approx \mathrm{N}\left(\frac{b + \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right) - \mathrm{N}\left(\frac{a - \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right),$$

  where $\mathrm{N}(\cdot)$ is the distribution function of a standard normal random variable $\tilde{z}$.

- The $\pm\frac{1}{2}$ continuity correction becomes less important as n increases when $a < b$.

- Moreover,

$$P\{\tilde{x} = x\} \approx \mathrm{N}\left(\frac{x + \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right) - \mathrm{N}\left(\frac{x - \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right).$$

- The normal approximation to the binomial distribution is good when $n\theta > 5$ and $n(1-\theta) > 5$.
- **Example.** Let $n = 100$ and $\theta = 0.4$ $(n\theta = 40,\ n\theta(1-\theta) = 24)$. Find $P\{\tilde{x} = 35\}$.
-

$$\begin{aligned} P\{\tilde{x} = 35\} &\approx \mathrm{N}\left(\frac{35 + \frac{1}{2} - 40}{\sqrt{24}}\right) - \mathrm{N}\left(\frac{35 - \frac{1}{2} - 40}{\sqrt{24}}\right) \\ &= \mathrm{N}(-0.91856) - \mathrm{N}(-1.1227) = 0.04838, \end{aligned}$$

- whereas

$$b(35; 100, 0.4) = \binom{100}{35} (0.4)^{35} (0.6)^{65} = 0.04913.$$

Thus, the approximation error is 0.00075.

6.8. Weak and strong laws of large numbers

- Let $\{\tilde{x}_i\}_{i=1}^{\infty}$ be a collection of random variables (or stochastic process) defined on the probability space $(\Omega, \mathcal{F}, P)$. Define the random variable "sum", indexed by n, as

$$\tilde{S}_n = \sum_{i=1}^{n} \tilde{x}_i,$$

and the random variable "average", indexed by n, as

$$\overline{\mathbf{x}}_n = \frac{\tilde{S}_n}{n} = \frac{\sum_{i=1}^{n} \tilde{x}_i}{n}.$$

- **Definition.** The collection $\{\tilde{x}_i\}_{i=1}^{\infty}$ of random variables satisfies the weak law of large numbers (WLLN) if $\overline{\mathbf{x}}_n - \mathrm{E}(\overline{\mathbf{x}}_n) \xrightarrow{p} 0$, i.e.,

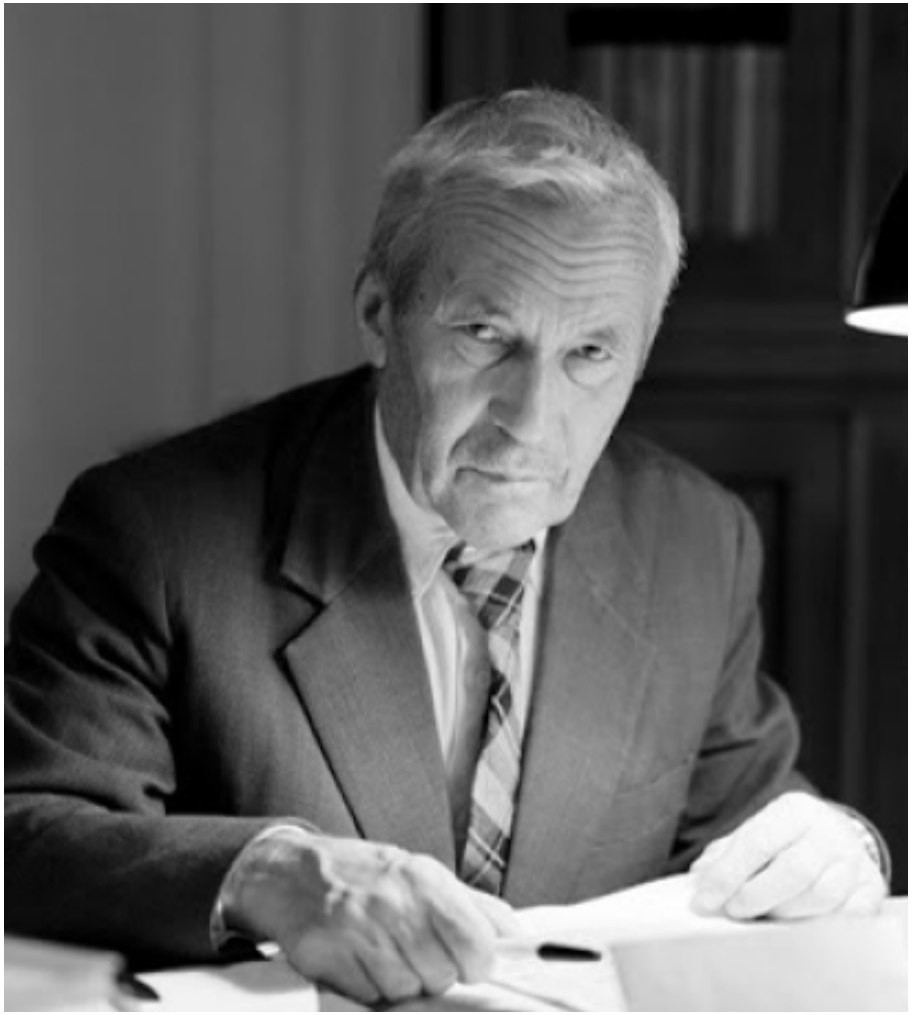
$$\lim_{n\to\infty} P\left\{\left|\overline{\mathbf{x}}_n - \mathrm{E}(\overline{\mathbf{x}}_n) - 0\right| \geq \varepsilon\right\} = 0 \text{ for all } \varepsilon > 0$$

or

$$\lim_{n\to\infty} P\left\{\left|\overline{\mathbf{x}}_n - \mathrm{E}(\overline{\mathbf{x}}_n)\right| \geq \varepsilon\right\} = 0, \text{ for all } \varepsilon > 0.$$

- *Note:* $\{\tilde{x}_i\}_{i=1}^{\infty}$ satisfies the WLLN if and only if $\overline{\mathbf{x}}_n - \mathrm{E}(\overline{\mathbf{x}}_n) \xrightarrow{d} 0$. Moreover, if $\overline{\mathbf{x}}_n - \mathrm{E}(\overline{\mathbf{x}}_n) \xrightarrow{m} 0$, then $\{\tilde{x}_i\}_{i=1}^{\infty}$ satisfies the WLLN.
- **Definition.** The collection $\{\tilde{x}_i\}_{i=1}^{\infty}$ of random variables satisfies the strong law of large numbers (SLLN) if $\overline{\mathbf{x}}_n - \mathrm{E}(\overline{\mathbf{x}}_n) \xrightarrow{a.s} 0$, i.e.,

$$P\left\{\lim_{n\to\infty}\left[\overline{\mathbf{x}}_n - \mathrm{E}(\overline{\mathbf{x}}_n)\right] = 0\right\} = 1.$$



Andrei Kolmogorov (1903 - 1987)

- **Kolmogorov's Theorem 1.** If

  (a) $\{\tilde{x}_i\}_{i=1}^{\infty}$ are independent random variables with $\tilde{x}_i \in \mathrm{L}^2$ for all i, and

  (b)

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}\left(\tilde{x}_i\right)}{i^2} < \infty,$$

  then $\{\tilde{x}_i\}_{i=1}^{\infty}$ satisfies the SLLN.

- *Note:* Under the assumption of independency,

$$\mathrm{Var}\left(\overline{\mathbf{x}}_n\right) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} \tilde{x}_i}{n}\right) = \frac{\sum_{i=1}^{n} \mathrm{Var}\left(\tilde{x}_i\right)}{n^2}.$$

- **Particular cases:**

- **(1)** $\{\tilde{x}_i\}_{i=1}^{\infty}$ are independent and $\text{Var}(\tilde{x}_i) \leq M$ for all i, where $M < \infty$ is a fixed bound.
  Clearly, in this case,

$$\sum_{i=1}^{\infty} \frac{\text{Var}(\tilde{x}_i)}{i^2} \leq \sum_{i=1}^{\infty} \frac{M}{i^2} = M \cdot \left(\sum_{i=1}^{\infty} \frac{1}{i^2} \right) = M \cdot \frac{\pi^2}{6} < \infty.$$

- **(2)** $\{\tilde{x}_i\}_{i=1}^{\infty}$ are independent random variables with $\text{Var}(\tilde{x}_i) = \sigma^2 < \infty$ for all i (this is a particular case of Case 1).

  In this case,

$$\text{Var}(\overline{\mathbf{x}}_n) = \frac{\sum_{i=1}^{n} \text{Var}(\tilde{x}_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

- **(3)** Both (a) and (b) hold and $\mathrm{E}(\tilde{x}_i) = \mu$ for all i.

  In this case,

$$\mathrm{E}(\overline{\mathbf{x}}_n) = \mathrm{E}\left(\frac{\sum_{i=1}^{n} \tilde{x}_i}{n}\right) = \frac{\sum_{i=1}^{n} \mathrm{E}(\tilde{x}_i)}{n} = \frac{n\mu}{n} = \mu, \quad \text{for all } n,$$

  so that $\overline{\mathbf{x}}_n - \mu \xrightarrow{a.s} 0$, which can be written as $\overline{\mathbf{x}}_n \xrightarrow{a.s} \mu$ since $P\left\{\lim_{n\to\infty}(\overline{\mathbf{x}}_n - \mu) = 0\right\} = 1$ is equivalent to $P\left\{\lim_{n\to\infty}\overline{\mathbf{x}}_n = \mu\right\} = 1$.

- Similarly, $\overline{\mathbf{x}}_n - \mu \xrightarrow{p} 0$ can be written as $\overline{\mathbf{x}}_n \xrightarrow{p} \mu$.

- *Note:* If $\{\tilde{x}_i\}_{i=1}^{\infty}$ are independent with the finite mean μ and the finite variance σ^2 for all i, then

$$\mathrm{E}(\overline{\mathbf{x}}_n) = \mu \quad \text{and} \quad \mathrm{Var}(\overline{\mathbf{x}}_n) = \frac{\sigma^2}{n}.$$

- **Kolmogorov's Theorem 2.** If $\{\tilde{x}_i\}_{i=1}^{\infty}$ are i.i.d. random variables with $\tilde{x}_i \in \mathrm{L}^1$ for all i, then $\{\tilde{x}_i\}_{i=1}^{\infty}$ satisfies the SLLN.

- *Notes:*

  (a) Theorem 2 implies that $\overline{\mathbf{x}}_n \xrightarrow{a.s} \mu$, where $\mu = \mathrm{E}(\overline{\mathbf{x}}_n) = \mathrm{E}(\tilde{x}_i)$ for all i.

  (b) Theorem 1 requires finite variance and $\sum_{i=1}^{\infty} \frac{\mathrm{Var}(\tilde{x}_i)}{i^2} < \infty$ whereas Theorem 2 does not (only finite mean is required in Theorem 2).

  (c) Theorem 2 requires i.i.d. whereas Theorem 1 does not (only independency is required in Theorem 1).

- Let us prove the WLLN for the previous particular case 1, i.e., when $\{\tilde{x}_i\}_{i=1}^{\infty}$ are independent and $\text{Var}(\tilde{x}_i) \leq M$ for all i, where $M < \infty$ is a fixed bound.

- First observe that the Chevyshev's inequality,

$$P\left\{\left|\tilde{y} - \mu_{\tilde{y}}\right| \geq k\sigma_{\tilde{y}}\right\} \leq \frac{1}{k^2}, \quad \text{for all } k > 0,$$

can be rewritten as

$$P\left\{\left|\tilde{y} - \mu_{\tilde{y}}\right| \geq \varepsilon\right\} \leq \frac{\sigma_{\tilde{y}}^2}{\varepsilon^2}, \quad \text{for all } \varepsilon > 0.$$

by making the change of variable $k\sigma_{\tilde{y}} = \varepsilon > 0$.

- **Proof.** Make $\tilde{y} = \overline{\mathbf{x}}_n$ so that

$$P\left\{\left|\overline{\mathbf{x}}_n - \mathrm{E}\left(\overline{\mathbf{x}}_n\right)\right| \geq \varepsilon\right\} \leq \frac{\mathrm{Var}\left(\overline{\mathbf{x}}_n\right)}{\varepsilon^2}, \quad \text{for all } \varepsilon > 0.$$

Take limits in both sides,

$$\lim_{n\to\infty} \underbrace{P\left\{\left|\overline{\mathbf{x}}_n - \mathrm{E}\left(\overline{\mathbf{x}}_n\right)\right| \geq \varepsilon\right\}}_{\geq 0} \leq \lim_{n\to\infty} \frac{\mathrm{Var}\left(\overline{\mathbf{x}}_n\right)}{\varepsilon^2}$$

$$= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathrm{Var}\left(\tilde{x}_i\right)}{n^2\varepsilon^2} \leq \lim_{n\to\infty} \frac{nM}{n^2\varepsilon^2} = \lim_{n\to\infty} \frac{M}{n\varepsilon^2} = 0. \qquad \textit{Q.E.D.}$$

- **The LLN for higher moments:** if $\left\{\tilde{x}_i\right\}_{i=1}^{\infty}$ are i.i.d. random variables with $\tilde{x}_i \in \mathrm{L}^k$ for all i and $k = 1, 2, ...$, then the process $\left\{\tilde{x}_i^k\right\}_{i=1}^{\infty}$ is i.i.d. and satisfies the SLLN, that is $m'_{k,n} \xrightarrow{a.s} \mu'_k$, where $m'_{k,n}$ is the average $\frac{\sum_{i=1}^{n} \tilde{x}_i^k}{n}$ and $\mu'_k = \mathrm{E}\left(\tilde{x}_i^k\right)$ is the kth (non-central) moment of $\tilde{x}_i$ for all i.

- **Example:**
- Let $\{\tilde{x}_i\}_{i=1}^{\infty}$ be i.i.d. Bernoulli random variables with parameter θ so that $\mathrm{E}(\tilde{x}_i) = \theta$ and $\mathrm{Var}(\tilde{x}_i) = \theta(1-\theta)$ for all i.
- Then, the random variable $\tilde{S}_n = \sum_{i=1}^{n} \tilde{x}_i$ is binomial, $\tilde{S}_n \sim \mathrm{B}(n, \theta)$.
- Moreover, the random variable

$$\overline{\mathbf{x}}_n = \frac{\tilde{S}_n}{n} = \frac{\sum_{i=1}^{n} \tilde{x}_i}{n}$$

is the percentage (or proportion) of successes in n trials and

$$\mathrm{E}(\overline{\mathbf{x}}_n) = \theta \quad \text{and} \quad \mathrm{Var}(\overline{\mathbf{x}}_n) = \frac{\theta(1-\theta)}{n}.$$

- Therefore, from any of the two Kolmogorov's theorems, we get

$$\overline{\mathbf{x}}_n \xrightarrow{a.s} \theta,$$

i.e., the percentage (or proportion) of successes converges (almost surely) to the probability of success in each trial when $n \longrightarrow \infty$.

- However, it is not true that the number of successes (almost surely) converges to the expected number of successes since the expected number of successes $\mathrm{E}\left(\tilde{S}_n\right) = n\theta$ tends to infinity as $n \longrightarrow \infty$. In fact, even the following is false:

$$\tilde{S}_n - \mathrm{E}\left(\tilde{S}_n\right) \xrightarrow{p} 0 \iff \tilde{S}_n - n\theta \xrightarrow{p} 0. \quad \text{(The fallacy of the LLN)}$$

- Note that, from Chevyshev's theorem, we have

$$P\left(\left|\tilde{S}_n - \mathrm{E}\left(\tilde{S}_n\right)\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}\left(\tilde{S}_n\right)}{\varepsilon^2} = \frac{n\theta(1-\theta)}{\varepsilon^2},$$

and taking limits we get

$$\lim_{n\to\infty} P\left(\left|\tilde{S}_n - n\theta\right| \geq \varepsilon\right) \leq \lim_{n\to\infty} \frac{n\theta(1-\theta)}{\varepsilon^2} = \infty,$$

which is a useless inequality!

- **Application:**
- Consider a set of n individuals (or objects). For each individual i there is a random variable $\tilde{z}_i$, $i = 1, ..., n$. The collection of random variables $\{\tilde{z}_i\}_{i=1}^n$ is i.i.d. and $P\{\tilde{z}_i \in A\} = \theta$ for all i.
- Let $\tilde{x}_n$ be the number of individuals for whom their respective random variables $\tilde{z}_i$, $i = 1, ..., n$, take a value in the set A ($\tilde{z}_i \in A$).
- Then, the expected number of individuals for which $\tilde{z}_i \in A$, $i = 1, ..., n$, is

$$\mathrm{E}(\tilde{x}_n) = n\theta.$$

- Let $\tilde{y}_n = \dfrac{\tilde{x}_n}{n}$ be the percentage (or proportion) of individuals for which $\tilde{z}_i \in A$, $i = 1, ..., n$.
- Then, the expected percentage (or proportion) of individuals for which $\tilde{z}_i \in A$, $i = 1, ..., n$, is

$$\mathrm{E}(\tilde{y}_n) = \theta.$$

- The law of large numbers tells us not only that $\mathrm{E}(\tilde{y}_n) = \theta$ but also that $\tilde{y}_n \xrightarrow{a.s} \theta$ (and, thus, $\tilde{y}_n \xrightarrow{p} \theta$) as $n \to \infty$.

6.9. The Central Limit Theorem (CLT)

- **Theorem (Lindeberg-Lévy).** Let $\{\tilde{x}_i\}_{i=1}^{\infty}$ be i.i.d. random variables with $\mathrm{E}(\tilde{x}_i) = \mu$ and $\mathrm{Var}(\tilde{x}_i) = \sigma^2$, with $0 < \sigma^2 < \infty$, for all i. Then

$$\tilde{z}_n \equiv \frac{\bar{\mathbf{x}}_n - \mathrm{E}(\bar{\mathbf{x}}_n)}{\sqrt{\mathrm{Var}\,(\bar{\mathbf{x}}_n)}} = \frac{\bar{\mathbf{x}}_n - \mu}{\sigma \big/ \sqrt{n}}$$

  converges in distribution to a standard normal random variable as $n \to \infty$. That is, $\tilde{z}_n \longrightarrow \mathrm{N}(0,1)$.

- **Proof:** See the handout.

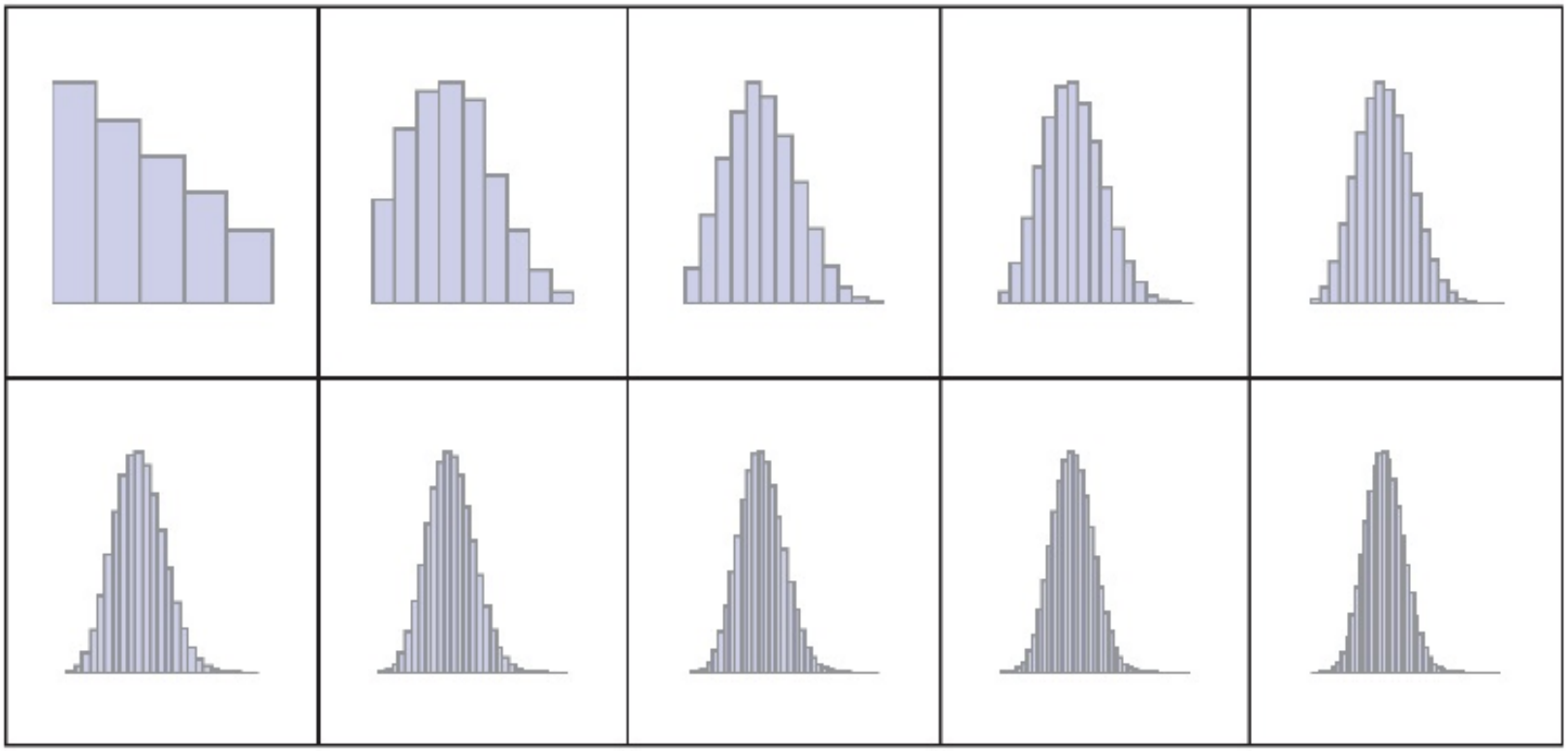
- *Note 1:* The theorem does not say that $\bar{\mathbf{x}}_n \longrightarrow \mathrm{N}\left(\mu, \sigma^2 \big/ n\right)$ since $\sigma^2 \big/ n \longrightarrow 0$ and, thus, $\bar{\mathbf{x}}_n \xrightarrow{d} \mu$, which is a consequence of any of the two Kolmogrov's theorems.

- *Note 2:* $\tilde{z}_n \equiv \dfrac{\bar{\mathbf{x}}_n - \mu}{\sigma \big/ \sqrt{n}} \longrightarrow \mathrm{N}(0,1)$ can be written as

$$\sqrt{n}\,(\bar{\mathbf{x}}_n - \mu) \longrightarrow \mathrm{N}(0, \sigma^2).$$

The evolution of the standardization of $\bar{\mathbf{x}}_n$.

x	1	2	3	4	5
$f_{\tilde{x}}(x)$	0.30	0.25	0.20	0.15	0.10



Jarl W. Lindeberg (1876 – 1932)

- **Other Central Limit Theorems (CLT's):**
- **Theorem (Liapunov).** Let $\{\tilde{x}_i\}_{i=1}^{\infty}$ be independent random variables with finite third moments, $\mathrm{E}(\tilde{x}_i) = \mu_i$, $\mathrm{Var}(\tilde{x}_i) = \sigma_i^2$ with $0 < \sigma_i^2 < \infty$, and $\mathrm{E}[|\tilde{x}_i - \mu_i|^3] = m_{3i}$. If

$$\lim_{j \to \infty} \frac{\left(\sum_{i=1}^{j} m_{3i} \right)^{1/3}}{\left(\sum_{i=1}^{j} \sigma_i^2 \right)^{1/2}} = 0,$$

then $\tilde{z}_n \equiv \dfrac{\bar{\mathbf{x}}_n - \mathrm{E}(\bar{\mathbf{x}}_n)}{\sqrt{\mathrm{Var}\,(\bar{\mathbf{x}}_n)}} \longrightarrow \mathrm{N}(0,1)$.

Aleksandr Liapunov (1857 – 1918)

- **Theorem (Lindeberg-Feller).** Let $\{\tilde{x}_i\}_{i=1}^{\infty}$ be independent random variables with distributions P_i, $\mathrm{E}(\tilde{x}_i) = \mu_i$, and $\mathrm{Var}(\tilde{x}_i) = \sigma_i^2$ with $0 < \sigma_i^2 < \infty$. Define $c_j = \left(\sum_{i=1}^{j} \sigma_i^2 \right)^{1/2}$. If

$$\lim_{j \to \infty} \frac{1}{c_j^2} \sum_{i=1}^{j} \int_{|x-\mu_i| > \varepsilon c_i} (x - \mu_i)^2 dP_i(x) = 0, \quad \text{for every } \varepsilon > 0,$$

then $\tilde{z}_n \equiv \dfrac{\bar{\mathbf{x}}_n - \mathrm{E}(\bar{\mathbf{x}}_n)}{\sqrt{\mathrm{Var}\,(\bar{\mathbf{x}}_n)}} \longrightarrow \mathrm{N}(0,1)$.

- *Note:* The theorems of Lindeberg-Lévy and Liapunov are special cases of the Lindeberg-Feller theorem since the assumptions of any of the first two Central Limit Theorems imply those of the Lindeberg-Feller.

William ("Vilim") Feller (1906 – 1970)

- **Multivariate Central Limit Theorem:**

  Let $\{\tilde{X}_i\}_{i=1}^{\infty}$ be a sequence of k-dimensional random vectors, $\tilde{X}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, ..., \tilde{x}_{ik})^{\top}$. If $\{c^{\intercal}\tilde{X}_i\}_{i=1}^{\infty}$ converges in distribution to a normal random variable for every k-dimensional vector of scalars, $c = (c_1, c_2, ..., c_k)^{\top}$ with $c \neq \underline{0}$, then the limiting distribution of $\{\tilde{X}_i\}_{i=1}^{\infty}$ is multivariate normal.

- Notice that showing convergence in distribution to a normal random variable for each random variable appearing in the random vector $\tilde{X}_i$ separately (that is, showing that $\{\tilde{x}_{ij}\}_{i=1}^{\infty}$ converges in distribution to a normal random variable for $j = 1, 2, ..., k$) is not sufficient for the convergence in distribution of the sequence $\{\tilde{X}_i\}_{i=1}^{\infty}$ of random vectors to a random vector having the multivariate normal distribution.

Proof that convergence in probability implies convergence in distribution.

Step 1. Let $\widetilde{y}$ and $\widetilde{z}$ be random variables on the same probability space $(\Omega, \mathcal{F}, P)$, x a real number and $\varepsilon > 0$; then

$$P\{\widetilde{z} \leq x\} = P\{\widetilde{z} \leq x, \widetilde{y} \leq x + \varepsilon\} + P\{\widetilde{z} \leq x, \widetilde{y} > x + \varepsilon\}$$

$$\leq P\{\widetilde{z} \leq x, \widetilde{y} \leq x + \varepsilon\} + P\{\widetilde{z} \leq x, \widetilde{y} \geq x + \varepsilon\} \leq P\{\widetilde{y} \leq x + \varepsilon\} + P\{\widetilde{z} \leq x, x \leq \widetilde{y} - \varepsilon\}$$

$$\leq P\{\widetilde{y} \leq x + \varepsilon\} + P\{\widetilde{z} - \widetilde{y} \leq -\varepsilon\} \leq P\{\widetilde{y} \leq x + \varepsilon\} + P\{|\widetilde{z} - \widetilde{y}| \geq \varepsilon\}$$

since

$$P\{\widetilde{z} \leq x, x \leq \widetilde{y} - \varepsilon\} = P\{\widetilde{z} \leq x \leq \widetilde{y} - \varepsilon\} \leq P\{\widetilde{z} \leq \widetilde{y} - \varepsilon\} = P\{\widetilde{z} - \widetilde{y} \leq -\varepsilon\}$$

and

$$P\{|\widetilde{z} - \widetilde{y}| \geq \varepsilon\} = P\{\widetilde{z} - \widetilde{y} \geq \varepsilon\} + P\{\widetilde{z} - \widetilde{y} \leq -\varepsilon\} \geq P\{\widetilde{z} - \widetilde{y} \leq -\varepsilon\}.$$

Summing up,

$$P\{\widetilde{z} \leq x\} \leq P\{\widetilde{y} \leq x + \varepsilon\} + P\{|\widetilde{z} - \widetilde{y}| \geq \varepsilon\}.$$

Step 2. Consider now the sequence of random variables $\{\widetilde{x}_n\}_{n=1}^{\infty}$ and the random variable $\widetilde{x}$ on the same probability space $(\Omega, \mathcal{F}, P)$ such that $\widetilde{x}_n \xrightarrow{p} \widetilde{x}$, that is,

$$\lim_{n\to\infty} P\{|\widetilde{x}_n - \widetilde{x}| \geq \varepsilon\} = 0, \quad \text{for all } \varepsilon > 0$$

Using step 1 we obtain, for every $\varepsilon > 0$ and $x \in \mathbb{R}$,

$$P\{\widetilde{x}_n \leq x\} \leq P\{\widetilde{x} \leq x + \varepsilon\} + P\{|\widetilde{x}_n - \widetilde{x}| \geq \varepsilon\}$$

and, equivalently,

$$P\{\widetilde{x} \leq x - \varepsilon\} \leq P\{\widetilde{x}_n \leq x\} + \underbrace{P\{|\widetilde{x} - \widetilde{x}_n| \geq \varepsilon\}}_{=P\{|\widetilde{x}_n - \widetilde{x}| \geq \varepsilon\}}.$$

Combining the previous two inequalities, we have

$$P\{\widetilde{x} \leq x - \varepsilon\} - P\{|\widetilde{x}_n - \widetilde{x}| \geq \varepsilon\} \leq P\{\widetilde{x}_n \leq x\} \leq P\{\widetilde{x} \leq x + \varepsilon\} + P\{|\widetilde{x}_n - \widetilde{x}| \geq \varepsilon\}.$$

Taking the limit when $n \longrightarrow \infty$, and using the fact that $\widetilde{x}_n \xrightarrow{p} \widetilde{x}$, so that $\lim_{n\to\infty} P(|\widetilde{x}_n - \widetilde{x}| \geq \varepsilon) = 0$ for all $\varepsilon > 0$, we obtain

$$P\{\widetilde{x} \leq x - \varepsilon\} \leq \lim_{n\to\infty} P\{\widetilde{x}_n \leq x\} \leq P\{\widetilde{x} \leq x + \varepsilon\},$$

which can be rewritten as

$$F(x-\varepsilon) \leq \lim_{n\to\infty} F_n(x) \leq F(x+\varepsilon),$$

where F_n is the distribution function of $\widetilde{x}_n$ and F is the distribution function of $\widetilde{x}$.

Assume that the distribution function F is continuous at x. Then,

$$\lim_{\varepsilon\to 0} F(x-\varepsilon) \leq \lim_{n\to\infty} F_n(x) \leq \lim_{\varepsilon\to 0} F(x+\varepsilon)$$

becomes

$$F(x) \leq \lim_{n\to\infty} F_n(x) \leq F(x),$$

which means that $\lim_{n\to\infty} F_n(x) = F(x)$ at every continuity point x of F. Therefore, $\widetilde{x}_n \xrightarrow{d} \widetilde{x}$. *Q.E.D.*

– – – – – – – – – – – – – – – – –

Proof that convergence in distribution to a constant implies convergence in probability.

Let $\widetilde{x}_n \xrightarrow{d} a$, where a is a constant. For any $\varepsilon > 0$ we have that

$$P\{|\widetilde{x}_n - a| < \varepsilon\} = P\{a-\varepsilon < \widetilde{x}_n < a+\varepsilon\} = \lim_{z\to(a+\varepsilon)^-} F_n(z) - F_n(a-\varepsilon). \quad (*)$$

Since $\widetilde{x}_n \xrightarrow{d} a$, F_n converges pointwise to

$$F(z) = \begin{cases} 0 & \text{for } z < a \\ \\ 1 & \text{for } z \geq a \end{cases}$$

at every continuity point of F, i.e., for all $z \neq a$. This implies that $F_n(a-\varepsilon)$ converges to 0 since $a-\varepsilon < a$, while $F_n(z)$ converges to 1 for all $z > a$. Hence, $\lim_{n\to\infty}\left(\lim_{z\to(a+\varepsilon)^-} F_n(z)\right) = 1$ since $a+\varepsilon > a$.

Using the previous facts, and taking the limit of $(*)$ when $n \longrightarrow \infty$, we get

$$\lim_{n\to\infty} P\{|\widetilde{x}_n - a| < \varepsilon\} = 1 - 0 = 1, \quad \text{for all } \varepsilon > 0,$$

so that $\widetilde{x}_n \xrightarrow{p} a$. *Q.E.D.*

ABSORBING MARKOV CHAINS

Definition 1. A Markov chain is **absorbing** if it has at least one absorbing state and if from every state it is possible to go to an absorbing state (not necessarily in one step).

Obviously, in an absorbing Markov chain, a state that is not absorbing is transient. When a Markov chain reaches an absorbing state, we shall say that it is absorbed.

The most obvious questions that can be asked about an absorbing Markov chain are: What is the probability that the chain will eventually reach an absorbing state? On the average, how many times will the chain be in each transient state? On the average, how long will it take for the chain to be absorbed? What is the probability that the chain will end up in a given absorbing state? The answers to all these questions depend, in general, on the state from which the chain starts as well as the transition probabilities. The next Theorems 1, 2, 3 and 4 provide the answers to those four questions.

Consider an arbitrary absorbing Markov chain with the state space $S = \{s_1, ..., s_N\}$. Renumber the N states so that the transient states come first. If there are q transient states and d absorbing states ($q + d = N$), the $N \times N$ transition matrix Π of the chain can be written in the following **canonical form**:

$$\Pi = \begin{array}{c} \\ \text{Trans.} \\ \text{Absorb.} \end{array} \begin{array}{c} \begin{array}{cc} \text{Trans.} & \text{Absorb.} \end{array} \\ \begin{pmatrix} Q & D \\ \widehat{0} & \mathrm{I} \end{pmatrix} \end{array}, \tag{1}$$

where I is a $d \times d$ identity matrix, $\widehat{0}$ is a matrix of zeroes (in this case a $d \times q$ matrix), D is a non-zero $q \times d$ matrix and Q is a $q \times q$ matrix. Thus, now the first q states are transient and the last d states are absorbing.

We know that the entry $\pi_{ij,t}$ of the matrix Π^t is the probability of being in state s_j after t steps when the chain is started in state s_i. A standard matrix algebra argument shows that Π^t has the following form:

$$\Pi^t = \begin{array}{c} \\ \text{Trans.} \\ \text{Absorb.} \end{array} \begin{array}{c} \begin{array}{cc} \text{Trans.} & \text{Absorb.} \end{array} \\ \begin{pmatrix} Q^t & * \\ \widehat{0} & \mathrm{I} \end{pmatrix} \end{array},$$

where the asterisk $*$ stands for the $q \times d$ matrix in the upper right-hand corner of Π^t. This submatrix can be written in terms of Q and D, but the expression is not needed at this time. The form of Π^t shows that the entries of Q^t give the probabilities for being in each of the transient states after t steps for each possible

transient starting state. In our first theorem we will prove that the probability of being in the transient states after t steps approaches zero. Thus, every entry of Q^t must approach zero as t goes to infinity.

Theorem 1. In an absorbing Markov chain, the probability that the chain will be absorbed is 1, i.e., $\lim_{t\to\infty} Q^t = \widehat{0}$, where $\widehat{0}$ is now a $q \times q$ matrix of zeroes

Proof. From each non-absorbing state s_i it is possible to reach an absorbing state. Let m_i be the minimum number of steps required to reach an absorbing state, starting from s_i. Let π_i be the probability that, starting from s_i, the chain will not reach an absorbing state in m_i steps. Then $\pi_i < 1$. Let m be the largest of the m_i and let π be the largest of the π_i, $i = 1, ..., q$. The probability of not being absorbed in m steps is less than or equal to π, in $2m$ steps less than or equal to π^2, etc. Since $\pi < 1$, these probabilities tend to 0. Therefore, the probability of not being absorbed in t steps is monotone decreasing in t and tends to 0 as t goes to infinity. Hence, $\lim_{t\to\infty} Q^t = \widehat{0}$. $Q.E.D.$

The next theorem shows that the (i, j)-entry n_{ij} of the $q \times q$ matrix $(\mathrm{I} - Q)^{-1}$ gives the expected number of times that the chain is in the transient state s_j if it is started in the transient state s_i:

Theorem 2. For an absorbing Markov chain the matrix $\mathrm{I} - Q$ has an inverse and

$$(\mathrm{I} - Q)^{-1} = \mathrm{I} + Q + Q^2 + Q^3 + ... \equiv \sum_{k=0}^{\infty} Q^k. \tag{2}$$

The (i, j)-entry n_{ij} of the matrix $(\mathrm{I} - Q)^{-1}$ is the expected number of times the chain is in the transient state s_j given that it starts in the transient state s_i. The initial state is counted when $i = j$.

Proof. In order to prove that the matrix $\mathrm{I} - Q$ has an inverse, we will use the fact that a square matrix G has an inverse if and only if $Gx = \underline{0}$ implies that $x = \underline{0}$, where $\underline{0}$ is a column vector of zeroes. Let $(\mathrm{I} - Q)\, x = \underline{0}$, that is, $x = Qx$. Then, iterating this we see that $x = Q^t x$ for all t. Since $\lim_{t\to\infty} Q^t = \widehat{0}$, we have that $\lim_{t\to\infty} Q^t x = \underline{0}$, which implies that $x = \underline{0}$. Thus, the inverse matrix $(\mathrm{I} - Q)^{-1}$ exists. Note next that

$$(\mathrm{I} - Q)\left(\mathrm{I} + Q + Q^2 + Q^3 + ... + Q^t\right) = \mathrm{I} - Q^{t+1}.$$

Letting t tend to infinity, from Theorem 1 we get

$$(\mathrm{I} - Q)\,(\mathrm{I} + Q + Q^2 + Q^3 + ...) = \mathrm{I} - \lim_{t\to\infty} Q^{t+1} = \mathrm{I} - \widehat{0} = \mathrm{I}$$

so that (2) holds.

Let s_i and s_j be two transient states and assume throughout the remainder of the proof that i and j are fixed. Let $\tilde{x}_t$ be a random variable that equals 1

if the chain, starting from s_i, is in state s_j after t steps, and equals 0 otherwise. For each t, this random variable depends upon both i and j. We have,

$$P\{\tilde{x}_t = 1\} = q_{ij,t}$$

and

$$P\{\tilde{x}_t = 0\} = 1 - q_{ij,t},$$

where $q_{ij,t}$ is the (i,j)-entry of the matrix Q^t. These equations hold for $t = 0$ since $Q^0 = \mathrm{I}$. Therefore, since $\tilde{x}_t$ is a 0-1 random variable, $\mathrm{E}(\tilde{x}_t) = q_{ij,t}$.

The expected number of times the chain is in state s_j in the first k periods, given that it starts in state s_i, is clearly

$$\mathrm{E}(\tilde{x}_0 + \tilde{x}_1 + \tilde{x}_2 + ... + \tilde{x}_k) = \sum_{t=0}^{k} q_{ij,t}.$$

Letting k tend to infinity we have,

$$\mathrm{E}(\tilde{x}_0 + \tilde{x}_1 + \tilde{x}_2 + ...) = \sum_{t=0}^{\infty} q_{ij,t} = n_{ij},$$

where the last equality comes from (2). *Q.E.D.*

Definition 2. Consider an absorbing Markov chain with transition matrix Π having the canonical form (1). The matrix $(\mathrm{I} - Q)^{-1}$ is called the **fundamental matrix** for this absorbing chain.

The following theorem states that the sum of the entries in the ith row of the fundamental matrix $(\mathrm{I} - Q)^{-1}$ is the expected number of periods before the chain is absorbed when it starts in state s_i:

Theorem 3. Let n_i be the expected number of periods before the chain is absorbed, given that the chain starts in the transient state s_i, and let n be the column vector of dimension q whose ith entry is n_i. Then, $n = (\mathrm{I} - Q)^{-1}\,\underline{1}$, where $\underline{1}$ is a column vector whose entries are ones.

Proof. If we add all the entries in the ith row of $(\mathrm{I} - Q)^{-1}$, we will obtain the expected number of times in any of the transient states for a given starting state s_i, that is, the expected time required before being absorbed. Thus, n_i is the sum of the entries in the ith row of $(\mathrm{I} - Q)^{-1}$. If we write this statement in matrix form, we obtain the theorem. *Q.E.D.*

Our last theorem gives the probability that the chain will end up in a given absorbing state.

Theorem 4. Let b_{ij} be the probability that an absorbing chain will be absorbed in the absorbing state s_j if it starts in the transient state s_i. Let B be the $q \times d$

matrix with entries b_{ij}. Then, $B = (\mathrm{I} - Q)^{-1} D$, where D is as in the canonical form (1).

Proof. Let $q_{im,t}$ be the (i,m)-entry of the matrix Q^t and d_{mj} the (m,j)-entry of the matrix D, $i = 1, ..., q$, $m = 1, ..., q$, and $j = 1, ..., d$. The probability b_{ij} may be computed as follows: find the probability of being absorbed by state s_j after spending exactly t steps in transient states and being in the transient state s_m in the last step before absorption. This probability is $q_{im,t} d_{mj}$. Then, add over all transient states before absorption and all number of steps,

$$b_{ij} = \sum_{t=0}^{\infty} \sum_{m=1}^{q} q_{im,t} d_{mj} = \sum_{m=1}^{q} \sum_{t=0}^{\infty} q_{im,t} d_{mj} = \sum_{m=1}^{q} n_{im} d_{mj},$$

where the last equality comes from Theorem 2 and the last expression is the (i,j)-entry of the $q \times d$ matrix $(\mathrm{I} - Q)^{-1} D$. $Q.E.D.$

An implication of the previous theorem is that

$$\lim_{t \to \infty} \Pi^t = \widehat{\Pi} = \begin{pmatrix} \widehat{0}_{q \times q} & B_{q \times d} \\ \\ \widehat{0}_{d \times q} & \mathrm{I}_{d \times d} \end{pmatrix},$$

which means that the limiting matrix $\widehat{\Pi}$ exists for absorbing chains.

Example 1. Consider the Markov chain with the following transition matrix (Case II in the class notes):

$$\Pi = \begin{pmatrix} Q_{1 \times 1} & D_{1 \times 2} \\ \\ \widehat{0}_{2 \times 1} & \mathrm{I}_{2 \times 2} \end{pmatrix} = \begin{pmatrix} 2/5 & 1/5 & 2/5 \\ \\ 0 & 1 & 0 \\ \\ 0 & 0 & 1 \end{pmatrix}.$$

Since the matrix Π has the canonical form (1), the Markov chain is absorbing. It has one transient state (s_1) and two absorbing states (s_2 and s_2). Moreover, for this matrix we found that

$$\lim_{t \to \infty} \Pi^t = \widehat{\Pi} = \begin{pmatrix} \widehat{0}_{1 \times 1} & B_{1 \times 2} \\ \\ \widehat{0}_{2 \times 1} & \mathrm{I}_{1 \times 2} \end{pmatrix} = \begin{pmatrix} 0 & 1/3 & 2/3 \\ \\ 0 & 1 & 0 \\ \\ 0 & 0 & 1 \end{pmatrix}.$$

This chain is absorbed with probability 1 (Theorem 1) and the expected number of times the chain will be in the unique transient state s_1, given that

it starts in s_1, is the first (and unique) entry of the 1×1 fundamental matrix $(\mathrm{I} - Q)^{-1}$,

$$n_1 = \left(1 - \frac{2}{5}\right)^{-1} = \frac{5}{3} = 1.666\,7,$$

as follows from Theorem 2. From Theorem 3 we can compute the expected number of periods before the chain is absorbed, given that the chain starts in the unique transient state s_1,

$$(\mathrm{I} - Q)^{-1}\,\underline{1} = \left(1 - \frac{2}{5}\right)^{-1} \cdot 1 = \frac{5}{3} = 1.6667,$$

which in this case coincides with n_1 since there is only one transient state.

Finally, using Theorem 4 we can compute the probabilities that the chain will be absorbed in the absorbing states s_2 and s_3 if it starts in the transient state s_1,

$$B = (\mathrm{I} - Q)^{-1}\,D = \left(1 - \frac{2}{5}\right)^{-1} \begin{pmatrix} 1/5 & , & 2/5 \end{pmatrix} = \begin{pmatrix} 1/3 & , & 2/3 \end{pmatrix}$$

so that the probabilities that the chain will be absorbed in the absorbing states s_2 and s_3 if it starts in the transient state s_1 are $1/3$ and $2/3$, respectively. This result agrees with the values of the entries of the 1×2 matrix B appearing in the limiting matrix $\widehat{\Pi}$.

IRREDUCIBLE MARKOV CHAINS

Definition 3. A Markov chain is called **irreducible** (also called **ergodic**) if it is possible to go from every state to every state (not necessarily in one step).

The most obvious questions that can be asked about an irreducible Markov chain are: If the chain is started in state s_i, how long will it take on average to return to s_i for the first time? On the average, how long will it take for the chain to reach state s_j for the first time if it started in state s_i? The next Theorems 5 and 6 provide the answers to those two questions.

Obviously, a Markov chain is irreducible if its transition matrix Π is irreducible, that is, if Π cannot be written in the following block upper-triangular form by just renumbering the states:

$$\begin{pmatrix} A & C \\ \widehat{0} & E \end{pmatrix},$$

where A and E are square matrices (possibly of different size so that the matrix $\widehat{0}$ of zeroes does not need to be square).

Remember that a stationary probabilistic row vector satisfies $p^* = p^*\Pi$, that is, p^* is a left-eigenvector associated with the eigenvalue 1. An irreducible chain has a unique stationary probabilistic row vector p^*, which is equivalent to say that the irreducible transition matrix Π has only one eigenvalue equal to 1 (see the class notes). Moreover, if the chain is irreducible then $p^* \gg 0$.

Definition 4. A Markov chain is called **regular** if its transition matrix Π is regular (i.e., there exists a natural number t for which the matrix Π^t has only strictly positive elements).

In other words, when a chain is regular, for some t, it is possible to go from any state to any state in exactly t steps. Therefore, regular Markov chains are irreducible. However, not all irreducible chains are regular. For example, a Markov chain with the transition matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

is irreducible but it is not regular (Π^t contains always zeroes for all natural t). In this example, it is possible to move from any state to any state but, if t is odd, then it is not possible to move from state s_1 to state s_1 in t steps, and if t is even, then it is not possible to move from state s_1 to state s_2 in t steps. Thus, the chain is not regular. For regular Markov chains the ergodic probabilistic vector $\widehat{p}$ exists and, obviously, $\widehat{p} = p^* \gg 0$.

Note that, for a given state s_j, if we start in any state an irreducible chain will eventually reach state s_j. In fact, it will be in state s_j infinitely often.

Definition 6. If an irreducible Markov chain is started in state s_i, the expected number of steps to reach state s_j for the first time is called the **mean first passage time** from s_i to s_j. It is denoted by m_{ij}. By convention, $m_{ii} = 0$.

To find the mean first passage time from s_i to s_j for an irreducible chain we can use the following method: form a new Markov chain by making s_j an absorbing state, that is, define $\pi_{jj} = 1$. If we start at any state other than s_j, this new process will behave exactly like the original chain up to the first time that state s_j is reached. Since the original chain was an irreducible chain, it was possible to reach s_j from any other state. Thus the new chain is an absorbing chain with a single absorbing state s_j that will eventually be reached from any other initial state s_i with $i \neq j$. Let $(\mathrm{I} - Q)^{-1}$ be the fundamental matrix for the new absorbing chain. The entries of $(\mathrm{I} - Q)^{-1}$ give then expected number of times in each state before absorption. In terms of the original chain, these quantities give the expected number of times in each of the states before reaching state s_j for the first time. The ith component of the q dimension column vector $(\mathrm{I} - Q)^{-1}\,\underline{1}$ gives the expected number of steps before absorption in the new chain, starting in state s_i. In terms of the old chain, this is the expected number of steps

required to reach state s_j for the first time starting at state s_i, i.e., it is the mean first passage time from s_i to s_j. This procedure for computing the mean passage time m_{ij} involves the construction of a new absorbing chain for each state s_j to be reached. We will present a more direct procedure in Theorem 6.

Assume that we start in state s_i and consider the length of time before we return to s_i for the first time. If the chain is irreducible, it is clear that we must return, since we either stay at s_i the first step or go to some other state s_j, and from any other state s_j we will eventually reach s_i. Therefore, all the states of an irreducible Markov chain are recurrent.

Definition 6. If an irreducible Markov chain is started in state s_i, the expected number of steps to return to s_i for the first time is the **mean recurrence time** for s_i. It is denoted by r_i.

Consider the mean first passage time from s_i to s_j and assume that $i \neq j$. This may also be computed as follows: take the expected number of steps required given the outcome of the first step, multiply by the probability that this outcome occurs, and add. If the first step is to s_j, the expected number of steps required is 1; if it is to some other state s_k, the expected number of steps required is m_{kj} plus 1 for the step already taken. Thus,

$$m_{ij} = \pi_{ij} + \sum_{k \neq j} \pi_{ik} \left(m_{kj} + 1 \right),$$

or, since $\sum_{k=1}^{N} \pi_{ik} = 1$,

$$m_{ij} = 1 + \sum_{k \neq j} \pi_{ik} m_{kj}. \tag{3}$$

Similarly, starting in s_i, it must take at least one step to return. Considering all possible first steps gives us the mean recurrence time for state s_i,

$$r_i = \sum_{k=1}^{N} \pi_{ik} \left(m_{ki} + 1 \right) = 1 + \sum_{k=1}^{N} \pi_{ik} m_{ki}. \tag{4}$$

Let us now define two matrices M and R. The (i, j)-entry m_{ij} of M is the mean first passage time to go from from s_i to s_j if $i \neq j$ and its diagonal entries are 0. The matrix M is called the mean first passage matrix. The matrix R is a matrix with all entries equal to 0 except the diagonal entries, which are $r_{ii} = r_i$, i.e., the mean recurrence times for each state. The matrix R is called the mean recurrence matrix. Let $\widehat{1}$ be an $N \times N$ matrix with all entries equal to 1. Using (3) for the case $i \neq j$ and (4) for the case $i = j$, we obtain the following matrix equation:

$$M + R = \Pi M + \widehat{1}$$

or

$$(\mathrm{I} - \Pi) M = \widehat{1} - R. \tag{5}$$

The previous equation with $m_{ii} = 0$ implies (3) and (4). The next theorem gives the mean recurrence time for a given state.

Theorem 5. For an irreducible Markov chain, the mean recurrence time for state s_i is $r_i = 1/p_i^*$, where p_i^* is the ith component of the unique stationary probabilistic row vector p^* for the transition matrix Π.

Proof. Multiplying both sides of (5) by p^*, and using the fact that

$$p^*(\mathrm{I} - \Pi) = \underline{0}^\intercal,$$

we get

$$p^*\widehat{1} - p^*R = \underline{0}^\intercal.$$

Here $p^*\widehat{1}$ is a row vector with all entries equal to 1 (since p^* is a probabilistic vector) and p^*R is a row vector with ith entry equal $p_i^* r_i$ (since R is a diagonal matrix). Thus,

$$(1, 1, ..., 1) = (p_1^* r_1, p_2^* r_2, ..., p_N^* r_N)$$

so that $r_i = 1/p_i^*$, as was to be proved. $Q.E.D.$

A direct corollary of the previous theorem is that, for an irreducible Markov chain, the entries of the stationary probabilistic row vector p^* are strictly positive. This is so because we know that the values of r_i are finite and, hence, $p_i^* = 1/r_i$ cannot be equal to 0.

We next present a more direct procedure to calculate the mean passage matrix M based on the use of the fundamental matrix for irreducible Markov chains. In order to state the main result in Theorem 6 we need two previous results (a proposition and a lemma). The proofs of these Proposition and Lemma rely only on the assumption that the transition matrix Π of the Markov chain has a only one eigenvalue equal to 1 or, equivalently, that the chain has only one stationary probabilistic row vector p^*. This assumption is satisfied by all irreducible Markov chains.

Proposition (Existence of the fundamental matrix). Assume that the Markov chain has only one stationary probabilistic row vector p^* and let Π^* be a stochastic matrix all of whose rows are the stationary probabilistic row vector p^*. The matrix $\mathrm{I} - \Pi + \Pi^*$ has an inverse $(\mathrm{I} - \Pi + \Pi^*)^{-1}$.

Proof. Let x be a column vector such that

$$(\mathrm{I} - \Pi + \Pi^*)\, x = \underline{0}. \tag{6}$$

To prove the proposition, it is sufficient to show that x must be the zero vector. Multiplying (6) by the stationary probabilistic vector p^* and using the fact that p^* satisfies $p^*(\mathrm{I} - \Pi) = \underline{0}^\intercal$ and $p^*\Pi^* = p^*$, we get

$$p^*\,(\mathrm{I} - \Pi + \Pi^*)\, x = p^* x = 0.$$

Therefore, since the entries of the vector $\Pi^* x$ are all $p^* x = 0$, we have that $\Pi^* x = \underline{0}$. Thus, (6) becomes simply $(\mathrm{I} - \Pi)\, x = \underline{0}$ or $(\Pi - \mathrm{I})\, x = \underline{0}$ so that x is a column vector belonging to the (right-)eigenspace of the matrix Π. Remember from the class notes that, since the irreducible transition matrix Π has only one eigenvalue equal to 1, the (right-)eigenspace associated with this eigenvalue consist of all the column vectors with identical values in its cells $x = (\alpha, \alpha, ..., \alpha)^\intercal$. This means that all the entries of the vector x must have the same sign. Since $p^* > 0$ and $p^* x = 0$, then $x = \underline{0}$.[1] *Q.E.D.*

Definition 7. Consider an irreducible Markov chain with the transition matrix Π and the stationary probabilistic row vector p^*. The matrix $(\mathrm{I} - \Pi + \Pi^*)^{-1}$ is called the **fundamental matrix** for this irreducible Markov chain.

Note that the fundamental matrix for an irreducible Markov chain, $(\mathrm{I} - \Pi + \Pi^*)^{-1}$, has not the same expression as the fundamental matrix for an absorbing Markov chain, $(\mathrm{I} - Q)^{-1}$.

Lemma (Properties of the fundamental matrix).

$$(\mathrm{I} - \Pi + \Pi^*)^{-1} \underline{1} = \underline{1} \text{ or, equivalently, } (\mathrm{I} - \Pi + \Pi^*)^{-1} \widehat{1} = \widehat{1} \qquad \text{(a)}$$

$$p^* (\mathrm{I} - \Pi + \Pi^*)^{-1} = p^*, \text{ or, equivalently, } \Pi^* (\mathrm{I} - \Pi + \Pi^*)^{-1} = \Pi^* \qquad \text{(b)}$$

and

$$(\mathrm{I} - \Pi + \Pi^*)^{-1} (\mathrm{I} - \Pi) = \mathrm{I} - \Pi^*. \qquad \text{(c)}$$

Proof. Since, $\mathrm{I}\underline{1} = \underline{1}$, $\Pi\underline{1} = \underline{1}$ and $\Pi^*\underline{1} = \underline{1}$,

$$\underline{1} = (\mathrm{I} - \Pi + \Pi^*)\, \underline{1}.$$

If we premultiply both sides of this equation by $(\mathrm{I} - \Pi + \Pi^*)^{-1}$, we obtain (a).

Similarly, since $p^*\mathrm{I} = p^*$, $p^*\Pi = p^*$ and $p^*\Pi^* = p^*$,

$$p^* = p^* (\mathrm{I} - \Pi + \Pi^*).$$

If we postmultiply both sides of this equation by $(\mathrm{I} - \Pi + \Pi^*)^{-1}$, we obtain (b).

Finally, since both Π and Π^* are stochastic matrices and Π^* has identical rows, we have that $\Pi\Pi^* = \Pi^*$ and $\Pi^*\Pi^* = \Pi^*$. Therefore,

$$(\mathrm{I} - \Pi + \Pi^*)(\mathrm{I} - \Pi^*) = \mathrm{I} - \Pi^* - \Pi + \Pi\Pi^* + \Pi^* - \Pi^*\Pi^*$$

$$= \mathrm{I} - \Pi^* - \Pi + \Pi^* + \Pi^* - \Pi^* = \mathrm{I} - \Pi.$$

If we premultiply both sides of this equation by $(\mathrm{I} - \Pi + \Pi^*)^{-1}$, we obtain (c). *Q.E.D.*

[1] Remember that $p^* > 0$ means that $p_i \geq 0$ for $i = 1, ..., N$ and $p_i > 0$ for at least one i. This has to be the case for a probabilistic vector since $\sum_{i=1}^{N} p_i = 1$.

Theorem 6. The mean first passage matrix M for an irreducible chain is determined from the fundamental matrix $(\mathrm{I}-\Pi+\Pi^*)^{-1}$ and the stationary probabilistic row vector p^* by

$$m_{ij}=\frac{z_{jj}-z_{ij}}{p_j^*}, \tag{7}$$

where z_{ij} is the (i,j)-entry of the fundamental matrix $(\mathrm{I}-\Pi+\Pi^*)^{-1}$ and p_j^* is the jth entry of the stationary probabilistic row vector p^*.

Proof. Premultiplying both sides of (5) by the fundamental matrix $(\mathrm{I}-\Pi+\Pi^*)^{-1}$ we get

$$(\mathrm{I}-\Pi+\Pi^*)^{-1}(\mathrm{I}-\Pi)M=(\mathrm{I}-\Pi+\Pi^*)^{-1}\widehat{1}-(\mathrm{I}-\Pi+\Pi^*)^{-1}R,$$

and from (a) and (c) in the Lemma, the previous equation becomes

$$(\mathrm{I}-\Pi^*)M=\widehat{1}-(\mathrm{I}-\Pi+\Pi^*)^{-1}R$$

or

$$M=\widehat{1}-(\mathrm{I}-\Pi+\Pi^*)^{-1}R+\Pi^*M.$$

From this equation and the definition of the diagonal matrix R, we see that

$$m_{ij}=1-z_{ij}r_j+(p^*M)_j, \tag{8}$$

where $(p^*M)_j$ is the jth entry of the row vector p^*M. Since $m_{jj}=0$, the previous expression becomes

$$0=1-z_{jj}r_j+(p^*M)_j$$

or

$$(p^*M)_j=z_{jj}r_j-1. \tag{9}$$

Plugging (9) into (8), we have

$$m_{ij}=(z_{jj}-z_{ij})r_j.$$

and, since from Theorem 5 $r_i=1/p_i^*$, we obtain (7). *Q.E.D.*

Example 2. Consider the Markov chain with the following regular transition matrix (Case I in the class notes):

$$\Pi=\begin{pmatrix} 3/10 & 0 & 7/10 \\ 4/5 & 1/5 & 0 \\ 0 & 3/5 & 2/5 \end{pmatrix}.$$

For this example we found that

$$p^* = \widehat{p} = \left(\frac{24}{73},\ \frac{21}{73},\ \frac{28}{73}\right) \gg 0.$$

Therefore, for instance, the mean recurrence time for s_2 is $r_2 = 1/p_2^* = 73/21 = 3.4762$. We can also compute the fundamental matrix for this chain,

$$(\mathrm{I} - \Pi + \Pi^*)^{-1} = \begin{pmatrix} 1 - \frac{3}{10} + \frac{24}{73} & 0 - 0 + \frac{21}{73} & 0 - \frac{7}{10} + \frac{28}{73} \\ 0 - \frac{4}{5} + \frac{24}{73} & 1 - \frac{1}{5} + \frac{21}{73} & 0 - 0 + \frac{28}{73} \\ 0 - 0 + \frac{24}{73} & 0 - \frac{3}{5} + \frac{21}{73} & 1 - \frac{2}{5} + \frac{28}{73} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} \frac{751}{730} & \frac{21}{73} & -\frac{231}{730} \\ -\frac{172}{365} & \frac{397}{365} & \frac{28}{73} \\ \frac{24}{73} & -\frac{114}{365} & \frac{359}{365} \end{pmatrix}^{-1} = \begin{pmatrix} \frac{4342}{5329} & -\frac{672}{5329} & \frac{1659}{5329} \\ \frac{2152}{5329} & \frac{4073}{5329} & -\frac{896}{5329} \\ -\frac{768}{5329} & \frac{1518}{5329} & \frac{4579}{5329} \end{pmatrix}.$$

So, for example, the mean first passage time from s_1 to s_2, is

$$m_{12} = \frac{z_{22} - z_{12}}{p_2^*} = \frac{\dfrac{4073}{5329} - \left(-\dfrac{672}{5329}\right)}{\dfrac{21}{73}} = \frac{65}{21} = 3.0952.$$

Example 3. Consider the Markov chain with the following transition matrix (Case III in the class notes):

$$\Pi = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 3/10 & 2/10 & 1/2 \end{pmatrix}.$$

This chain is not irreducible since, if the chain starts in either state s_1 or state s_2, it will never reach state s_3. The set $\{s_1, s_2\}$ is ergodic, whereas the state s_3 is transient. Note that the previous matrix it is not irreducible since by permuting states s_1 and s_3 we obtain the following block upper-triangular matrix:

$$\begin{pmatrix} A & C \\ \widehat{0} & E \end{pmatrix} = \begin{pmatrix} 1/2 & 2/10 & 3/10 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

For this example we found that the matrix Π has the eigenvalues $1, -1$ and 0.5 so that it has a unique eigenvalue equal to 1. Therefore, the chain has a unique stationary probabilistic row vector, which is

$$p^* = \left(\frac{1}{2},\ \frac{1}{2},\ 0\right).$$

However, the stationary probabilistic vector does not have all its entries strictly positive.

Finally, this Markov chain is not absorbing either: it has no absorbing states and its transition matrix Π cannot be rewritten as in the canonical form (1) since the matrix E is not the identity matrix I. Nevertheless, the chain ends up being absorbed by the ergodic set $\{s_1, s_2\}$ and, after being absorbed, it oscillates between states s_1 and s_2 forever.

The Poisson distribution as a limit of binomial distributions

$$\lim_{\substack{n\to\infty \\ n\theta=\lambda\in(0,\infty)}} b(x;n,\theta) = \lim_{\substack{\theta\to 0 \\ n\theta=\lambda\in(0,\infty)}} b(x;n,\theta) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{for } x = 0,\, 1,\, \ldots$$

Proof: Since $\theta = \lambda/n$,

$$b(x;n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}\left(\frac{\lambda}{n}\right)^x(1-\theta)^{n-x}$$

$$= \frac{n(n-1)(n-2)\cdot\cdots\cdot(n-x+1)}{x!}\frac{\lambda^x}{n^x}\left[(1-\theta)^{-n/\lambda}\right]^{-\lambda}(1-\theta)^{-x}$$

$$= 1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdot\cdots\cdot\left(1-\frac{x-1}{n}\right)\frac{\lambda^x}{x!}\left[(1-\theta)^{-1/\theta}\right]^{-\lambda}(1-\theta)^{-x}.$$

Therefore,

$$\lim_{\substack{n\to\infty \\ \theta\to 0 \\ n\theta=\lambda\in(0,\infty)}} b(x;n,\theta) = 1\cdot\frac{\lambda^x}{x!}\cdot\lim_{\theta\to 0}\left[(1-\theta)^{-1/\theta}\right]^{-\lambda}\cdot 1$$

$$= \frac{\lambda^x}{x!}\cdot\lim_{\theta\to 0}\left[(1-\theta)^{-1/\theta}\right]^{-\lambda} = \frac{\lambda^x}{x!}\left[\lim_{\theta\to 0}(1-\theta)^{-1/\theta}\right]^{-\lambda} = \frac{\lambda^x}{x!}e^{-\lambda},$$

where the last equality follows from

$$\lim_{\theta\to 0}(1-\theta)^{-1/\theta} = \lim_{\theta\to 0}\left(e^{\ln\left[(1-\theta)^{-1/\theta}\right]}\right) = \lim_{\theta\to 0}\left[e^{-\frac{1}{\theta}\ln(1-\theta)}\right] = e^{-\lim\limits_{\theta\to 0}\left[\frac{1}{\theta}\ln(1-\theta)\right]} = e$$

since, using the l'Hôpital's rule, we can show that $\lim\limits_{\theta\to 0}\left[\frac{1}{\theta}\ln(1-\theta)\right] = -1$. *Q.E.D.*

– – – – – – – – – – – – – – – – – –

The random variable $\tilde{x}$ has a Poisson distribution if its probability function is

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ with } \lambda > 0, \quad \text{for } x = 0,\, 1,\, \ldots$$

We write $\widetilde{x} \sim \mathrm{P}(\lambda)$.

Thus, if $\widetilde{x}_n \sim \mathrm{B}(n, \theta)$, then $x_n \to \mathrm{P}(\lambda)$ when $n \to \infty$ and $n\theta = \lambda \in (0, \infty)$ (or when $\theta \to 0$ and $n\theta = \lambda \in (0, \infty)$).

– – – – – – – – – – – – – – – – – –

Mean and variance of the Poisson distribution:

$$\mu = n\theta = \lambda,$$

and

$$\sigma^2 = n\theta(1-\theta) = \lambda,$$

since $n\theta = \lambda$ and $\theta \to 0$.

– – – – – – – – – – – – – – – – – –

Moment-generating function of the Poisson distribution, $M_{\tilde{x}}(t)$.

We compute the MGF $M_{\tilde{x}}(t)$ as the limit of the moment-generating function of the binomial distribution, $[1 + \theta(e^t - 1)]^n$.

$$M_{\tilde{x}}(t) = \lim_{\substack{n\to\infty \\ n\theta=\lambda\in(0,\infty)}} \left[1 + \theta(e^t - 1)\right]^n = \lim_{n\to\infty} \left[1 + \frac{\lambda}{n}(e^t - 1)\right]^n$$

$$= \lim_{n\to\infty} \left(\left[1 + \frac{1}{\dfrac{n}{\lambda(e^t-1)}} \right]^{\frac{n}{\lambda(e^t-1)}} \right)^{\lambda(e^t-1)} = \lim_{z\to\infty} \left[\left(1 + \frac{1}{z}\right)^z \right]^{\lambda(e^t-1)}$$

$$= \left[\lim_{z\to\infty} \left(1 + \frac{1}{z}\right)^z \right]^{\lambda(e^t-1)} = e^{\lambda(e^t-1)},$$

where $z = \dfrac{n}{\lambda(e^t-1)} \to \infty$ as $n \to \infty$.

– – – – – – – – – – – – – – – – – –

Exercise: Check that $M_{\tilde{x}}(t) = \sum_{x=0}^{\infty} e^{tx} p(x;\, \lambda) = e^{\lambda(e^t-1)}$ and use $M_{\tilde{x}}(t)$ to find μ and σ^2.

– – – – – – – – – – – – – – – – – –

The standard normal distribution as a limit of standardized binomial distributions

Theorem (De Moivre-Laplace). If $\tilde{x}_n$ is a random variable having a binomial distribution with parameters n and $\theta \in (0,1)$, i.e., its probability function is $b(x;n,\theta)$, then the moment generating function of the standardization of $\tilde{x}_n$,

$$\tilde{\tilde{z}}_n = \frac{\tilde{\tilde{x}}_n - \mathrm{E}\left(\tilde{\tilde{x}}_n\right)}{\sqrt{\mathrm{Var}\left(\tilde{\tilde{x}}_n\right)}} = \frac{\tilde{x}_n - n\theta}{\sqrt{n\theta(1-\theta)}}$$

tends to that of the standard normal distribution as $n \to \infty$. Hence, $\tilde{z}_n \to \mathrm{N}(0,1)$.

Proof: Let $n\theta \equiv \mu$ and $n\theta(1-\theta) \equiv \sigma^2$. Then we know that

$$M_{\tilde{z}_n}(t) = M_{\frac{\tilde{x}_n - \mu}{\sigma}}(t) = e^{-\frac{\mu t}{\sigma}} M_{\tilde{x}_n}\left(\frac{t}{\sigma}\right) = e^{-\frac{\mu t}{\sigma}}[1 + \theta(e^{t/\sigma} - 1)]^n.$$

Then, taking logarithms and substituting the MacLaurin series (or Taylor expansion around zero) of $e^{t/\sigma}$, we get

$$\ln M_{\tilde{z}_n}(t) = -\frac{\mu t}{\sigma} + n\ln[1 + \theta(e^{t/\sigma} - 1)]$$

$$= -\frac{\mu t}{\sigma} + n\ln\left[1 + \theta\underbrace{\left\{\overbrace{\frac{1}{1!}\left(\frac{t}{\sigma}\right) + \frac{1}{2!}\left(\frac{t}{\sigma}\right)^2 + \frac{1}{3!}\left(\frac{t}{\sigma}\right)^3 + \ldots}^{e^{t/\sigma}-1}\right\}}_{x}\right]$$

and using the infinite series $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \ldots$, which converges for $|x| < 1$, to expand this logarithm, it follows that

$$\ln M_{\tilde{z}_n}(t) = -\frac{\mu t}{\sigma} + n\theta\left[\frac{t}{\sigma} + \frac{1}{2}\left(\frac{t}{\sigma}\right)^2 + \frac{1}{6}\left(\frac{t}{\sigma}\right)^3 + \ldots\right]$$

$$-\frac{n\theta^2}{2}\left[\frac{t}{\sigma} + \frac{1}{2}\left(\frac{t}{\sigma}\right)^2 + \frac{1}{6}\left(\frac{t}{\sigma}\right)^3 + \ldots\right]^2$$

$$+\frac{n\theta^3}{3}\left[\frac{t}{\sigma} + \frac{1}{2}\left(\frac{t}{\sigma}\right)^2 + \frac{1}{6}\left(\frac{t}{\sigma}\right)^3 + \ldots\right]^3 - \ldots$$

Collecting powers of t, we obtain

$$\ln M_{\tilde{z}_n}(t) = \underbrace{\left(-\frac{\mu}{\sigma} + \frac{n\theta}{\sigma}\right)}_{=0} t + \left(\frac{n\theta}{2\sigma^2} - \frac{n\theta^2}{2\sigma^2}\right) t^2$$

$$+ \left(\frac{n\theta}{6\sigma^3} - \frac{n\theta^2}{2\sigma^3} + \frac{n\theta^3}{3\sigma^3}\right) t^3 + \dots$$

$$= \frac{1}{\sigma^2}\left(\frac{n\theta - n\theta^2}{2}\right) t^2 + \frac{n}{\sigma^3}\left(\frac{\theta - 3\theta^2 + 2\theta^3}{6}\right) t^3 + \dots$$

since $\mu = n\theta$. Then substituting $\sigma^2 = n\theta(1-\theta) = n\theta - n\theta^2$, we find that

$$\ln M_{\tilde{z}_n}(t) = \frac{1}{2}t^2 + \frac{n}{\sigma^3}\left(\frac{\theta - 3\theta^2 + 2\theta^3}{6}\right) t^3 + \dots,$$

where for $r > 2$ the coefficient of t^r is a constant times $\dfrac{n}{\sigma^r}$, which approaches 0 when $n \to \infty$ since

$$\frac{n}{\sigma^r} = \frac{n}{n^{r/2}[\theta(1-\theta)]^{r/2}}.$$

It follows that

$$\lim_{n\to\infty} [\ln M_{\tilde{z}_n}(t)] = \frac{1}{2}t^2$$

and, since the limit of the logarithm equals the logarithm of the limit (provided that the limit exists), we conclude that

$$\lim_{n\to\infty} M_{\tilde{z}_n}(t) = e^{\frac{1}{2}t^2},$$

which is the moment-generating function of a standard normal random variable. *Q.E.D.*

Central Limit Theorem (Lindeberg-Levy)

Theorem. Let $\{\widetilde{x}_i\}_{i=1}^{\infty}$ be i.i.d. random variables with $\mathrm{E}(\widetilde{x}_i) = \mu$ and $\mathrm{Var}(\widetilde{x}_i) = \sigma^2$, with $0 < \sigma^2 < \infty$, for all i. Then

$$\widetilde{z}_n \equiv \frac{\overline{\boldsymbol{x}}_n - \mathrm{E}(\overline{\boldsymbol{x}}_n)}{\sqrt{\mathrm{Var}(\overline{\boldsymbol{x}}_n)}} = \frac{\overline{\boldsymbol{x}}_n - \mu}{\sigma/\sqrt{n}}$$

converges in distribution to a standard normal random variable as $n \to \infty$. That is, $\widetilde{z}_n \longrightarrow \mathrm{N}(0,1)$.

Proof. We will assume in this proof that the common distribution of all the random variables in the sequence $\{\widetilde{x}_i\}_{i=1}^{\infty}$ has a moment-generating function $M_{\widetilde{x}}(t)$ that is well-defined (i.e., finite) in a neighborhood of $t = 0$. Then,

$$M_{\widetilde{z}_n}(t) = M_{\frac{\overline{\boldsymbol{x}}_n - \mu}{\sigma/\sqrt{n}}}(t) = e^{-\frac{\sqrt{n}\cdot\mu t}{\sigma}} M_{\overline{\boldsymbol{x}}_n}\left(\frac{\sqrt{n}\cdot t}{\sigma}\right) = e^{-\frac{\sqrt{n}\cdot\mu t}{\sigma}} M_{n\cdot\overline{\boldsymbol{x}}_n}\left(\frac{t}{\sigma\sqrt{n}}\right).$$

Since $n \cdot \overline{\boldsymbol{x}}_n = \widetilde{x}_1 + \widetilde{x}_2 + \ldots + \widetilde{x}_n$ and $\{\widetilde{x}_i\}_{i=1}^{\infty}$ are i.i.d., it follows that

$$M_{\widetilde{z}_n}(t) = e^{-\frac{\sqrt{n}\cdot\mu t}{\sigma}} \left[M_{\widetilde{x}}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n,$$

where $M_{\widetilde{x}}(t) \equiv M_{\widetilde{x}_i}(t)$, for all i. Hence,

$$\ln M_{\widetilde{z}_n}(t) = -\frac{\sqrt{n}\cdot\mu t}{\sigma} + n \ln M_{\widetilde{x}}\left(\frac{t}{\sigma\sqrt{n}}\right).$$

Expanding $M_{\widetilde{x}}\left(\dfrac{t}{\sigma\sqrt{n}}\right)$ as a power series in t, we obtain

$$\ln M_{\widetilde{z}_n}(t) = -\frac{\sqrt{n}\cdot\mu t}{\sigma} + n \ln\left[1 + \mu_1' \frac{t}{\sigma\sqrt{n}} + \mu_2' \frac{t^2}{2!\cdot\sigma^2 n} + \mu_3' \frac{t^3}{3!\cdot\sigma^3 n^{3/2}} + \ldots\right],$$

where μ_1', μ_2', μ_3', ... are the (non-central) moments of $\widetilde{x}_i$, for all i.

If n is sufficiently large, we can use the expansion of $\ln(1+x)$ as a power series in x, $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \ldots$, getting

$$\begin{aligned}\ln M_{\widetilde{z}_n}(t) = -\frac{\sqrt{n}\cdot\mu t}{\sigma} + n\Bigg\{ &\left[\mu_1' \frac{t}{\sigma\sqrt{n}} + \mu_2' \frac{t^2}{2\sigma^2 n} + \mu_3' \frac{t^3}{6\sigma^3 n^{3/2}} + \ldots\right] \\ &-\frac{1}{2}\left[\mu_1' \frac{t}{\sigma\sqrt{n}} + \mu_2' \frac{t^2}{2\sigma^2 n} + \mu_3' \frac{t^3}{6\sigma^3 n^{3/2}} + \ldots\right]^2 \\ &+\frac{1}{3}\left[\mu_1' \frac{t}{\sigma\sqrt{n}} + \mu_2' \frac{t^2}{2\sigma^2 n} + \mu_3' \frac{t^3}{6\sigma^3 n^{3/2}} + \ldots\right]^3 - \ldots\Bigg\}\end{aligned}$$

Then collecting powers of t, we obtain

$$\ln M_{\widetilde{z}_n}(t) = \left(-\frac{\sqrt{n}\cdot\mu}{\sigma} + \frac{\sqrt{n}\cdot\mu_1'}{\sigma} \right) t + \left(\frac{\mu_2'}{2\sigma^2} - \frac{(\mu_1')^2}{2\sigma^2} \right) t^2$$

$$+ \left(\frac{\mu_3'}{6\sigma^3\sqrt{n}} - \frac{\mu_1'\mu_2'}{2\sigma^3\sqrt{n}} + \frac{(\mu_1')^3}{3\sigma^3\sqrt{n}} \right) t^3 + \ldots,$$

and, since $\mu_1' = \mu$ and $\mu_2' - (\mu_1')^2 = \sigma^2$, this reduces to

$$\ln M_{\widetilde{z}_n}(t) = \frac{1}{2}t^2 + \left(\frac{\mu_3'}{6} - \frac{\mu_1'\mu_2'}{2} + \frac{\mu_1'^3}{3} \right) \frac{t^3}{\sigma^3\sqrt{n}} + \ldots$$

Finally, observing that the coefficient of t^3 is a constant times $\dfrac{1}{\sqrt{n}}$ and, in general, the coefficient of t^r is a constant times $\dfrac{1}{n^{(r-2)/2}}$, we get

$$\lim_{n\to\infty} \left[\ln M_{\widetilde{z}_n}(t) \right] = \frac{1}{2}t^2$$

and, hence,

$$\lim_{n\to\infty} M_{\widetilde{z}_n}(t) = e^{\frac{1}{2}t^2},$$

which is the moment-generating function of a standard normal random variable $\widetilde{z}$. Therefore, since $\lim_{n\to\infty} M_{\widetilde{z}_n}(t) = M_{\widetilde{z}}(t)$, it follows that

$$\widetilde{z}_n \xrightarrow{d} \widetilde{z} \quad \text{or} \quad \widetilde{z}_n \to \mathrm{N}(0,1). \qquad \textit{Q.E.D.}$$

Note: An almost identical proof could be made using the characteristic function instead of the moment-generating function. Notice that the characteristic function is finite everywhere for all random variables.

Exercises. Probability and Statistics. IDEA.
6. Stochastic Processes and Limiting Distributions

1. Consider the stochastic process $\{\tilde{x}_t\}_{t=0}^T$ of random variables on the probability space $(\Omega, \mathcal{F}, P)$ adapted to the filtration $\mathbb{F}$. Prove the following two theorems:

   (a) Let $\{\tilde{x}_t\}_{t=0}^T$ be a submartingale and g a convex, increasing function from $\mathbb{R}$ to $\mathbb{R}$. If $g(\tilde{x}_t)$ is integrable for all t, then $\{g(\tilde{x}_t)\}_{t=0}^T$ is a submartingale.

   (b) Let $\{\tilde{x}_t\}_{t=0}^T$ be a martingale and g a convex function from $\mathbb{R}$ to $\mathbb{R}$. If $g(\tilde{x}_t)$ is integrable for all t, then $\{g(\tilde{x}_t)\}_{t=0}^T$ is a submartingale. Thus, if $r \geq 1$, $\{\tilde{x}_t\}_{t=0}^T$ is a martingale and $|\tilde{x}_t|^r$ is integrable for all t, then $\{|\tilde{x}_t|^r\}_{t=0}^T$ is a submartingale.

2. Imagine that you flip a balanced coin 16 times.

   (a) Find the probability of obtaining exactly 6 heads. Solve this problem using both the binomial distribution and the corresponding approximation by the standard normal. Compare the results.

   (b) Find the probability of obtaining a number of heads strictly higher than 4 and smaller or equal than 7. Solve this problem using both the binomial distribution and the corresponding approximation by the standard normal. Compare the results.

3. Find the probability that exactly 10 individuals pass a (very demanding) exam if 3000 individuals have taken the exam and the probability of passing is just 0.005. Solve this problem using both the binomial distribution and the corresponding approximation by the Poisson distribution. Compare the results.

4. Consider a Markov chain $\{\tilde{x}_t\}_{t=0}^\infty$ on the probability space $(\Omega, \mathcal{F}, P)$. The state space is $S = \{s_1, s_2, s_3\}$ and the time-invariant transition matrix is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1/6 & 1/6 & 2/3 \\ 0 & 0 & 1 \end{pmatrix},$$

   where the value appearing in the (i, j)-entry of the transition matrix is the conditional probability $\pi_{ij} = P\{\tilde{x}_{t+1} = s_j | \tilde{x}_t = s_i\}$ for all t.

   (a) Find the stationary probabilistic vector (or vectors, if there are more than one), the ergodic probabilistic vector (if it exists), the transient states (if they exist), the recurrent states (if they exist), the absorbing states (if they exist) and the ergodic sets of the Markov chain.

   (b) Assume that the Markov chain starts in state s_2, $\tilde{x}_0 = s_2$. Compute

$$\lim_{t\to\infty} P\{\tilde{x}_t = s_1 | \tilde{x}_0 = s_2\}, \lim_{t\to\infty} P\{\tilde{x}_t = s_2 | \tilde{x}_0 = s_2\}, \text{ and } \lim_{t\to\infty} P\{\tilde{x}_t = s_3 | \tilde{x}_0 = s_2\}.$$

   (c) We flip simultaneously three balanced coins. If we obtain an odd number (one or three) of heads, the initial state is s_1 ($\tilde{x}_0 = s_1$); if we get two heads,

the initial state is s_2 ($\tilde{x}_0 = s_2$); and if we get no heads, the initial state is s_3 ($\tilde{x}_0 = s_3$). What is the unconditional (i.e., before flipping the coins) distribution of states in the long run?

(d) Find the expected number of periods (or steps) before the Markov chain is absorbed if it has started in state s_2?

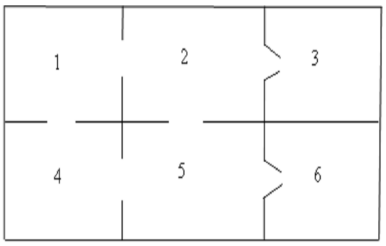
5. Consider the following time-invariant transition matrices:

$$(a)\,\mathrm{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},\ (b)\ \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},\ (c)\ \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},\ (d)\ \begin{pmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \end{pmatrix}$$

(i) Find the stationary probabilistic vectors, the ergodic probabilistic vectors (if they exist), the transient states (if they exist), the recurrent states (if they exist), the absorbing states (if they exist) and the ergodic sets of the Markov chains having the previous transition matrices.

(ii) Consider the Markov chain with the transition matrix (d). Find the mean recurrence time for each of the two states of the Markov chain and the mean first passage time from state s_1 to s_2 and from state s_2 to s_1.

6. A mouse is in the following labyrinth:



In each period the mouse moves randomly from one cell to an adjacent one. All the adjacent cells are equally likely. However, if the mouse enters into cells 3 or 6, it cannot exit from them.

(a) Find the probability that the mouse will enter into cell 6 if it was initially in cell 1.

(b) Find the expected number of periods the mouse will be in cell 2 given that it was initially in cell 5.

(c) Find the expected number of periods before the mouse is trapped if it was initially in cell 4.

7. Two players a and b have two dollars each at the beginning of the game. They bet on repeated flips of a coin. At each flip, the loser pays the winner one dollar, and the game continues until either player is "ruined".

(a) Find the expected duration of the game

(b) Find the probability that a will win the game after flipping the coin exactly four times.

(c) If we know that player a has three dollars in a given moment of the game, what is the probability that player b ends up winning the game?

8. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Assume that the sample space Ω is the closed interval of real numbers $[0,1]$, $\mathcal{F}$ is the $\sigma-$algebra of Borel sets in the interval $[0,1]$, and the probability P on $(\Omega, \mathcal{F})$ is the Lebesgue measure on $[0,1]$. Consider the sequence of random variables $\{\tilde{x}_n\}_{n=1}^{\infty}$ on $(\Omega, \mathcal{F}, P)$ defined as follows:

$$\tilde{x}_1(\omega) = \begin{cases} 1 & \text{if } \omega \in [0,1] \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{x}_2(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[0, \frac{1}{2}\right] \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{x}_3(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\frac{1}{2}, \frac{1}{2} + \frac{1}{3}\right] \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{x}_4(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[0, \frac{1}{12}\right] \cup \left[\frac{1}{2} + \frac{1}{3}, 1\right] \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{x}_5(\omega) = \begin{cases} 1 & \text{if } \omega \in \left[\frac{1}{12}, \frac{1}{12} + \frac{1}{5}\right] \\ \\ 0 & \text{otherwise,} \end{cases}$$

and so on. In other words, the subset of $[0,1]$ over which $\tilde{x}_n$ is equal to 1 has the length $\frac{1}{n}$ and keeps moving to the right until it reaches the right end point of $[0,1]$, at which point it moves back to 0 and starts again.

Does $\tilde{x}_n$ converges almost surely to the constant 0 as $n \to \infty$? Does $\tilde{x}_n$ converges in probability to the constant 0 as $n \to \infty$? Does $\tilde{x}_n$ converges in distribution to the constant 0 as $n \to \infty$?

Hint: Remember that $\sum_{i=1}^{\infty} \frac{1}{i} = \infty$.

9. Consider a Markov chain $\{\tilde{x}_t\}_{t=0}^{\infty}$ on the probability space $(\Omega, \mathcal{F}, P)$. The state space is $S = \{s_1, s_2, s_3\}$ and the time-invariant transition matrix is

$$\Pi = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 3/10 & 0 & 7/10 \end{pmatrix},$$

where the value appearing in the (i,j)-entry of the matrix Π is the conditional probability $\pi_{ij} = P\{\tilde{x}_{t+1} = s_j | \tilde{x}_t = s_i\}$. for all t.

(a) Find the stationary probabilistic vector (or vectors, if there are more than one), the ergodic probabilistic vector (if it exists), the transient states (if they exist), the recurrent states (if they exist), the absorbing states (if they exist) and the ergodic sets of the Markov chain.

(b) Compute the matrix $\widehat{\Pi} = \lim_{t\to\infty} \Pi^t$. *Hint:* Recall that $\Pi \cdot \widehat{\Pi} = \lim_{t\to\infty} \Pi^{t+1} = \widehat{\Pi}$ and, equivalently, $\widehat{\Pi} \cdot \Pi = \lim_{t\to\infty} \Pi^{t+1} = \widehat{\Pi}$.

(c) We roll a dice and, if we get either 1 or 5 dots, the Markov chain starts in state s_1; if we get an even number of dots, it starts in state s_2; and if we get 3 dots, it starts in state s_3. What is the unconditional (i.e., before rolling the dice) distribution of states in the long run?

10. Consider a Markov chain $\{\tilde{x}_t\}_{t=0}^{\infty}$ on the probability space $(\Omega, \mathcal{F}, P)$. The state space is $S = \{s_1, s_2, s_3, s_4\}$ and the time-invariant transition matrix is

$$\Pi = \begin{pmatrix} 1/5 & 4/5 & 0 & 0 \\ 7/10 & 3/10 & 0 & 0 \\ 0 & 0 & 3/5 & 2/5 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix},$$

where the value appearing in the (i,j)-entry of the matrix Π is the conditional probability $\pi_{ij} = P\{\tilde{x}_{t+1} = s_j | \tilde{x}_t = s_i\}$. for all t.

(a) Find the stationary probabilistic vector (or vectors, if there are more than one), the ergodic probabilistic vector (if it exists), the transient states (if they exist), the recurrent states (if they exist), and the ergodic sets of the Markov chain.

(b) Compute the matrix $\widehat{\Pi} = \lim_{t\to\infty} \Pi^t$. Find $\lim_{t\to\infty} P\{\tilde{x}_t = s_2 | \tilde{x}_0 = s_2\}$, $\lim_{t\to\infty} P\{\tilde{x}_t = s_1 | \tilde{x}_0 = s_3\}$, and $\lim_{t\to\infty} P\{\tilde{x}_t = s_3 | \tilde{x}_0 = s_4\}$. Assuming that the initial distribution of the Markov chain at $t = 0$ (i.e., the distribution of $\tilde{x}_0$) is such that all the four states are equally likely, what is the unconditional distribution of states in the long run?

11. Consider four different Markov chains on the probability space $(\Omega, \mathcal{F}, P)$ taking values in the state space $S = \{s_1, s_2\}$. These chains are governed by the following four time-invariant transition matrices:

$$\Pi_1 = \begin{pmatrix} 1/2 & 1/2 \\ 1/9 & 8/9 \end{pmatrix}, \ \Pi_2 = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}, \ \Pi_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \ \Pi_4 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix},$$

where the value appearing in the (i,j)-entry of each matrix is the conditional probability $P\{\tilde{x}_{t+1} = s_j | \tilde{x}_t = s_i\}$ for all t.

(a) Find the stationary probabilistic vector (or vectors, if there are more than one), the transient states (if they exist), the recurrent states (if they exist), the absorbing states (if they exist) and the ergodic sets for each of the four Markov chains above.

(b) Find the matrices $\widehat{\Pi}_k = \lim_{t\to\infty} \Pi_k^t$, $k = 1,2,3,4$ (if they exist) and the corresponding ergodic probabilistic vectors (if they exist).

(c) We roll a dice to decide the initial state of all those chains. If we get a number of dots smaller or equal than 4, the Markov chains start in state s_1; otherwise they start in state s_2. Find the unconditional (i.e., before rolling the dice) distributions of states in the long run (if they exist) for the four Markov chains above.

(d) Consider the Markov chain with the transition matrix Π_1. Find the mean recurrence time for each of the two states of the Markov chain and the mean first passage time from state s_1 to s_2 and from state s_2 to s_1.

(e) Consider the two Markov chains with the transition matrices Π_2 and Π_4. For these two chains, find the expected number of times the chain is in state s_1 given that it starts in state s_1. Moreover, for the same two chains, find the expected number of periods before the chain is absorbed when it starts in state s_1.

12. Consider a factory with a very large number of workers. All workers have the same very small probability of suffering an accident during a year and all months are equally likely for accidents to occur. We know that the probability of <u>at least</u> one accident occurring in this factory during a year is 0.9, i.e., $P\{\tilde{x} \geq 1\} = 0.9$, where $\tilde{x}$ is the number of accidents in the factory during a year.

    (a) Find the expected number of accidents during a year, $\mathrm{E}(\tilde{x})$.

    (b) Find the probability of three or more accidents occurring during a period of 6 months.

    (c) Suppose now that we know that there have been two or more accidents during a given year. Find the probability that the number $\tilde{x}$ of accidents during this year is larger or equal than one and smaller or equal than three, $P\{1 \leq \tilde{x} \leq 3 \,|\, \tilde{x} \geq 2\}$.

13. Let $\tilde{x} \sim \mathrm{N}(0,1)$ and $\tilde{x}_n = -\tilde{x}$ for $n = 1,2,3,...$ Does $\tilde{x}_n \xrightarrow{d} \tilde{x}$? Does $\tilde{x}_n \xrightarrow{p} \tilde{x}$?

14. Consider a random variable $\tilde{x}$ having the Poisson distribution with parameter λ.

    (a) Use both the probability function of the Poisson distribution and the Taylor expansion of e^y around $y = 0$ to find the moment generating function $M_{\tilde{x}}(t)$ of the random variable $\tilde{x}$.

(b) Use the moment generating function found in part (a) to compute the mean μ and the variance σ^2 of $\tilde{x}$.

15. Assume that in a given period three firms are created. In every subsequent period, each firm faces the probability of getting bankrupt and disappearing forever. This probability is 1/5 for all firms and for all periods. The mortality of a firm is independent of the other firms' mortality. Find the expected number of periods before all three firms disappear.

16. Consider a Markov chain $\{\tilde{x}_t\}_{t=0}^{\infty}$ on the probability space $(\Omega, \mathcal{F}, P)$. The state space is $S = \{s_1, s_2, s_3\}$ and the time-invariant transition matrix is

$$\Pi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3/5 & 2/5 \\ 0 & 1/2 & 1/2 \end{pmatrix},$$

where the value appearing in the (i, j)-entry of the transition matrix Π is the conditional probability $\pi_{ij} = P\{\tilde{x}_{t+1} = s_j | \tilde{x}_t = s_i\}$ for all t.

(a) Find the stationary probabilistic vector (or vectors, if there are more than one), the ergodic probabilistic vector (if it exists), the transient states (if they exist), the recurrent states (if they exist), the absorbing states (if they exist), and the ergodic sets of the Markov chain.

(b) Find the probability in the long run of being in state s_3 if the chain starts in state s_2,

$$\lim_{t \to \infty} P\{\tilde{x}_t = s_3 | \tilde{x}_0 = s_2\}.$$

(c) We roll a dice and, if we get either 2 or 6 dots, the Markov chain starts in state s_1; if we get an odd number of dots, it starts in state s_2; and if we get 4 dots, it starts in state s_3. Find the unconditional (i.e., before rolling the dice) probability of being in state s_2 in the long run,

$$\lim_{t \to \infty} P\{\tilde{x}_t = s_2\}.$$

17. Show that, if $\tilde{x}$ has a gamma distribution with parameters α and β, $\tilde{x} \sim \Gamma(\alpha, \beta)$, then the standardization of $\tilde{x}$ converges in distribution to a standard normal random variable when $\alpha \to \infty$ and β remains constant. *Hint:* Use Lévy's convergence theorem concerning the equivalence between convergence in distribution and convergence of the corresponding moment-generating function.

18. Use Chebyshev's inequality to prove the weak law of large numbers for a sequence $\{\tilde{x}_i\}_{i=1}^{\infty}$ of independent random variables with identical finite variance, $\mathrm{Var}(\tilde{x}_i) = \sigma^2 < \infty$ for all i.

7. Sampling

7.1. Samples, random samples, and statistics

- Let $\tilde{X} = (\tilde{x}_1, ..., \tilde{x}_n)$ be a random vector with distribution $P_{\tilde{X}} : \mathcal{B}(\mathbb{R}^n) \longrightarrow [0, 1]$.
- In statistical inference, a sample is a random vector $\tilde{X} = (\tilde{x}_1, ..., \tilde{x}_n)$.
- To observe a sample of the random vector $\tilde{X}$ is to observe a realization (a value) $X = (x_1, ..., x_n) \in \mathbb{R}^n$ of the random vector $\tilde{X} = (\tilde{x}_1, ..., \tilde{x}_n)$.

- **Definition.** A random sample of size n is a collection of random variables $\{\tilde{x}_i\}_{i=1}^n$ (or $(\tilde{x}_1, ..., \tilde{x}_n)$) that are i.i.d. Its common distribution $P_{\tilde{x}} = P_{\tilde{x}_i}$, for all i, is called the population (or parent) distribution.

- We say that $\{\tilde{x}_i\}_{i=1}^n$ is a random sample of size n from a population $\tilde{x}$ with distribution $P_{\tilde{x}}$.

- Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population with distribution $P_{\tilde{x}}$, and consider the vector $\tilde{X} = (\tilde{x}_1, ..., \tilde{x}_n)$, then

$$P_{\tilde{X}} = P_{\tilde{x}} \times P_{\tilde{x}} \times ... \times P_{\tilde{x}} \equiv (P_{\tilde{x}})^n$$

- The random vector $\tilde{u} = h\left(\tilde{x}_1, ..., \tilde{x}_n\right)$, where $h : \left(\mathbb{R}^n, \mathcal{B}\left(\mathbb{R}^n\right)\right) \longrightarrow \left(\mathbb{R}^m, \mathcal{B}\left(\mathbb{R}^m\right)\right)$ is a Borel measurable function, is called a **statistic** of the sample $\left\{\tilde{x}_i\right\}_{i=1}^n$.
- The value $u = h\left(x_1, ..., x_n\right) \in \mathbb{R}^m$, where $\left(x_1, ..., x_n\right) \in \mathbb{R}^n$ is a realization of (or a value taken on by) the sample $\left(\tilde{x}_1, ..., \tilde{x}_n\right)$, is the value of the statistic $\tilde{u}$.

- **Two examples of statistics:**

  Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample, then

$$\bar{\mathbf{x}}_n = \bar{\mathbf{x}} = \frac{\sum_{i=1}^{n} \tilde{x}_i}{n}$$

  is the **sample mean** (or mean of the random sample), and

$$\mathbf{s}_n^2 = \mathbf{s}^2 = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}}_n)^2}{n-1}$$

  is the **sample variance** (or variance of the random sample).

- We can omit the subindex n in $\bar{\mathbf{x}}_n$ and $\mathbf{s}_n^2$ when the sample size is fixed.

- Statistics are random variables (or vectors if $m > 1$) and the distribution of a statistic is called "sampling distribution".
- We should distinguish the random variables "statistic", sample mean, or sample variance, $\tilde{u}$, $\bar{\mathbf{x}}_n$, and $\mathbf{s}_n^2$, from the values u, $\overline{x}_n$, and s_n^2 taken by the corresponding random variables.
- If statistics are used to estimate the parameter vector $\theta \in \mathbb{R}^K$ characterizing the distribution $P_{\tilde{X}}(\cdot;\theta)$ of the vector $\tilde{X}$ or the population distribution $P_{\tilde{x}}(\cdot;\theta)$, then they are called "estimators".

7.2. The distribution of the sample mean

- **Theorem.** If $\{\tilde{x}_i\}_{i=1}^n$ is a random sample from a population with the mean μ and the variance σ^2, with $0 < \sigma^2 < \infty$, then

(a)

$$\mathrm{E}(\bar{\mathbf{x}}_n) = \mu \quad \text{and} \quad \mathrm{Var}(\bar{\mathbf{x}}_n) = \frac{\sigma^2}{n}.$$

(b) Strong law of large numbers:

$$\bar{\mathbf{x}}_n \xrightarrow{a.s} \mu.$$

(c) Central limit theorem:

$$\tilde{z}_n \equiv \frac{\bar{\mathbf{x}}_n - \mathrm{E}(\bar{\mathbf{x}}_n)}{\sqrt{\mathrm{Var}(\bar{\mathbf{x}}_n)}} = \frac{\bar{\mathbf{x}}_n - \mu}{\sigma / \sqrt{n}} \overset{a}{\sim} \mathrm{N}(0,1) \quad (\text{or } \tilde{z}_n \longrightarrow \mathrm{N}(0,1))$$

or, equivalently,

$$\sqrt{n}(\bar{\mathbf{x}}_n - \mu) \longrightarrow \mathrm{N}(0, \sigma^2).$$

- **Theorem.** If $\bar{\mathbf{x}}_n$ is the mean of a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from a normal population $\tilde{x}$ with the mean μ and the variance σ^2, its sampling distribution is a normal distribution with the mean μ and the variance σ^2 / n.
- **Proof.**

$$M_{\bar{\mathbf{x}}_n}(t) = M_{\frac{\sum_{i=1}^n \tilde{x}_i}{n}}(t) = M_{\sum_{i=1}^n \tilde{x}_i}\left(\frac{t}{n}\right) = \left[M_{\tilde{x}}\left(\frac{t}{n}\right)\right]^n$$

$$= \left[e^{\mu\frac{t}{n}+\frac{1}{2}\sigma^2\frac{t^2}{n^2}}\right]^n = e^{\mu t+\frac{1}{2}\frac{\sigma^2}{n}t^2},$$

which is the moment-generating function of a normal distribution with the mean μ and the variance σ^2 / n. Therefore,

$\bar{\mathbf{x}}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$. *Q.E.D.*

- Note that

$$\bar{\mathbf{x}}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \Longleftrightarrow \frac{\bar{\mathbf{x}}_n - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1) \Longleftrightarrow \sqrt{n}\,(\bar{\mathbf{x}}_n - \mu) \sim \mathrm{N}(0, \sigma^2).$$

7.3. The distribution of the variance of a random sample and the chi-square distribution

- Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population with the mean μ and the finite variance σ^2.
- $\mathbf{s^2} = \dfrac{\sum_{i=1}^n (\tilde{x}_i - \bar{\mathbf{x}})^2}{n-1}$ (sample variance or variance of the random sample).
- $\hat{\mathbf{s}}^{\mathbf{2}} = \dfrac{\sum_{i=1}^n (\tilde{x}_i - \bar{\mathbf{x}})^2}{n} = \dfrac{(n-1)\mathbf{s^2}}{n}$.
- $\breve{\mathbf{s}}^{\mathbf{2}} = \dfrac{\sum_{i=1}^n (\tilde{x}_i - \mu)^2}{n}$.

- **Proposition.** (a) $\mathrm{E}\left(\mathbf{s^2}\right) = \sigma^2$, (b) $\mathrm{E}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \left(\frac{n-1}{n}\right)\sigma^2$, and (c) $\mathrm{E}\left(\breve{\mathbf{s}}^{\mathbf{2}}\right) = \sigma^2$.
- **Proof. (a)**

$$\mathrm{E}\left(\mathbf{s^2}\right) = \mathrm{E}\left[\frac{\sum_{i=1}^{n}\left(\tilde{x}_i - \bar{\mathbf{x}}\right)^2}{n-1}\right] = \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}\left[\left(\tilde{x}_i - \mu\right) - \left(\bar{\mathbf{x}} - \mu\right)\right]^2\right)$$

$$= \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}\left[\left(\tilde{x}_i - \mu\right)^2 - 2\left(\tilde{x}_i - \mu\right)\left(\bar{\mathbf{x}} - \mu\right) + \left(\bar{\mathbf{x}} - \mu\right)^2\right]\right)$$

$$= \frac{1}{n-1}\mathrm{E}\left(\sum_{i=1}^{n}\left(\tilde{x}_i - \mu\right)^2 - 2n\left(\bar{\mathbf{x}} - \mu\right)^2 + n\left(\bar{\mathbf{x}} - \mu\right)^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathrm{E}\left[\left(\tilde{x}_i - \mu\right)^2\right] - n\mathrm{E}\left[\left(\bar{\mathbf{x}} - \mu\right)^2\right]\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2.$$

(b)

$$\mathrm{E}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \mathrm{E}\left[\frac{(n-1)\mathbf{s}^{\mathbf{2}}}{n}.\right] = \left(\frac{n-1}{n}\right)\mathrm{E}\left(\mathbf{s}^{\mathbf{2}}\right) = \left(\frac{n-1}{n}\right)\sigma^2.$$

(c)

$$\mathrm{E}\left(\breve{\mathbf{s}}^{\mathbf{2}}\right) = \mathrm{E}\left[\frac{\sum_{i=1}^{n}\left(\tilde{x}_i-\mu\right)^2}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\tilde{x}_i-\mu\right)^2 = \frac{1}{n}n\sigma^2 = \sigma^2. \quad \textit{Q.E.D.}$$

- *Note*:

$$\mathrm{E}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2 = \mathrm{E}\left(\mathbf{s}^{\mathbf{2}}\right).$$

However, $\mathrm{E}\left(\hat{\mathbf{s}}_n^{\mathbf{2}}\right)$ converges to σ^2 when $n \longrightarrow \infty$.

- **Proposition.** If $\{\tilde{z}_i\}_{i=1}^n$ are i.i.d. random variables and $\tilde{z}_i \sim \mathrm{N}(0,1)$, for $i = 1, 2, ..., n$, then

$$\tilde{y} = \sum_{i=1}^{n} \tilde{z}_i^2 \sim \chi_n^2.$$

- **Proof.** Remember that $\tilde{z}_i^2 \sim \chi_1^2$ so that

$$M_{\tilde{z}_i^2}(t) = (1-2t)^{-1/2}, \quad \text{for } t < 1/2.$$

Therefore, from independency,

$$M_{\tilde{y}}(t) = \left[(1-2t)^{-1/2}\right]^n = (1-2t)^{-n/2}, \quad \text{for } t < 1/2,$$

which is the moment-generating function of a random variable whose distribution is χ_n^2. Thus, $\tilde{y} \sim \chi_n^2$. *Q.E.D.*

- **Proposition.** If $\{\tilde{x}_i\}_{i=1}^n$ are independent random variables and $\tilde{x}_i \sim \chi^2_{\nu_i}$, for $i = 1, 2, ..., n$, then

$$\tilde{y} = \sum_{i=1}^{n} \tilde{x}_i \sim \chi^2_{(\nu_1+\nu_2+...+\nu_n)}.$$

- **Proof.** From independency,

$$M_{\tilde{y}}(t) = \prod_{i=1}^{n} (1-2t)^{-\nu_i/2} = (1-2t)^{-(\nu_1+\nu_2+...+\nu_n)/2} \quad \text{for } t < 1/2.$$

Therefore, $\tilde{y} \sim \chi^2_{(\nu_1+\nu_2+...+\nu_n)}$. *Q.E.D.*

- **Proposition.** If $\tilde{x}_1$ and $\tilde{x}_2$ are independent random variables and $\tilde{x}_1 \sim \chi^2_{\nu_1}$ and $\tilde{x}_1 + \tilde{x}_2 \sim \chi^2_{\nu}$, with $\nu > \nu_1$, then $\tilde{x}_2 \sim \chi^2_{(\nu-\nu_1)}$.
- **Proof.** From independency,

$$M_{\tilde{x}_1+\tilde{x}_2}(t) = M_{\tilde{x}_1}(t) \cdot M_{\tilde{x}_2}(t)$$

or

$$(1-2t)^{-\nu/2} = (1-2t)^{-\nu_1/2} M_{\tilde{x}_2}(t), \quad \text{for } t < 1/2.$$

Therefore,

$$M_{\tilde{x}_2}(t) = (1-2t)^{-(\nu-\nu_1)/2}, \quad \text{for } t < 1/2.$$

Thus, $\tilde{x}_2 \sim \chi^2_{(\nu-\nu_1)}$. *Q.E.D.*

- **Theorem.** If $\bar{\mathbf{x}}$ and $\mathbf{s^2}$ are the mean and the variance of a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from a normal population with mean μ and variance σ^2, then

  (1) $\bar{\mathbf{x}}$ and $\mathbf{s^2}$ are independent.

  (2)

$$\frac{(n-1)\mathbf{s^2}}{\sigma^2} \equiv \frac{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

- **Proof. (1)**

  (a) $\tilde{x}_n - \bar{\mathbf{x}} = -\sum_{i=1}^{n-1} (\tilde{x}_i - \bar{\mathbf{x}})$ since $\sum_{i=1}^{n} \tilde{x}_i = n\bar{\mathbf{x}}$ (or $\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}}) = 0$);

  (b) $\mathbf{s^2} = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}})^2}{n-1}$ is thus a function of the random vector $\widetilde{w} = (\tilde{x}_1 - \bar{\mathbf{x}}, \tilde{x}_2 - \bar{\mathbf{x}}, ..., \tilde{x}_{n-1} - \bar{\mathbf{x}})$;

  (c) The n-dimensional random vector $(\bar{\mathbf{x}}, \widetilde{w})^\intercal$ has a multivariate normal distribution according to the General Proposition in Section 4.10 of Chapter 4 since

$$(\bar{\mathbf{x}}, \widetilde{w})^{\mathsf{T}} = \begin{pmatrix} \bar{\mathbf{x}} \\ \tilde{x}_1 - \bar{\mathbf{x}} \\ \tilde{x}_2 - \bar{\mathbf{x}} \\ \vdots \\ \tilde{x}_{n-1} - \bar{\mathbf{x}} \end{pmatrix}_{n \times 1} =$$

$$\begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \vdots & \frac{1}{n} & \frac{1}{n} \\ 1 - \frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \vdots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & -\frac{1}{n} & \vdots & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \vdots & 1 - \frac{1}{n} & -\frac{1}{n} \end{pmatrix}_{n \times n} \times \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix}_{n \times 1}$$

and $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)^\intercal$ is multivariate normal because it is a vector of independent, normally distributed variables;

(d) $\text{Cov}(\bar{\mathbf{x}}, \tilde{x}_i - \bar{\mathbf{x}}) = 0$ for all i (Exercise) so that $\bar{\mathbf{x}}$ and $\widetilde{w}$ are uncorrelated and, thus, they are independent. Therefore, $\bar{\mathbf{x}}$ and $\mathbf{s^2}$ are independent.

- **(2)** The following equality holds:

$$\sum_{i=1}^{n} (\tilde{x}_i - \mu)^2 = \sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}})^2 + n (\bar{\mathbf{x}} - \mu)^2 \qquad (*)$$

since it is equivalent to

$$\sum_{i=1}^{n} \left(\tilde{x}_i^2 - 2\tilde{x}_i\mu + \mu^2\right) = \sum_{i=1}^{n} \left(\tilde{x}_i^2 - 2\tilde{x}_i\bar{\mathbf{x}} + \bar{\mathbf{x}}^2\right) + n \left(\bar{\mathbf{x}}^2 - 2\bar{\mathbf{x}}\mu + \mu^2\right).$$

As $\sum_{i=1}^{n} \tilde{x}_i = n\bar{\mathbf{x}}$, the previous expression becomes

$$\sum_{i=1}^{n} \tilde{x}_i^2 - 2n\bar{\mathbf{x}}\mu + n\mu^2 = \sum_{i=1}^{n} \tilde{x}_i^2 \underbrace{-2n\bar{\mathbf{x}}^2 + n\bar{\mathbf{x}}^2 + n\bar{\mathbf{x}}^2}_{=0} - 2n\bar{\mathbf{x}}\mu + n\mu^2.$$

Divide $(*)$ by σ^2,

$$\sum_{i=1}^{n} \left(\frac{\tilde{x}_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}}_n)^2}{\sigma^2} + \left(\frac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}}\right)^2.$$

- We know that $\frac{\tilde{x}_i - \mu}{\sigma} \sim \mathrm{N}(0,1)$ so that $\sum_{i=1}^{n} \left(\frac{\tilde{x}_i - \mu}{\sigma} \right)^2 \sim \chi^2_n$ and $\frac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1)$ so that $\left(\frac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}} \right)^2 \sim \chi^2_1$. Therefore, $\frac{\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}})^2}{\sigma^2} \equiv \frac{(n-1)\mathbf{s^2}}{\sigma^2} \sim \chi^2_{n-1}$ since $\left(\frac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}} \right)^2$ and $\frac{(n-1)\mathbf{s^2}}{\sigma^2}$ are independent as follows from part (a). *Q.E.D.*

- Note that only $n-1$ terms of the sum $\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}})^2$ are allowed to vary freely since they are constrained by the relation $\sum_{i=1}^{n} (\tilde{x}_i - \bar{\mathbf{x}}) = 0$. This is why we say that the statistic $\frac{(n-1)\mathbf{s^2}}{\sigma^2}$ has $n-1$ degrees of freedom. In general, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

- **Proposition.** Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then,

  (a) $\text{Var}\left(\mathbf{s^2}\right) = \dfrac{2\sigma^4}{n-1}$, (b) $\text{Var}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \dfrac{2(n-1)\sigma^4}{n^2}$, and

  (c) $\text{Var}\left(\check{\mathbf{s}}^{\mathbf{2}}\right) = \dfrac{2\sigma^4}{n}$.

- **Proof. (a)** Since

$$\frac{\sum_{i=1}^{n}\left(\tilde{x}_i - \bar{\mathbf{x}}\right)^2}{\sigma^2} = \frac{(n-1)\mathbf{s^2}}{\sigma^2} \sim \chi^2_{n-1},$$

we get

$$\text{Var}\left[\frac{(n-1)\mathbf{s^2}}{\sigma^2}\right] = \frac{(n-1)^2}{\sigma^4}\text{Var}\left(\mathbf{s^2}\right) = 2\,(n-1)$$

and, thus, $\text{Var}\left(\mathbf{s^2}\right) = \dfrac{2\sigma^4}{n-1}$.

(b)

$$\text{Var}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \text{Var}\left[\frac{(n-1)\mathbf{s}^{\mathbf{2}}}{n}\right] = \left(\frac{n-1}{n}\right)^2 \text{Var}\left(\mathbf{s}^{\mathbf{2}}\right)$$

$$= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}.$$

(c) Since $\sum_{i=1}^{n}\left(\frac{\tilde{x}_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$, we get

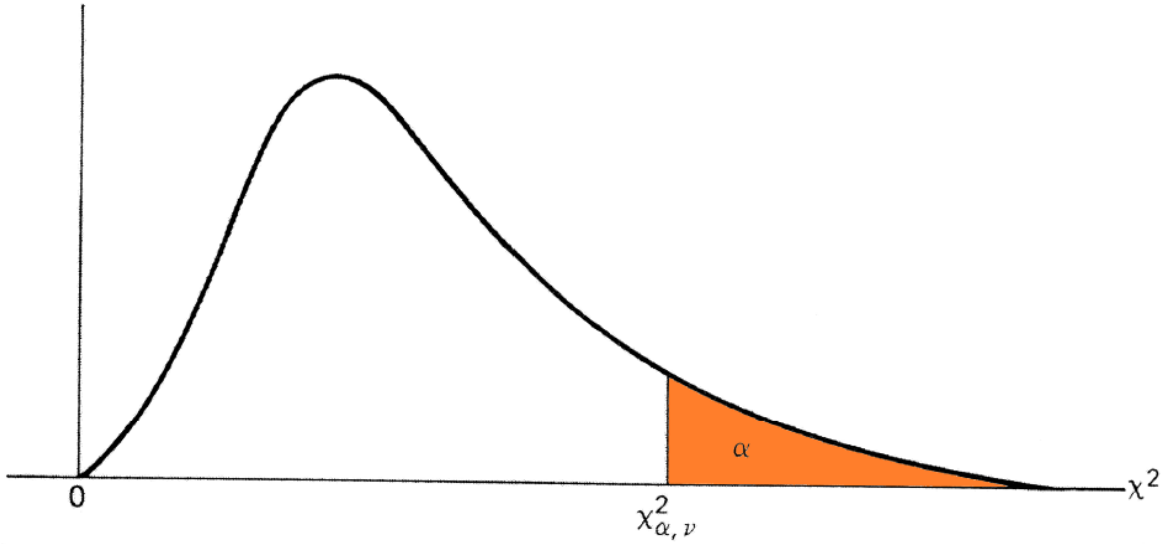
$$\text{Var}\left[\sum_{i=1}^{n}\left(\frac{\tilde{x}_i-\mu}{\sigma}\right)^2\right] = \frac{n^2}{\sigma^4}\text{Var}\left[\frac{\sum_{i=1}^{n}(\tilde{x}_i-\mu)^2}{n}\right] = \frac{n^2}{\sigma^4}\text{Var}\left(\breve{\mathbf{s}}^{\mathbf{2}}\right) = 2n$$

and, thus, $\text{Var}\left(\breve{\mathbf{s}}^{\mathbf{2}}\right) = \frac{2\sigma^4}{n}$. *Q.E.D.*

- In the table of the chi-square we find the value $\chi^2_{\alpha,\nu}$ such that

$$P\left\{\tilde{x} \geq \chi^2_{\alpha,\nu}\right\} = P_{\tilde{x}}\left[\chi^2_{\alpha,\nu}, \infty\right) = \alpha,$$

when $\tilde{x} \sim \chi^2_\nu$.



- If $\tilde{x}_\nu \sim \chi^2_\nu$, then $\frac{\tilde{x}_\nu - \nu}{\sqrt{2\nu}} \longrightarrow \mathrm{N}(0,1)$ as $\nu \longrightarrow \infty$. (Exercise).
- Thus, $P\left\{\tilde{x}_\nu \geq \chi^2_{\alpha,\nu}\right\} = P\left\{\frac{\tilde{x}_\nu - \nu}{\sqrt{2\nu}} \geq \frac{\chi^2_{\alpha,\nu} - \nu}{\sqrt{2\nu}}\right\} = \alpha \approx 1 - \mathrm{N}\left(\frac{\chi^2_{\alpha,\nu} - \nu}{\sqrt{2\nu}}\right)$ for ν large.
- Hence, if $\tilde{z} \sim \mathrm{N}(0,1)$ and $P\left\{\tilde{z} \geq z_\alpha\right\} = \alpha$, then $\frac{\chi^2_{\alpha,\nu} - \nu}{\sqrt{2\nu}} \approx z_\alpha$ and, thus, $\chi^2_{\alpha,\nu} \approx \nu + \left(\sqrt{2\nu}\right) z_\alpha$ for ν large.

- The **Wishart distribution** is the multivariate generalization of the Chi-square distribution.
- If $\left(\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n\right)$ is a random sample from a k-dimensional multivariate normal population $\tilde{X} \sim \mathrm{MN}(\underline{0}, \Sigma)$, where Σ is a $k \times k$ symmetric postive definite matrix, then the $k \times k$ random matrix $\tilde{S} = \sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^{\intercal}$ is said to follow a Wishart distribution with n degrees of freedom and scale matrix Σ, $\tilde{S} \sim \mathrm{W}_n(\Sigma)$.
- Thus, the Wishart distribution is a distribution for symmetric positive definite matrices.
- Note that if $k = 1$ and $\Sigma = 1$, then $\tilde{X} \sim \mathrm{N}(0, 1)$ and $\tilde{S} \sim \chi^2_n$.

7.4. The t distribution

William S. Gosset ("Student") (1876 – 1937)

- **Definition.** A random variable $\tilde{x}$ has the (Student's) t distribution (or is t) with $\nu > 0$ degrees of freedom if its density is

$$f_{\tilde{x}}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\cdot\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } -\infty < x < \infty.$$

- We write $\tilde{x} \sim \mathrm{t}_\nu$.
- Note that the t density is symmetric with respect to zero.
- **Theorem.** If $\tilde{y}$ and $\tilde{z}$ are independent random variables with $\tilde{y} \sim \chi^2_\nu$ and $\tilde{z} \sim \mathrm{N}(0,1)$, then

$$\tilde{x} = \frac{\tilde{z}}{\sqrt{\tilde{y}\,/\nu}} \sim \mathrm{t}_\nu.$$

- **Proof.** See the handout.

- Let $\tilde{x} \sim \mathrm{t}_\nu$. Then $\tilde{x} \in \mathrm{L}^k$ if and only if $\nu > k$. Thus, the t distribution has not well-defined (i.e., finite) moment-generating function in a neighborhood of 0. In fact, $M_{\tilde{x}}(t)$ is finite only at $t = 0$.

- Moments: If $\nu > k$, then

$$\mu'_k = \begin{cases} \dfrac{\nu^{k/2}\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{\nu-k}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu}{2}\right)} & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd.} \end{cases}$$

Mean:

$$\text{If } \nu > 1, \text{ then } \mathrm{E}(\tilde{x}) = 0.$$

Variance:

$$\text{If } \nu > 2, \text{ then } \mathrm{Var}(\tilde{x}) = \frac{\nu}{\nu - 2}.$$

- When $\nu = 1$, the t distribution is the Cauchy distribution with parameters $\alpha = 0$ and $\beta = 1$. Recall that the density of the Cauchy distribution is

$$f_{\tilde{x}}(x) = \frac{\beta/\pi}{(x-\alpha)^2 + \beta^2} \quad \text{for } -\infty < x < \infty,$$

  and that the Cauchy distribution has not well-defined mean (and thus has indefinite/infinite higher odd/even central moments).

- Therefore, the density of a t_1 random variable is

$$f_{\tilde{x}}(x) = \frac{1}{\pi(1+x^2)} \quad \text{for } -\infty < x < \infty. \qquad \longleftarrow \text{Exercise}$$

- If $\tilde{x}_\nu \sim t_\nu$, then $\tilde{x}_\nu \longrightarrow N(0,1)$ as $\nu \longrightarrow \infty$. (Exercise).

- **Theorem.** If $\bar{\mathbf{x}}$ and $\mathbf{s^2}$ are the mean and the variance of a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from a normal population with mean μ and variance σ^2, then

$$\frac{\bar{\mathbf{x}}-\mu}{\mathbf{s}\Big/\sqrt{n}} \sim \mathrm{t}_{n-1},$$

where $\mathbf{s}$ is the sample standard deviation (or standard deviation of the random sample), $\mathbf{s} = \left(\mathbf{s^2}\right)^{1/2}$.

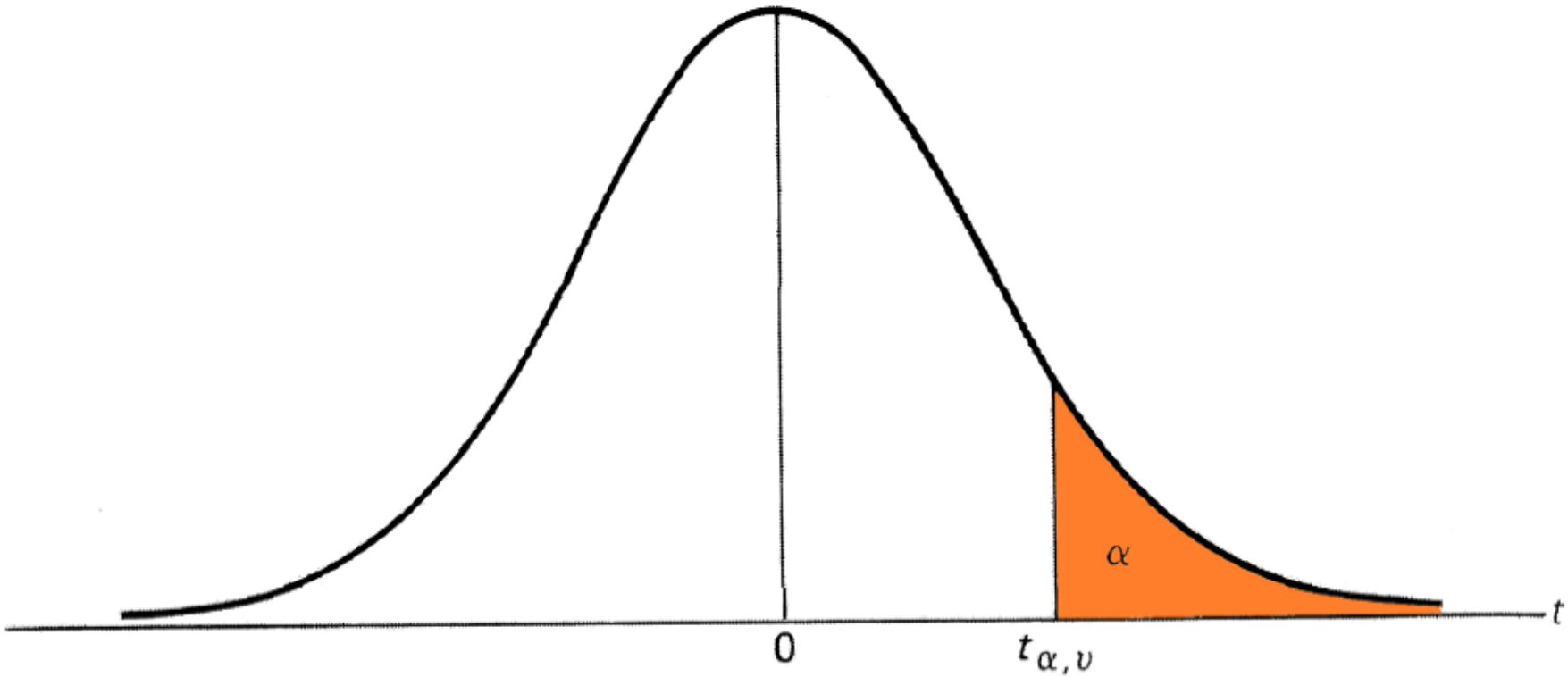
- **Proof.** Let $\tilde{y} = \dfrac{(n-1)\mathbf{s^2}}{\sigma^2} \sim \chi^2_{n-1}$ and $\tilde{z} = \dfrac{\bar{\mathbf{x}}-\mu}{\sigma\Big/\sqrt{n}} \sim \mathrm{N}(0,1)$. The random variables $\tilde{z}$ and $\tilde{y}$ are independent. Then,

$$\frac{\dfrac{\bar{\mathbf{x}}-\mu}{\sigma\Big/\sqrt{n}}}{\sqrt{\dfrac{(n-1)\mathbf{s^2}}{\sigma^2}\Big/(n-1)}} = \frac{\bar{\mathbf{x}}-\mu}{\mathbf{s}\Big/\sqrt{n}} \sim \mathrm{t}_{n-1}. \qquad \textit{Q.E.D.}$$

- In the table of the t_ν we find the value $t_{\alpha,\nu}$ such that

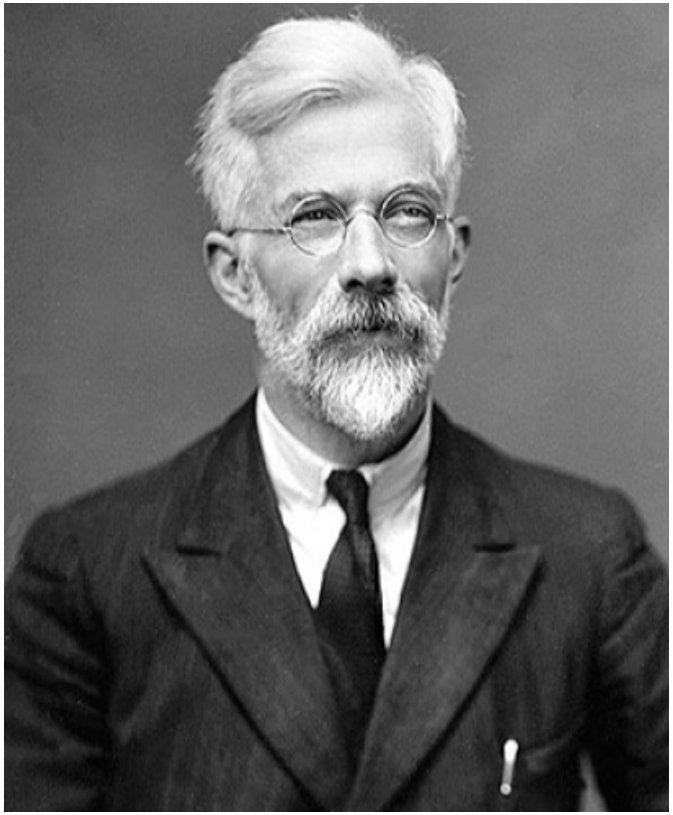
$$P\left\{\tilde{x} \geq t_{\alpha,\nu}\right\} = P_{\tilde{x}}\left[t_{\alpha,\nu}, \infty\right) = \alpha,$$

when $\tilde{x} \sim t_\nu$.



- Note that $P\left\{\tilde{x} \leq -t_{\alpha,\nu}\right\} = \alpha$ or $P\left\{\tilde{x} \geq -t_{\alpha,\nu}\right\} = 1-\alpha$. Moreover, $P\left\{\tilde{x} \leq t_{\alpha,\nu}\right\} = 1-\alpha$.

7.5. The F distribution



Ronald Fisher (1890 - 1962)

George W. Snedecor (1881 - 1974)

- **Definition.** A random variable $\tilde{y}$ has the (Fisher-Snedecor's) F distribution (or is F) with $\nu_1 > 0$ and $\nu_2 > 0$ degrees of freedom if its density is

$$f_{\tilde{y}}(y) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\cdot\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} y^{\frac{\nu_1}{2}-1} \left(1+\frac{\nu_1}{\nu_2}y\right)^{-\frac{1}{2}(\nu_1+\nu_2)} & \text{for } y > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

- We write $\tilde{y} \sim \mathrm{F}_{\nu_1,\nu_2}$.
- Note that $\mathrm{F}_{\nu_1,\nu_2} \neq \mathrm{F}_{\nu_2,\nu_1}$ but $\tilde{x} = \dfrac{1}{\tilde{y}} \sim \mathrm{F}_{\nu_2,\nu_1}$ (Exercise).
- **Theorem.** If $\tilde{u}$ and $\tilde{v}$ are independent random variables with $\tilde{u} \sim \chi^2_{\nu_1}$ and $\tilde{v} \sim \chi^2_{\nu_2}$, then

$$\tilde{y} = \frac{\tilde{u}/\,\nu_1}{\tilde{v}/\,\nu_2} \sim \mathrm{F}_{\nu_1,\nu_2}.$$

- **Proof.** See the handout.

- The F distribution has well-defined (i.e., finite) moment-generating function $M_{\tilde{y}}(t)$ if and only if $t \leq 0$.
- Let $\tilde{y} \sim \mathrm{F}_{\nu_1,\nu_2}$. Then, $\tilde{y} \in \mathrm{L}^k$ if and only if $\nu_2 > 2k$.
- Moments:

$$\mu'_k = \left(\frac{\nu_2}{\nu_1}\right)^k \frac{\Gamma\left(\frac{\nu_1}{2}+k\right)\Gamma\left(\frac{\nu_2}{2}-k\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}, \quad \text{for } \nu_2 > 2k.$$

Mean:

$$\text{If } \nu_2 > 2, \text{ then } \mathrm{E}(\tilde{y}) = \frac{\nu_2}{\nu_2 - 2}.$$

Variance:

$$\text{If } \nu_2 > 4, \text{ then } \mathrm{Var}(\tilde{y}) = \frac{2\nu_2^2\,(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2\,(\nu_2-4)}.$$

- **Theorem.** Let $\mathbf{s_1^2}$ and $\mathbf{s_2^2}$ be the variances of two independent random samples of size n_1 and n_2 (i.e., the $n_1 + n_2$ random variables constituting the two random samples are independent) from two normal populations with variances σ_1^2 and σ_2^2, respectively. Then,

$$\tilde{y} = \frac{\mathbf{s_1^2}/\sigma_1^2}{\mathbf{s_2^2}/\sigma_2^2} = \frac{\sigma_2^2 \cdot \mathbf{s_1^2}}{\sigma_1^2 \cdot \mathbf{s_2^2}} \sim \mathrm{F}_{n_1-1,n_2-1}.$$

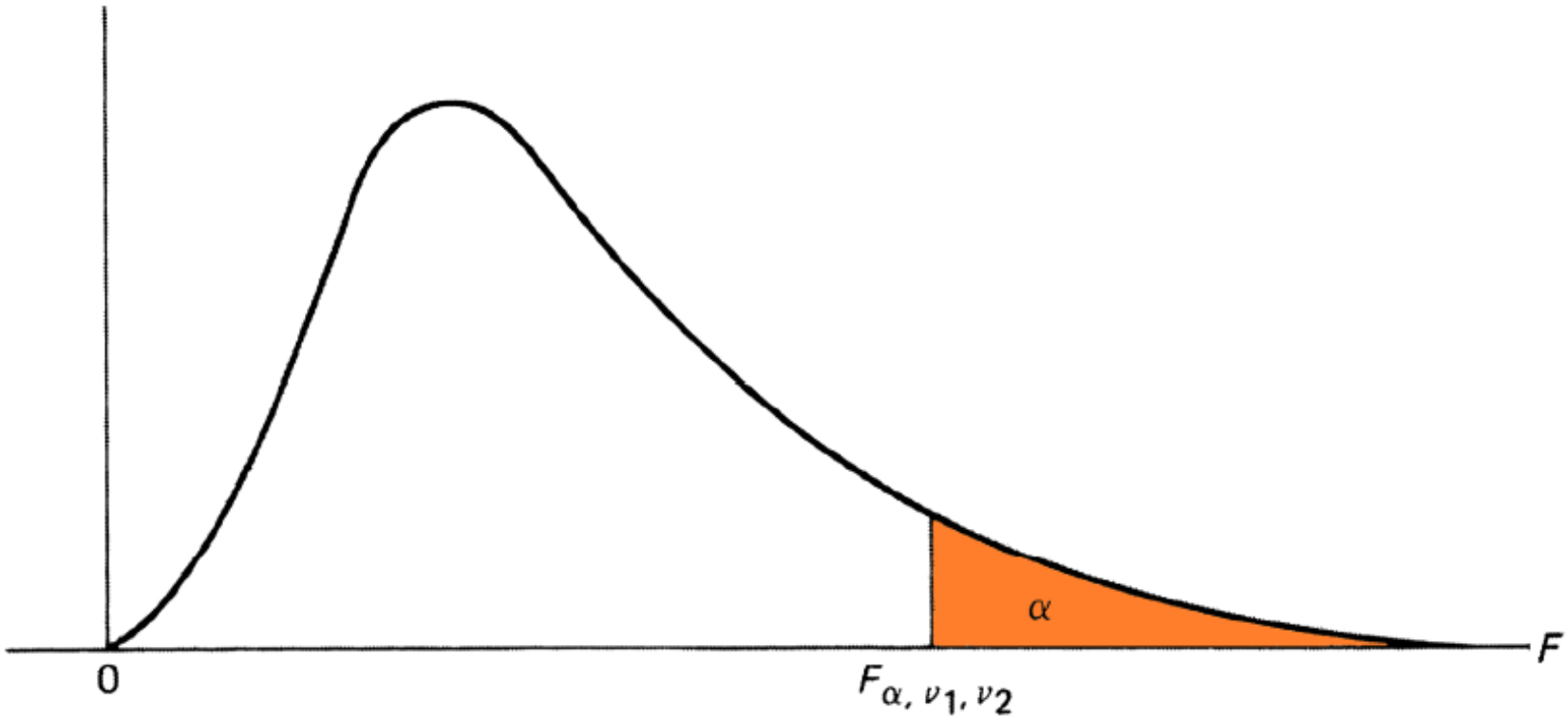
- **Proof.** Let $\tilde{u} = \dfrac{(n_1-1)\mathbf{s_1^2}}{\sigma_1^2} \sim \chi^2_{n_1-1}$ and $\tilde{v} = \dfrac{(n_2-1)\mathbf{s_2^2}}{\sigma_2^2} \sim \chi^2_{n_2-1}$.
The random variables $\tilde{u}$ and $\tilde{v}$ are independent. Then

$$\frac{\dfrac{(n_1-1)\mathbf{s_1^2}}{\sigma_1^2} \Big/ (n_1-1)}{\dfrac{(n_2-1)\mathbf{s_2^2}}{\sigma_2^2} \Big/ (n_2-1)} = \frac{\mathbf{s_1^2}/\sigma_1^2}{\mathbf{s_2^2}/\sigma_2^2} \sim \mathrm{F}_{n_1-1,n_2-1}. \qquad \textit{Q.E.D.}$$

- In the table of the F_{ν_1,ν_2} we find the value F_{α,ν_1,ν_2} such that

$$P\left\{\tilde{y} \geq F_{\alpha,\nu_1,\nu_2}\right\} = P_{\tilde{y}}\left[F_{\alpha,\nu_1,\nu_2}, \infty\right) = \alpha,$$

when $\tilde{y} \sim \mathrm{F}_{\nu_1,\nu_2}$.



- Note that $F_{\alpha,\nu_1,\nu_2} = \dfrac{1}{F_{1-\alpha,\nu_2,\nu_1}}$ (think about it).

7.6. Order statistics

- The jth order statistic $\tilde{x}_{(j)}$ of a random sample $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ of size n is the random variable of the sample that takes the jth smallest value.
- See the handout for the distribution of the order statistics.

The t Distribution

Theorem. If $\widetilde{y}$ and $\widetilde{z}$ are independent random variables with $\widetilde{y} \sim \chi^2_\nu$ and $\widetilde{z} \sim \mathrm{N}(0,1)$, then

$$\widetilde{x} = \frac{\widetilde{z}}{\sqrt{\widetilde{y}\,/\nu}} \sim \mathrm{t}_\nu.$$

Proof. Since $\tilde{z}$ and $\tilde{y}$ are independent, their joint density is given by

$$f_{\widetilde{y},\widetilde{z}}(y,z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\nu/2}} y^{\frac{\nu}{2}-1} e^{-y/2}$$

for $y > 0$, $-\infty < z < \infty$ and $f_{\widetilde{y},\widetilde{z}}(y,z) = 0$ elsewhere. Then, we solve $x = \dfrac{z}{\sqrt{y/\nu}}$ for z, getting $z = x\sqrt{y/\nu}$. Thus, we use the following change of variable:

$$g^{-1}: \begin{cases} z = x\sqrt{y/\nu} \\ \\ y = y \end{cases} \Rightarrow |J_{g^{-1}}| = \sqrt{y/\nu} \;\; ; \qquad g: \begin{cases} x = \dfrac{z}{\sqrt{y/\nu}} \in (-\infty,\infty) \\ \\ y = y > 0 \end{cases}$$

Therefore, the joint density of $\tilde{y}$ and $\tilde{x}$ is given by (check it!):

$$f_{\widetilde{y},\widetilde{x}}(y,x) = \begin{cases} \dfrac{1}{\sqrt{2\pi\nu}\cdot\Gamma\left(\frac{\nu}{2}\right)2^{\nu/2}} y^{\frac{\nu-1}{2}} e^{-\frac{y}{2}\left(1+\frac{x^2}{\nu}\right)} & \text{for } y>0,\ x\in(-\infty,\infty) \\ \\ 0 & \text{elsewhere} \end{cases}$$

and integrating out y with the aid of the change of variable $w = \frac{y}{2}\left(1+\frac{x^2}{\nu}\right)$, which implies that $\frac{dy}{dw} = \dfrac{2}{1+\frac{x^2}{\nu}}$, we get the following density (check it!):

$$f_{\tilde{x}}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\cdot\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad \text{for } x\in(-\infty,\infty),$$

which is the t density with ν degrees of freedom. *Q.E.D.*

– – – – – – – – – – – – – –

The F Distribution

Theorem. If $\widetilde{u}$ and $\widetilde{v}$ are independent random variables with $\widetilde{u} \sim \chi^2_{\nu_1}$ and $\widetilde{v} \sim \chi^2_{\nu_1}$, then

$$\widetilde{y} = \frac{\widetilde{u}/\,\nu_1}{\widetilde{v}/\,\nu_2} \sim \mathrm{F}_{\nu_1,\nu_2}.$$

Proof. Since $\tilde{u}$ and $\tilde{v}$ are independent, the joint density of $\tilde{u}$ and $\tilde{v}$ is given by

$$f_{\widetilde{u},\widetilde{v}}(u,v) = \frac{1}{2^{\nu_1/2}\cdot\Gamma(\frac{\nu_1}{2})}u^{\frac{\nu_1}{2}-1}e^{-\frac{u}{2}}\frac{1}{2^{\nu_2/2}\Gamma(\frac{\nu_2}{2})}v^{\frac{\nu_2}{2}-1}e^{-v/2}$$

$$= \frac{1}{2^{(\nu_1+\nu_2)/2}\cdot\Gamma(\frac{\nu_1}{2})\cdot\Gamma(\frac{\nu_2}{2})}u^{\frac{\nu_1}{2}-1}v^{\frac{\nu_2}{2}-1}e^{-\frac{(u+v)}{2}},$$

for $u>0$, $v>0$ and $f_{\widetilde{u},\widetilde{v}}(u,v)=0$, elsewhere. Then, we solve $y=\dfrac{u/\nu_1}{v/\nu_2}$ for u, getting $u=\dfrac{\nu_1}{\nu_2}vy$. Thus, we use the following change of variable

$$g^{-1}:\begin{cases} u=\dfrac{\nu_1}{\nu_2}vy \\ \\ v=v \end{cases} \Rightarrow |J_{g^{-1}}| = \frac{\nu_1}{\nu_2}v \;\; ; \qquad g:\begin{cases} y=\dfrac{u\,/\nu_1}{v\,/\nu_2}>0 \\ \\ v=v>0 \end{cases}$$

Therefore, the joint density of $\tilde{y}$ and $\tilde{v}$ is given by

$$f_{\widetilde{y},\widetilde{v}}(y,v) = \frac{\left(\dfrac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{2^{(\nu_1+\nu_2)/2}\cdot\Gamma\left(\dfrac{\nu_1}{2}\right)\cdot\Gamma\left(\dfrac{\nu_2}{\nu_1}\right)}y^{\frac{\nu_1}{2}-1}v^{\frac{\nu_1+\nu_2}{2}-1}e^{-\frac{v}{2}(\frac{\nu_1 y}{\nu_2}+1)}$$

for $y>0$ and $v>0$; and $f_{\widetilde{y},\widetilde{v}}(y,v)=0$ elsewhere (check it!). Now integrating out v by using the change of variable $w=\dfrac{v}{2}\left(\dfrac{\nu_1 y}{\nu_2}+1\right)$, we finally get the following density (check it!):

$$f_{\widetilde{y}}(y) = \begin{cases} \dfrac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\dfrac{\nu_1}{2}\right)\cdot\Gamma\left(\dfrac{\nu_2}{2}\right)}\left(\dfrac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}y^{\frac{\nu_1}{2}-1}\left(1+\dfrac{\nu_1}{\nu_2}y\right)^{-\frac{1}{2}(\nu_1+\nu_2)} & \text{for } y>0 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

which is the F density with ν_1 and ν_2 degrees of freedom. *Q.E.D.*

The Chi-square distribution

Values of $\chi^2_{\alpha,\nu}$

ν	$\alpha = .995$	$\alpha = .99$	$\alpha = .975$	$\alpha = .95$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$	ν
1	.0000393	.000157	.000982	.00393	3.841	5.024	6.635	7.879	1
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597	2
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838	3
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860	4
5	.412	.554	.831	1.145	11.070	12.832	15.086	16.750	5
6	.676	.872	1.237	1.635	12.592	14.449	16.812	18.548	6
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156	18
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401	21
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796	22
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181	23
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558	24
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928	25
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290	26
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645	27
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993	28
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336	29
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672	30

The t distribution

Values of $t_{\alpha,\nu}$

ν	$\alpha = .10$	$\alpha = .05$	$\alpha = .025$	$\alpha = .01$	$\alpha = .005$	ν
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
inf.	1.282	1.645	1.960	2.326	2.576	inf.

The F distribution

Values of $F_{.05,\nu_1,\nu_2}$

ν_1 = Degrees of freedom for numerator

ν_2 = Degrees of freedom for denominator	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Values of $F_{.01,\nu_1,\nu_2}$

ν_1 = Degrees of freedom for numerator

ν_2 = Degrees of freedom for denominator	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

ORDER STATISTICS

Definition. The jth order statistic $\tilde{x}_{(j)}$ of a random sample $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ of size n is the random variable of the sample that takes the jth smallest value.

Thus, the order statistics $\tilde{x}_{(1)}, \tilde{x}_{(2)}, ..., \tilde{x}_{(n)}$ of a random sample $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ of size n are the same random variables of the sample relabelled in the same ascending order as the values they take. Thus, the values of the order statistics $\tilde{x}_{(1)}, \tilde{x}_{(2)}, ..., \tilde{x}_{(n)}$ satisfy $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$.

We can define the values of the order statistics as

$$x_{(1)} = \min\{x_1, x_2, ..., x_n\},$$

$$x_{(2)} = \min\left\{\{x_1, x_2, ..., x_n\} / \{x_{(1)}\}\right\},$$

$$\dots$$

$$x_{(j)} = \min\left\{\{x_1, x_2, ..., x_n\} / \{x_{(1)}, x_{(2)}, ..., x_{(j-1)}\}\right\},$$

$$\dots$$

$$x_{(n)} = \min\left\{\{x_1, x_2, ..., x_n\} / \{x_{(1)}, x_{(2)}, ..., x_{(n-1)}\}\right\} = \max\{x_1, x_2, ..., x_n\}.$$

There are some statistics that are easily defined in terms of the order statistics:

(1) The **minimum** $\tilde{x}_{\min}$ of the random sample is $\tilde{x}_{(1)}$ and the **maximum** $\tilde{x}_{\max}$ of the random sample is $\tilde{x}_{(n)}$, where n is the sample size.

(2) **The sample median**, which we will denote by $\tilde{x}_{\text{med}}$, is the number such that one half of the observations are smaller than $\tilde{x}_{\text{med}}$ and one-half are greater. Note that $\tilde{x}_{\text{med}}$ is random since each realization of the random sample will have its own realization of the median. In terms of order statistics, $\tilde{x}_{\text{med}}$ is defined by

$$\tilde{x}_{\text{med}} = \begin{cases} \tilde{x}_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \dfrac{\tilde{x}_{\left(\frac{n}{2}\right)} + \tilde{x}_{\left(\frac{n}{2}+1\right)}}{2} & \text{if } n \text{ is even.} \end{cases}$$

The median is a measure of location (or central tendency) that might be considered an alternative to the sample mean. One advantage of the sample median over the sample mean is that it is less affected by extreme observations (or outliers).

(3) **The sample range**, $\widetilde{R} = \tilde{x}_{(n)} - \tilde{x}_{(1)} = \tilde{x}_{\max} - \tilde{x}_{\min}$, is the distance between the largest and the smallest observations. It is a measure of the dispersion (or variability) in the sample and should reflect the dispersion in the population.

(4) **The p% sample percentile**. The value of the $p\%$ sample percentile, which we will denote by $x_{p\%}$ with $p \in [0, 100]$, is the value such that $pn/100$ of the observed values are smaller than $x_{p\%}$ and $(1-p)n/100$ of the observed values are

greater. In terms of order statistics, the statistic $\tilde{x}_{p\%}$ called $p\%$ sample percentile is defined by

$$\tilde{x}_{p\%} = \tilde{x}_{(j)},$$

where the integer j (between 1 and n) is the nearest to the value

$$\frac{pn}{100} + \frac{1}{2}. \tag{1}$$

If there were two integers j and $j+1$ that are at the same distance from (1), then the $p\%$ percentile is

$$\tilde{x}_{p\%} = \frac{\tilde{x}_{(j)} + \tilde{x}_{(j+1)}}{2}.$$

Note that the 50% percentile is the sample median, the 25% percentile is called the lower quartile, the 75% percentile is called the upper quartile, the 20% percentile is called the lower quintile, and the 80% percentile is called the upper quintile.

(5) **The interquartile range**, which is the distance between the lower and upper quartiles, $\tilde{x}_{75\%} - \tilde{x}_{25\%}$. This is also used as a measure of dispersion of the sample.

MEDIAN OF A POPULATION

The **median** m of a population $\tilde{x}$ having the distribution function F is the real number m for which $P\{\tilde{x} \leq m\} = F(m) \geq 1/2$ and $P\{\tilde{x} \geq m\} \geq 1/2$. If there are two values satisfying this definition, we take the average of them.

If the population $\tilde{x}$ has a continuous distribution function F, then the median m satisfies $F(m) = P\{\tilde{x} \leq m\} = 1/2$ (or, equivalently, $1 - F(m) = P\{\tilde{x} \geq m\} = 1/2$ since $P\{\tilde{x} = m\} = 0$).

The distribution of a random variable $\tilde{x}$ is symmetric with respect to (or around) $x^0 \in \mathbb{R}$ if $P\{\tilde{x} \leq x^0 - z\} = P\{\tilde{x} \geq x^0 + z\}$ for all $z \in \mathbb{R}$. Thus, a distribution is symmetric around x^0 if its probability function (density function) f satisfies $f(x^0 - z) = f(x^0 + z)$ for all z. Note that, when $\tilde{x} \in \mathrm{L}^3$, if the distribution is symmetric then its coefficient of skewness (or asymmetry) is equal to zero.

The **mode** of $\tilde{x}$ (or of its distribution) is the value (or the values) of the random variable for which its probability function (density function) f is at its global maximum. Sometimes, by modifying the density function on a set with zero Lebesgue measure, we can make

$$\max_x f(x) = \sup_x f(x).$$

A distribution is unimodal if it has a single mode.

If a random variable $\tilde{x}$ with well-defined mean μ (i.e., $\tilde{x} \in \mathrm{L}^1$) has a symmetric distribution, then the median and the mean are equal, $m = \mu$.

If a random variable $\tilde{x}$ with well-defined mean μ (i.e., $\tilde{x} \in \mathrm{L}^1$) has a symmetric and unimodal distribution, then the mode, the median, and the mean are all equal.

DISTRIBUTION OF THE MAXIMUM AND THE MINIMUM

Proposition 1. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population $\tilde{x}$ having the distribution function (cdf) F. The distribution functions of the minimum $\tilde{x}_{\min} \equiv \tilde{x}_{(1)}$ and the maximum $\tilde{x}_{\max} \equiv \tilde{x}_{(n)}$ of the sample are

$$F_{\tilde{x}_{\min}}(x) = 1 - [1 - F(x)]^n \tag{2}$$

and

$$F_{\tilde{x}_{\max}}(x) = [F(x)]^n . \tag{3}$$

Proof. For the minimum $\tilde{x}_{\min}$,

$$F_{\tilde{x}_{\min}}(x) = F_{\tilde{x}_{(1)}}(x) = P\left\{\tilde{x}_{(1)} \leq x\right\} = 1 - P\left\{\tilde{x}_{(1)} > x\right\}$$

$$= 1 - P\left\{\tilde{x}_1 > x, \tilde{x}_2 > x, ..., \tilde{x}_n > x\right\}$$

$$= 1 - P\left\{\tilde{x}_1 > x\right\} \cdot P\left\{\tilde{x}_2 > x\right\} \cdot ... \cdot P\left\{\tilde{x}_n > x\right\}$$

$$= 1 - \left[P\left\{\tilde{x} > x\right\}\right]^n = 1 - \left[1 - P\left\{\tilde{x} \leq x\right\}\right]^n = 1 - [1 - F(x)]^n .$$

For the maximum $\tilde{x}_{\max}$,

$$F_{\tilde{x}_{\max}}(x) = F_{\tilde{x}_{(n)}}(x) = P\left\{\tilde{x}_{(n)} \leq x\right\} = P\left\{\tilde{x}_1 \leq x, \tilde{x}_2 \leq x, ..., \tilde{x}_n \leq x\right\}$$

$$= P\left\{\tilde{x}_1 \leq x\right\} \cdot P\left\{\tilde{x}_2 \leq x\right\} \cdot ... \cdot P\left\{\tilde{x}_n \leq x\right\}$$

$$\left[P\left\{\tilde{x} \leq x\right\}\right]^n = [F(x)]^n . \qquad Q.E.D.$$

Observe that

$$\lim_{n\to\infty} F_{\tilde{x}_{\min}}(x) = \lim_{n\to\infty} \left(1 - [1 - F(x)]^n\right) = \begin{cases} 0 & \text{if } F(x) = 0 \\ \\ 1 & \text{if } F(x) > 0 \end{cases}$$

and

$$\lim_{n\to\infty} F_{\tilde{x}_{\max}}(x) = \lim_{n\to\infty} [F(x)]^n = \begin{cases} 0 & \text{if } F(x) < 1 \\ \\ 1 & \text{if } F(x) = 1. \end{cases}$$

Let $M_F = \sup\left\{x \in \overline{\mathbb{R}} \,|F(x) < 1)\right\}$, then the extended random variable $\tilde{x}_{\max}$ converges in distribution as $n \to \infty$ to the constant M_F whether it is finite or infinite. Let $m_F = \inf\left\{x \in \overline{\mathbb{R}} \,|F(x) > 0)\right\}$, then the extended random variable $\tilde{x}_{\min}$ converges in distribution as $n \to \infty$ to the constant m_F whether it is finite or infinite. Thus, the distributions of both the minimum and the maximum of the random sample tend to be degenerate as the sample size goes to infinity.

DISTRIBUTION FUNCTION OF THE ORDER STATISTICS

Proposition 2. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population $\tilde{x}$ having the distribution function F. Then, the distribution function $F_{\tilde{x}_{(j)}}$ of the jth order statistic $\tilde{x}_{(j)}$ is

$$F_{\tilde{x}_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}. \tag{4}$$

Proof. Fix $x \in \mathbb{R}$, and let $\tilde{y}$ be a random variable that counts the number of random variables in $\{\tilde{x}_i\}_{i=1}^n$ that are smaller than or equal to x. For each of the random variables in $\{\tilde{x}_i\}_{i=1}^n$, call the event $\{\tilde{x}_i \leq x\}$ a "success" and $\{\tilde{x}_i > x\}$ a "failure". Then $\tilde{y}$ is the number of successes in n trials. Thus, $\tilde{y}$ is binomial with parameters n and $F(x)$, $\tilde{y} \sim \mathrm{B}(n, F(x))$. The event $\{\tilde{x}_{(j)} \leq x\}$ is equivalent to the event $\{\tilde{y} \geq j\}$, that is, at least j of the sample values are smaller than or equal to x. Then,

$$F_{\tilde{x}_{(j)}}(x) = P\left\{\tilde{x}_{(j)} \leq x\right\} = P\{\tilde{y} \geq j\} = \sum_{k=j}^{n} P\{\tilde{y} = k\} = \sum_{k=j}^{n} b(y; n, F(x)),$$

which becomes (4). *Q.E.D.*

Observe that equation (4) becomes (2) when $j = 1$. To see this, note that the binomial probability function $b(\cdot; n, F(x))$ satisfies

$$\sum_{k=0}^{n} b(x; n, F(x)) = \sum_{k=0}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} = 1$$

so that

$$\sum_{k=1}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} = 1 - \binom{n}{0} [F(x)]^0 [1 - F(x)]^{n-0} = 1 - [1 - F(x)]^n.$$

Therefore,

$$F_{\tilde{x}_{\min}}(x) = F_{\tilde{x}_{(1)}}(x) = \sum_{k=1}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} = 1 - [1 - F(x)]^n.$$

Moreover, when $j = n$, equation (4) becomes (3),

$$F_{\tilde{x}_{\max}}(x) = F_{\tilde{x}_{(n)}}(x) = \sum_{k=n}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}$$

$$= \binom{n}{n} [F(x)]^n [1 - F(x)]^0 = [F(x)]^n.$$

PROBABILITY FUNCTION OF THE ORDER STATISTICS

Proposition 3. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a discrete population $\tilde{x}$ having the probability function (pmf) $f : \tilde{x}(\Omega) \longrightarrow [0,1]$, where $\{x_1, x_2, ...\}$ are the different values of the range $\tilde{x}(\Omega)$ of $\tilde{x}$. Then, the probability function, $f_{\tilde{x}_{(j)}} : \tilde{x}(\Omega) \longrightarrow [0,1]$, of the jth order statistics $\tilde{x}_{(j)}$ is

$$f_{\tilde{x}_{(j)}}(x_i) =$$

$$\sum_{k=j}^{n} \binom{n}{k} \left(\underbrace{\left[\sum_{x \leq x_i} f(x) \right]^k}_{P\{\tilde{x} \leq x_i\}} \underbrace{\left[\sum_{x > x_i} f(x) \right]^{n-k}}_{P\{\tilde{x} > x_i\}} - \underbrace{\left[\sum_{x < x_i} f(x) \right]^k}_{P\{\tilde{x} < x_i\}} \underbrace{\left[\sum_{x \geq x_i} f(x) \right]^{n-k}}_{P\{\tilde{x} \geq x_i\}} \right)$$

for all $x_i \in \tilde{x}(\Omega)$.

Proof. It is immediate from using (4) and noticing that the probability function $f_{\tilde{x}_{(j)}}$ of $\tilde{x}_{(j)}$ satisfies

$$f_{\tilde{x}_{(j)}}(x_i) = F_{\tilde{x}_{(j)}}(x_i) - \lim_{x \to x_i^-} F_{\tilde{x}_{(j)}}(x), \quad \text{for all } x_i \in \tilde{x}(\Omega). \tag{5}$$

Therefore,

$$f_{\tilde{x}_{(j)}}(x_i) =$$

$$\sum_{k=j}^{n} \binom{n}{k} \left([F(x_i)]^k [1 - F(x_i)]^{n-k} - \left[\lim_{x \to x_i^-} F(x) \right]^k \left[1 - \lim_{x \to x_i^-} F(x) \right]^{n-k} \right),$$

for all $x_i \in \tilde{x}(\Omega)$. Since $F(x_i) = \sum_{x \leq x_i} f(x) = P\{\tilde{x} \leq x_i\}$ and $\lim_{x \to x_i^-} F(x) = \sum_{x < x_i} f(x) = P\{\tilde{x} < x_i\}$, we get the desired result. *Q.E.D.*

The previous proposition implies that the value of the probability function of the jth order statistics evaluated at the lowest value, say x_1, of the range $\tilde{x}(\Omega)$ of the discrete population $\tilde{x}$ is

$$f_{\tilde{x}_{(j)}}(x_1) = F_{\tilde{x}_{(j)}}(x_1) = \sum_{k=j}^{n} \binom{n}{k} [F(x_1)]^k [1 - F(x_1)]^{n-k}$$

$$= \sum_{k=j}^{n} \binom{n}{k} [f(x_1)]^k [1 - f(x_1)]^{n-k}$$

since $\lim_{x \to x_1^-} F(x) = 0$ and $F(x_1) = f(x_1)$.

Corollary 1. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a discrete population $\tilde{x}$ having the probability function (pmf) $f : \tilde{x}(\Omega) \longrightarrow [0,1]$, where $\{x_1, x_2, ...\}$

are the different values of the range $\tilde{x}(\Omega)$ of $\tilde{x}$. Then, the probability functions $f_{\tilde{x}_{\min}}$ and $f_{\tilde{x}_{\max}}$ of the minimum and of the maximum of the sample are

$$f_{\tilde{x}_{\min}}(x_i) = \underbrace{\left[\sum_{x \geq x_i} f(x)\right]^n}_{P\{\tilde{x} \geq x_i\}} - \underbrace{\left[\sum_{x > x_i} f(x)\right]^n}_{P\{\tilde{x} > x_i\}} \text{ for all } x_i \in \tilde{x}(\Omega),$$

and

$$f_{\tilde{x}_{\max}}(x_i) = \underbrace{\left[\sum_{x \leq x_i} f(x)\right]^n}_{P\{\tilde{x} \leq x_i\}} - \underbrace{\left[\sum_{x < x_i} f(x)\right]^n}_{P\{\tilde{x} < x_i\}} \text{ for all } x_i \in \tilde{x}(\Omega).$$

Proof. Combine (5) for $j = 1$ and for $j = n$ with (2) and (3), respectively, to get

$$f_{\tilde{x}_{\min}}(x_i) = \left[1 - \lim_{x \to x_i^-} F(x)\right]^n - [1 - F(x_i)]^n$$

and

$$f_{\tilde{x}_{\max}}(x_i) = [F(x_i)]^n - \left[\lim_{x \to x_i^-} F(x)\right]^n,$$

for all $x_i \in \tilde{x}(\Omega)$, and the result immediately follows. $Q.E.D.$

DENSITY FUNCTION OF THE ORDER STATISTICS

Proposition 4. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from an absolutely continuous population $\tilde{x}$ having the density function (pdf) f and the distribution function F. Then, the density $f_{\tilde{x}_{(j)}}$ of the jth order statistic is

$$f_{\tilde{x}_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x) [F(x)]^{j-1} [1 - F(x)]^{n-j}. \tag{6}$$

Proof. One way of proving this proposition is to compute the derivative of (4) with respect to x. This is a little bit painful.

We will use an alternative approach in our proof. We can assume safely that $x_{(1)} < x_{(2)} < ... < x_{(n)}$ since $P\{\tilde{x}_i = \tilde{x}_j\} = 0$ for all $i \neq j$ when the population is continuous. Then, we divide the real axis into the following three intervals: $(-\infty, x)$, $[x, x+h]$, and $(x+h, \infty)$. Then, the probability that the jth order statistic $\tilde{x}_{(j)}$ falls into the interval $[x, x+h]$ is equal to the probability that $j-1$ of the sample values fall into the interval $(-\infty, x)$, one falls into the

interval $[x, x+h]$, and $n-j$ fall into the interval $(x+h, \infty)$. Therefore, since the distribution function F of $\tilde{x}$ is continuous, we have

$$P\left\{\tilde{x}_{(j)} \in [x, x+h]\right\} =$$

$$\frac{n!}{(j-1)!1!(n-j)!}\left[P\left\{\tilde{x} \in (-\infty, x)\right\}\right]^{j-1} \cdot P\left\{\tilde{x} \in [x, x+h]\right\} \cdot \left[P\left\{\tilde{x} \in (x+h, \infty)\right\}\right]^{n-j}$$

$$= \frac{n!}{(j-1)!(n-j)!}\left[F(x)\right]^{j-1} \cdot P\left\{\tilde{x} \in [x, x+h]\right\} \cdot \left[1-F(x+h)\right]^{n-j}, \tag{7}$$

according to the formula of the multinomial distribution.

We know that, for an absolutely continuous random variable $\tilde{y}$,

$$P\left\{\tilde{y} \in [x, x+h]\right\} = F_{\tilde{y}}(x+h) - F_{\tilde{y}}(x),$$

where $F_{\tilde{y}}$ is the (continuous) distribution function of $\tilde{y}$. Moreover, using the mean value theorem, there exists a value $\varepsilon_{\tilde{y}} \in [x, x+h]$ such that

$$F_{\tilde{y}}(x+h) - F_{\tilde{y}}(x) = F'_{\tilde{y}}(\varepsilon_{\tilde{y}})h = f_{\tilde{y}}(\varepsilon_{\tilde{y}})h, \quad \text{a.e.},$$

where $f_{\tilde{y}}$ is the density of $\tilde{y}$.

Thus, (7) becomes

$$f_{\tilde{x}_{(j)}}(\varepsilon_{\tilde{x}_{(j)}})h = \frac{n!}{(j-1)!(n-j)!}\left[F(x)\right]^{j-1} \cdot f(\varepsilon_{\tilde{x}})h \cdot \left[1-F(x+h)\right]^{n-j},$$

for some $\varepsilon_{\tilde{x}_{(j)}} \in [x, x+h]$ and $\varepsilon_{\tilde{x}} \in [x, x+h]$. If we let $h \to 0$ and divide by h, the previous expression becomes

$$f_{\tilde{x}_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!}\left[F(x)\right]^{j-1} \cdot f(x) \cdot \left[1-F(x)\right]^{n-j}. \qquad Q.E.D.$$

Corollary 2. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from an absolutely continuous population $\tilde{x}$ having the density f and the distribution function F. The density functions $f_{\tilde{x}_{\min}}$ and $f_{\tilde{x}_{\max}}$ of the minimum and the maximum of the sample are

$$f_{\tilde{x}_{\min}}(x) = nf(x)\left[1-F(x)\right]^{n-1} \tag{8}$$

and

$$f_{\tilde{x}_{\max}}(x) = nf(x)\left[F(x)\right]^{n-1}. \tag{9}$$

Proof. Just compute the derivative with respect to x of the distribution functions obtained in Proposition 1 or, alternatively, evaluate (6) at $j=1$ and at $j=n$. *Q.E.D.*

DENSITY FUNCTION OF THE MEDIAN

Consider a random sample $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ of size n, where n is an odd number. Obviously, the sample median $\tilde{x}_{\text{med}}$ is the $((n+1)/2)$th order statistic. Then, using equation (6) for this statistic, we get immediately the following:

Proposition 4. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n (with n being an odd number) from an absolutely continuous population $\tilde{x}$ having the density f and the distribution function F. Then, the density $f_{\tilde{x}_{\text{med}}}$ of the sample median $\tilde{x}_{\text{med}}$ is

$$f_{\tilde{x}_{\text{med}}}(x) = \frac{n!}{\left[\left(\frac{n-1}{2}\right)!\right]^2} f(x) \left[F(x)\right]^{(n-1)/2} \left[1 - F(x)\right]^{(n-1)/2}.$$

We next provide with a nice proposition, which we state without a proof.

Proposition 5. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from an absolutely continuous population $\tilde{x}$ having the median m and the density f. Assume that $f(m) > 0$ and that f is continuously differentiable in a neighborhood of m. Then, as $n \to \infty$,

$$2\sqrt{n} f(m) \left(\tilde{x}_{\text{med}} - m\right) \longrightarrow \text{N}(0,1), \tag{10}$$

or, equivalently,

$$\sqrt{n} \left(\tilde{x}_{\text{med}} - m\right) \longrightarrow \text{N}\left(0, \frac{1}{4\left[f(m)\right]^2}\right). \tag{11}$$

Corollary 3. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a normal population $\tilde{x}$ with the mean μ and the variance σ^2, $\tilde{x} \sim \text{N}(\mu, \sigma^2)$. Then, as $n \to \infty$,

$$\sqrt{\frac{2}{\pi}} \left(\frac{\tilde{x}_{\text{med}} - \mu}{\sigma / \sqrt{n}}\right) \longrightarrow \text{N}(0,1), \tag{12}$$

or, equivalently,

$$\sqrt{n} \left(\tilde{x}_{\text{med}} - \mu\right) \longrightarrow \text{N}\left(0, \frac{\pi\sigma^2}{2}\right). \tag{13}$$

Proof. First, note that $m = \mu$ since the normal distribution is symmetric. Moreover, the normal density evaluated at μ is equal to $f(\mu) = 1/\sigma\sqrt{2\pi}$. Therefore, the expression (10) becomes (12), while (11) becomes (13). *Q.E.D.*

If the population is normal, $\tilde{x} \sim \text{N}(\mu, \sigma^2)$, the previous corollary tells us that, for n large, $\text{E}(\tilde{x}_{\text{med}}) \approx \mu$. Remember that the sample mean $\bar{\boldsymbol{x}}$, satisfies $\text{E}(\bar{\boldsymbol{x}}) = \mu$. Thus, for n large, both $\tilde{x}_{\text{med}}$ and $\bar{\boldsymbol{x}}$ have approximately the same mean, which is the population mean μ. However, for n large, the sample median $\tilde{x}_{\text{med}}$ has a larger variance than the sample mean $\bar{\boldsymbol{x}}$,

$$\text{Var}(\tilde{x}_{\text{med}}) \approx \frac{\pi\sigma^2}{2n} > \text{E}(\bar{\boldsymbol{x}}) = \frac{\sigma^2}{n},$$

as follows from (13) and the fact that $\pi/2 > 1$.

EXTREME VALUE THEORY

We have already seen that, if $\{\tilde{x}_1, \tilde{x}_2, ...\}$ are i.i.d. random variables with common cdf F, then $\tilde{x}^n_{\max} = \max\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ converges in distribution as $n \to \infty$ to the constant $M(F) = \sup\left\{x \in \overline{\mathbb{R}} \,|F(x) < 1)\right\}$ whether it is finite or infinite. Thus, the distribution of the maximum of the random sample tends to be degenerate as the sample size goes to infinity.

Extremal types theorem. Assume that for suitable constants $c_n > 0$ and b_n, the random variable $\tilde{z}_n = \dfrac{\tilde{x}^n_{\max} - b_n}{c_n}$ converges in distribution to a non-degenerate random variable, $F_{\tilde{z}_n} \xrightarrow{w} G$. Then, the limiting distribution function G is of one of the following three classes:
(i)

$$G(x) = e^{-e^{-\left(\frac{x-m}{s}\right)}}, \quad \text{for all } x \in \mathbb{R},$$

for some $m \in \mathbb{R}$ and $s > 0$.
(ii)

$$G(x) = \begin{cases} 0 & \text{for } x < m \\ \\ e^{-\left(\frac{x-m}{s}\right)^{-\alpha}} & \text{for } x \geq m. \end{cases}$$

for some $m \in \mathbb{R}$, $s > 0$, and $\alpha > 0$.
(iii)

$$G(x) = \begin{cases} e^{-\left(-\frac{x-m}{s}\right)^{\alpha}} & \text{for } x < m \\ \\ 1 & \text{for } x \geq m \end{cases}$$

for some $m \in \mathbb{R}$, $s > 0$, and $\alpha > 0$.

Note that the classes (i) and (ii) correspond to the generalized extreme value type I (Gumbel) and type II (Fréchet) distributions, respectively. Concerning the class III, recall that the generalized extreme value type III (Weibull) distribution has the following distribution function:

$$F(y; m_0, s, \alpha) = P(\tilde{y} \leq y) = \begin{cases} 0 & \text{for } y < m_0 \\ \\ 1 - e^{-\left(\frac{y-m_0}{s}\right)^{\alpha}} & \text{for } y \geq m_0. \end{cases}$$

for $m \in \mathbb{R}$, $s > 0$, and $\alpha > 0$. Therefore, the distribution function of the random variable $\tilde{x} = -\tilde{y}$ will be

$$F_{\tilde{x}}(x) = P(\tilde{x} \leq x) = P(-\tilde{y} \leq x) = P(\tilde{y} \geq -x) = 1 - P(\tilde{y} \leq -x) = 1 - F(-x; m_0, s, \alpha)$$

so that

$$F_{\tilde{x}}(x) = 1 - F(-x; m_0, s, \alpha) = \begin{cases} 1 - 0 = 1 & \text{for } -x < m_0 \text{ or } x \geq -m_0 \\ \\ 1 - \left[1 - e^{-\left(\frac{-x-m_0}{s}\right)^{\alpha}}\right] = e^{-\left(\frac{-x-m_0}{s}\right)^{\alpha}} & \text{for } -x \geq m_0 \text{ or } x < -m_0. \end{cases}$$

Making $m = -m_0 \in \mathbb{R}$, the previous distribution function becomes

$$F_{\tilde{x}}(x) = \begin{cases} 1 & \text{for } x \geq m \\ \\ e^{-\left(\frac{-x+m}{s}\right)^{\alpha}} = e^{-\left(-\frac{x-m}{s}\right)^{\alpha}} & \text{for } x < m, \end{cases}$$

which coincides with the distribution function of the class (iii). Thus, the distribution of the class (iii) is the distribution of the negative of a random variable that has the generalized extreme value type III (or Weibull) distribution. This distribution is called **negative Weibull.**

Exercises. Probability and Statistics. IDEA.
7. Sampling

1. Verify the following computing formula for the value of the sample variance:

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n\,(n-1)}.$$

2. If $\tilde{x}_1, \tilde{x}_2, ...,$ and $\tilde{x}_n$ are independent random variables having identical Bernoulli distributions with the parameter θ, then the sample mean $\bar{\boldsymbol{x}}_{\mathbf{n}}$ is the proportion of successes in n trials.

   (a) Verify that $\mathrm{E}(\bar{\boldsymbol{x}}_{\mathbf{n}}) = \theta$ and $\mathrm{Var}(\bar{\boldsymbol{x}}_{\mathbf{n}}) = \dfrac{\theta\,(1-\theta)}{n}$.

   (b) Using the moment generating function method, prove that, if $\tilde{x}_1, ..., \tilde{x}_n$ are independently distributed random variables having a Bernoulli distribution with parameter θ, then $\tilde{y}_n = \sum_{i=1}^{n} \tilde{x}_i$ has a binomial distribution with parameters n and θ.

   (c) Use parts (a) and (b), and the central limit theorem to prove that

$$\frac{\tilde{y}_n - n\theta}{\sqrt{n\theta\,(1-\theta)}} \longrightarrow \mathrm{N}\,(0,1)$$

3. A random sample of size $n = 100$ is taken from a population with the mean $\mu = 75$ and the variance $\sigma^2 = 256$. Use Chebyshev's inequality to provide a lower bound for the probability that the value of the sample mean $\bar{\boldsymbol{x}}$ falls between 67 and 83?

4. Use the central limit theorem to find an approximate value for the probability of Exercise 3.

5. A random sample of size 100 is taken from a normal population with $\sigma = 25$. Find the probability that the mean of the sample will differ from the mean of the population by 3 or more either way.

6. Find approximate values for the probability that a random variable $\tilde{x}$ having a chi-square distribution with 50 degrees of freedom will take on a value greater than 68.

   (a) by treating $\dfrac{\tilde{x} - \nu}{\sqrt{2\nu}}$ with $\nu = 50$ as a random variable having the standard normal distribution;

   (b) by treating $\sqrt{2\tilde{x}} - \sqrt{2\nu}$ with $\nu = 50$ as a random variable having the standard normal distribution.

   Also, judge the merits of these approximations, given that the actual value of the probability (rounded to five decimals) is 0.04596.

   (c) Justify the approximations given in (a) and (b). *Hint:* You should prove that, for all ν and for every positive real valued random variable $\tilde{x}$, the

probability that the random variable $\sqrt{2\tilde{x}} - \sqrt{2\nu}$ takes on a value less than k equals the probability that the random variable $\dfrac{\tilde{x} - \nu}{\sqrt{2\nu}}$ takes on a value less than $k + \dfrac{k^2}{2\sqrt{2\nu}}$.

7. Show that the density of a F distribution with 4 and 4 degrees of freedom is given by

$$g(y) = \begin{cases} 6y\,(1+y)^{-4} & \text{for } y > 0 \\ \\ 0 & \text{elsewhere} \end{cases}$$

and use this density to find the probability that for independent random samples of size 5 from normal populations having the same variance, $\dfrac{\mathbf{s_1^2}}{\mathbf{s_2^2}}$ will take on a value less than 1/2 or greater than 2.

8. If $\tilde{x}$ has an F distribution with ν_1 and ν_2 degrees of freedom, show that $\tilde{y} = \dfrac{1}{\tilde{x}}$ has the F distribution with ν_2 and ν_1 degrees of freedom.

9. If $\tilde{x}$ has an F distribution with ν_1 and ν_2 degrees of freedom, show that when $\nu_2 \to \infty$ the distribution of $\nu_1\tilde{x}$ approaches the chi-square distribution with ν_1 degrees of freedom.

10. Show that if $\tilde{x}$ has the t distribution with ν degrees of freedom, then $\tilde{x}^2$ has the F distribution with 1 and ν degrees of freedom.

11. The claim that the variance of a normal population is $\sigma^2 = 25$ is to be rejected if the variance of a random sample of size 16 exceeds 54.668 or is less than 12.102. What is the probability that this claim will be rejected even though $\sigma^2 = 25$?

12. Let $\boldsymbol{s}_n$ be the standard deviation of a random sample of size n from a normal population. Use the approximation given in part (b) of Exercise 6, to prove that for large n the variance of the sampling distribution of $\boldsymbol{s}_n$ is approximately $\dfrac{\sigma^2}{2(n-1)}$.

13. Consider an experiment consisting of randomly selecting n values from a finite set of distinct numbers $C = \{c_1, c_2, ..., c_N\}$ with $N \geq n$. Let us assume that the selection is without replacement and $\tilde{x}_1$ is the first number drawn, $\tilde{x}_2$ is the second number drawn,..., and $\tilde{x}_n$ is the nth number drawn.

    (a) Find the joint probability function $f(x_1, x_2, ..., x_n)$ for each ordered n-tuple of distinct values selected from the set C.

    (b) Find the probability of each subset of n of the N elements of the set C, regardless of the order in which the values are obtained.

    (c) Find the marginal probability function of each $\tilde{x}_i$, $f(x_i)$ for $x_i = c_1, c_2, ..., c_N$.

    (d) Find the mean μ and the variance σ^2 of each $\tilde{x}_i$.

(e) Find the marginal probability function of any two of the random variables $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$, $g(x_i, x_j)$, for each ordered pair (x_i, x_j) of distinct values of the set C.

For the rest of this exercise let us assume that we do not know the values $\{c_1, c_2, ..., c_N\}$, but we know that the mean and the variance of each $\tilde{x}_i$ are μ and σ^2, respectively.

(f) Find the covariance of any two of the random variables $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$, $\text{Cov}(\tilde{x}_i, \tilde{x}_j)$ with $i \neq j$. *Note:* Your answer should be a function of σ^2 and N only since $\text{Cov}(\tilde{x}_i, \tilde{x}_j)$ turns out to be independent of n and μ.

(g) Let $\bar{\boldsymbol{x}}$ be the average of the random variables $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$, $\bar{\boldsymbol{x}} = \dfrac{1}{n}\sum_{i=1}^{n} \tilde{x}_i$.

Find $\text{Var}(\bar{\boldsymbol{x}})$. *Note:* Your answer should be a function of σ^2, N and n only since $\text{Var}(\bar{\boldsymbol{x}})$ turns out to be independent of μ.

(h) Compute $\underset{N\to\infty}{\text{Var}}(\bar{\boldsymbol{x}})$. Discuss this result.

14. The density function of the random vector $(\tilde{x}, \tilde{y})$ is

$$f(x,y) = \begin{cases} \dfrac{1}{\pi} e^{-\frac{1}{2}(x^2+y^2)} & \text{for } 0 < x < \infty, \quad -\infty < y < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

(a) Prove that $f(x, y)$ is the joint density of two independent random variables $\tilde{x}$ and $\tilde{y}$, where $\tilde{x}$ is the absolute value of a standard normal random variable and $\tilde{y}$ is standard normal.

(b) Find $\text{E}(\tilde{x})$, $\text{E}(\tilde{y})$, $\text{Var}(\tilde{x})$, $\text{Var}(\tilde{y})$, and $\text{Cov}(\tilde{x}, \tilde{y})$.

(c) Find the density function of the random variable $\tilde{v} = \dfrac{\tilde{y}}{\tilde{x}}$. Show all the computations.

(d) Prove that the random variable $\tilde{v}$ has a Student's t distribution with one degree of freedom, $\tilde{v} \sim \text{t}_1$.

15. Assume that $\bar{\boldsymbol{x}}$ is the mean of a random sample $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ of size n from a population having the mean μ and the finite variance σ^2. Note that $\bar{\boldsymbol{x}}$ is a random variable.

Find

(a) $\text{Cov}(\tilde{x}_i, \tilde{x}_j)$, for $i \neq j$.

(b) $\text{Cov}(\bar{\boldsymbol{x}}, \tilde{x}_i)$, for $i = 1, ..., n$.

(c) $\text{Cov}(\tilde{x}_i, \tilde{x}_i - \bar{\boldsymbol{x}})$, for $i = 1, ..., n$.

(d) $\text{Cov}(\tilde{x}_i, \tilde{x}_j - \bar{\boldsymbol{x}})$, for $i \neq j$.

(e) $\text{Var}(\tilde{x}_i - \bar{\boldsymbol{x}})$, for $i = 1, ..., n$.

(f) $\text{Cov}(\tilde{x}_i - \bar{\boldsymbol{x}}, \tilde{x}_j - \bar{\boldsymbol{x}})$, for $i \neq j$.

(*g*) $\mathrm{Cov}\left(\bar{\boldsymbol{x}}, \tilde{x}_i - \bar{\boldsymbol{x}}\right)$, for $i = 1, ..., n$.

Hint: You can use the results of Exercise 29 of List 3.

16. (*a*) Prove that, as the number ν of degrees of freedom goes to infinity, the limit of the t density is the standard normal density.

(*b*) Prove that the t distribution with one degree of freedom is a Cauchy distribution with parameters $\alpha = 0$ and $\beta = 1$. Recall that the density of the Cauchy distribution with parameters α and β is

$$f(x) = \frac{\beta/\pi}{(x-\alpha)^2 + \beta^2} \quad \text{for} - \infty < x < \infty.$$

(*c*) Assume that the random variable $\tilde{x} : (\Omega, \mathcal{F}, P) \longrightarrow (\mathbb{R}, \mathcal{B})$ has the t distribution with one degree of freedom. Compute the conditional expectation of $\tilde{x}$ given that $\tilde{x} \in [0, 1]$. Recall that

$$\mathrm{E}\left(\tilde{x} \mid \tilde{x} \in [0,1]\right) \equiv \mathrm{E}\left(\tilde{x} \mid \mathbb{I}_A = 1\right),$$

where $\mathbb{I}_A$ is the indicator function of the set $A = \{\omega \in \Omega \mid \tilde{x}(\omega) \in [0,1]\}$. Moreover, if $\psi(x) = \arctan(x)$, where the function $\arctan(\cdot) : \mathbb{R} \longrightarrow (-\pi/2, \pi/2)$ is the inverse function of $\tan(\cdot) : (-\pi/2, \pi/2) \longrightarrow \mathbb{R}$, then the derivative of ψ is $\psi'(x) = \dfrac{1}{1+x^2}$. Recall also that $\tan\left(\dfrac{\pi}{4}\right) = 1$, $\tan\left(-\dfrac{\pi}{4}\right) = -1$, $\lim\limits_{\theta \to \pi/2} \tan(\theta) = \infty$, $\lim\limits_{\theta \to -\pi/2} \tan(\theta) = -\infty$ and $\tan(0) = 0$.

17. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample of size n from a population $\tilde{x}$ having the uniform density on $(0, 1)$.

(*a*) Find the density $f_{\tilde{x}_{(j)}}$ of the jth order statistic. Could you identify the distribution of the jth order statistic $\tilde{x}_{(j)}$? Find the expectation of $\tilde{x}_{(j)}$. Plot the density of the 2nd order statistic $\tilde{x}_{(2)}$ when the sample size is $n = 7$ and find the expectation of $\tilde{x}_{(2)}$ in this case.

(*b*) Find the expectation of the sample median $\tilde{x}_{\mathrm{med}}$. Find the density of the sample median when the sample size is $n = 7$. Plot this density.

18. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample of size n from a population $\tilde{x}$ having the exponential density with the parameter θ.

(*a*) Find the densities, $f_{\tilde{x}_{\min}}$ and $f_{\tilde{x}_{\max}}$, and the distribution functions, $F_{\tilde{x}_{\min}}$ and $F_{\tilde{x}_{\max}}$, of the minimum and the maximum of the sample, respectively. Could you identify the distribution of the sample minimum $\tilde{x}_{\min}$? Find the expectation of $\tilde{x}_{\min}$. Plot the density of the sample maximum $\tilde{x}_{\max}$ when $\theta = 3$ and $n = 11$ and find the expectation of both $\tilde{x}_{\min}$ and $\tilde{x}_{\max}$ in this case.

(*b*) Find the median and the mode of the population distribution and compare them with the population mean. Find the density of the sample median $\tilde{x}_{\mathrm{med}}$ when n is an odd number. Plot the density of $\tilde{x}_{\mathrm{med}}$ when $\theta = 3$ and $n = 11$. For this case, find the expectation of $\tilde{x}_{\mathrm{med}}$ and compare it with the expectation

of the sample mean, the population mean, the population median, the mode of the population distribution and the mode of the distribution of $\tilde{x}_{\text{med}}$.

19. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a collection of independent random variables having the Poisson distribution with the parameters λ_i, for $i = 1, 2, ..., n$.

    (a) Use the formula for the convolution pmf to prove that the sum $\tilde{S} = \sum_{i=1}^{n} \tilde{x}_i$ has a Poisson distribution with the parameter $\sum_{i=1}^{n} \lambda_i$.

    (b) Find the probability function of the average $\bar{\boldsymbol{x}} = \dfrac{\sum_{i=1}^{n} \tilde{x}_i}{n}$.

    (c) Assume now that $\lambda_i = \lambda$, for $i = 1, 2, ..., n$, which implies that $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ is a random sample. Find the probability function of the sample mean $\bar{\boldsymbol{x}}$.

    (d) Under the assumption of part (c) prove that $\sqrt{n}\,(\bar{\boldsymbol{x}}_n - \lambda) \longrightarrow \mathrm{N}(0, \lambda)$, as $n \longrightarrow \infty$, i.e., $\sqrt{n}\,(\bar{\boldsymbol{x}}_n - \lambda)$ is asymptotically distributed as a normal random variable with zero mean and variance λ, where $\bar{\boldsymbol{x}}_{\boldsymbol{n}} = \dfrac{\sum_{i=1}^{n} \tilde{x}_i}{n}$.

20. Consider a random sample of size n from an absolutely continuous population. Compute the derivative of the distribution function $F_{\tilde{x}_{(j)}}(x)$ of the jth order statistic given in the expression (4) of the handout "Order Statistics" to obtain the corresponding density $f_{\tilde{x}_{(j)}}(x)$, which is given in the expression (6).

21. We say that a random variable $\tilde{x}$ has a distribution symmetric with respect to x^0 if $P\{\tilde{x} \leq x^0 - z\} = P\{\tilde{x} \geq x^0 + z\}$ for all $z \in \mathbb{R}$.

    Let us define the random variable $\tilde{y} = \tilde{x} - x^0$.

    (a) Prove that $\tilde{x}$ has a distribution symmetric with respect to x^0 if and only if the random variable $\tilde{y}$ has a distribution symmetric with respect to zero.

    (b) Prove that the random variable $\tilde{y}$ has a distribution symmetric with respect to zero if and only if $\tilde{y}$ has the same distribution as the random variable $-\tilde{y}$.

    (c) Assume that the random variable $\tilde{y}$ is discrete (absolutely continuous). Prove that $\tilde{y}$ has a distribution symmetric with respect to zero if and only if its probability (density) function satisfies $f_{\tilde{y}}(-y) = f_{\tilde{y}}(y)$ for all y, i.e., the function $f_{\tilde{y}}$ is symmetric with respect to zero.

    (d) Assume that the random variable $\tilde{x}$ is discrete (absolutely continuous). Prove that $\tilde{x}$ has a distribution symmetric with respect to x^0 if and only if its probability (density) function satisfies $f_{\tilde{x}}(x^0 - z) = f_{\tilde{x}}(x^0 + z)$ for all z, i.e., the function $f_{\tilde{x}}$ is symmetric with respect to x^0.

    For the rest of the exercise assume that the random variable $\tilde{x}$ has a distribution symmetric with respect to x^0.

    (e) Prove that $\mathrm{E}(\tilde{x}) = x^0$ when $\tilde{x} \in \mathrm{L}^1$.

    (f) Prove that the coefficient of asymmetry (or skewness) of $\tilde{x}$ is equal to zero when $\tilde{x} \in \mathrm{L}^3$.

8. Estimation

8.1. Point estimation

- Point estimation is the use of the value $\widehat{\theta}$ of a statistic $\widehat{\boldsymbol{\theta}} = g\left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$ to make a guess for the unknown parameter vector $\theta \in \Theta \subset \mathbb{R}^K$ characterizing the distribution $P_{\tilde{X}}\left(\cdot; \theta\right)$ of the random vector (or sample) $\tilde{X} = \left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$.
- Note that, since the parameter θ is unknown, the estimator $\widehat{\boldsymbol{\theta}}$ is an statistic that cannot depend on θ but only on the sample $\left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$.
- The statistic $\widehat{\boldsymbol{\theta}} = g\left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$ we use for estimation is called **estimator**. The value $\widehat{\theta} = g\left(x_1, x_2, \ldots, x_n\right)$ taken by an estimator for a particular sample value $X = \left(x_1, x_2, \ldots, x_n\right)$ is called an **estimate**.
- For instance, $\mathbf{s}^2 \neq \sigma^2$ in general. Questions: what is the relationship between $\mathbf{s}^2$ and σ^2?, how "close" are they?, does $\mathbf{s}_n^2$ converge to σ^2?, etc.

8.2. The mean square error of an estimator and the relative efficiency of estimators

- **Definition.** The mean square error (MSE) of an estimator $\widehat{\boldsymbol{\theta}}$ for the population parameter $\theta \in \mathbb{R}$ is

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}-\theta\right)^2\right].$$

- If $\theta \in \mathbb{R}^K$, then the MSE is

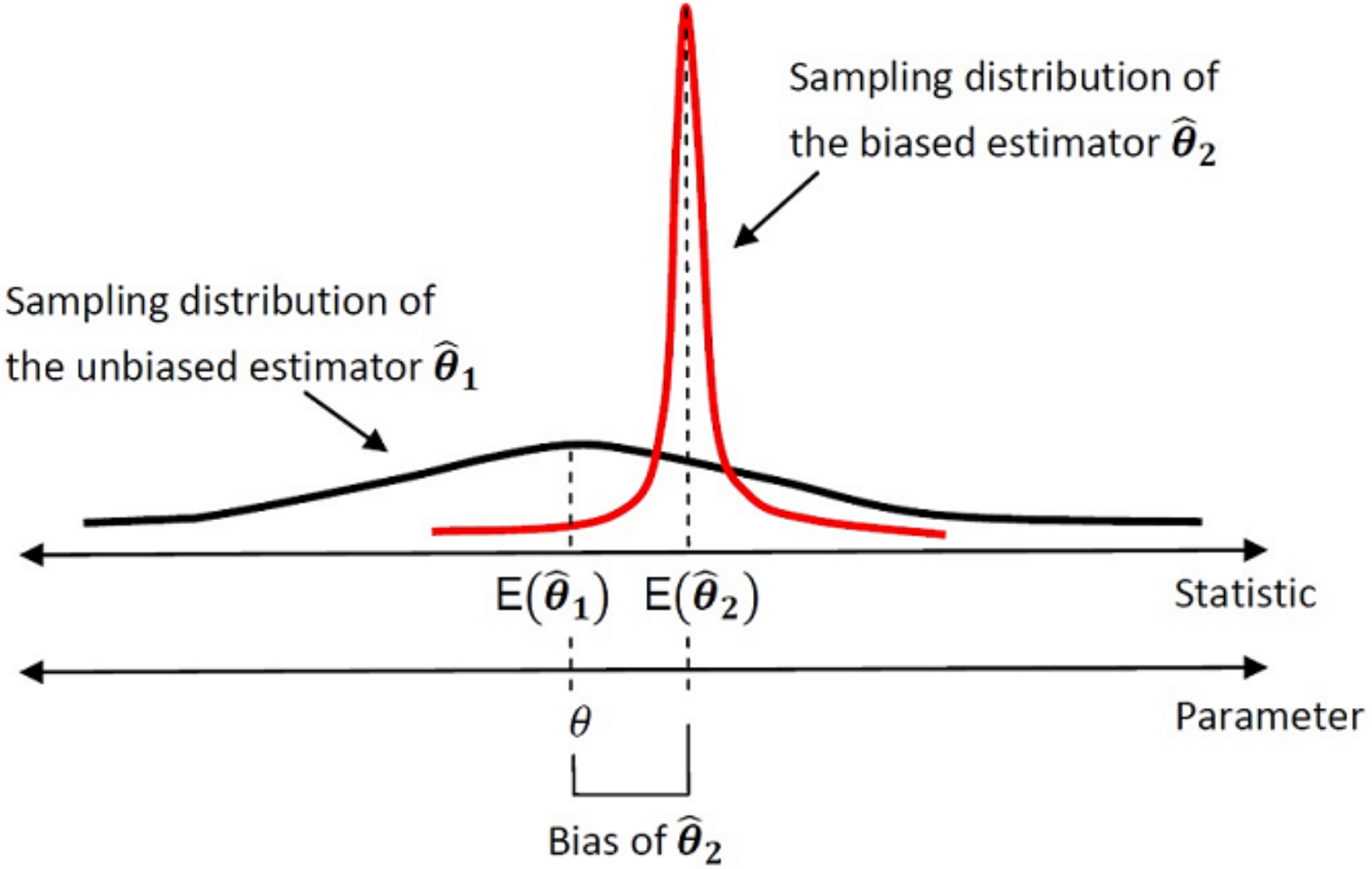
$$\mathrm{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\theta\right\|^2\right].$$

- Note that

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}-\theta\right)^2\right]=\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}-\theta\right)+\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}}-\theta\right)\right]^2=\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)+\left[b_{\widehat{\boldsymbol{\theta}}}\left(\theta\right)\right]^2,$$

where $b_{\widehat{\boldsymbol{\theta}}}\left(\theta\right)=\mathrm{E}\left(\widehat{\boldsymbol{\theta}}-\theta\right)=\mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right)-\theta$ is the bias of the estimator $\widehat{\boldsymbol{\theta}}$.

A biased estimator with small variance may be preferable to an unbiased estimator with large variance



- **Definition.** We say that $\widehat{\boldsymbol{\theta}}_1$ is a better (or a more efficient) estimator for $\theta \in \mathbb{R}$ than $\widehat{\boldsymbol{\theta}}_2$ if

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_1 - \theta\right)^2\right] \leq \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_2 - \theta\right)^2\right].$$

- If the two previous estimators for θ were unbiased, $b_{\widehat{\boldsymbol{\theta}}_1}(\theta) = b_{\widehat{\boldsymbol{\theta}}_2}(\theta) = 0$, then $\widehat{\boldsymbol{\theta}}_1$ is better (or more efficient) than $\widehat{\boldsymbol{\theta}}_2$ if

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_1\right) \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_2\right).$$

- **Definition.** We say that $\widehat{\boldsymbol{\theta}}^*$ is the best (or the most efficient) estimator in the class C of estimators for $\theta \in \mathbb{R}$ if

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}^* - \theta\right)^2\right] \leq \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\right], \quad \text{for all } \widehat{\boldsymbol{\theta}} \in C.$$

8.3. The sample mean and sample variance as unbiased estimators

- Consider a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from a population $\tilde{x}$ with distribution $P_{\tilde{x}}$ having the mean μ and the finite variance σ^2.
- Sample mean:

$$\bar{\mathbf{x}} = \bar{\mathbf{x}}_n = \frac{\sum_{i=1}^{n} \tilde{x}_i}{n}.$$

- Sample variance:

$$\mathbf{s^2} = \mathbf{s}_n^2 = \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}_n\right)^2}{n-1}.$$

- Reminder:

(a)

$$\mathrm{E}\left(\bar{\mathbf{x}}\right) = \mu \quad \text{and} \quad \mathrm{Var}\left(\bar{\mathbf{x}}\right) = \frac{\sigma^2}{n}.$$

(b) Strong law of large numbers:

$$\bar{\mathbf{x}}_n \xrightarrow{a.s} \mu.$$

(c) Central limit theorem:

$$\tilde{z}_n \equiv \frac{\bar{\mathbf{x}}_n - \mathrm{E}(\bar{\mathbf{x}}_n)}{\sqrt{\mathrm{Var}\left(\bar{\mathbf{x}}_n\right)}} = \frac{\bar{\mathbf{x}}_n - \mu}{\sigma \big/ \sqrt{n}} \overset{a}{\sim} \mathrm{N}(0,1) \qquad (\text{or} \quad \tilde{z}_n \longrightarrow \mathrm{N}(0,1))$$

or, equivalently,

$$\sqrt{n}\left(\bar{\mathbf{x}}_n - \mu\right) \longrightarrow \mathrm{N}(0, \sigma^2).$$

(d) If the population is normal, then

$$\bar{\mathbf{x}}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \iff \frac{\bar{\mathbf{x}}_n - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1) \iff \sqrt{n}\left(\bar{\mathbf{x}}_n - \mu\right) \sim \mathrm{N}(0, \sigma^2).$$

(e) $\mathbf{s}^2$ is an unbiased estimator for σ^2 :

$$\mathrm{E}\left(\mathbf{s}^2\right) = \sigma^2.$$

(f) If the population mean μ is known, then the statistic

$$\breve{\mathbf{s}}^2 = \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \mu\right)^2}{n}$$

is an unbiased estimator for σ^2, $\mathrm{E}\left(\breve{\mathbf{s}}^2\right) = \mathrm{E}\left(\mathbf{s}^2\right) = \sigma^2$.

(g) If $\hat{\mathbf{s}}^2 = \dfrac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}\right)^2}{n} = \dfrac{(n-1)\mathbf{s}^2}{n}$, then

$$\mathrm{E}\left(\hat{\mathbf{s}}^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2 = \mathrm{E}\left(\mathbf{s}^2\right).$$

However,

$$\lim_{n\to\infty} \mathrm{E}\left(\hat{\mathbf{s}}_n^2\right) = \mathrm{E}\left(\mathbf{s}_n^2\right) = \sigma^2.$$

(h) If the population is normal, then

$$\frac{(n-1)\mathbf{s}^2}{\sigma^2} \equiv \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}\right)^2}{\sigma^2} \sim \chi^2_{n-1}.$$

(i) If the population is normal, then

$$\mathrm{Var}\left(\mathbf{s}^2\right) = \frac{2\sigma^4}{n-1}, \quad \mathrm{Var}\left(\hat{\mathbf{s}}^2\right) = \frac{2(n-1)\sigma^4}{n^2}, \quad \text{and} \quad \mathrm{Var}\left(\breve{\mathbf{s}}^2\right) = \frac{2\sigma^4}{n}.$$

(j) If the population is normal, then

$$\frac{\bar{\mathbf{x}}_n - \mu}{\mathbf{s} \big/ \sqrt{n}} \sim \mathrm{t}_{n-1}$$

(k) If $\mathbf{s}_1^2$ and $\mathbf{s}_2^2$ are the variances of two independent random samples of size n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2, respectively, then,

$$\frac{\mathbf{s}_1^2 \big/ \sigma_1^2}{\mathbf{s}_2^2 \big/ \sigma_2^2} = \frac{\sigma_2^2 \cdot \mathbf{s}_1^2}{\sigma_1^2 \cdot \mathbf{s}_2^2} \sim \mathrm{F}_{n_1-1, n_2-1}.$$

In particular, if the two population variances are equal, $\sigma_1^2 = \sigma_2^2$, then

$$\frac{\mathbf{s}_1^2}{\mathbf{s}_2^2} \sim \mathrm{F}_{n_1-1, n_2-1}.$$

- Assume that the population is normal. Then,

$$\mathrm{E}\left(\mathbf{s^2}\right) = \sigma^2 \quad \text{so that } b_{\mathbf{s^2}}\left(\sigma^2\right) = 0,$$

while

$$\mathrm{E}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2$$

so that the bias of $\hat{\mathbf{s}}^{\mathbf{2}}$ as an estimator for σ^2 is

$$b_{\hat{\mathbf{s}}^{\mathbf{2}}}\left(\sigma^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

- Moreover,

$$\mathrm{Var}\left(\mathbf{s^2}\right) = \frac{2\sigma^4}{n-1} \quad \text{and} \quad \mathrm{Var}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \frac{2(n-1)\sigma^4}{n^2}.$$

- If the population is normal, $\mathbf{s^2}$ is an unbiased estimator for σ^2, while $\hat{\mathbf{s}}^{\mathbf{2}}$ is a biased estimator. However, $\mathrm{Var}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) < \mathrm{Var}\left(\mathbf{s^2}\right)$. In fact, it can be proved that $\hat{\mathbf{s}}^{\mathbf{2}}$ is an estimator for σ^2 more efficient than $\mathbf{s^2}$,

$$\mathrm{E}\left[\left(\hat{\mathbf{s}}^{\mathbf{2}} - \sigma^2\right)^2\right] < \mathrm{E}\left[\left(\mathbf{s^2} - \sigma^2\right)^2\right]. \qquad \text{(Exercise)}$$

8.4. The Cramér-Rao lower bound for unbiased estimators

Harald Cramér (1893 - 1985)

C. R. Rao (1920 - 2023)

- **Theorem (Cramér-Rao).** Let $\tilde{X}$ be a n-dimensional vector of random variables (not necessarily independent), the joint density of which is given by $h(X;\theta)$, where θ is a K-dimensional vector of parameters in some parameter space Θ. Let $\widehat{\boldsymbol{\theta}} = g\left(\tilde{X}\right)$ be an unbiased estimator for θ with a finite covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\theta}}\right)$. Furthermore, assume that $h(X;\theta)$ and $g(X)$ satisfy

- (a)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} h dX = \int_{\mathbb{R}^n} \frac{\partial h}{\partial\theta} dX \leftarrow \text{This is a } K\text{-dimensional column vector}$$

- (b)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} ghdX = \int_{\mathbb{R}^n} g\frac{\partial h}{\partial\theta} dX \leftarrow \text{This is a } K\text{-dimensional column vector}$$

- (c)

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} \frac{\partial h}{\partial \theta^T} dX = \int_{\mathbb{R}^n} \frac{\partial^2 h}{\partial \theta \partial \theta^T} dX \leftarrow \text{This is a } K \times K\text{-dimensional matrix}$$

- (d)

$$\text{Var}\left(\frac{\partial \ln h\left(\tilde{X}; \theta\right)}{\partial \theta} \right) \text{ is a positive definite } K \times K \text{ matrix.}$$

- Then, for all $\theta \in \Theta$,

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-\text{E}\left[\frac{\partial^2 \ln h\left(\tilde{X}; \theta\right)}{\partial \theta \partial \theta^T} \right] \right)^{-1} \equiv CR \qquad \text{(CR-1)}$$

or, equivalently,

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(\text{E}\left[\frac{\partial \ln h\left(\tilde{X}; \theta\right)}{\partial \theta} \cdot \frac{\partial \ln h\left(\tilde{X}; \theta\right)}{\partial \theta^T} \right] \right)^{-1} \equiv CR. \qquad \text{(CR-2)}$$

- *Note:* $A \geqq B$ means that $A - B$ is a positive semi-definite matrix, which implies that the elements along the diagonal of $A - B$ are non-negative. Thus, the Cramér-Rao inequality provides a lower bound for $\text{Var}\left(\widehat{\boldsymbol{\theta}}_k\right)$, for all $k = 1, ..., K$.
- **Proof.** See the handout for the case $K = 1$.

- **About the assumptions:**
  - Assumption (d) is very mild.
  - Assumptions (a), (b) and (c) mean that the operations of differentiation and integration can be interchanged. This implies, among other things, that the domain of positive density of $\tilde{X}$ is independent of θ. Sometimes such a domain depends on the parameters (e.g. the uniform distribution).
- The matrix

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2 \ln h\left(\tilde{X};\theta\right)}{\partial\theta\partial\theta^T}\right] = \mathrm{E}\left[\frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta}\cdot\frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta^T}\right]$$

  is called the Fisher information (matrix). Note that the Cramér-Rao lower bound (CR) is equal to $\left[I(\theta)\right]^{-1}$.
- Therefore, if $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator for θ and $\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) = CR$, then $\widehat{\boldsymbol{\theta}}$ is the best (or the most efficient) estimator within the class of unbiased estimators for θ. In this case, we say that the estimator $\widehat{\boldsymbol{\theta}}$ is an efficient estimator for θ.

- If the random vector $\tilde{X}$ were discrete instead of absolutely continuous, the result would be the same. In this case $h(X;\theta)$ would be the probability function of the random vector $\tilde{X}$ and we should change the integrals in both the statement and the proof by sums over the range of $\tilde{X}$. Obviously, this range should not depend on the parameters we are estimating in order to fulfil the assumptions (a), (b) and (c).

- **Corollary.** If in the previous theorem the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is a random sample from a population $\tilde{x}$ having the density (or probability function) $f(x;\theta)$, then

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-n\mathrm{E}\left[\frac{\partial^2 \ln f(\tilde{x};\theta)}{\partial\theta\partial\theta^T}\right]\right)^{-1} \tag{i}$$

or, equivalently,

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(n\mathrm{E}\left[\frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta} \cdot \frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta^T}\right]\right)^{-1}. \tag{ii}$$

- **Proof.** See the handout for the case $K = 1$.

- **Single parameter case:** $K = 1$ $(\theta \in \mathbb{R})$.
- **(a)** For $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$, not necessarily a random sample, we have

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-\text{E}\left[\frac{\partial^2 \ln h\left(\tilde{X};\theta\right)}{\partial\theta^2}\right]\right)^{-1} \quad \text{(Theorem CR-1 for } K = 1\text{)}$$

or, equivalently,

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(\text{E}\left[\left(\frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta}\right)^2\right]\right)^{-1}. \quad \text{(Theorem CR-2 for } K = 1\text{)}$$

- **(b)** For the random sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$, we have

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-n\text{E}\left[\frac{\partial^2 \ln f\left(\tilde{x};\theta\right)}{\partial\theta^2}\right]\right)^{-1} \quad \text{(Corollary (i) for } K = 1\text{)}$$

or, equivalently,

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(n\text{E}\left[\left(\frac{\partial \ln f\left(\tilde{x};\theta\right)}{\partial\theta}\right)^2\right]\right)^{-1}. \quad \text{(Corollary (ii) for } K = 1\text{)}$$

8.5. Asymptotic properties of estimators: consistent estimators

- Let $\tilde{x}_n$ be a random variable with $\tilde{x}_n \sim P_n$, for $n = 1, 2, ...$ The limit of the expectation (or limiting expectation) of $\tilde{x}_n$ is

$$\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right) = \lim_{n\to\infty} \int_{\mathbb{R}} x dP_n(x),$$

whereas the asymptotic expectation is

$$\mathrm{AE}\left(\tilde{x}_n\right) = \int_{\mathbb{R}} x d\widehat{P},$$

where $\tilde{x}_n \longrightarrow \widehat{P}$ as $n \longrightarrow \infty$. Thus, $\widehat{P}$ is the limiting distribution of $\left\{\tilde{x}_n\right\}_{n=1}^{\infty}$.

- **Theorem.** If $\mathrm{E}\left(\left|\tilde{x}_n\right|^r\right) < M$ for all n, where $M < \infty$ is a fixed bound, then

$$\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n^s\right) = \mathrm{AE}\left(\tilde{x}_n^s\right), \quad \text{for all } s < r.$$

  In particular, if $\mathrm{E}\left(\tilde{x}_n^2\right) < M$ for all n, where $M < \infty$ is a fixed bound, then $\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right) = \mathrm{AE}\left(\tilde{x}_n\right)$.

- Some statisticians refer to $\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right)$ as the asymptotic expectation.

- Similarly, we can define the limit of the variance (or limiting variance) and the asymptotic variance as

$$\lim_{n\to\infty} \mathrm{Var}\left(\tilde{x}_n\right) = \lim_{n\to\infty} \mathrm{E}\left(\left[\tilde{x}_n - \mathrm{E}\left(\tilde{x}_n\right)\right]^2\right)$$

  and

$$\mathrm{AVar}\left(\tilde{x}_n\right) = \mathrm{AE}\left(\left[\tilde{x}_n - \mathrm{AE}\left(\tilde{x}_n\right)\right]^2\right),$$

  respectively.

- The subindex n in estimators will typically refer to the sample size.

- **Definition.** We say that the estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is "unbiased in the limit" or "asymptotically unbiased" if

$$\lim_{n\to\infty} \mathrm{E}\left(\widehat{\boldsymbol{\theta}}_n\right) = \theta \qquad \text{or} \qquad \mathrm{AE}\left(\widehat{\boldsymbol{\theta}}_n\right) = \theta,$$

respectively.

Moreover, the "limiting bias" or the "asymptotic bias" of the estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is

$$\lim_{n\to\infty} \mathrm{E}\left(\widehat{\boldsymbol{\theta}}_n\right) - \theta \qquad \text{or} \qquad \mathrm{AE}\left(\widehat{\boldsymbol{\theta}}_n\right) - \theta,$$

respectively.

- **Definition.** An estimator $\widehat{\boldsymbol{\theta}}_n^*$ is "the best (or the most efficient) in the limit" or "asymptotically the best (or the most efficient)" in the class C of estimators for $\theta \in \mathbb{R}$ if

$$\lim_{n\to\infty} \frac{\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n^* - \theta\right)^2\right]}{\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right]} \leq 1 \text{ or } \frac{\mathrm{AE}\left[\phi(n)\left(\widehat{\boldsymbol{\theta}}_n^* - \theta\right)^2\right]}{\mathrm{AE}\left[\phi(n)\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right]} \leq 1, \text{ for all } \widehat{\boldsymbol{\theta}}_n \in C,$$

  respectively.

- The function $\phi(n)$ is selected in order to make the previous relative asymptotic efficiency well defined in $\mathbb{R}$ (i.e., to make $\mathrm{AE}\left[\phi(n)\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right] \neq 0$). Usually, we use $\phi(n) = n$.

- **Definition.** Let $\widehat{\theta}_n$ be an "unbiased in the limit" or an "asymptotically unbiased" estimator for $\theta \in \mathbb{R}$. Then, $\widehat{\theta}_n$ is an "efficient in the limit" or an "asymptotically efficient" estimator for θ if

$$\lim_{n\to\infty} \underbrace{\left(\frac{CR_n}{\text{Var}\left(\widehat{\theta}_n\right)} \right)}_{\leq 1 \text{ (Cramér-Rao)}} = 1 \quad \text{or} \quad \frac{CR^*}{\text{AVar}\left(\sqrt{\phi(n)}\,\widehat{\theta}_n\right)} = 1,$$

respectively, where CR^* is the following "adjusted" asymptotic Cramer-Rao lower bound (or inverse of the Fisher information):

$$CR^* = \left(-\text{AE}\left[\frac{1}{\phi(n)} \frac{\partial^2 \ln h\left(\tilde{X};\theta\right)}{\partial \theta^2} \right] \right)^{-1} =$$

$$\left(\text{AE}\left[\left(\frac{1}{\sqrt{\phi(n)}} \frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial \theta} \right)^2 \right] \right)^{-1}.$$

- If $\phi(n) = n$ and $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is a random sample from a population $\tilde{x}$ having the density function (probability function) $f(x;\theta)$, then

$$CR^* = \left(-\text{AE}\left[\frac{\partial^2 \ln f(\tilde{x};\theta)}{\partial\theta^2}\right]\right)^{-1} = \left(\text{AE}\left[\left(\frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta}\right)^2\right]\right)^{-1}.$$

- *Note:* An estimator may be the best (or the most efficient) in the limit (or asymptotically) in the class of unbiased estimators in the limit (or asymptotically) even if it is not efficient in the limit (or asymptotically). In other words, the Cramér-Rao lower bound does not need to be reached in the limit (or asymptotically) by an statistic that is the best (or the most efficient) unbiased estimator in the limit (or asymptotically).

- **Definition.** An estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is said to be (weakly) consistent if

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \theta \quad \left(\text{or } \operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\theta}}_n = \theta \right),$$

  which is equivalent to $\widehat{\boldsymbol{\theta}}_n \xrightarrow{d} \theta$.
- **Definition.** An estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is said to be strongly consistent if

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s} \theta.$$

- **Theorem.** If $\lim_{n\to\infty} \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right] = 0$, then $\widehat{\boldsymbol{\theta}}_n$ is a (weakly) consistent estimator for $\theta \in \mathbb{R}$.
- **Proof.** Obvious since $\widehat{\boldsymbol{\theta}}_n \xrightarrow{m} \theta$ implies that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \theta$.
- Note that, since

$$\lim_{n\to\infty} \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right] = \lim_{n\to\infty} \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_n\right) + \lim_{n\to\infty} \left[b_{\widehat{\boldsymbol{\theta}}_n}(\theta)\right]^2,$$

  then $\widehat{\boldsymbol{\theta}}_n \xrightarrow{m} \theta$ if and only if $\lim_{n\to\infty} \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_n\right) = \lim_{n\to\infty} b_{\widehat{\boldsymbol{\theta}}_n}(\theta) = 0$.

- Let $\bar{\mathbf{x}}_n$ and $\mathbf{s}_n^2$ be the mean and variance of a random sample of size n.
- Since $\mathrm{E}\left(\bar{\mathbf{x}}_n\right) = \mu$ (or $b_{\bar{\mathbf{x}}_n}\left(\mu\right) = 0$) for all n, and $\mathrm{Var}\left(\bar{\mathbf{x}}_n\right) = \sigma^2 / n$ so that $\lim_{n\to\infty} \mathrm{Var}\left(\bar{\mathbf{x}}_n\right) = 0$, the sample mean $\bar{\mathbf{x}}_n$ is a (weakly) consistent estimator for the population mean μ.
- In fact, $\bar{\mathbf{x}}_n \xrightarrow{a.s} \mu$ because of any of the strong law of large numbers.
- Assume that the population is normal. Then $\mathrm{E}\left(\mathbf{s}_n^2\right) = \sigma^2$ $\left(\text{or } b_{\mathbf{s}_n^2}\left(\sigma^2\right) = 0\right)$ for all n, and $\mathrm{Var}\left(\mathbf{s}_n^2\right) = 2\sigma^4 / \left(n-1\right)$ so that $\lim_{n\to\infty} \mathrm{Var}\left(\mathbf{s}_n^2\right) = 0$. Hence, the sample variance $\mathbf{s}_n^2$ is a (weakly) consistent estimator for the population variance σ^2.

- Consider now the statistic

$$\hat{\mathbf{s}}_n^2 = \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}_n\right)^2}{n} = \frac{(n-1)\mathbf{s}_n^2}{n}.$$

Then,

$$\mathrm{E}\left(\hat{\mathbf{s}}_n^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2$$

and the bias is

$$b_{\hat{\mathbf{s}}_n^2}\left(\sigma^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} \longrightarrow 0, \text{ as } n \to \infty.$$

Moreover,

$$\mathrm{Var}\left(\hat{\mathbf{s}}_n^2\right) = \frac{2(n-1)\sigma^4}{n^2} \to 0, \text{ as } n \to \infty.$$

- This means that $\hat{\mathbf{s}}_n^2 \xrightarrow{m} \sigma^2$ and, thus, $\hat{\mathbf{s}}_n^2$ is a (weakly) consistent estimator for σ^2.

8.6. Sufficient estimators

- The statistic $\widehat{\boldsymbol{\theta}}$ is a sufficient estimator (or statistic) for the parameter θ if it uses all the information relevant for the estimation of the population parameter $\theta \in \Theta \subset \mathbb{R}^K$, that is, if all the knowledge we can gain about θ by actually specifying the individual sample values and their order can just as well be obtained by observing only the value of the statistic $\widehat{\boldsymbol{\theta}}$.

- Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (a sample) whose joint distribution $P_{\tilde{X}}(\cdot; \theta)$ depends on the parameter θ. Then, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}}$, the conditional distribution of $\tilde{X}$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$, $P_{\tilde{X}|\widehat{\boldsymbol{\theta}}}\left(\cdot \left| \widehat{\theta}; \theta \right.\right)$ should not depend on the value θ. Note that, if $P_{\tilde{X}|\widehat{\boldsymbol{\theta}}}\left(\cdot \left| \widehat{\theta}; \theta \right.\right)$ depends on θ, particular values of $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ yielding a particular value of $\widehat{\boldsymbol{\theta}}$ are more likely for some values of θ than for others, and the knowledge of these sample values will help in the estimation for θ.

- **Definition.** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution $P_{\tilde{X}}(\cdot; \theta)$ depends on the parameter $\theta \in \Theta \subset \mathbb{R}^K$. The statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for θ if, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}}$, the conditional distribution of the sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$, $P_{\tilde{X}|\widehat{\theta}}\left(\cdot \left| \widehat{\theta}; \theta \right.\right) : \mathcal{B}(\mathbb{R}^n) \longrightarrow [0, 1]$, is independent of θ.

- *Note:* If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ has a discrete (absolutely continuous) distribution, then sufficiency means that the conditional probability function (density) of $\tilde{X}$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$,

$$f_{\tilde{X}|\widehat{\theta}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}; \theta \right.\right) = \frac{f_{(\tilde{X}, \widehat{\boldsymbol{\theta}})}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right)}{f_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\theta}; \theta\right)},$$

for all $\widehat{\theta}$ with $f_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\theta}; \theta\right) \neq 0$, does not depend on the parameter θ.

Note that

$$f_{\left(\tilde{X},\widehat{\theta}\right)}\left(\underbrace{x_1,x_2,...,x_n}_{X\in\mathbb{R}^n},\widehat{\theta};\theta\right)=\begin{cases} f_{\tilde{X}}\left(x_1,x_2,...,x_n;\theta\right) & \text{if } \widehat{\theta}=t\left(X\right), \\ 0 & \text{otherwise.}\end{cases} \quad (*)$$

and, for $\tilde{X}$ absolutely continuous,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\int_{\mathbb{R}^n} f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)dX=\int_A f_{\tilde{X}}\left(x_1,x_2,...,x_n;\theta\right)dX,$$

$$\text{with } A=\left\{X=\left(x_1,x_2,...,x_n\right)\in\mathbb{R}^n\,\middle|\,t(X)=\widehat{\theta}\right\},$$

whereas, if $\tilde{X}$ is discrete,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\sum_{X\in\prod_{i=1}^{n}\tilde{x}_i(\Omega)} f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)=\sum_{X\in B} f_{\tilde{X},}\left(x_1,x_2,...,x_n;\theta\right),$$

$$\text{with } B=\left\{X=\left(x_1,x_2,...,x_n\right)\in\prod_{i=1}^{n}\tilde{x}_i\left(\Omega\right)\,\middle|\,t(X)=\widehat{\theta}\right\}.$$

- **Factorization theorem (Fisher-Neyman).** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution is discrete (absolutely continuous) and depends on the parameter $\theta \in \Theta \subset \mathbb{R}^K$. Then, the statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for the parameter θ if and only if the probability function (density) of $\tilde{X}$ can be factorized as follows:

$$f_{\tilde{X}}(x_1, x_2, ..., x_n; \theta) = \underbrace{h(x_1, x_2, ..., x_n)}_{\text{independent of } \theta} \cdot g\left(\widehat{\theta}, \theta\right),$$

where $\widehat{\theta} = t(x_1, x_2, ..., x_n)$.

- **Proof.** (Sufficiency $\Longrightarrow$ Factorization). If

$$f_{\tilde{X}|\widehat{\theta}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}\,; \theta\right.\right) = \frac{f_{\left(\tilde{X}, \widehat{\theta}\right)}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right)}{f_{\widehat{\theta}}\left(\widehat{\theta}; \theta\right)}$$

is independent of θ, for all values $\widehat{\theta}$ taken by $\widehat{\theta} = t\left(\tilde{X}\right)$, then

$$\underbrace{f_{\left(\tilde{X}, \widehat{\theta}\right)}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right)}_{f_{\tilde{X}}(x_1, x_2, ..., x_n; \theta) \leftarrow \text{from } (*)}$$

$$= \underbrace{f_{\tilde{X}|\widehat{\theta}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}\,; \theta\right.\right)}_{\text{independent of } \theta \rightarrow\ m\left(X, \widehat{\theta}\right) = m(X, t(X)) = h(X)} \times \underbrace{f_{\widehat{\theta}}\left(\widehat{\theta}; \theta\right)}_{g\left(\widehat{\theta}, \theta\right)}.$$

- (Factorization $\Longrightarrow$ Sufficiency). If

$$f_{\tilde{X}}\left(x_1, x_2, ..., x_n; \theta\right) = h\left(x_1, x_2, ..., x_n\right) \cdot g\left(\widehat{\theta}, \theta\right),$$

then, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}} = t\left(\tilde{X}\right)$,

$$f_{\tilde{X}\left|\widehat{\theta}\right.}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}\,; \theta\right.\right) = \frac{\overbrace{h\left(x_1, x_2, ..., x_n\right) \cdot g\left(\widehat{\theta}, \theta\right)}^{f_{\left(\tilde{X}, \widehat{\boldsymbol{\theta}}\right)}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right) \leftarrow \text{from } (*)}}{f_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\theta}; \theta\right)}. \quad (**)$$

If the random vector $\tilde{X}$ is absolutely continuous,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\int_{A}\underbrace{h\left(x_1,x_2,...,x_n\right)\cdot g\left(\widehat{\theta},\theta\right)}_{f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)}dX=g\left(\widehat{\theta},\theta\right)\cdot\underbrace{\int_{A}h\left(x_1,x_2,...,x_n\right)dX}_{c_1\left(\widehat{\theta}\right)},$$

where $A=\left\{X=\left(x_1,x_2,...,x_n\right)\in\mathbb{R}^n\,\middle|\,t(X)=\widehat{\theta}\right\}$ or, if $\tilde{X}$ is discrete,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\sum_{X\in B}\underbrace{h\left(x_1,x_2,...,x_n\right)\cdot g\left(\widehat{\theta},\theta\right)}_{f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)}=g\left(\widehat{\theta},\theta\right)\cdot\underbrace{\sum_{X\in B}h\left(x_1,x_2,...,x_n\right)}_{c_2\left(\widehat{\theta}\right)},$$

where $B=\left\{X=\left(x_1,x_2,...,x_n\right)\in\prod_{i=1}^{n}\tilde{x}_i\left(\Omega\right)\,\middle|\,t(X)=\widehat{\theta}\right\}$ and $c_i\left(\widehat{\theta}\right)$, $i=1,2$, only depend on the value $\widehat{\theta}$ of the statistic. Therefore, from $(**)$ we get

$$f_{\tilde{X}\left|\widehat{\theta}\right.}\left(x_1,x_2,...,x_n\,\middle|\,\widehat{\theta};\theta\right)=\frac{h\left(x_1,x_2,...,x_n\right)}{c_i\left(\widehat{\theta}\right)},\quad i=1,2,$$

which is independent of θ. *Q.E.D.*

8.7. The Rao-Blackwell theorem

C. R. Rao (1920 - 2023) - David Blackwell (1919 - 2010)

- **Theorem (Rao-Blackwell).** Let $\widehat{\boldsymbol{\theta}}$ be an estimator for the parameter θ. Suppose that $\tilde{t} = t\left(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\right)$ is a sufficient statistic for θ and let $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$. Then,

$$\mathrm{E}\left[\left(\boldsymbol{\theta}^* - \theta\right)^2\right] \leq \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\right]. \tag{1}$$

- **Proof.** Note that, from the sufficiency of $\tilde{t}$, $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$ does not depend on the parameter θ but only on the sample $\left(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\right)$ so that $\boldsymbol{\theta}^*$ is an estimator. Then,

$$\mathrm{E}\left[\left(\boldsymbol{\theta}^* - \theta\right)^2\right] = \mathrm{E}\left[\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}}\middle|\, \tilde{t}\right) - \theta\right]^2\right] = \mathrm{E}\left[\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}} - \theta\middle|\, \tilde{t}\right)\right]^2\right]$$

$$\leq \mathrm{E}\left[\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\middle|\, \tilde{t}\right]\right] = \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\right],$$

where the inequality follows from Jensen's inequality applied to a quadratic, and thus convex, function and the last equality follows from Adam's law. *Q.E.D.*

- The previous theorem tells us that an improvement in efficiency of an estimator can be obtained by taking its conditional expectation with respect to a sufficient statistic. This process of improvement is called Rao-Blackwellization. Thus, $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$ is the Rao-Blackwellized estimator.
- In other words, the theorem says that, if an estimator is not a function of a sufficient statistic, it can be improved in terms of MSE.
- **Corollary.** Let $\widehat{\boldsymbol{\theta}}$ be an unbiased estimator for the parameter θ. Suppose that $\tilde{t}$ is a sufficient statistic for θ and let $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$. Then, (a) $\boldsymbol{\theta}^*$ is an unbiased estimator for θ, and
(b) $\mathrm{Var}\left(\boldsymbol{\theta}^*\right) \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)$.
- **Proof. (a)** $\mathrm{E}\left(\boldsymbol{\theta}^*\right) = \mathrm{E}\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)\right] = \mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right) = \theta$, where the second equality follows from Adam's law and the last from the unbiasedness of $\widehat{\boldsymbol{\theta}}$. **(b)** From the Rao-Blackwell theorem, we know that the inequality (1) holds. Since both $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ are unbiased estimators for θ, then inequality (1) simply becomes $\mathrm{Var}\left(\boldsymbol{\theta}^*\right) \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)$. *Q.E.D.*

8.8. The method of moments

- **Definition.** The kth (non-central) sample moment of the value $X = (x_1, x_2, ..., x_n)$ taken by a random sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is given by

$$m'_k(X) = \frac{\sum_{i=1}^{n} x_i^k}{n}.$$

- The population moments depend on the parameter $\theta \in \Theta \subset \mathbb{R}^K$, $\mu'_k(\theta) \equiv \mathrm{E}\left(\tilde{x}^k\right)$, $k = 1, 2, ...$
- The Method of Moments (MM) consists of solving the following system of equations:

$$m'_k(X) = \mu'_k(\theta), \text{ for } k = 1, 2, ..., K,$$

for the K parameters of the population distribution.

- The solution is the estimate $\widehat{\theta}_{MM} = t(X)$ and the statistic $\widehat{\boldsymbol{\theta}}_{MM} = t\left(\tilde{X}\right)$ is the Method of Moments estimator for θ.

- More generally, we start with the first moment and keep equating sample moments with the corresponding population moments in ascending order until we have as many equations as parameters and then we solve for the parameter vector.
- Alternatively, we can use central moments and solve for the parameter vector θ in the system of equations

$$m_1'(X) = \mu_1'(\theta) \quad \text{and} \quad m_k(X) = \mu_k(\theta), \text{ for } k = 2, 3, ..., K,$$

where

$$m_k(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{n} \quad \text{and} \quad \mu_k(\theta) = \mathrm{E}\left[\left(\tilde{x}^k - \mu\right)^k\right],$$

and $\bar{x}$ is the value of the sample mean, $\bar{x} = m_1'(X)$, and μ is the population mean, $\mu = \mu_1'(\theta)$.

- *Intuition behind the MM estimator:* Find the parameters of the distribution that match the empirical moments to the population moments of the distribution.

- We could generate sample moments of order larger than K,

$$m'_k(X) = \mu'_k(\theta), \text{ for } k = 1, 2, ..., J, \quad \text{where } J \geq K.$$

However, the previous system of J equations and K unknowns does not have solution generically when $J > K$.

- In this case, the Generalized Method of Moments (GMM) estimate is given by

$$\widehat{\theta}_{GMM} = \arg\min_{\theta \in \Theta} \left\{ \left[m'(X) - \mu'(\theta) \right]^{\mathsf{T}} W \left[m'(X) - \mu'(\theta) \right] \right\},$$

where $\left[m'(X) - \mu'(\theta) \right]^{\mathsf{T}} = \left(..., m'_j(X) - \mu'_j(\theta), ... \right)_{1 \times J}$ and $W_{J \times J}$ is a weighting matrix.

- Note that if $J = K$ and $W = \mathrm{I}$, then the GMM estimator coincides with the MM estimator.

8.9. Maximum likelihood estimation

- Let $X = (x_1, x_2, ..., x_n)$ be the value of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ having the density (probability function) $h(X; \theta)$ where θ is a K-dimensional vector of parameters in some parameter space Θ. The likelihood function $L : \Theta \longrightarrow \mathbb{R}$ is given by

$$L(\theta; X) \equiv h(X; \theta).$$

- Note that $h(X; \theta)$ is the value of the joint density (probability function) of the random vector $\tilde{X}$ at the observed sample value.
- The maximum likelihood (ML) estimate is given by

$$\widehat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta; X).$$

- The solution to the previous maximization problem is the estimate $\widehat{\theta}_{ML} = t(X)$ and the statistic $\widehat{\boldsymbol{\theta}}_{ML} = t(\tilde{X})$ is the Maximum Likelihood estimator for θ.

- *Intuition behind the ML estimator:* Find the parameters of the distribution for which the observed data is most probable.
- If the parameter space Θ is compact and $L(\theta; X)$ is continuous on Θ, then $\widehat{\theta}_{ML}$ exists.
- Sometimes, by modifying the density function on a set with zero Lebesgue measure, we can make

$$\max_{\theta \in \Theta} L(\theta; X) = \sup_{\theta \in \Theta} L(\theta; X),$$

  where $\sup_{\theta \in \Theta} L(\theta; X)$ exists if the likelihood function $L(\cdot; X)$ is bounded and Θ is compact.
- **Invariance property.** If $\widehat{\boldsymbol{\theta}}_{ML}$ is a ML estimator for θ and $g : \Theta \longrightarrow \Theta' \subset \mathbb{R}^K$ is a one-to-one correspondence (or bijection), then $g\left(\widehat{\boldsymbol{\theta}}_{ML}\right)$ is a ML estimator for $g(\theta) \in \Theta'$.

- **Proof.** Let $L(\theta; X)$ be the likelihood function on the parameter space Θ, $\widehat{L}(\theta'; X)$ be the likelihood function on the parameter space Θ', $h(X; \theta)$ be the density (probability function) of the random vector $\tilde{X}$, and $\widehat{h}(X; \theta')$ be the density function (probability function) of $\tilde{X}$ under the reparameterization dictated by the function g, according to which $\theta' = g(\theta)$ and $\theta = g^{-1}(\theta')$ with $\theta \in \Theta$ and $\theta' \in \Theta'$. Then, since $h(X; \theta) = h(X; g^{-1}(\theta')) = \widehat{h}(X; \theta')$ and $L(\theta; X) = h(X; \theta)$, we have that

$$\widehat{L}(\theta'; X) = \widehat{h}(X; \theta') = h(X; g^{-1}(\theta')) = L(\underbrace{g^{-1}(\theta')}_{=\theta}; X).$$

Therefore,

$$\max_{\theta' \in \Theta'} \widehat{L}(\theta'; X) = \max_{g^{-1}(\theta') \in \Theta} L(g^{-1}(\theta'); X) = \max_{\theta \in \Theta} L(\theta; X).$$

The previous equality tells us that, if $\widehat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta; X)$ and $\widehat{\theta}'_{ML} = \arg\max_{\theta' \in \Theta'} \widehat{L}(\theta'; X)$, then $\widehat{\theta}_{ML} = g^{-1}\left(\widehat{\theta}'_{ML}\right)$ or $\widehat{\theta}'_{ML} = g\left(\widehat{\theta}_{ML}\right)$.

Q.E.D.

- **Example:** Let $\widehat{\theta}_{ML}$ be the Maximum Likelihood estimator for the parameter θ of the exponential distribution. The variance of the exponential distribution is $\sigma^2 = g(\theta) = \theta^2$ and g is a one-to-one correspondence since $\theta > 0$, i.e., $\Theta = \mathbb{R}_{++}$. Therefore, $\left(\widehat{\theta}_{ML}\right)^2$ is the Maximum Likelihood estimator for the population variance σ^2.

- If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a random sample from a population with density (probability function) $f(x;\theta)$, then

$$L(\theta; X) = h(X;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

- Moreover, in this case,

$$\widehat{\theta}_{ML} = \arg\max_{\theta\in\Theta} L(\theta; X) = \arg\max_{\theta\in\Theta} \ln L(\theta; X) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \ln f(x_i;\theta).$$

8.10. Bayesian estimation

- $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a random vector (sample) and $\tilde{w} = g(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a statistic, which may be a random vector.
- The random vector $\tilde{w} = g(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ has the density (probability function) $f_{\tilde{w}|\boldsymbol{\theta}}(w|\theta)$, where the parameter θ is the realization of the random vector $\boldsymbol{\theta}: (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$ with $\boldsymbol{\theta}(\Omega) = \Theta \subset \mathbb{R}^K$.
- $f_{\tilde{w}|\boldsymbol{\theta}}(w|\theta)$ is also called the likelihood of $\tilde{w}$ for $\boldsymbol{\theta} = \theta$.
- The density (probability function) of $\boldsymbol{\theta}$ is $f_{\boldsymbol{\theta}}$ (this is called the prior density or prior probability function).

- The posterior density (probability function) of $\boldsymbol{\theta}$ given $\tilde{w} = w$ is

$$f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right) = \frac{f_{(\boldsymbol{\theta},\tilde{w})}\left(\theta, w\right)}{f_{\tilde{w}}\left(w\right)} = \frac{f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)}{f_{\tilde{w}}\left(w\right)}, \text{ for } f_{\tilde{w}}\left(w\right) \neq 0,$$

where $f_{\boldsymbol{\theta}}\left(\theta\right)$ is the prior density (probability function) of $\boldsymbol{\theta}$, $f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)$ is the likelihood of $\tilde{w}$ for $\boldsymbol{\theta} = \theta$, and

$$f_{\tilde{w}}\left(w\right) = \int_{\Theta} f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right) d\theta \tag{1}$$

or

$$f_{\tilde{w}}\left(w\right) = \sum_{\theta \in \Theta} f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right). \tag{2}$$

- Note that the formula for the posterior density/probability function,

$$f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right) = \frac{f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)}{f_{\tilde{w}}\left(w\right)}, \text{ for } f_{\tilde{w}}\left(w\right) \neq 0,$$

also holds if $f_{\boldsymbol{\theta}}\left(\theta\right)$ is a density/probability function and the likelihood $f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)$ is a conditional probability function/density (see Exercise 13 of List 2). Then, $f_{\tilde{w}}\left(w\right)$ would be a probability function/density obtained through the formula (1)/(2).

- We define the loss function $L\left(\theta, \widehat{\theta}\right)$. For $K = 1$, we usually use

$$L\left(\theta, \widehat{\theta}\right) = \left(\theta - \widehat{\theta}\right)^2 \qquad \text{(quadratic loss function)}$$

or, for $K > 1$,

$$L\left(\theta, \widehat{\theta}\right) = \left(\theta - \widehat{\theta}\right)^{\mathsf{T}} W \left(\theta - \widehat{\theta}\right),$$

where W is a $K \times K$ weighting matrix.

- Then, the Bayesian estimate is given by

$$\widehat{\theta}_B = \underset{\widehat{\theta} \in \Theta}{\arg\min}\, \mathrm{E}\left[L\left(\boldsymbol{\theta}, \widehat{\theta}\right) \middle| \tilde{w} = w\right] = \underset{\widehat{\theta} \in \Theta}{\arg\min} \int_{\Theta} L\left(\theta, \widehat{\theta}\right) f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta \,|w\right) d\theta$$

or

$$\widehat{\theta}_B = \underset{\widehat{\theta} \in \Theta}{\arg\min}\, \mathrm{E}\left[L\left(\boldsymbol{\theta}, \widehat{\theta}\right) \middle| \tilde{w} = w\right] = \underset{\widehat{\theta} \in \Theta}{\arg\min} \sum_{\theta \in \Theta} L\left(\theta, \widehat{\theta}\right) f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta \,|w\right).$$

- In general,

$$\widehat{\theta}_B = \arg\min_{\widehat{\theta}\in\Theta} \mathrm{E}\left[L\left(\boldsymbol{\theta},\widehat{\theta}\right)\middle|\,\tilde{w}=w\right] = \arg\min_{\widehat{\theta}\in\Theta} \int_{\Theta} L\left(\theta,\widehat{\theta}\right) dP_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right).$$

- The solution is the estimate $\widehat{\theta}_B = h(w) = h\left(g(X)\right) = t(X)$ and the statistic $\widehat{\boldsymbol{\theta}}_B = t\left(\tilde{X}\right)$ is the Bayesian estimator for the random parameter $\boldsymbol{\theta}$ (or for the parameter value θ).

- *Intuition behind the Bayesian estimator:* A Bayesian estimator minimizes the posterior expected value of the loss function.

- **Proposition.** Let $K = 1$. If $L\left(\theta,\widehat{\theta}\right) = \left(\theta-\widehat{\theta}\right)^2$, then $\widehat{\theta}_B = \mathrm{E}\left[\boldsymbol{\theta}\,|\tilde{w}=w\right]$.

- **Proof.** To solve

$$\underset{\widehat{\theta}\in\Theta\subset\mathbb{R}}{\arg\min}\ \mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)^2\middle|\,\tilde{w}=w\right], \tag{1}$$

we compute the first derivative of $\mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)^2\middle|\,\tilde{w}=w\right]$ w.r.t. $\widehat{\theta}$ and equate it to zero,

$$-2\,\mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)\middle|\,\tilde{w}=w\right]=0.$$

Solving for $\widehat{\theta}$ in the previous expression, we get

$$\widehat{\theta}_B=\mathrm{E}\left[\boldsymbol{\theta}\,|\tilde{w}=w\right]. \tag{2}$$

Note that $\mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)^2\middle|\,\tilde{w}=w\right]$ is a strictly convex function of $\widehat{\theta}$.

To see this, we compute its second derivative w.r.t. $\widehat{\theta}$,

$$-2\,\mathrm{E}\left[-1|\,\tilde{w}=w\right]=(-2)(-1)=2>0.$$

Therefore, the value $\widehat{\theta}_B$ obtained in (2) is a solution for (1) and, hence, it is a Bayesian estimate for $\boldsymbol{\theta}$. *Q.E.D.*

- **Conjugate families of distributions.**

- There are cases where the prior distribution of the random vector $\boldsymbol{\theta}$ and the posterior distribution of $\boldsymbol{\theta}$ given any value of the statistic (or any value of the random sample) both belong to the same family of distributions. The prior and posterior are then called conjugate distributions for the corresponding likelihood.

- **Examples:**

- **1)** Prior: λ has the gamma distribution with the parameters α and β. Likelihood: $\tilde{x}$ has the Poisson distribution with the random parameter λ. Posterior of λ given the value of a random sample of size n from the population $\tilde{x}$, $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$: gamma distribution with the parameters $\alpha + \sum_{i=1}^{n} x_i = \alpha + n\bar{x}$, where $\bar{x}$ is the value of the sample mean, and $\frac{\beta}{1+n\beta}$. Thus, the family of gamma distributions is conjugate for the Poisson likelihood.

- **2)** Prior: $\boldsymbol{\mu}$ has the normal distribution with the mean μ_0 and the variance σ_0 (or precision $\tau_0 = 1/\sigma_0^2$). Likelihood: $\tilde{x}$ has the normal distribution with the random mean $\boldsymbol{\mu}$ and the known variance σ^2 (or precision $\tau = 1/\sigma^2$), which means that the mean $\bar{\mathbf{x}}$ of a random sample of size n from the population $\tilde{x}$ is normal with the random mean $\boldsymbol{\mu}$ and the known variance σ^2/n. Posterior of $\boldsymbol{\mu}$ given $\bar{\mathbf{x}} = \bar{x}$ (or given the sample value $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$): normal distribution with the mean μ_1 and the variance σ_1^2 (or precision $\tau_1 = 1/\sigma_1^2$), where

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \left(\frac{\tau_0}{\tau_0 + n\tau}\right)\mu_0 + \underbrace{\left(\frac{n\tau}{\tau_0 + n\tau}\right)}_{=\left(\frac{\tau}{\tau_0+n\tau}\right)\sum_{i=1}^{n} x_i}\bar{x}$$

and

$$\tau_1 = \frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} = \tau_0 + n\tau.$$

Thus, the family of normal distributions is conjugate for the normal likelihood.

- **3)** Prior: $\boldsymbol{\theta}$ has the beta distribution with the parameters α and β. Likelihood: $\tilde{x}$ has the Bernoulli distribution with the random parameter $\boldsymbol{\theta}$. Posterior of $\boldsymbol{\theta}$ given the value of a random sample of size n from the population $\tilde{x}$, $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$: beta distribution with the parameters $\sum_{i=1}^{n} x_i + \alpha = n\bar{x} + \alpha$ and $n - \sum_{i=1}^{n} x_i + \beta = n(1 - \bar{x}) + \beta$, where $\bar{x}$ is the value of the sample mean (i.e., the percentage of successes). Thus, the family of beta distributions is conjugate for the Bernoulli likelihood.

  OR, equivalently,

- **3')** Prior: $\boldsymbol{\theta}$ has the beta distribution with the parameters α and β. Likelihood: $\tilde{y}$ has the binomial distribution with the parameters $\boldsymbol{\theta}$ and n, where n is known. Posterior of $\boldsymbol{\theta}$ given $\tilde{y} = y$: beta distribution with the parameters $y + \alpha$ and $n - y + \beta$. In fact, the family of beta distributions is conjugate for the binomial likelihood.

- **Sufficiency in Bayesian estimation.**
- **Definition.** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution $P_{\tilde{X}|\boldsymbol{\theta}}(\cdot|\theta)$ depends on the random parameter $\boldsymbol{\theta}: (\Omega, \mathcal{F}) \longrightarrow \left(\mathbb{R}^K, \mathcal{B}\left(\mathbb{R}^K\right)\right)$ with $\boldsymbol{\theta}(\Omega) = \Theta \subset \mathbb{R}^K$. The statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for the random parameter $\boldsymbol{\theta}$ if, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}}$ and all values θ taken by $\boldsymbol{\theta}$, the conditional distribution of the sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$ and $\boldsymbol{\theta} = \theta$, $P_{\tilde{X}|\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}}\left(\cdot \left| \widehat{\theta}, \theta \right.\right) : \mathcal{B}(\mathbb{R}^n) \longrightarrow [0,1]$, is independent of θ.
- If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ has a discrete (absolutely continuous) distribution, then sufficiency means that the conditional probability function (density) of $\tilde{X}$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$ and $\boldsymbol{\theta} = \theta$,

$$f_{\tilde{X}|\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}, \theta \right.\right) = \frac{f_{(\tilde{X},\widehat{\boldsymbol{\theta}})}\left(x_1, x_2, ..., x_n, \widehat{\theta}\,|\theta\right)}{f_{\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}}\left(\widehat{\theta}\,|\theta\right)},$$

for all $\widehat{\theta}$ and θ with $f_{\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}}\left(\widehat{\theta}\,|\theta\right) \neq 0$, does not depend on θ.

- **Factorization theorem.** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution is discrete (absolutely continuous) and depends on the random parameter $\boldsymbol{\theta} : (\Omega, \mathcal{F}) \longrightarrow \left(\mathbb{R}^K, \mathcal{B}\left(\mathbb{R}^K\right)\right)$ with $\boldsymbol{\theta}(\Omega) = \Theta \subset \mathbb{R}^K$. Then, the statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for the random parameter $\boldsymbol{\theta}$ if and only if, for all values θ taken by $\boldsymbol{\theta}$, the probability function (density) of $\tilde{X}$ can be factorized as follows:

$$f_{\tilde{X}}(x_1, x_2, ..., x_n | \theta) = h(x_1, x_2, ..., x_n) \cdot g\left(\widehat{\theta}, \theta\right),$$

where $\widehat{\theta} = t(x_1, x_2, ..., x_n)$.

8.11. Interval estimation: the pivotal method

- Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution depends on the parameter $\theta \in \Theta \subset \mathbb{R}$.
- Interval estimation is the use of a sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ to calculate an interval $\left(a\left(\tilde{X}\right), b\left(\tilde{X}\right)\right)$ of probable values of the unknown parameter θ.
- **Pivotal method:** Find a real-valued random variable $\tilde{w} = g\left(\tilde{X}, \theta\right)$, called the pivotal variable (or pivot), such that its distribution is independent of the parameter θ we want to estimate.
- A $(1-\alpha)100\%$ confidence (Borel) set $\widehat{C}_{1-\alpha}$ for $g\left(\tilde{X}, \theta\right)$ is given by

$$P_{g\left(\tilde{X},\theta\right)}\left(\widehat{C}_{1-\alpha}\right) = P\left\{g\left(\tilde{X}, \theta\right) \in \widehat{C}_{1-\alpha}\right\} = 1-\alpha.$$

- If $\widehat{C}_{1-\alpha}$ is an interval (c, d), then a $(1-\alpha)100\%$ confidence interval for $g\left(\tilde{X}, \theta\right)$ is given by

$$P\left\{c < g\left(\tilde{X}, \theta\right) < d\right\} = 1 - \alpha. \qquad (*)$$

- Since there are many confidence intervals, we usually make

$$P\left\{g\left(\tilde{X}, \theta\right) \geq d\right\} = P\left\{g\left(\tilde{X}, \theta\right) \leq c\right\} = \frac{\alpha}{2}.$$

- Then, we solve for θ in $(*)$ to get a random $(1-\alpha)100\%$ confidence set $H\left(\tilde{X}, c, d\right)$ for θ,

$$P\left\{\theta \in H(\tilde{X}, c, d)\right\} = 1 - \alpha.$$

- If the set $H(\tilde{X}, c, d)$ is a random interval $\left(a\left(\tilde{X}\right), b\left(\tilde{X}\right)\right)$, then

$$P\left\{a\left(\tilde{X}\right) < \theta < b\left(\tilde{X}\right)\right\} = 1 - \alpha.$$

- This random interval is called the $(1-\alpha)100\%$ confidence interval for the parameter θ.

- Usually, the $(1-\alpha)100\%$ confidence interval for θ is presented like this:

$$a(X) < \theta < b(X),$$

where X is a particular value of the sample. Therefore, we are just providing a "realization" of the random $(1-\alpha)100\%$ confidence interval.

- **Examples:**
- Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample from a normal population with mean μ and variance σ^2
- **(1)** If σ^2 is known, the random variable $\dfrac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0, 1)$. Therefore,

$$P\left\{-z_{\alpha/2} < \frac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}\right\} = 1 - \alpha,$$

where $z_{\alpha/2} = 2.576$ if $1 - \alpha = 0.99$; $z_{\alpha/2} = 1.96$ if $1 - \alpha = 0.95$; and $z_{\alpha/2} = 1.645$ if $1 - \alpha = 0.9$. Then,

$$P\left\{\bar{\mathbf{x}} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{\mathbf{x}} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha.$$

Thus, the $(1 - \alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

where $\overline{x}$ is the value of the sample mean $\bar{\mathbf{x}}$.

- **(2)** If σ^2 is unknown, the random variable $\dfrac{\bar{\mathbf{x}} - \mu}{\mathbf{s}/\sqrt{n}} \sim \mathrm{t}_{n-1}$. Therefore,

$$P\left\{-t_{\alpha/2,n-1} < \frac{\bar{\mathbf{x}} - \mu}{\mathbf{s}/\sqrt{n}} < t_{\alpha/2,n-1}\right\} = 1 - \alpha.$$

Then,

$$P\left\{\bar{\mathbf{x}} - t_{\alpha/2,n-1} \cdot \frac{\mathbf{s}}{\sqrt{n}} < \mu < \bar{\mathbf{x}} + t_{\alpha/2,n-1} \cdot \frac{\mathbf{s}}{\sqrt{n}}\right\} = 1 - \alpha.$$

Thus, the $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},$$

where $\overline{x}$ is the value of the sample mean $\bar{\mathbf{x}}$ and s is the value of the sample standard deviation $\left(\mathbf{s} = \left(\mathbf{s^2}\right)^{1/2}\right)$.

- **(3)** The random variable $\dfrac{(n-1)\mathbf{s^2}}{\sigma^2} \sim \chi^2_{n-1}$. Therefore,

$$P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)\mathbf{s^2}}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha.$$

Then,

$$P\left\{\frac{(n-1)\mathbf{s^2}}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\mathbf{s^2}}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right\} = 1-\alpha.$$

Thus, the $(1-\alpha)100\%$ confidence interval for the population variance σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}},$$

where s^2 is the value of the sample variance $\mathbf{s^2}$.

- Let $\mathbf{s_1^2}$ and $\mathbf{s_2^2}$ be the variances of independent random samples of size n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2, respectively.
- **(4)** The random variable $\dfrac{\mathbf{s_1^2}/\sigma_1^2}{\mathbf{s_2^2}/\sigma_2^2} \sim \mathrm{F}_{n_1-1,n_2-1}$. Therefore,

$$P\left\{F_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{\mathbf{s_1^2}/\sigma_1^2}{\mathbf{s_2^2}/\sigma_2^2} < F_{\alpha/2,n_1-1,n_2-1}\right\} = 1-\alpha.$$

Then,

$$P\left\{\frac{\mathbf{s_1^2}}{\mathbf{s_2^2}} \cdot \frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\mathbf{s_1^2}}{\mathbf{s_2^2}} \cdot \frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}\right\} = 1-\alpha.$$

Thus, the $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2 is given by

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}$$

or, since

$$F_{\alpha/2,n_2-1,n_1-1}=\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}},$$ (see the handout)

the confidence interval becomes

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{s_1^2}{s_2^2}\cdot F_{\alpha/2,n_2-1,n_1-1},$$

where s_1 and s_2 are the values of the standard deviations of the two random samples.

8.11. Confidence intervals for means, for differences between means, for proportions, for differences between proportions, for variances, and for ratios of two variances

- See the handout.

Cramér-Rao lower bound

Theorem (Cramér-Rao). Let $\tilde{X}$ be a n-dimensional vector of random variables (not necessarily independent), the joint density of which is given by $h(X;\theta)$, where θ is a K-dimensional vector of parameters in some parameter space Θ. Let $\widehat{\boldsymbol{\theta}} = g\left(\tilde{X}\right)$ be an unbiased estimator for θ with a finite covariance matrix $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$. Furthermore, assume that $h(X;\theta)$ and $g(X)$ satisfy

(a)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} h dX = \int_{\mathbb{R}^n} \frac{\partial h}{\partial\theta} dX \;\leftarrow \text{This is a } K\text{-dimensional column vector}$$

(b)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} g h dX = \int_{\mathbb{R}^n} g\frac{\partial h}{\partial\theta} dX \;\leftarrow \text{This is a } K\text{-dimensional column vector}$$

(c)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} \frac{\partial h}{\partial\theta^T} dX = \int_{\mathbb{R}^n} \frac{\partial^2 h}{\partial\theta\partial\theta^T} dX \;\leftarrow \text{This is a } K\times K\text{-dimensional matrix}$$

(d)

$$\mathrm{Var}\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right) \text{ is a positive definite } K\times K \text{ matrix.}$$

Then, for all $\theta\in\Theta$,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta\partial\theta^T}\right]\right)^{-1} \tag{CR-1}$$

or, equivalently,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(\mathrm{E}\left[\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\cdot\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta^T}\right]\right)^{-1}. \tag{CR-2}$$

Proof for the case $K = 1$. We have $1 = \int_{\mathbb{R}^n} h(X;\theta)dX$ for all $\theta\in\Theta$, where $X = (x_1, x_2, \ldots, x_n)\in\mathbb{R}^n$. Therefore, differentiating both sides with respect to θ and using assumption (a), we get

$$0 = \int_{\mathbb{R}^n} \frac{\partial h(X;\theta)}{\partial\theta} dX = \int_{\mathbb{R}^n} \frac{\partial \ln h(X;\theta)}{\partial\theta} h(X;\theta) dX \tag{1}$$

Define $\tilde{z} = \dfrac{\partial \ln h(\tilde{X};\theta)}{\partial \theta}$, then (1) means that $\mathrm{E}(\tilde{z}) = 0$. Furthermore, $\mathrm{Var}(\tilde{z}) = \mathrm{E}(\tilde{z}^2) = \mathrm{E}\left[\left(\dfrac{\partial \ln h(\tilde{X};\theta)}{\partial \theta}\right)^2\right]$. Since $\widehat{\boldsymbol{\theta}} = g\left(\tilde{X}\right)$ is an unbiased estimator for θ, we have

$$\theta = \mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right) \iff \theta = \int_{\mathbb{R}^n} \underbrace{g(X)}_{=\widehat{\theta}} h(X;\theta)dX.$$

Therefore, differentiating both sides with respect to θ and using assumption (b), we get

$$1 = \int_{\mathbb{R}^n} g(X)\frac{\partial h(X;\theta)}{\partial \theta}dX = \int_{\mathbb{R}^n} g(X)\frac{\partial \ln h(X;\theta)}{\partial \theta}h(X;\theta)dX. \tag{2}$$

Note that (2) means that $\mathrm{E}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right) = 1$. Thus, we have

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right) = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right) - \underbrace{\mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right)}_{=\theta}\underbrace{\mathrm{E}(\tilde{z})}_{=0} = 1.$$

The Cauchy-Schwarz inequality implies that the absolute value of the correlation coefficient is smaller than 1 so that

$$\underbrace{\left[\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right)\right]^2}_{=1} \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \cdot \mathrm{Var}\left(\tilde{z}\right).$$

Note that $[\mathrm{Var}\left(\tilde{z}\right)]^{-1}$ is well defined if assumption (d) holds. Therefore, we obtain the inequality (CR-2) since

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq [\mathrm{Var}\left(\tilde{z}\right)]^{-1} = \left(\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial \theta}\right)^2\right]\right)^{-1}. \tag{3}$$

To get the inequality (CR-1) observe that

$$\frac{\partial \ln h(X;\theta)}{\partial \theta} = \frac{1}{h(X;\theta)}\frac{\partial h(X;\theta)}{\partial \theta} \tag{4}$$

so that

$$\left(\frac{\partial \ln h(X;\theta)}{\partial \theta}\right)^2 = \frac{1}{[h(X;\theta)]^2}\left(\frac{\partial h(X;\theta)}{\partial \theta}\right)^2$$

and

$$\mathrm{E}\left[\left(\frac{\partial \ln h(X;\theta)}{\partial \theta}\right)^2\right] = \mathrm{E}\left[\frac{1}{[h(X;\theta)]^2}\left(\frac{\partial h(X;\theta)}{\partial \theta}\right)^2\right]. \tag{5}$$

Differentiating (4) with respect to θ, we get

$$\frac{\partial^2 \ln h(X;\theta)}{\partial \theta^2} = \frac{-1}{[h(X;\theta)]^2}\left(\frac{\partial h(X;\theta)}{\partial \theta}\right)^2 + \frac{1}{h(X;\theta)}\frac{\partial^2 h(X;\theta)}{\partial \theta^2}.$$

Therefore, taking the expectation of the previous equation and using (5), we get

$$\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] + \mathrm{E}\left[\frac{1}{h(\tilde{X};\theta)}\frac{\partial^2 h(\tilde{X};\theta)}{\partial\theta^2}\right]$$

$$= -\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] + \int_{\mathbb{R}^n} \frac{1}{h(X;\theta)}\frac{\partial^2 h(X;\theta)}{\partial\theta^2} h(X;\theta)dX$$

$$= -\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] + \int_{\mathbb{R}^n} \frac{\partial^2 h(X;\theta)}{\partial\theta^2} dX, \tag{6}$$

and using assumption (c), we get

$$\int_{\mathbb{R}^n} \frac{\partial^2 h(X;\theta)}{\partial\theta^2} dX = \frac{\partial}{\partial\theta} \underbrace{\int_{\mathbb{R}^n} \frac{\partial h(X;\theta)}{\partial\theta} dX}_{=0} = 0, \tag{7}$$

where the last equality follows from differentiating both sides of the first equality in (1) with respect to θ. Therefore, combining (6) and (7) we obtain

$$-\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right] = \mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right]. \tag{8}$$

Plugging (8) into (3) we obtain the inequality (CR-1). *Q.E.D.*

Corollary. If in the previous theorem the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is a random sample from a population $\tilde{x}$ having the density (or probability function) $f(x;\theta)$, then

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-n\mathrm{E}\left[\frac{\partial^2 \ln f(\tilde{x};\theta)}{\partial\theta\partial\theta^T}\right]\right)^{-1} \tag{i}$$

or, equivalently,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(n\mathrm{E}\left[\frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta} \cdot \frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta^T}\right]\right)^{-1}. \tag{ii}$$

Proof for the case $K = 1$. Since $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ are independent, we have

$$h(X;\theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta),$$

so that

$$\ln h(X;\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta),$$

which implies that

$$\frac{\partial^2 \ln h(X;\theta)}{\partial\theta^2} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i;\theta)}{\partial\theta^2},$$

and, hence,

$$\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right] = n\mathrm{E}\left[\frac{\partial^2 \ln f(\widetilde{x};\theta)}{\partial\theta^2}\right],$$

which combined with (CR-1) proves inequality (i) in the statement of the Corollary.

Let us prove inequality (ii). We have that

$$\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] = \mathrm{E}\left[\left(\frac{\partial\left[\sum_{i=1}^n \ln f(\widetilde{x}_i;\theta)\right]}{\partial\theta}\right)^2\right] = \mathrm{E}\left[\left(\sum_{i=1}^n \frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right)^2\right]$$

$$= \mathrm{E}\left[\sum_{i=1}^n \left(\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right)^2 + \sum_{j\neq i}\sum_{i=1}^n \left(\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(\widetilde{x}_j;\theta)}{\partial\theta}\right)\right]$$

$$= \sum_{i=1}^n \left[\mathrm{E}\left(\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right)^2\right] + \sum_{j\neq i}\sum_{i=1}^n \mathrm{E}\left[\underbrace{\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(\widetilde{x}_j;\theta)}{\partial\theta}}_{\text{independent random variables}}\right]$$

$$= n\mathrm{E}\left[\left(\frac{\partial \ln f(\widetilde{x};\theta)}{\partial\theta}\right)^2\right] + \sum_{j\neq i}\sum_{i=1}^n \underbrace{\mathrm{E}\left[\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right]}_{=0}\cdot\underbrace{\mathrm{E}\left[\frac{\partial \ln f(\widetilde{x}_j;\theta)}{\partial\theta}\right]}_{=0}$$

since we know from (1) that

$$0 = \mathrm{E}\left[\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right] = \mathrm{E}\left[\sum_{i=1}^n \frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right] = n\mathrm{E}\left[\frac{\partial \ln f(\widetilde{x};\theta)}{\partial\theta}\right].$$

Therefore, we have proved that

$$\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] = n\mathrm{E}\left[\left(\frac{\partial \ln f(\widetilde{x};\theta)}{\partial\theta}\right)^2\right], \tag{9}$$

which combined with (CR-2) proves inequality (ii) in the statement of the Corollary. *Q.E.D.*

Single parameter case: $K = 1$ ($\theta \in \mathbb{R}$)**.**

(a) For $\tilde{X} = (\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n)$, not necessarily a random sample, we have

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right]\right)^{-1} \qquad \text{(Theorem CR-1 for } K = 1)$$

or, equivalently,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right]\right)^{-1}. \qquad \text{(Theorem CR-2 for } K = 1)$$

(b) For the random sample $\tilde{X} = (\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n)$, we have

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-n\mathrm{E}\left[\frac{\partial^2 \ln f(\widetilde{x};\theta)}{\partial\theta^2}\right]\right)^{-1} \qquad \text{(Corollary (i) for } K = 1)$$

or, equivalently,

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(n\mathrm{E}\left[\left(\frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta}\right)^2\right]\right)^{-1}. \qquad \text{(Corollary (ii) for } K = 1)$$

Interval Estimation

Confidence intervals for means

$$\frac{\overline{\boldsymbol{x}}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1) \Rightarrow P\left\{-z_{\alpha/2} < \frac{\overline{\boldsymbol{x}}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\} = 1-\alpha,$$

$$\text{where } \int_{z_{\alpha/2}}^{\infty} n(x;0,1)dx = \frac{\alpha}{2}.$$

Then,

$$P\left\{\overline{\boldsymbol{x}} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \overline{\boldsymbol{x}} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha.$$

Proposition 1. If $\overline{x}$ is the value of the mean of a random sample of size n from a normal population with the known variance σ^2, a $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}.$$

Remark 1. By virtue of the central limit theorem, this result can be also used for random samples from non-normal populations with the known variance σ^2, provided that n is sufficiently large ($n \geq 30$). If σ is unknown but $n \geq 30$, we can replace σ by the value s of the sample standard deviation $\boldsymbol{s} = (\boldsymbol{s^2})^{1/2}$.

– – – – – – – – – – – – – – –

$$\frac{\overline{\boldsymbol{x}}-\mu}{\boldsymbol{s}/\sqrt{n}} \sim \mathrm{t}_{n-1} \Rightarrow P\left\{-t_{\alpha/2,n-1} < \frac{\overline{\boldsymbol{x}}-\mu}{\boldsymbol{s}/\sqrt{n}} < t_{\alpha/2,n-1}\right\} = 1-\alpha.$$

Then,

$$P\left\{\overline{\boldsymbol{x}} - t_{\alpha/2,n-1}\cdot\frac{\boldsymbol{s}}{\sqrt{n}} < \mu < \overline{\boldsymbol{x}} + t_{\alpha/2,n-1}\cdot\frac{\boldsymbol{s}}{\sqrt{n}}\right\} = 1-\alpha.$$

Proposition 2. If $\overline{x}$ and s are the values of the mean and the standard deviation of a random sample of size n from a normal population with the unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}}.$$

Remark 2. Note that Proposition 2 applies to <u>small</u> samples ($n < 30$) from a normal population. For $n \geq 30$, the confidence interval of Proposition 2 and the one of Proposition 1 with σ replaced by s will generally yield nearly the same result.

Confidence intervals for differences between means

Let $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ be the means of independent random samples of size n_1 and n_2 from normal populations having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. Then, $\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$. It follows that

$$\frac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1).$$

From Proposition 1 we get,

Proposition 3. If $\overline{x}_1$ and $\overline{x}_2$ are the values of the means of independent random samples of size n_1 and n_2 from normal populations with the known variances σ_1^2 and σ_2^2, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Remark 1 also applies to Proposition 3 for large samples ($n_1 \geq 30$ and $n_2 \geq 30$).

If we had two small independent random samples from normal populations having the same variance σ^2, which is unknown, then we must construct the "pooled" estimator

$$\boldsymbol{s}_p^{\boldsymbol{2}} = \frac{(n_1 - 1)\boldsymbol{s}_1^{\boldsymbol{2}} + (n_2 - 1)\boldsymbol{s}_2^{\boldsymbol{2}}}{n_1 + n_2 - 2}$$

which is, indeed, an unbiased estimator for σ^2. The independent random variables $\dfrac{(n_1 - 1)\boldsymbol{s}_1^{\boldsymbol{2}}}{\sigma^2}$ and $\dfrac{(n_2 - 1)\boldsymbol{s}_2^{\boldsymbol{2}}}{\sigma^2}$ have chi-square distributions with $n_1 - 1$ and $n_2 - 1$ degrees of freedom and their sum

$$\frac{(n_1 - 1)\boldsymbol{s}_1^{\boldsymbol{2}}}{\sigma^2} + \frac{(n_2 - 1)\boldsymbol{s}_2^{\boldsymbol{2}}}{\sigma^2} = \frac{(n_1 + n_2 - 2)\boldsymbol{s}_p^{\boldsymbol{2}}}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Since

$$\frac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad \text{and} \quad \frac{(n_1 + n_2 - 2)\boldsymbol{s}_p^{\boldsymbol{2}}}{\sigma^2}$$

are independent, it follows that

$$\frac{\dfrac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\sqrt{\boldsymbol{s}_p^{\boldsymbol{2}}/\sigma^2}} = \frac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\boldsymbol{s}_p \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim \mathrm{t}_{n_1+n_2-2}.$$

From Proposition 2 we get,

Proposition 4. If $\overline{x}_1$ and $\overline{x}_2$ are the values of the means of two independent random samples of size n_1 and n_2 from normal populations with unknown but equal variances, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned}(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2,n_1+n_2-2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &< \mu_1 - \mu_2 \\ &< (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2,n_1+n_2-2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.\end{aligned}$$

Confidence intervals for proportions

Let $\tilde{x} \sim \mathrm{B}(n,\theta)$. Then $\dfrac{\tilde{x} - n\theta}{\sqrt{n\theta(1-\theta)}} \to \mathrm{N}(0,1)$ as $n \to \infty$.

Let $\hat{\boldsymbol{\theta}} = \dfrac{\tilde{x}}{n}$ be the sample proportion. Then

$$\frac{\hat{\boldsymbol{\theta}} - \theta}{\sqrt{\dfrac{\theta(1-\theta)}{n}}} \to \mathrm{N}(0,1) \quad \text{as } n \to \infty.$$

Using Proposition 1 and Remark 1, we get,

Proposition 5. An approximate $(1-\alpha)100\%$ confidence interval (for large n) for the binomial parameter θ is given by

$$\hat{\theta} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} < \theta < \hat{\theta} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},$$

where $\hat{\theta} = \dfrac{x}{n}$.

Confidence intervals for differences between proportions

Using Proposition 3 we get,

Proposition 6. An approximate $(1-\alpha)100\%$ confidence interval (for large n) for the difference between two binomial parameters, $\theta_1 - \theta_2$, is given by

$$\begin{aligned}(\hat{\theta}_1 - \hat{\theta}_2) - z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} &< \theta_1 - \theta_2 \\ &< (\hat{\theta}_1 - \hat{\theta}_2) + z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}.\end{aligned}$$

where $\hat{\theta}_1 = \dfrac{x_1}{n_1}$ and $\hat{\theta}_2 = \dfrac{x_2}{n_2}$.

Confidence intervals for variances

If $\dfrac{(n-1)\boldsymbol{s^2}}{\sigma^2} \sim \chi^2_{n-1}$, then

$$P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)\boldsymbol{s^2}}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha.$$

Proposition 7. If s^2 is the value of the variance of a random sample of size n from a normal population, a $(1-\alpha)100\%$ confidence interval for the population variance σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}.$$

Confidence intervals for ratios of two variances

If $\dfrac{\boldsymbol{s_1^2}/\sigma_1^2}{\boldsymbol{s_2^2}/\sigma_2^2} \sim \mathrm{F}_{n_1-1,n_2-1}$, then

$$P\left\{F_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{\boldsymbol{s_1^2}/\sigma_1^2}{\boldsymbol{s_2^2}/\sigma_2^2} < F_{\alpha/2,n_1-1,n_2-1}\right\} = 1-\alpha.$$

Lemma. $F_{1-\alpha,\nu_1,\nu_2} = \dfrac{1}{F_{\alpha,\nu_2,\nu_1}}$.

Proof. Exercise 8 of List 7 says that

$$\tilde{x} \sim \mathrm{F}_{\nu_1,\nu_2} \iff \tilde{y} \equiv \frac{1}{\tilde{x}} \sim \mathrm{F}_{\nu_2,\nu_1}.$$

Let $P\{\tilde{x} \geq F_{1-\alpha,\nu_1,\nu_2}\} = 1-\alpha$. Therefore,

$$P\left\{\frac{1}{\tilde{y}} \geq F_{1-\alpha,\nu_1,\nu_2}\right\} = 1-\alpha \iff P\left\{\tilde{y} \geq \frac{1}{F_{1-\alpha,\nu_1,\nu_2}}\right\} = \alpha.$$

Since $P\{\tilde{y} \geq F_{\alpha,\nu_2,\nu_1}\} = \alpha$, we arrive at the desired conclusion. *Q.E.D.*

Proposition 8. If s_1^2 and s_2^2 are the values of the variances of independent random samples of size n_1 and n_2 from two normal populations, a $(1-\alpha)100\%$ confidence interval for the ratio σ_1^2/σ_2^2 of the two population variances is given by

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2,n_2-1,n_1-1}. \qquad \text{(Check it!)}$$

Exercises. Probability and Statistics. IDEA.
8. Estimation

1. Show that if $\hat{\boldsymbol{\theta}}$ is an unbiased estimator for θ and $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ does not equal 0, then $\hat{\boldsymbol{\theta}}^{\mathbf{2}}$ is not an unbiased estimator for θ^2.

2. If $\tilde{x}_1$, $\tilde{x}_2$,..., and $\tilde{x}_n$ are independent Bernoulli random variables with the same parameter θ, show that the sample proportion $\hat{\boldsymbol{\theta}} = \dfrac{\tilde{y}}{n}$, where $\tilde{y} = \sum_{i=1}^{n} \tilde{x}_i$, is a minimum variance unbiased estimator for the parameter θ. Note that $\tilde{y}$ is a binomial random variable with the parameters n and θ.

3. Let $\bar{\boldsymbol{x}}_{\mathbf{1}}$ be the mean of a random sample of size n from a normal population with the mean μ and the variance σ_1^2, and $\bar{\boldsymbol{x}}_{\mathbf{2}}$ be the mean of a random sample of size n from a normal population with the mean μ and the variance σ_2^2. If the two random samples are independent, show that

   (a) $w \cdot \bar{\boldsymbol{x}}_{\mathbf{1}} + (1-w) \cdot \bar{\boldsymbol{x}}_{\mathbf{2}}$, where $0 \leq w \leq 1$, is an unbiased estimator for μ;

   (b) the variance of the estimator is a minimum when $w = \dfrac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$.

4. If $\bar{\boldsymbol{x}}_{\mathbf{1}}$ and $\bar{\boldsymbol{x}}_{\mathbf{2}}$ are the means of independent random samples of size n_1 and n_2 from a normal population with the mean μ and the variance σ^2, show that the variance of the unbiased estimator $w \cdot \bar{\boldsymbol{x}}_{\mathbf{1}} + (1-w) \cdot \bar{\boldsymbol{x}}_{\mathbf{2}}$ is a minimum when $w = \dfrac{n_1}{n_1 + n_2}$.

5. Show that the mean of a random sample of size n from an exponential population $\tilde{x}$ is a weakly consistent estimator for its parameter θ.

6. With reference to Exercise 5 show that the sample mean is a sufficient estimator for the exponential parameter θ. Show it in two ways: (a) without using the factorization theorem and (b) making use of the factorization theorem.

7. With reference to Exercise 2 show that the statistic $\hat{\boldsymbol{\theta}} = \dfrac{\tilde{y}}{n}$ is a sufficient estimator for θ. Show it in two ways: (a) without using the factorization theorem and (b) making use of the factorization theorem.

8. Use the method of moments to find an estimator for the parameter θ of the uniform density on the interval $(0, \theta)$.

9. If x_1, x_2,..., and x_n are the values of a random sample of size n from a population $\tilde{x}$ having the density

$$f(x;\theta) = \begin{cases} \dfrac{2(\theta - x)}{\theta^2} & \text{for } 0 < x < \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

   find an estimator for θ by the method of moments.

10. (a) Given a random sample of size n from a normal population with the mean μ and the variance σ^2, find joint maximum likelihood estimators for these two parameters.

    (b) Given a random sample of size n from a normal population with the known mean μ, find the maximum likelihood estimator for σ.

11. If x_1, x_2,..., and x_n are the values of a random sample of size n from a geometric population $\tilde{x}$, find an estimate of its parameter θ using

    (a) the method of moments;

    (b) the method of maximum likelihood.

12. Given a random sample of size n from a population $\tilde{x}$ having the density

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } x > \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

    find an estimator for the parameter θ by the method of maximum likelihood.

13. Given independent random samples $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$, and $\{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n\}$ from two normal populations having the means $\mu_1 = \alpha + \beta$ and $\mu_2 = \alpha - \beta$ and the common variance $\sigma^2 = 1$, find simultaneous maximum likelihood estimators for α and β.

14. (a) Prove that, if $\tilde{x}$ is a binomial random variable with the parameters $\boldsymbol{\theta}$ and n, where n is known, and the prior distribution of its parameter $\boldsymbol{\theta}$ is a beta distribution with parameters α and β, then the posterior distribution of $\boldsymbol{\theta}$ given $\tilde{x} = x$ is a beta distribution with the parameters $x + \alpha$ and $n - x + \beta$.

    (b) Under the same distributional assumptions as in (a), find the Bayesian estimate $\widehat{\theta}$ for the parameter value θ when the loss function is $L\left(\theta, \widehat{\theta}\right) = \left(\theta - \widehat{\theta}\right)^2$ and we observe one realization x of the random variable $\tilde{x}$.

    (c) Is the previous Bayesian estimator $\hat{\boldsymbol{\theta}}$ a strongly consistent estimator for the true parameter value $\theta \in (0,1)$, i.e., does $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \theta$ when $n \longrightarrow \infty$?

15. (a) Prove that, if $\bar{\boldsymbol{x}}$ is the mean of a random sample of size n from a normal population with the known variance σ^2, and the prior distribution of the mean $\boldsymbol{\mu}$ is a normal distribution with the mean μ_0 and the variance σ_0^2, then the posterior distribution of $\boldsymbol{\mu}$ given $\bar{\boldsymbol{x}} = \bar{x}$ is a normal distribution with the mean μ_1 and the variance σ_1^2, where

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \text{ and } \frac{1}{\sigma_1^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}.$$

    Write the previous formulas in terms of the precisions of the previous distributions, $\tau \equiv 1/\sigma^2$, $\tau_0 \equiv 1/\sigma_0^2$, and $\tau_1 \equiv 1/\sigma_1^2$.

(b) Under the same distributional assumptions as in (a), find the Bayesian estimate $\hat{\mu}$ for the value μ of the population mean when the loss function is $L(\mu, \hat{\mu}) = (\mu - \hat{\mu})^2$ and we observe the value $\bar{x}$ of the mean $\bar{\boldsymbol{x}}$ of a random sample of size n.

16. If a random sample of size $n = 20$ from a normal population with the variance $\sigma^2 = 225$ has the mean $\bar{x} = 64.3$, construct a 95% confidence interval for the population mean μ.

17. A paint manufacturer wants to determine the average drying time of a new interior wall paint. The drying time of a given area is normally distributed. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutes and a standard deviation of 8.4 minutes, construct a 95% confidence interval for the true mean μ.

18. Construct a 94% confidence interval for the actual difference between the average lifetimes of two kinds of light bulbs, given that a random sample of 40 light bulbs of one kind lasted on the average 418 hours of continuous use and 50 light bulbs of another kind lasted on the average 402 hours. The population standard deviations are known to be $\sigma_1 = 26$ and $\sigma_2 = 22$.

19. A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of Brand A had an average nicotine content of 3.1 milligrams with the standard deviation of 0.5 milligram, while eight cigarettes of Brand B had an average nicotine content of 2.7 milligrams with a standard deviation of 0.7 milligram. Assuming that the two sets of data are random samples from normal populations with equal variances, construct a 95% confidence interval for the true difference in the average nicotine content of the two brands of cigarettes.

20. If x is a value of a random variable having an exponential distribution with parameter θ, find k so that the interval $(0, kx)$ is a 95% confidence interval for the parameter θ.

21. If x_1 and x_2 are the values of a random sample of size 2 from a population having a uniform density on the interval $(0, \theta)$, find k so that

$$0 < \theta < k(x_1 + x_2)$$

is a $(1 - \alpha)\,100\%$ confidence interval for θ with $\alpha < 1/2$.

22. Measurements of the blood pressure of 25 elderly women have a mean of $\bar{x} = 140$ mm of mercury. If these data can be looked upon as a random sample from a normal population with $\sigma = 10$ mm of mercury, construct a 95% confidence interval for the population mean μ.

23. A study is being made to estimate the proportion of voters in a sizeable community who favor the construction of a nuclear plant. If it is found that only 140 of 400 voters selected at random favor the project, find a 95% confidence interval for the proportion of all voters in this community who favor the project.

24. If it is found that 132 of 200 voters in District A favor a given candidate for election to the United States Senate and 90 of 150 voters in District B favor the same candidate, find a 99% confidence interval for $\theta_1 - \theta_2$, the difference between the actual proportions of voters from the two districts who favor the candidate.

25. In 16 test runs the gasoline consumption of an experimental engine had a standard deviation of 2.2 gallons. Assume that the observed data can be looked upon as a random sample from a normal population. Construct a 99% confidence interval for σ^2, measuring the true variability of the gasoline consumption of this engine.

26. With reference to Exercise 19, find a 98% confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$.

27. Let $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3\}$ be a random sample from a Bernoulli population $\tilde{x}$ with parameter θ. Recall that θ is the probability of success in a Bernoulli trial. Consider the statistic $\tilde{y} = \dfrac{1}{6}\left(\tilde{x}_1 + 2\tilde{x}_2 + 3\tilde{x}_3\right)$.

    (a) Find the probability function $f(x_1, x_2, x_3, y)$ of the random vector $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{y})$.

    (b) Find the probability function $f_{\tilde{y}}(y)$ of the random variable $\tilde{y}$.

    (c) Find the marginal probability function $h(x_2, y)$ of the random vector $(\tilde{x}_2, \tilde{y})$.

    (d) Find the covariance between the random variables $\tilde{x}_2$ and $\tilde{y}$, $\mathrm{Cov}\left(\tilde{x}_2, \tilde{y}\right)$.

    (e) Find the conditional probability function of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ given $\tilde{y} = \frac{1}{2}$ evaluated at $X = (1, 1, 0)$, $f_{\tilde{X}|\tilde{y}}\left(1, 1, 0 \,\middle|\, \frac{1}{2}\right)$. Is the statistic $\tilde{y}$ a sufficient estimator for θ?

    (f) Is the statistic $\tilde{y}$ an unbiased estimator for θ? Is the statistic $\tilde{y}$ an efficient estimator for θ? (i.e., is $\tilde{y}$ and unbiased estimator whose variance equals the Cramér-Rao lower bound?) Is the statistic $\tilde{y}$ a minimum variance unbiased estimator for θ? To answer this question compare the variance of $\tilde{y}$ with the variance of the sample mean. Is the Cramér-Rao lower bound reached by the sample mean?

28. Let $\tilde{x}_n$ be a random variable defined as follows:

$$\tilde{x}_n = \begin{cases} \tilde{z} & \text{with probability } (n-1)/n \\ \\ n & \text{with probability } 1/n, \end{cases}$$

    where $\tilde{z} \sim \mathrm{N}(0, 1)$. Compute $\operatorname{plim}_{n\to\infty} \tilde{x}_n$, $\lim_{n\to\infty} \mathrm{E}(\tilde{x}_n)$, and $\mathrm{AE}(\tilde{x}_n)$. Note that $\tilde{x}_n$ is a mixture.

29. Let $\tilde{y}_n$ be a random variable defined as follows:

$$\tilde{y}_n = \begin{cases} 0 & \text{with probability } (n-1)/n \\ \\ n^2 & \text{with probability } 1/n. \end{cases}$$

Compute $\underset{n\to\infty}{\text{plim}}\ \tilde{y}_n$, $\lim_{n\to\infty} \text{E}(\tilde{y}_n)$, and $\text{AE}(\tilde{y}_n)$.

30. We have obtained x "successes" in n identical and independent trials.

    (a) Find the value $\hat{\theta}_{ML}$ of the maximum likelihood estimator of the probability θ of success in each trial.

    (b) Prove that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ for θ is the most efficient (or the best) estimator in the class of unbiased estimators for θ.

    (c) Does the estimator $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ converge in mean square to θ, $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{m} \theta$ as $n \to \infty$? Does $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{p} \theta$? Does $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{d} \theta$? Does $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{a.s.} \theta$?

31. In the experiment of randomly and independently choosing n real numbers from the closed interval $[0, \beta]$, we have obtained the numbers $x_1, x_2, ..., x_n$. Let x^* be the largest of these numbers, $x^* = \max\{x_1, x_2, ..., x_n\}$.

    (a) Prove that x^* is the value of the maximum likelihood estimator for the parameter β.

    (b) Prove that the statistic $\tilde{x}^* = \max\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ is a weakly consistent estimator for β, that is, $\tilde{x}^* \xrightarrow{p} \beta$ as $n \to \infty$ (or $\underset{n\to\infty}{\text{plim}}\ \tilde{x}^* = \beta$).

    (c) Use the method of moments to find an estimator $\hat{\boldsymbol{\beta}}_{\mathbf{MM}}$ for the parameter β.

32. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population $\tilde{x}$. Assume that the population distribution is normal with the density

$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } x \in (-\infty, \infty)$$

    and that the population variance is known. Prove that the maximum likelihood estimator for the population mean is also a sufficient estimator. You can use the factorization theorem for your proof.

33. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample of size n from a population $\tilde{x}$ having the Poisson distribution with a random parameter $\boldsymbol{\lambda}$. The prior distribution of $\boldsymbol{\lambda}$ is a gamma distribution with the parameters α and β.

    (a) Prove that the posterior distribution of $\boldsymbol{\lambda}$ given $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$ is a gamma distribution with the parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\dfrac{\beta}{1 + n\beta}$.

    (b) Find the Bayesian estimate $\hat{\lambda}_B$ for the parameter value λ when the loss function is quadratic,

$$L\left(\lambda, \widehat{\lambda}\right) = \left(\lambda - \widehat{\lambda}\right)^2,$$

    and we observe that $\bar{\boldsymbol{x}} = \bar{x}$, where $\bar{\boldsymbol{x}}$ is the mean of a random sample with size n and $\bar{x}$ is the realization of this sample mean.

    (c) Compute the following biases associated with the Bayesian estimator $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$:

(i) $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda} \middle| \boldsymbol{\lambda} = \lambda\right]$, which is the bias when the true parameter value is λ, i.e., when $\boldsymbol{\lambda} = \lambda$. Note that $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda} \middle| \boldsymbol{\lambda} = \lambda\right] = \mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \lambda \middle| \boldsymbol{\lambda} = \lambda\right]$.

(ii) $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda} \middle| \bar{\boldsymbol{x}} = \bar{x}\right]$, which is the posterior bias given a realization of the sample mean.

(iii) $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right]$, which is the prior bias.

(d) Answer the following questions about the mean square error associated with the Bayesian estimator $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$:

(i) Compute $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2 \middle| \boldsymbol{\lambda} = \lambda\right]$, which is the mean square error when the true parameter value is λ. What is the limiting bias of $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$ as an estimator for the parameter value λ? Is $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$ a weakly consistent estimator for λ? Note that $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2 \middle| \boldsymbol{\lambda} = \lambda\right] = \mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \lambda\right)^2 \middle| \boldsymbol{\lambda} = \lambda\right]$.

(ii) Compute $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2 \middle| \bar{\boldsymbol{x}} = \bar{x}\right]$, which is the posterior mean square error given a realization of the sample mean.

(iii) Compute $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2\right]$, which is the prior mean square error. Does $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} = \boldsymbol{\lambda}$?

34. (An application of Bayesian estimation) Imagine that you have a coin that might be unbalanced. Your prior distribution of the probability $\tilde{\theta}$ of getting a head when you flip once this coin is given by the uniform density on the interval $(0, 1)$. Then, you run the experiment of flipping the coin 7 times and you obtain 5 heads. What is the posterior distribution of the probability $\tilde{\theta}$ after observing the outcome of that experiment? Find the posterior expected probability of $\tilde{\theta}$ after observing the outcome of that experiment, i.e., compute the conditional expectation $\mathrm{E}\left(\tilde{\theta} \,|\tilde{x} = 5\right)$, where $\tilde{x}$ is the random number of heads when flipping the coin 7 times? Compare this posterior expectation with the prior (or unconditional) expectation.

35. Let $\{\tilde{x}_1, \tilde{x}_1, ..., \tilde{x}_n\}$ be a random sample of size n from a normal population $\tilde{x}$ with mean μ and variance σ^2. Consider the following two statistics:

(a) the sample variance,

$$\boldsymbol{s}^2 = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \overline{\boldsymbol{x}})^2}{n-1}$$

and

(*b*)

$$\hat{\boldsymbol{s}}^2 = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \overline{\boldsymbol{x}})^2}{n},$$

where $\overline{\boldsymbol{x}}$ is the sample mean.

Is the statistic given in (a) a more efficient estimator for the population variance σ^2 than the statistic given in (b)?

36. Consider a random sample $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ of size n from a normal population $\tilde{x}$ with the known mean μ and the unknown variance $V > 0$.

(*a*) Find the Cramér-Rao lower bound on the variance of an unbiased estimator for the population variance V.

(*b*) Does the variance of the sample variance $\boldsymbol{s}_n^{\boldsymbol{2}}$ reach the Cramér-Rao lower bound? Is $\boldsymbol{s}_n^{\boldsymbol{2}}$ an efficient in the limit estimator for V?

Note: Recall that the parameter σ appearing in the normal density is the standard deviation of the population distribution and, thus, $\sigma \equiv V^{1/2} > 0$. Therefore, the normal density can be written as

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{(2\pi)^{1/2} \cdot V^{1/2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{V}} \equiv f(x; V), \quad -\infty < x < \infty.$$

37. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population $\tilde{x}$. Assume that the population distribution is Poisson with the parameter $\lambda > 0$.

(*a*) Find the the method of moments estimator for the parameter λ.

(*b*) Find the maximum likelihood estimator for the parameter λ.

(*c*) Prove that the sample mean is a sufficient estimator for the parameter λ. You can use the factorization theorem for your proof.

(*d*) Use the Cramér-Rao lower bound to check whether the sample mean is a minimum variance unbiased estimator for the parameter λ.

(*e*) The coefficient of skewness (or asymmetry) of a Poisson distribution is $\lambda^{-1/2}$. Find the maximum likelihood estimator for the coefficient of skewness of $\tilde{x}$.

38. Assume that the random variable $\tilde{x}$ has the **Laplace distribution**, whose density is:

$$f(x; \mu, \theta) = \frac{1}{2\theta} e^{-|x-\mu|/\theta}, \quad -\infty < x < \infty.$$

Note that in the exponent of the previous density we have the negative of the absolute value $|x - \mu|$ so that $-|x - \mu| = -(x - \mu)$ if $x \geq \mu$, whereas $-|x - \mu| = x - \mu$ if $x \leq \mu$.

(*a*) Draw the density of $\tilde{x}$ for the particular case where $\mu = -3$ and $\theta = 2$.

(*b*) Find the distribution function (cdf) of $\tilde{x}$. Is the cdf continuous? Is it differentiable everywhere? Is it twice differentiable everywhere?

(c) Find the moment-generating function of $\tilde{x}$? Draw it for the particular case where $\mu = -3$ and $\theta = 2$.

(d) Find the mean and the variance σ^2 of $\tilde{x}$.

(e) Find the density of the random variable $\tilde{z} = |\tilde{x} - \mu|$. Find the mean of $\tilde{z}$.

(f) Let $\mu = -3$ and $\theta = 2$. Find the probability that $\tilde{x} \in (-4, 1)$ given that we know that the event $\tilde{x} \in (-5, -3)$ has occurred. *Note:* You can use for the computations of this part the distribution function you have found in part (b).

Assume for the rest of the exercise that you have a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from the population $\tilde{x}$.

(g) Find the method of moments estimator for the parameter vector (μ, θ).

Assume from now on that the value of the parameter μ is known.

(h) Find the value of the Cramér-Rao lower bound for an unbiased estimator for the parameter θ of the distribution of $\tilde{x}$.

(i) Find the method of moments estimators $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MM}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{MM}}$ for the variance σ^2 of $\tilde{x}$ and for the parameter θ, respectively.

(j) Prove that the method of moments estimator $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MM}}$ is an unbiased estimator for σ^2. Is it consistent? (i.e., check if $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MM}} = \sigma^2$ holds).

(k) Is $\hat{\boldsymbol{\theta}}_{\mathbf{MM}}$ an unbiased estimator for the parameter θ?

(l) Find the maximum likelihood estimators $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{ML}}$ for the parameter θ and for the variance σ^2 of $\tilde{x}$, respectively.

(m) Prove that the maximum likelihood estimator $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{ML}}$ is a biased estimator for the variance σ^2 of $\tilde{x}$. Is it unbiased in the limit?

(n) Prove that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ is an unbiased estimator for θ. Is it efficient? Is it sufficient?

39. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ with $n > 1$ be a random sample from a Poisson population $\tilde{x}$ with parameter λ. Consider the following unbiased estimator for λ : $\hat{\boldsymbol{\lambda}} = \tilde{x}_1$. Improve this estimator in terms of mean square error using the Rao-Blackwell theorem and the result in part (c) of the previous Exercise 37.

40. The probability function of the random variable $\tilde{x}$ is

$$f(x;\theta) = \frac{1}{\theta}\left(\frac{\theta}{\theta+1}\right)^x, \quad \text{for } x = 1, 2, 3, \ldots$$

with $\theta > 0$. Note that the distribution of $\tilde{x}$ is discrete but it has infinite support.

(a) Find the moment-generating function $M_{\tilde{x}}(t)$ of the random variable $\tilde{x}$ and use it to find the mean $\mathrm{E}(\tilde{x})$, the second non-central moment $\mathrm{E}(\tilde{x}^2)$, and the variance $\mathrm{Var}(\tilde{x})$.

Assume that we observe a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from the population $\tilde{x}$ and let $\bar{\boldsymbol{x}}$ be the sample mean of that random sample.

(b) Find the method of moments estimator $\hat{\boldsymbol{\theta}}_{\mathbf{MM}}$ for the parameter θ.

(c) Find the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ for the parameter θ. Check that the second-order condition for a maximum holds.

(d) Find the maximum likelihood estimator $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{ML}}$ for the population variance $\mathrm{Var}(\tilde{x})$.

(e) Is the statistic $\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{x}} - 1$ an unbiased estimator for θ whose variance is equal to the Cramér-Rao lower bound?

(f) Is the statistic $\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{x}} - 1$ a (weakly) consistent estimator for θ?

(g) Is the sample mean $\bar{\boldsymbol{x}}$ a sufficient statistic for θ? Is the statistic $\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{x}} - 1$ a sufficient statistic for θ? Use the factorization theorem for this part.

41. Let us assume that the random variable $\tilde{x}$ has the following density, which is characterized by the parameter $\alpha > 0$:

$$f(x;\alpha) = \begin{cases} \dfrac{10\alpha - 2x}{k\alpha^2} & \text{for } 0 < x < 5\alpha \\ \\ 0 & \text{elsewhere.} \end{cases}$$

(a) Prove that $k = 25$ for all $\alpha > 0$.

From now on, set $k = 25$.

(b) Find the mean and the variance (as functions of α) of the random variable $\tilde{x}$.

For the remaining of this exercise, assume that you observe only a single value x from the population $\tilde{x}$ (i.e., you have access to a sample of size $n = 1$) and that you want to estimate the value of the parameter α using that single observation.

(c) Find the method of moments estimate $\hat{\alpha}_{MM}$ for the parameter α.

(d) Find the bias and the variance of the method of moments estimator $\hat{\boldsymbol{\alpha}}_{\mathbf{MM}}$ for the parameter α.

(e) Find and draw the likelihood function for the parameter α.

(f) Find the maximum likelihood estimate $\hat{\alpha}_{ML}$ for the parameter α.

(g) Find the bias and the variance of the maximum likelihood estimator $\hat{\boldsymbol{\alpha}}_{\mathbf{ML}}$ for the parameter α.

(h) Compare the mean square errors of the previous two estimators, $\hat{\boldsymbol{\alpha}}_{\mathbf{MM}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{ML}}$.

(i) Does the Cramér–Rao lower bound theorem apply to either of the two previous estimators, $\hat{\boldsymbol{\alpha}}_{\mathbf{MM}}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{ML}}$?

(j) Assume that the value α of the parameter is the realization of a random variable $\tilde{\alpha}$ whose prior density is $f_{\tilde{\alpha}}(\alpha)$. Explain how you would obtain the posterior density of $\tilde{\alpha}$ after observing the single observation x from the population $\tilde{x}$.

(k) If the loss function is $\mathrm{L}(\alpha, \hat{\alpha}) = (\alpha - \hat{\alpha})^2$, use your answer to part (j) to explain how you would compute the Bayesian estimator $\hat{\boldsymbol{\alpha}}_{\mathbf{B}}$ for the parameter α.

Note on parts (j) and (k): Because an explicit functional form for the prior density $f_{\tilde{\alpha}}(\alpha)$ is not provided, you are not required to explicitly derive the posterior density of $\tilde{\alpha}$ or the Bayesian estimator. Instead, provide a description of the procedure you would use to obtain them.

9. Hypothesis Testing

9.1. Statistical hypotheses and their tests

- A statistical hypothesis is a conjecture about the distribution of a random vector.
- A hypothesis can be simple (if the conjecture contains a single distribution) or composite (if the conjecture is composed of several simple hypotheses).
- A test of a hypothesis is a rule that allows to reject or not the hypothesis depending on each foreseeable result of a random experiment.

- The hypothesis we want to test is called the "null hypothesis" H_0.
- A null hypothesis is tested against an alternative hypothesis H_1.
- Let $\theta \in \Theta \subset \mathbb{R}^K$ be an unknown parameter vector characterizing the distribution of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$.
- **Examples:**
- Both the null and the alternative hypotheses are simple:

$$H_0 : \theta = \theta_0 \ , \quad H_1 : \theta = \theta_1 .$$

- The null hypothesis is simple but the alternative is composite:

$$H_0 : \theta = \theta_0 \ , \quad H_1 : \theta \neq \theta_0 .$$

- Both the null and the alternative hypotheses are composite:

$$H_0 : \theta \in B_0 \ , \quad H_1 : \theta \in B_1 \left(\equiv B_0^c \right) .$$

- Tests are usually based on the value X of a sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ or on the value t of a statistic $\tilde{t} = g\left(\tilde{X}\right)$. This statistic is called the **test statistic**.

- For a test based on a statistic $\tilde{t} = g\left(\tilde{X}\right)$, we construct a function $\widehat{\phi}$ called the test function (or simply "the test") as follows:

$$\widehat{\phi}(\tilde{t}) = \begin{cases} 1 & \text{if } \tilde{t} \in C_{\widehat{\phi}} \\ \\ 0 & \text{if } \tilde{t} \notin C_{\widehat{\phi}} \end{cases}$$

and we reject H_0 if $\widehat{\phi}(t) = 1$ $\left(\text{or if } t \in C_{\widehat{\phi}}\right)$ and we do not reject H_0 if $\widehat{\phi}(t) = 0$ $\left(\text{or if } t \notin C_{\widehat{\phi}}\right)$.

- Note that the previous test function can we written as a function of the sample:

$$\phi(\tilde{X}) = \begin{cases} 1 & \text{if } \tilde{X} \in C_\phi \\ \\ 0 & \text{if } \tilde{X} \notin C_\phi \end{cases}$$

  and we reject H_0 if $\phi(X) = 1$ $\left(\text{or if } X \in C_\phi\right)$ and we do not reject H_0 if $\phi(X) = 0$ $\left(\text{or if } X \notin C_\phi\right)$.

- Obviously, $\widehat{\phi}(\tilde{t}) = \widehat{\phi}(g(\tilde{X})) = \phi(\tilde{X})$, i.e., $\phi = \widehat{\phi}(g)$ and, since $\tilde{t} = g(\tilde{X})$, then $C_\phi = g^{-1}\left(C_{\widehat{\phi}}\right)$.

- The set C_ϕ (or $C_{\widehat{\phi}}$) where H_0 is rejected is called the critical region, whereas its complement C_ϕ^c (or $C_{\widehat{\phi}}^c$) is the acceptance region.

- Note that the function ϕ (or $\widehat{\phi}$) is the indicator function of the critical region C_ϕ (or $C_{\widehat{\phi}}$), $\phi = \mathbb{I}_{C_\phi}$ and $\widehat{\phi} = \mathbb{I}_{C_{\widehat{\phi}}}$.

- Assume that the parameter value characterizing the distribution of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is θ. The distribution of $\tilde{X}$ when the parameter is θ is denoted by $P_{\tilde{X}}(\cdot; \theta)$.
- Let $\mathrm{E}_\theta(\cdot)$ be the expectation when $\tilde{X} \sim P_{\tilde{X}}(\cdot; \theta) : \mathcal{B} \longrightarrow [0, 1]$. Let $P(\cdot; \theta) : \mathcal{F} \longrightarrow [0, 1]$ be a probability that induces the distribution $P_{\tilde{X}}(\cdot; \theta)$ for $\tilde{X}$.
- The rejection of the null hypothesis $H_0 : \theta \in B_0$ when it is true is called a **type I error (**or **false positive)**. The probability of committing a type I error under the test function ϕ when the value of the distribution parameter is $\theta \in B_0$ is denoted by $\alpha_\phi(\theta)$.
- The non-rejection (or acceptance) of the null hypothesis $H_0 : \theta \in B_0$ when it is false is called a **type II error (**or **false negative).** The probability of committing a type II error under the test function ϕ when the value of the distribution parameter is $\theta \in B_1$ is denoted by $\beta_\phi(\theta)$.

- Observe that, if $\theta \in B_0$, then

$$\alpha_\phi(\theta) = P\left\{\tilde{X} \in C_\phi; \theta\right\} = P_{\tilde{X}}\left(C_\phi; \theta\right) = \int_{C_\phi} dP_{\tilde{X}}(X; \theta)$$

$$= \int_{\mathbb{R}^n} \mathbb{I}_{C_\phi}(X)\, dP_{\tilde{X}}(X; \theta) = \mathrm{E}_\theta(\phi\left(\tilde{X}\right))$$

and, if $\theta \notin B_0$ (or $\theta \in B_1$), then

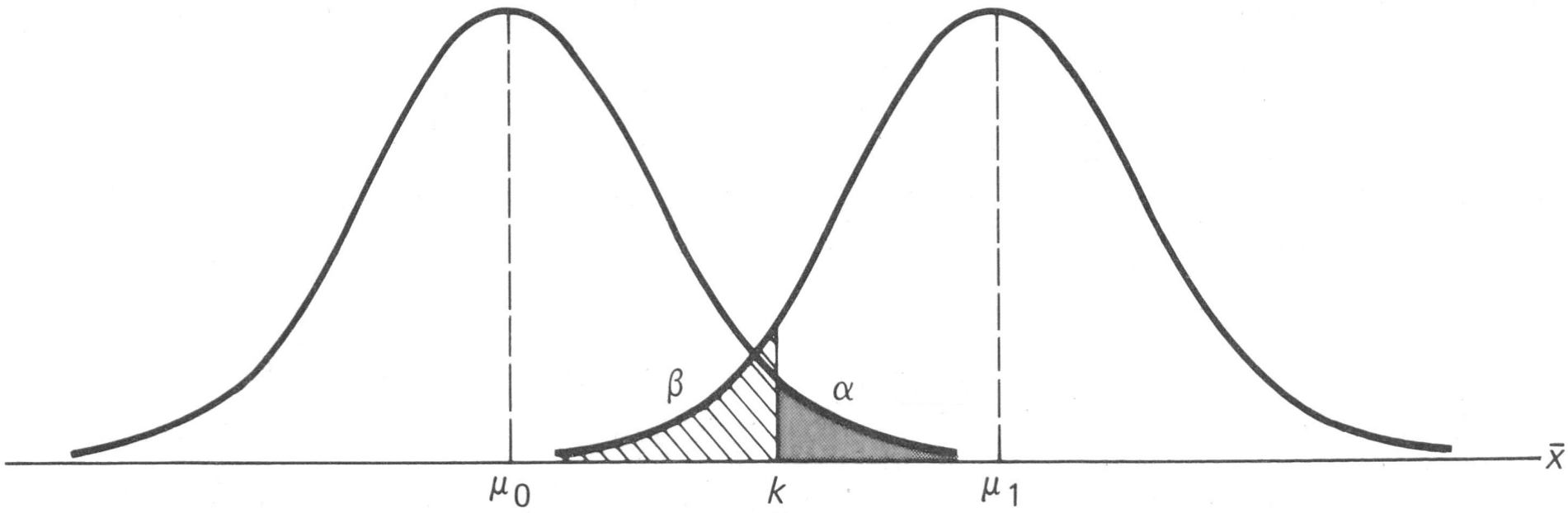
$$\beta_\phi(\theta) = P\left\{\tilde{X} \notin C_\phi; \theta\right\} = P_{\tilde{X}}\left(C_\phi^c; \theta\right) = 1 - P_{\tilde{X}}\left(C_\phi; \theta\right)$$

$$= 1 - \int_{C_\phi} dP_{\tilde{X}}(X; \theta) = 1 - \int_{\mathbb{R}^n} \mathbb{I}_{C_\phi}(X)\, dP_{\tilde{X}}(X; \theta) = 1 - \mathrm{E}_\theta(\phi\left(\tilde{X}\right)).$$

- Note that $\alpha_\phi : B_0 \longrightarrow [0, 1]$, whereas $\beta_\phi : B_1 \longrightarrow [0, 1]$.

- If the null hypothesis is simple, $H_0 : \theta = \theta_0$, then $\alpha_\phi \equiv \alpha_\phi(\theta_0)$ is called the **size** of the critical region or "the level of significance" of the test ϕ.
- When the null hypothesis is composite, $H_0 : \theta \in B_0$, the **size** of the critical region or the level of significance of the test ϕ is the supremum of the possible probabilities of a type I error, $\sup_{\theta \in B_0} \alpha_\phi(\theta)$.
- If $\theta \in B_1$, then $1 - \beta_\phi(\theta) = P_{\tilde{X}}\left(C_\phi; \theta\right) = P\left\{\tilde{X} \in C_\phi; \theta\right\}$ is the probability of rejecting the null hypothesis H_0 when it is false. Note that $1 - \beta_\phi(\theta) = \mathrm{E}_\theta(\phi(\tilde{X}))$ when $\theta \in B_1$.
- $1 - \beta_\phi(\theta)$ is called the **power** of the test ϕ when the distribution parameter is $\theta \in B_1$.
- The ideal test will be that for which $\alpha_\phi(\theta) = 0$ when $\theta \in B_0$ and $\beta_\phi(\theta) = 0$ when $\theta \in B_1$ or, equivalently, $\alpha_\phi(\theta) = 0$ when $\theta \in B_0$ and $1 - \beta_\phi(\theta) = 1$ when $\theta \in B_1$.

- However, α_ϕ and β_ϕ are related.



Critical region for testing $\mu = \mu_0$ against $\mu = \mu_1$.

If we move the critical value k, we will modify $\alpha = \alpha_\phi(\mu_0)$ and $\beta = \beta_\phi(\mu_1)$ in opposite directions.

9.2. The power function of a test

- Let $\theta \in \Theta \subset \mathbb{R}^K$ be a parameter vector characterizing the distribution of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$.
- **Definition.** The power function $\pi_\phi : \Theta \longrightarrow [0,1]$ of a test ϕ of a statistical hypothesis $H_0 : \theta \in B_0$ against an alternative hypothesis $H_1 : \theta \in B_1 = B_0^c$ is given by

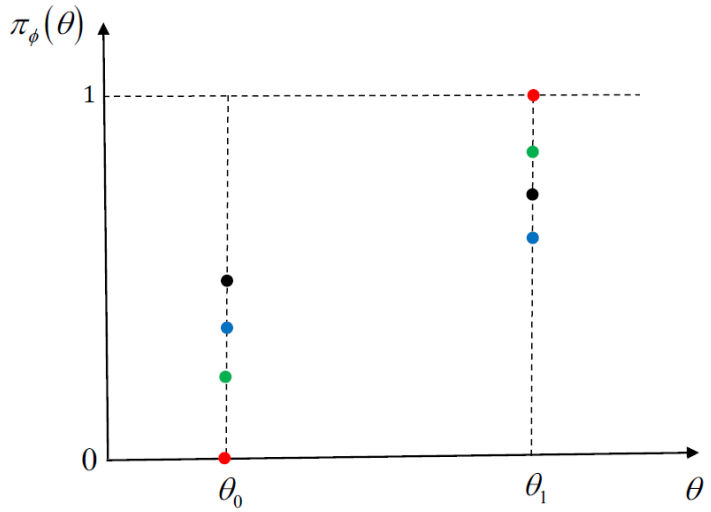
$$\pi_\phi(\theta) = \begin{cases} \alpha_\phi(\theta) & \text{for values of } \theta \text{ under } H_0 \ (\theta \in B_0) \\ 1 - \beta_\phi(\theta) & \text{for values of } \theta \text{ under } H_1 \ (\theta \in B_1). \end{cases}$$

or

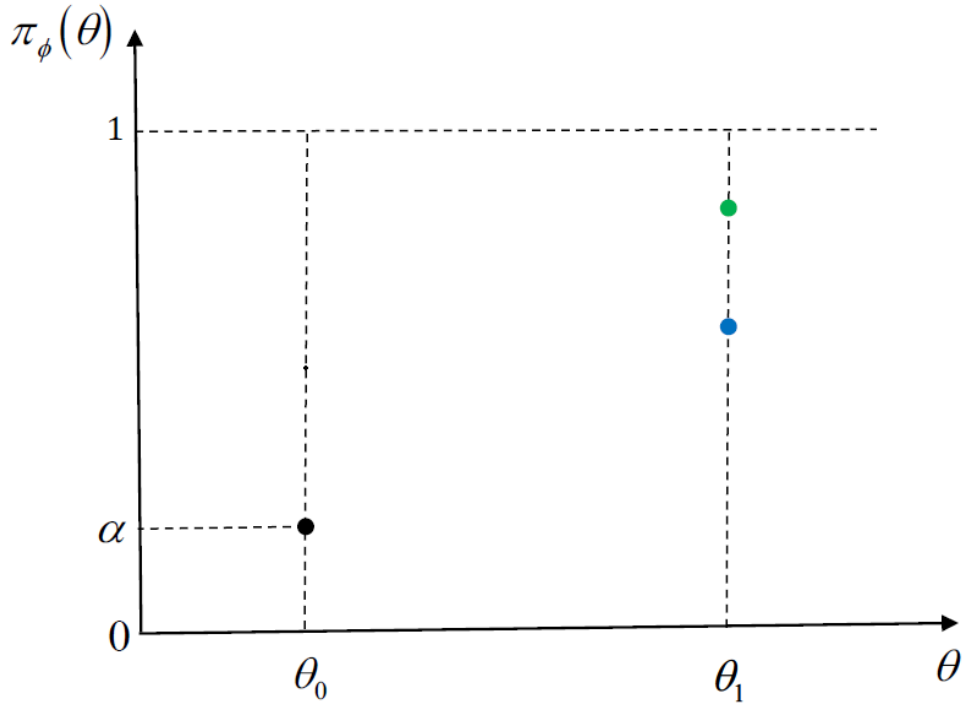
$$\begin{aligned} \pi_\phi(\theta) &= P_{\tilde{X}}\left(C_\phi; \theta\right) = P\left\{\tilde{X} \in C_\phi; \theta\right\} \\ &= \mathrm{E}_\theta(\phi\left(\tilde{X}\right)) \equiv \mathrm{E}_\theta(\phi), \quad \text{for all } \theta \in \Theta. \end{aligned}$$

- The power function gives us the probability of rejecting the null hypothesis for all possible values of the unknown parameter θ.

$$H_0 : \theta = \theta_0 \,, \; H_1 : \theta = \theta_1$$

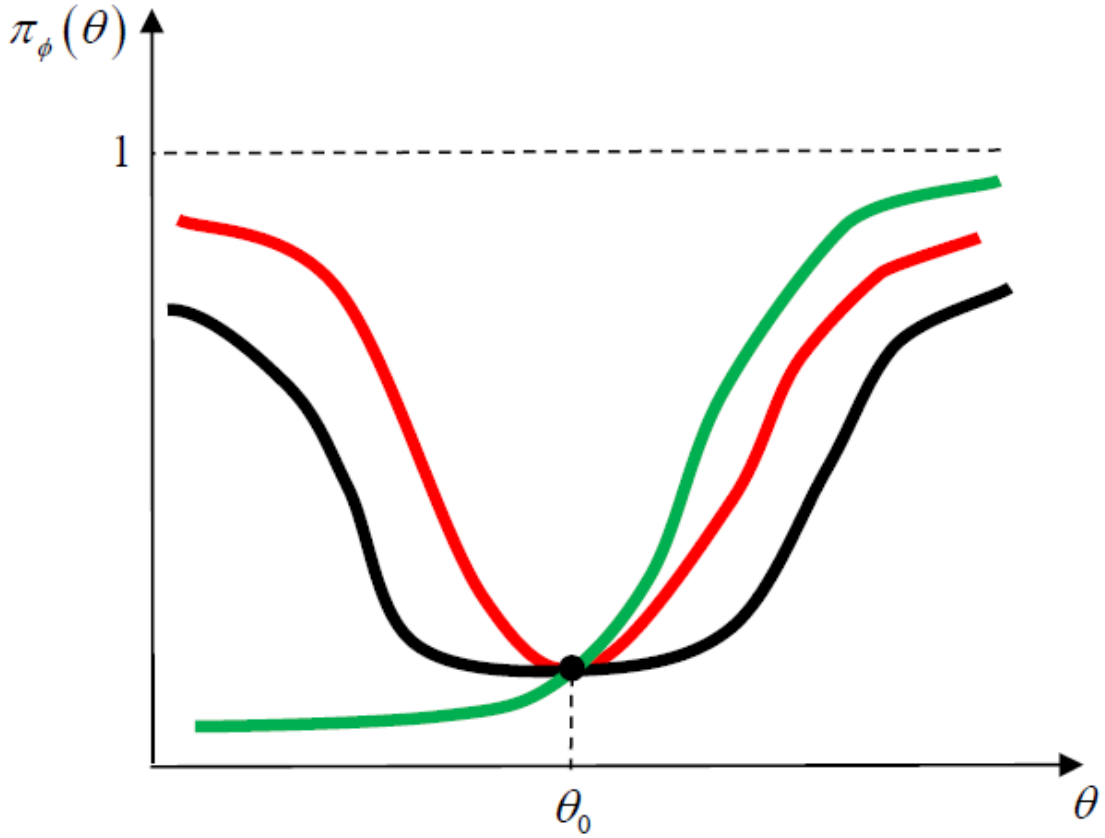


- The test with red power function is the ideal test. The test with green power function is better than the tests with either blue or black power functions since the test with green power function rejects the null hypothesis with a lower probability when it is true (which is desirable) and with a higher probability when it is false (which is also desirable).
- The tests with blue and black power function cannot be ranked since the test with blue power function rejects with a lower probability the null hypothesis when it is true (which is good) but it rejects also the null hypothesis with a lower probability when it is false (which is bad).



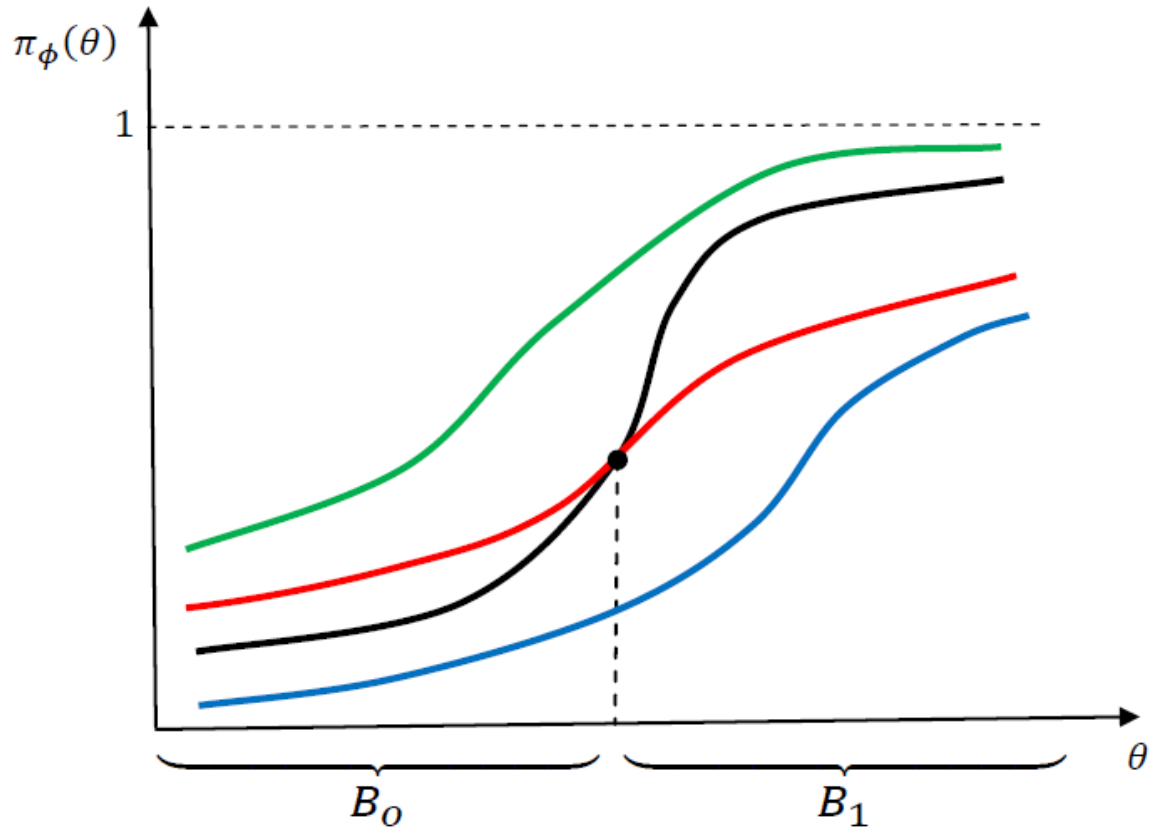
If we fix the probability $\alpha_\phi(\theta_0) \equiv \alpha_\phi$ of committing a type I error (or level of significance or size of the critical region) at the level α for all tests under consideration and then we look for the highest possible power $\pi_\phi(\theta_1) \equiv \pi_\phi = 1 - \beta_\phi(\theta_1) \equiv 1 - \beta_\phi$ (i.e., for the lowest probability $\beta_\phi(\theta_1) \equiv \beta_\phi$ of committing a type II error), then the test with the green power function is better than the test with the blue power function.

$$H_0 : \theta = \theta_0 \,, \; H_1 : \theta \neq \theta_0$$



The test with the red power function is better than that with the black power function. The test with the green power function cannot be ranked with the other two tests.

$$H_0 : \theta \in B_0 \,, \; H_1 : \theta \in B_1 = B_0^c$$



The test with the black power function is better than that with the red power function. The tests with the green and blue power functions cannot be ranked with the other tests.

- A related function is the operating characteristic function of the test,

$$OC_{\phi}(\theta) = P\left\{\tilde{X} \notin C_{\phi}; \theta\right\} = P\left\{\tilde{X} \in C_{\phi}^{c}; \theta\right\} = P_{\tilde{X}}\left(C_{\phi}^{c}; \theta\right)$$

$$1 - P_{\tilde{X}}\left(C_{\phi}; \theta\right) = 1 - \pi_{\phi}(\theta) = 1 - \mathrm{E}_{\theta}(\phi(\tilde{X}))$$

or

$$OC_{\phi}(\theta) = \begin{cases} 1 - \alpha_{\phi}(\theta) & \text{for values of } \theta \text{ under } H_0 \ \ (\theta \in B_0) \\ \beta_{\phi}(\theta) & \text{for values of } \theta \text{ under } H_1 \ \ (\theta \in B_1). \end{cases}$$

- The operating characteristic function gives us the probability of not rejecting the null hypothesis for all possible values of the unknown parameter θ.

9.3. Uniformly most powerful tests

- **Definition:** A test function $\phi(\tilde{X})$ is a Uniformly Most Powerful (UMP) test if for any other test function $\phi'(\tilde{X})$ such that

$$\sup_{\theta \in B_0} \alpha_{\phi'}(\theta) \equiv \alpha' \leq \alpha \equiv \sup_{\theta \in B_0} \alpha_{\phi}(\theta),$$

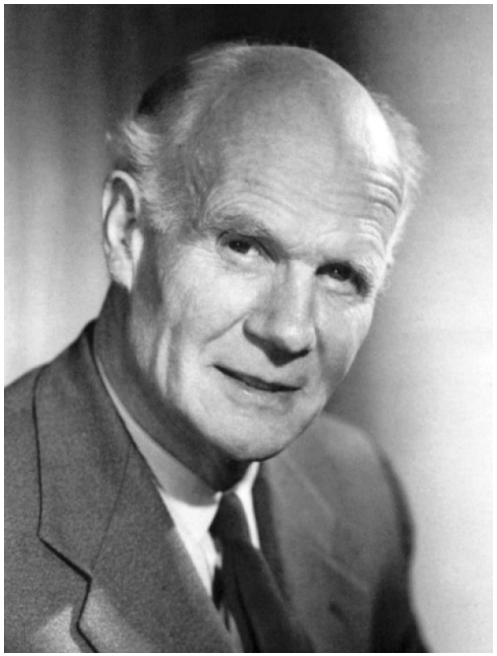
  we have that for all $\theta \in B_1$,

$$\pi_{\phi'}(\theta) \equiv 1 - \beta_{\phi'}(\theta) \leq 1 - \beta_{\phi}(\theta) \equiv \pi_{\phi}(\theta).$$

- That is, a UMP test with size (or level of significance) α is a test that has the greatest power among all possible tests with size no larger than α.

9.4. The Neyman-Pearson lemma

Jerzy Neyman (1894 - 1981)



Egon Pearson (1895 - 1980)

- Assume that both $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ are simple hypotheses and that the test ϕ is based on the value X of the sample $\tilde{X}$. Then, the probability of committing a type I error is

$$\alpha_\phi(\theta_0) = P\left\{\tilde{X} \in C_\phi; \theta_0\right\} \equiv P\left\{\tilde{X} \in C_\phi; H_0\right\} = P_{\tilde{X}}\left(C_\phi; H_0\right),$$

the probability of committing a type II error is

$$\beta_\phi(\theta_1) = P\left\{\tilde{X} \notin C_\phi; \theta_1\right\} \equiv P\left\{\tilde{X} \notin C_\phi; H_1\right\} = P_{\tilde{X}}\left(C_\phi^c; H_1\right),$$

and the power of the test is

$$\pi_\phi(\theta_1) = P\left\{\tilde{X} \in C_\phi; \theta_1\right\} \equiv P\left\{\tilde{X} \in C_\phi; H_1\right\} = P_{\tilde{X}}\left(C_\phi; H_1\right).$$

- We usually fix the probability $\alpha_\phi(\theta_0) \equiv \alpha_\phi$ of committing a type I error (or level of significance or size of the critical region) at the (low) level α (the most commonly used levels are 0.05, 0.01 or 0.001) and then we look for the highest possible power $\pi_\phi(\theta_1) \equiv \pi_\phi = 1 - \beta_\phi(\theta_1) \equiv 1 - \beta_\phi$ (i.e., for the lowest probability $\beta_\phi(\theta_1) \equiv \beta_\phi$ of committing a type II error).

- Let $h_0(X)$ be the density (probability function) of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ under the null hypothesis, whereas $h_1(X)$ is the density (probability function) of the random vector $\tilde{X}$ under the alternative hypothesis.
- If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ constitutes a random sample, then

$$h_j(X) = h_j(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_j(x_i), \quad j = 0, 1,$$

  where f_0 (f_1) is the pdf or pmf of the population under the null (alternative) hypothesis.
- The value of the ratio

$$\frac{h_0(X)}{h_1(X)}$$

  should be small if H_0 is false and large if H_0 is true. Therefore, we should expect that the critical region of the test contains the sample values X for which the previous ratio is small.
- *Note:* we assume values of X for which $h_0(X) / h_1(X)$ is well-defined in $\overline{\mathbb{R}}_+$, i.e., when it is not equal to $0/0$, which means that we must have $h_0(X) > 0$ or $h_1(X) > 0$ or both.

- Note that, for a given sample value X, $h_0(X) = L_0$, where L_0 is the value of the likelihood function under the null hypothesis, and $h_1(X) = L_1$, where L_1 is the value of the likelihood function under the alternative hypothesis.
- **Neyman-Pearson Lemma.** In testing the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative hypothesis $H_1 : \theta = \theta_1$, the test

$$\phi_k\left(\tilde{X}\right) = \begin{cases} 1 & \text{if } \tilde{X} \in C_{\phi_k} \\ \\ 0 & \text{if } \tilde{X} \notin C_{\phi_k}, \end{cases}$$

where

$$C_{\phi_k} = C_k \equiv \left\{ \frac{h_0(X)}{h_1(X)} \leq k \right\},$$

with $k > 0$, is a UMP test with level of significance α_{ϕ_k}, i.e., $\phi_k\left(\tilde{X}\right)$ is the most powerful test with significance level no larger than α_{ϕ_k}, where

$$\alpha_{\phi_k} = P_{\tilde{X}}\left(C_{\phi_k}; \theta_0\right) \equiv P_{\tilde{X}}\left(C_{\phi_k}; H_0\right).$$

- The lemma says that the power of any other test ϕ with $\alpha_\phi \leq \alpha_{\phi_k}$ does not exceed the power of ϕ_k. That is,

$$\alpha_{\phi_k} \geq \alpha_\phi \implies 1 - \beta_{\phi_k} \geq 1 - \beta_\phi.$$

- In particular, if we fix the level of significance at the level α,

$$\alpha = \alpha_\phi = \alpha_{\phi_k} \implies 1 - \beta_{\phi_k} \geq 1 - \beta_\phi.$$

- **Proof.** We want to prove that

$$\left(1 - \beta_{\phi_k}\right) - \left(1 - \beta_\phi\right) = \mathrm{E}_1(\phi_k) - \mathrm{E}_1(\phi) = \mathrm{E}_1\left(\phi_k - \phi\right) \geq 0$$

if ϕ is any other test such that

$$\alpha_{\phi_k} - \alpha_\phi = \mathrm{E}_0\left(\phi_k\right) - \mathrm{E}_0\left(\phi\right) = \mathrm{E}_0\left(\phi_k - \phi\right) \geq 0,$$

where $\mathrm{E}_0\left(\cdot\right) \equiv \mathrm{E}_{\theta_0}\left(\cdot\right)$ and $\mathrm{E}_1\left(\cdot\right) \equiv \mathrm{E}_{\theta_1}\left(\cdot\right)$.

Assume first that $X \in C_k$. Then,

$$\phi_k - \phi = 1 - \phi \geq 0 \quad \text{and} \quad h_1 \geq \frac{h_0}{k} \geq 0$$

so that

$$(\phi_k - \phi)\, h_1 \geq \frac{1}{k}\,(\phi_k - \phi)\, h_0 \geq 0.$$

Assume now that $X \notin C_k$. Then,

$$\phi_k - \phi = 0 - \phi \leq 0 \quad \text{and} \quad 0 \leq h_1 < \frac{h_0}{k}$$

so that

$$0 \geq (\phi_k - \phi)\, h_1 \geq \frac{1}{k}\,(\phi_k - \phi)\, h_0.$$

Therefore,

$$\left(1 - \beta_{\phi_k}\right) - \left(1 - \beta_{\phi}\right) = \mathrm{E}_1\left(\phi_k - \phi\right) \geq \frac{1}{k}\underbrace{\mathrm{E}_0\left(\phi_k - \phi\right)}_{\left(\alpha_{\phi_k} - \alpha_{\phi}\right) \geq 0} \geq 0. \qquad \textit{Q.E.D.}$$

9.5. The Monotone Likelihood Ratio property

- The Karlin-Rubin theorem is an extension of the Neyman-Pearson lemma for composite hypothesis. Before stating that theorem, we need to introduce a new property for pdfs/pmfs.
- **Definition:** The family of density (or probability) functions $f(x;\theta)$, with $\theta \in \Theta \subset \mathbb{R}$, for univariate distributions ($x \in \mathbb{R}$) has a monotone likelihood ratio (MLR) if, for every $\theta_2 > \theta_1$,

$$\frac{f(x;\theta_2)}{f(x;\theta_1)}$$

  is a non-decreasing function of x.
- *Note:* we will only consider again values of x for which $f(x;\theta_2)/f(x;\theta_1)$ is not equal to $0/0$, which means that we must have $f(x;\theta_2) > 0$ or $f(x;\theta_1) > 0$ or both.
- Thus, if $f(x;\theta)$ has a MLR, then for any pair $\theta_2 > \theta_1$, the greater is x the more likely that it is drawn from the distribution with the parameter value θ_2 rather than θ_1.

- **Application of the MLR property:**
- Assume that a worker exerts either a high or a low effort $\tilde{e}$ in her job. Thus, $\tilde{e}$ can take only two values e_H and e_L with $e_H > e_L$. The prior probabilities of these two values are $f_{\tilde{e}}(e_H)$ and $f_{\tilde{e}}(e_L)$, respectively. Let us assume that the amount of output $\tilde{y}$ produced by this worker arises from a stochastic technology that depends on the amount of effort exerted by the worker so that $g_{\tilde{y}|\tilde{e}}(y|e)$ is the conditional probability that the amount of output is y when the effort is e.
- Using Bayes' theorem, we can compute the conditional probability that $\tilde{e} = e_H$ given $\tilde{y} = y$,

$$f_{\tilde{e}|\tilde{y}}(e_H|y) = \frac{f_{\tilde{e}}(e_H) \cdot g_{\tilde{y}|\tilde{e}}(y|e_H)}{f_{\tilde{e}}(e_H) \cdot g_{\tilde{y}|\tilde{e}}(y|e_H) + f_{\tilde{e}}(e_L) \cdot g_{\tilde{y}|\tilde{e}}(y|e_L)}$$

$$= \frac{f_{\tilde{e}}(e_H)}{f_{\tilde{e}}(e_H) + f_{\tilde{e}}(e_L) \cdot \dfrac{1}{\dfrac{g_{\tilde{y}|\tilde{e}}(y|e_H)}{g_{\tilde{y}|\tilde{e}}(y|e_L)}}}.$$

- If $g_{\tilde{y}|\tilde{e}}\left(y\left|e\right.\right)$ has a strict MLR, i.e.,

$$\frac{g_{\tilde{y}|\tilde{e}}\left(y\left|e_H\right.\right)}{g_{\tilde{y}|\tilde{e}}\left(y\left|e_L\right.\right)}$$

  is strictly increasing in y, then $f_{\tilde{e}|\tilde{y}}\left(e_H\left|y\right.\right)$ is strictly increasing in y. Note that here we view the amount of effort as the parameter for the distribution of output.

- This means that the higher is the (observable) amount of output, the higher is the probability that the worker has exerted a high amount of effort (which is not observable).

- The MLR property plays a crucial role in contract theory, economics of uncertainty, and mechanism design.

- **Proposition 1.** If the family of density (or probability) functions $f(x;\theta)$, with $\theta \in \Theta \subset \mathbb{R}$ and $x \in \mathbb{R}$, has a MLR then, for every $\theta_2 > \theta_1$, the following holds:

  (a) First-order stochastic dominance:

$$F(x;\theta_2) \leq F(x;\theta_1) \text{ for all } x \in \mathbb{R}. \tag{1}$$

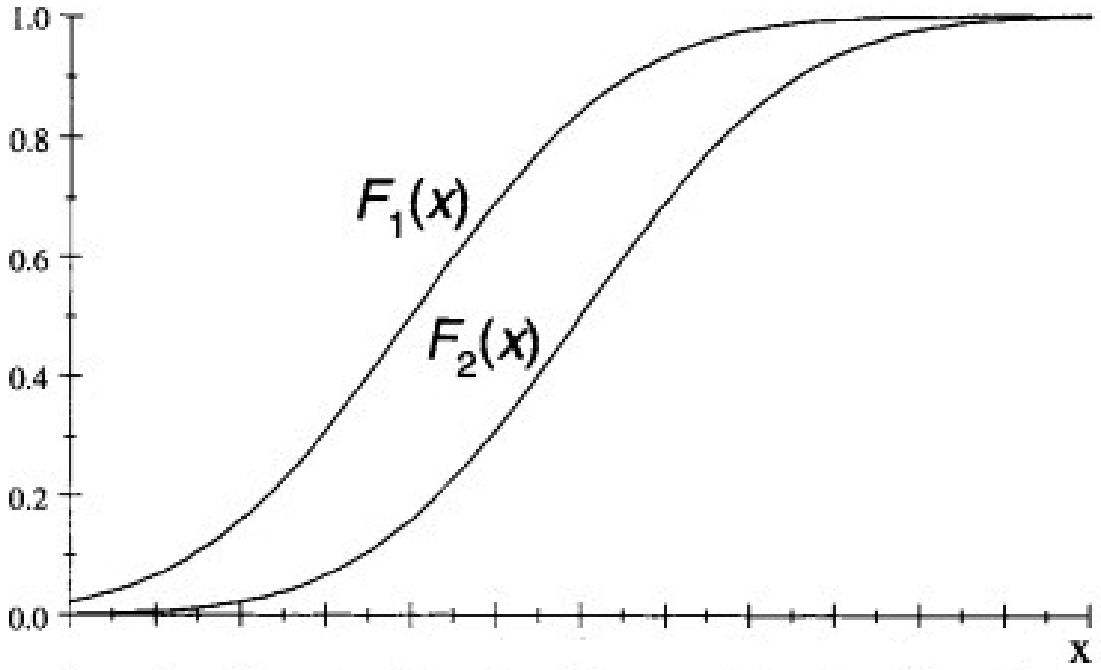
  (b)

$$P\{\tilde{x} > x;\theta_2\} \geq P\{\tilde{x} > x;\theta_1\} \text{ for all } x \in \mathbb{R}. \tag{2}$$

  (c) Non-increasing hazard rate:

$$\frac{f(x;\theta_1)}{1-F(x;\theta_1)} \geq \frac{f(x;\theta_2)}{1-F(x;\theta_2)} \text{ for all } x \in \mathbb{R}. \tag{3}$$

- **Proof:** See the handout.
- When (1) in the previous proposition holds, we say that the distribution with parameter θ_2 first-order "stochastically dominates" the distribution with parameter θ_1. Stochastic dominance is a key concept in the theory of decision under uncertainty.



- The distribution characterized by the distribution function F_2 first-order stochastically dominates the distribution characterized by the distribution function F_1.
- Note that the ordering induced on the set of distributions by the first-order stochastic dominance relation is very partial as any two arbitrary distributions cannot be typically ranked according to this criterion, i.e., we usually have $F_1(x) > F_2(x)$ for some values of x and $F_1(x) < F_2(x)$ for some other values of x.

- The following proposition tells us that, when choosing ex-ante between two distributions of consumption, which can be ranked according to the first-order stochastic dominance relation, all the expected utility maximizing individuals with non-decreasing utility (i.e., individuals who weakly prefer more to less) will unanimously weakly prefer the distribution of consumption that first-order stochastically dominates the other.
- **Proposition 2 (Rothschild and Stiglitz).** Consider two random variables $\tilde{x}$ and $\tilde{y}$ having distributions whose supports are subsets of the bounded interval $[a, b]$ and their associated distribution functions are $F_{\tilde{x}}$ and $F_{\tilde{y}}$, respectively. Then, $F_{\tilde{x}}(z) \leq F_{\tilde{y}}(z)$ for all $z \in \mathbb{R}$ if and only if

$$\mathrm{E}\left[u(\tilde{x})\right] \equiv \int_{\mathbb{R}} u(z) dF_{\tilde{x}}(z) \geq \mathrm{E}\left[u(\tilde{y})\right] \equiv \int_{\mathbb{R}} u(z) dF_{\tilde{y}}(z),$$

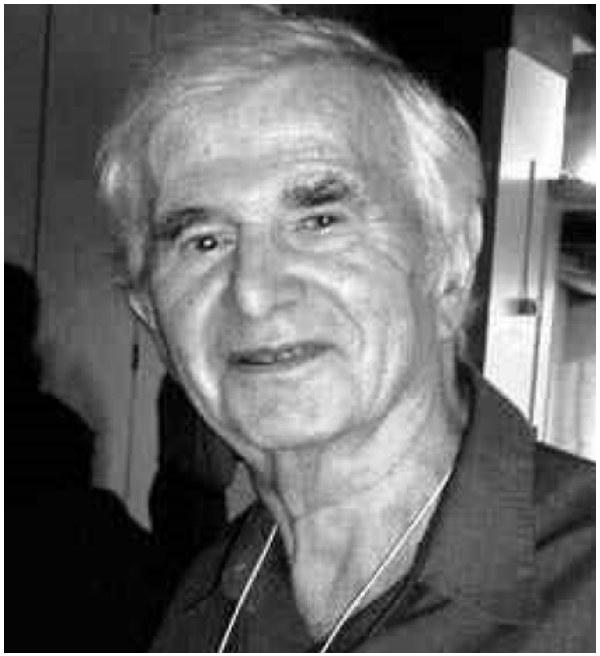
  for all Bernoulli utility functions u that are non-decreasing and continuous on $[a, b]$.
- **Proof:** See the handout.

- When (3) in Proposition 1 holds, we say that the "hazard rate" is non-increasing in θ.
- Imagine the following situation in survival/reliability analysis: the parameter θ measures the health/quality of an individual/component and $f(x;\theta)$ is the (unconditional) probability/density of dying/failure at age x. The hazard rate

$$\eta(x;\theta) = \frac{f(x;\theta)}{1 - F(x;\theta)}$$

  is the conditional probability/density of dying/failing given that the individual/component is at age x. Obviously, this requires to have survived until the age x, which is given by the "survival function" $1 - F(x;\theta)$.
- Under MLR, the higher is the health/quality θ of an individual/component the lower is the hazard rate, i.e., $\eta(x;\theta)$ is non-increasing in θ.
- The MLR property implies first-order stochastic dominance and decreasing hazard rate. The converse is not true.

9.6. The Karlin-Rubin theorem



Samuel Karlin (1924 - 2007)

Herman Rubin (1926 - 2018)

- **Karlin-Rubin theorem.** Assume that (i) $\tilde{t} = t\left(\tilde{X}\right)$ is a test statistic that is a real-valued (or univariate) sufficient estimator for the parameter $\theta \in \Theta \subset \mathbb{R}$ characterizing the distribution of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ and (ii) the family $g(t;\theta)$ of density (or probability) functions for $\tilde{t}$ has a MLR. Then, in testing the null hypothesis $H_0 : \theta \leq \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$, the test

$$\widehat{\phi}(\tilde{t}) = \begin{cases} 1 & \text{if } \tilde{t} > t_0 \\ \\ 0 & \text{if } \tilde{t} \leq t_0, \end{cases}$$

  is a UMP test with size $\alpha = P\{\tilde{t} > t_0; \theta_0\} = \mathrm{E}_{\theta_0}\left(\widehat{\phi}(\tilde{t})\right)$.
- **Proof.** See the handout.
- In all the tests we consider in this chapter, the critical (acceptance) region can be defined by either a weak or a strict inequality since, if the test statistic $\tilde{t}$ is absolutely continuous then $P\{\tilde{t} = t_0; \theta_0\} = 0$, while if $\tilde{t}$ is discrete we can choose t_0 so that $P\{\tilde{t} = t_0; \theta_0\} = 0$.

- Note that, in the proof of the Karlin-Rubin theorem, we invert the ratio of pdfs/pmfs of the Neyman-Pearson lemma and, therefore, we had to change de direction of the inequality in the test.

- Analogously, under the condition of the Karlin-Rubin theorem, the test that rejects $H_0 : \theta \geq \theta_0$ in favor of $H_1 : \theta < \theta_0$ if and only if $\tilde{t} < t_0$ is a UMP test with significance level $\alpha = P\left\{\tilde{t} < t_0; \theta_0\right\}$.

- The Karlin-Rubin theorem also covers the particular case of testing $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ and, thus, the case of testing $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta < \theta_0$. Obviously, the particular case of testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is also covered by the Karlin-Rubin theorem.

- Finally, if we he have a non-increasing likelihood ratio rather than non-decreasing, then we only have to change the direction of the inequalities in the previous tests.

- Obviously, if we use a single observation x from a population distribution characterized by the parameter θ, then $\tilde{x}$ is (trivially) a sufficient estimator for θ. Thus, if the family of density (or probability) functions $f(x;\theta)$, with $\theta \in \Theta \subset \mathbb{R}$, has a MLR then, in testing the null hypothesis $H_0 : \theta \leq \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$, the test

$$\phi(\tilde{x}) = \begin{cases} 1 & \text{if } \tilde{x} > x_0 \\ \\ 0 & \text{if } \tilde{x} \leq x_0, \end{cases}$$

  is a UMP test with size $\alpha = P\{\tilde{x} > x_0; \theta_0\} = \mathrm{E}_{\theta_0}(\phi(\tilde{x}))$.
- For most of the two-sided tests (i.e., $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$) no UMP test exists.
- This is so because the test that is UMP for $\theta < \theta_0$ is not the same as the test that is UMP for $\theta > \theta_0$. A UMP test must be most powerful across every value in H_1.
- A UMP test does not exist in general when we are testing for the value of a parameter vector, $\theta \in \Theta \subset \mathbb{R}^K$.

- If we use the sample mean (or the sum) of a random sample as a (sufficient) estimator $\tilde{t}$ for the parameter θ of the exponential distribution or for the parameter μ of the normal distribution (with σ known), then the pdf of $\tilde{t}$ has a MLR.
- If we use the number (or the percentage) of successes in independent Bernoulli experiments as a (sufficient) estimator $\tilde{t}$ for the parameter θ of the binomial distribution (with n known) or for the parameter λ of the Poisson distribution, then the pmf of $\tilde{t}$ has a MLR.
- The family of exponential pdfs or pmfs, which are the ones with the functional form

$$g(t;\theta) = c(\theta)h(t)e^{w(\theta)t},$$

has a MLR if $w(\theta)$ is non-decreasing.
- Obviously, if $\theta_2 > \theta_1$ and $w(\theta)$ is non-decreasing, then

$$\frac{g(t;\theta_2)}{g(t;\theta_1)} = \frac{c(\theta_2)h(t)e^{w(\theta_2)t}}{c(\theta_1)h(t)e^{w(\theta_1)t}} = \frac{c(\theta_2)}{c(\theta_1)}e^{[w(\theta_2)-w(\theta_1)]t}$$

is non-decreasing in t.

9.7. Likelihood ratio tests

- Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a random vector (or sample) with density (probability function) $h(X;\theta)$ where $\theta \in \Theta \subset \mathbb{R}^K$. Consider the following null and alternative hypotheses: $H_0 : \theta \in B_0$ and $H_1 : \theta \in B_1 = B_0^c$ (i.e., $B_0 \cup B_1 = \Theta$).
- If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a random sample, then

$$h(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

  where f is the density (or probability function) of the common population of the i.i.d. random variables $\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n$.
- Let $L(\theta; X) \equiv h(X;\theta)$ be the likelihood function, which is equal to $\prod_{i=1}^{n} f(x_i;\theta)$ if $\tilde{X}$ is a random sample.

- The value

$$\lambda = \frac{\sup_{\theta \in B_0} L(\theta; X)}{\sup_{\theta \in \Theta} L(\theta; X)} \in [0, 1],$$

which is the value of the likelihood ratio statistic

$$\tilde{\lambda} = \frac{\sup_{\theta \in B_0} L(\theta; \tilde{X})}{\sup_{\theta \in \Theta} L(\theta; \tilde{X})},$$

should be small if H_0 is false and large if H_0 is true.

- Sometimes, we can find the value of the parameter θ that maximizes the likelihood under the null hypothesis,

$$\widehat{\theta} = \underset{\theta \in B_0}{\arg\max}\ L\left(\theta; X\right).$$

- Similarly, we could find the maximum likelihood estimate,

$$\widehat{\theta}_{ML} = \underset{\theta \in \Theta}{\arg\max}\ L\left(\theta; X\right).$$

- Each of the previous maximization problems has solution if B_0 (resp. Θ) is compact and the likelihood function $L\left(\cdot; X\right)$ is continuous on B_0 (resp. Θ).

- Then, the value of the likelihood ratio test statistic becomes

$$\lambda = \frac{L\left(\widehat{\theta}; X\right)}{L\left(\widehat{\theta}_{ML}; X\right)} \in [0, 1] .$$

- Recall also that, sometimes, by modifying the density function on a set with zero Lebesgue measure, we can make

$$\max_{\theta \in B_0} L(\theta; X) = \sup_{\theta \in B_0} L(\theta; X) \qquad \text{and} \qquad \max_{\theta \in \Theta} L(\theta; X) = \sup_{\theta \in \Theta} L(\theta; X) ,$$

where each of the previous suprema exists if the likelihood function $L(\cdot; X)$ is bounded and B_0 (resp. Θ) is compact.

- *Note:* If both the null and the alternative hypotheses were simple, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, then the value of the likelihood ratio test statistic becomes

$$\lambda = \frac{L(\theta_0; X)}{\max\{L(\theta_0; X), L(\theta_1; X)\}} = \frac{L_0}{\max\{L_0, L_1\}}$$

$$= \frac{1}{\max\left\{1, \frac{1}{L_0 / L_1}\right\}} = \frac{1}{\max\left\{1, \frac{1}{h_0(X) / h_1(X)}\right\}} \in [0, 1],$$

  which yields a test equivalent to the one appearing in the Neyman-Pearson lemma.

- Thus, the likelihood ratio test applies the logic of the Neyman-Pearson lemma to test hypotheses that are not necessarily simple and for which the lemma does not necessarily apply!

- **Definition.** The test

$$\widehat{\phi}\left(\tilde{\lambda}\right) = \begin{cases} 1 & \text{if } \tilde{\lambda} \leq k \\ \\ 0 & \text{if } \tilde{\lambda} > k, \end{cases}$$

where $k \in (0,1)$, is a likelihood ratio test of the null hypothesis $H_0 : \theta \in B_0$ against the alternative $H_1 : \theta \in B_1$.

- If we fix the level of significance equal to α, we should find the value k such that

$$\sup_{\theta \in B_0} P\left\{\tilde{\lambda} \in [0,k]\,;\theta\right\} = \sup_{\theta \in B_0} P_{\tilde{\lambda}}\left([0,k]\,;\theta\right) = \sup_{\theta \in B_0} \int_{[0,k]} dP_{\tilde{\lambda}}(\lambda;\theta) = \alpha.$$

- The problem is that we need to find the distribution $P_{\tilde{\lambda}}(\cdot;\theta)$ of the test statistic $\tilde{\lambda}$. This is a difficult task!
- Assume now that $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a random sample. Under the null hypothesis,

$$-2 \ln \tilde{\lambda} \longrightarrow \chi^2_M \quad \text{as } n \to \infty. \qquad \text{(Wilks' Theorem)}$$

  where M is the difference between the dimensions of Θ and B_0.
- In this case, for n large, we should reject (not reject) the null hypothesis if

$$-2 \ln \lambda \geq (<) \chi^2_{\alpha,M}.$$

9.8. The Holy Trinity: likelihood ratio, Wald, and score tests.

- Let $L(\theta; X)$ be the likelihood function with $\theta \in \Theta \subset \mathbb{R}$, where X is the value of the sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$, and let $\widehat{\theta}_{ML}$ be the maximum likelihood estimator for θ,

$$\widehat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta; \tilde{X}) = \arg\max_{\theta \in \Theta} \left[\ln L(\theta; \tilde{X}) \right].$$

- Let us just consider the simple null hypothesis $H_0 : \theta = \theta_0$ against the composite alternative hypothesis $H_1 : \theta \neq \theta_0$.

- **The likelihood ratio (LR) test statistic** is

$$\tilde{\lambda} = \frac{L\left(\theta_0; \tilde{X}\right)}{\sup_{\theta \in \Theta} L\left(\theta; \tilde{X}\right)} = \frac{L\left(\theta_0; \tilde{X}\right)}{\max_{\theta \in \Theta} L\left(\theta; \tilde{X}\right)} = \frac{L\left(\theta_0; \tilde{X}\right)}{L\left(\widehat{\theta}_{ML}; \tilde{X}\right)},$$

  and the null hypothesis is rejected when the value λ of this test statistic is low.

- Note that the LR test requires to evaluate the likelihood function both at the null hypothesis and at its unconstrained maximum.

- **Wald test statistic.** It is based on the (squared) standardized difference between the hypothesized value θ_0 and the unconstrained maximizing argument $\widehat{\boldsymbol{\theta}}_{ML}$ of the likelihood function:

$$\widetilde{W} = \frac{\left(\widehat{\boldsymbol{\theta}}_{ML} - \theta_0\right)^2}{\text{Var}\left(\widehat{\boldsymbol{\theta}}_{ML}\right)} = \left(\frac{\widehat{\boldsymbol{\theta}}_{ML} - \theta_0}{\sigma_{\widehat{\boldsymbol{\theta}}_{ML}}}\right)^2,$$

  and the null hypothesis is rejected when the value W of this test statistic is high, i.e., when θ_0 is relatively far from $\widehat{\boldsymbol{\theta}}_{ML}$.

- Note that the Wald test only requires to find the unconstrained maximizing argument $\widehat{\boldsymbol{\theta}}_{ML}$ of the likelihood function.

Abraham Wald (1902 - 1950)

- **Score (or Rao or Lagrange multiplier) test statistic.** It is based on the square of the slope (relative to its curvature) of the log-likelihood function evaluated at the hypothesized value θ_0,

$$\widetilde{S} = \left. \frac{\left(\frac{\partial \ln L\left(\theta; \tilde{X}\right)}{\partial \theta} \right)^2}{-\mathrm{E}\left(\frac{\partial^2 \ln L\left(\theta; \tilde{X}\right)}{\partial \theta^2} \right)} \right|_{\theta=\theta_0},$$

  and the null hypothesis is rejected when the value S of this test statistic is high, i.e., when the slope of $\ln L\left(\theta; \tilde{X}\right)$ evaluated at θ_0 is far from zero so that $\ln L\left(\theta; \tilde{X}\right)$ evaluated at θ_0 is relatively far from its unconstrained maximum.

- Note that the denominator of $\widetilde{S}$, $-\mathrm{E}\left(\frac{\partial^2 \ln L\left(\theta; \tilde{X}\right)}{\partial \theta^2} \right)$, is the Fisher information $I(\theta)$ since $L\left(\theta; \tilde{X}\right) = h\left(\tilde{X}; \theta\right)$, where $h(X; \theta)$ is the density (probability function) of the random vector $\tilde{X}$. This denominator is a measure of the curvature of the function $\ln h\left(\cdot; \tilde{X}\right)$.

- The Fisher information measures also the amount of information that the random vector $\tilde{X}$ contains about the unknown parameter θ.
- If $h\left(\tilde{X};\cdot\right) = L\left(\cdot;\tilde{X}\right)$ (and, thus, its logarithm) is sharply peaked with respect to changes in θ, it is easy to infer the "correct" value of θ from observing the realization of $\tilde{X}$, i.e., the data provides a lot of information about the parameter θ. If $h\left(X;\cdot\right)$ is flat and spread-out, then it would more difficult to infer the actual "true" value of θ from observing the realization of $\tilde{X}$.
- The derivative of $\ln h\left(\tilde{X};\theta\right) = \ln L\left(\theta;\tilde{X}\right)$ with respect to θ is called the score, $\dfrac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta} = \dfrac{\partial \ln L\left(\theta;\tilde{X}\right)}{\partial\theta}$.
- In fact, the Fisher information is equal to the variance of the score,

$$I\left(\theta\right) = \operatorname{Var}\left(\frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta}\right) = \operatorname{Var}\left(\frac{\partial \ln L\left(\theta;\tilde{X}\right)}{\partial\theta}\right).$$

- This is so because, under the assumptions needed for the proof of the Cramér-Rao lower bound, we know that

$$\mathrm{E}\left[\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right] = 0,$$

so that

$$\mathrm{Var}\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right) = \mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right],$$

and in the the handout about the Cramér Rao lower bound we proved that

$$\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right].$$

- Therefore, the test statistic $\widetilde{S}$ is the square of the standardized score.

- Note also that the score test only requires to evaluate the derivatives of the likelihood function at the hypothesized value θ_0.
- The score test is also called Lagrange multiplier test. To see why, let us consider the following (trivial) constrained maximization problem:

$$\max_{\theta \in \Theta} \ln L\left(\theta; \tilde{X}\right)$$

subject to

$$\theta = \theta_0.$$

- The corresponding Lagrangean is

$$\mathrm{L}(\theta, \ell) = \ln L\left(\theta; \tilde{X}\right) + \ell\left(\theta_0 - \theta\right),$$

  where ℓ is the Lagrange multiplier.

- The F.O.C. with respect to θ is

$$\frac{\partial \mathrm{L}(\theta, \ell)}{\partial \theta} = 0 \iff \frac{\partial \ln L\left(\theta; \tilde{X}\right)}{\partial \theta} = \ell.$$

- Therefore, the score test can be understood as a test of the magnitude of the Lagrange multiplier associated with the constraint $\theta = \theta_0$ at the constrained maximum.

- Note that, if the constraint were non-binding at the constrained maximum, the Lagrange multiplier ℓ should not differ from zero. Conversely, if the constraint $\theta = \theta_0$ is very binding then ℓ should be very different from zero.

- Note that in this section we assume that the set of parameter values under the null hypothesis has zero dimension $(\theta = \theta_0)$ and that the parameter space $\Theta \subset \mathbb{R}$ has dimension equal to one.
- If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a random sample, then $-2 \ln \tilde{\lambda} = -2 \left[\ln L\left(\theta_0; \tilde{X}\right) - \ln L\left(\theta_{ML}; \tilde{X}\right) \right]$, $\widetilde{W}$, and $\widetilde{S}$ are all asymptotically distributed as a χ^2_M under the null hypothesis, where $M = 1$.
- Thus, for large random samples, the null hypothesis is rejected with a level of significance α when

$$-2 \ln \lambda \geq \chi^2_{\alpha,1} \ ,$$

$$W \geq \chi^2_{\alpha,1} \ ,$$

or

$$S \geq \chi^2_{\alpha,1} \ .$$

- The LR, Wald and score tests constitute the so-called "Holy Trinity" of tests.

9.9. Acceptance intervals

- Assume that the parameter value characterizing the distribution of the random vector (or sample) $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is $\theta \in \Theta \subset \mathbb{R}$.
- Consider a real-valued test statistic $\tilde{t} = g\left(\tilde{X}\right)$ whose distribution is $P_{\tilde{t}}\left(\cdot; \theta\right)$.

- Consider now the following test, based on the test statistic $\tilde{t} = g\left(\tilde{X}\right)$, of the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$:

$$\widehat{\phi}(\tilde{t}) = \begin{cases} 1 & \text{if } \tilde{t} \in C_{\widehat{\phi}} \\ \\ 0 & \text{if } \tilde{t} \notin C_{\widehat{\phi}} \end{cases}$$

  and fix the level of significance $\alpha_{\widehat{\phi}}$ $(= \alpha_{\widehat{\phi}}(\theta_0))$ equal to α (recall that the most commonly used levels are 0.05, 0.01 or 0.001).

- Then, if $\theta = \theta_0$, the critical region $C_{\widehat{\phi}}$ of size α must satisfy the following:

$$P\left\{g\left(\tilde{X}\right) \in C_{\widehat{\phi}}; \theta_0\right\} = P\left\{\tilde{t} \in C_{\widehat{\phi}}; \theta_0\right\} = P_{\tilde{t}}\left(C_{\widehat{\phi}}; \theta_0\right) = \alpha.$$

  or

$$P\left\{g\left(\tilde{X}\right) \notin C_{\widehat{\phi}}; \theta_0\right\} = P\left\{\tilde{t} \notin C_{\widehat{\phi}}; \theta_0\right\} = P_{\tilde{t}}\left(C^c_{\widehat{\phi}}; \theta_0\right) = 1 - \alpha.$$

- If we want that the acceptance region (which is the complement of the critical region) be an interval (a, b) then, under the null hypothesis, we must have

$$P\left\{a < g\left(\tilde{X}\right) < b; \theta_0\right\} = P_{\tilde{t}}\left\{(a, b); \theta_0\right\} = 1 - \alpha.$$

- Moreover, for two-tailed tests we make

$$P\left\{g\left(\tilde{X}\right) \leq a; \theta_0\right\} = P_{\tilde{t}}\left\{(-\infty, a]; \theta_0\right\} = \frac{\alpha}{2}$$

and

$$P\left\{g\left(\tilde{X}\right) \geq b; \theta_0\right\} = P_{\tilde{t}}\left\{[b, \infty); \theta_0\right\} = \frac{\alpha}{2}.$$

- Therefore, we reject the null hypothesis if

$$t = g(X) \notin (a, b)$$

and we do not reject the null hypothesis if

$$t = g(X) \in (a, b) \ .$$

- In applications, we use a test statistic $\tilde{t} = g\left(\tilde{X}, \theta\right)$ whose distribution $P_{\tilde{t}}(\cdot; \theta)$ under the null hypothesis is known and it is independent of the parameter θ. The random variable $\tilde{t}$ is thus a pivotal variable (see the next examples).

- **Examples (Two-tailed tests):**
- Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a random sample from a normal population with mean μ and variance σ^2.
- **(1)** If σ^2 is known, the random variable $\tilde{z} = \dfrac{\bar{\mathbf{x}} - \mu_0}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1)$ is a test statistic for the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
- If the level of significance is α, then the acceptance interval is $(-z_{\alpha/2}, z_{\alpha/2})$. We reject the null hypothesis if $|z| \geq z_{\alpha/2}$.

- **(2)** The random variable $\tilde{t} = \dfrac{\bar{\mathbf{x}} - \mu_0}{\mathbf{s}/\sqrt{n}} \sim \mathrm{t}_{n-1}$ is a test statistic for the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.
- If the level of significance is α, then the acceptance interval is $(-t_{\alpha/2,n-1}, t_{\alpha/2,n-1})$. We reject the null hypothesis if $|t| \geq t_{\alpha/2,n-1}$.
- **(3)** The random variable $\tilde{\chi}^2 = \dfrac{(n-1)\mathbf{s^2}}{\sigma_0^2} \sim \chi^2_{n-1}$ is a test statistic for the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$.
- If the level of significance is α, then the acceptance interval is $\left(\chi^2_{1-\frac{\alpha}{2},n-1}, \chi^2_{\alpha/2,n-1}\right)$. We reject the null hypothesis if $\chi^2 \notin \left(\chi^2_{1-\frac{\alpha}{2},n-1}, \chi^2_{\alpha/2,n-1}\right)$.

- Let $\mathbf{s_1^2}$ and $\mathbf{s_2^2}$ be the variances of independent random samples of size n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2, respectively.
- **(4)** The random variable $\widetilde{F} = \dfrac{\mathbf{s_1^2}}{\mathbf{s_2^2}} \sim \mathrm{F}_{n_1-1,n_2-1}$ is a test statistic for the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$.
- If the level of significance is α, then the acceptance interval is

$$\left(F_{1-\frac{\alpha}{2},n_1-1,n_2-1}, F_{\alpha/2,n_1-1,n_2-1}\right) = \left(\frac{1}{F_{\alpha/2,n_2-1,n_1-1}}, F_{\alpha/2,n_1-1,n_2-1}\right).$$

We reject the null hypothesis if

$$F \notin \left(\frac{1}{F_{\alpha/2,n_2-1,n_1-1}}, F_{\alpha/2,n_1-1,n_2-1}\right).$$

- **Examples (One-tailed tests):**
- Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a random sample from a normal population with mean μ and variance σ^2.
- **(1)** If σ^2 is known, the random variable $\dfrac{\bar{\mathbf{x}} - \mu_0}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1)$ is a test statistic for the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu > (<)\mu_0$.
- If the level of significance is α, then the acceptance interval is $(-\infty, z_\alpha)$ (or $(-z_\alpha, \infty)$) and the critical interval is $[z_\alpha, \infty)$ (or $(-\infty, -z_\alpha]$).

- **(2)** The random variable $\frac{\bar{\mathbf{x}} - \mu_0}{\mathbf{s}/\sqrt{n}} \sim t_{n-1}$ is a test statistic for the null hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu > (<)\mu_0$.
- If the level of significance is α, then the acceptance interval is $(-\infty, t_{\alpha,n-1})$ (or $(-t_{\alpha,n-1}, \infty)$) and the critical interval is $[t_{\alpha,n-1}, \infty)$ (or $(-\infty, -t_{\alpha,n-1}]$).
- **(3)** The random variable $\frac{(n-1)\mathbf{s}^2}{\sigma_0^2} \sim \chi^2_{n-1}$ is a test statistic for the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 > (<)\sigma_0^2$.
- If the level of significance is α, then the acceptance interval is $(0, \chi^2_{\alpha,n-1})$ (or $(\chi^2_{1-\alpha,n-1}, \infty)$) and the critical intervals is $[\chi^2_{\alpha,n-1}, \infty)$ (or $(0, \chi^2_{1-\alpha,n-1}]$).

- Let $\mathbf{s_1^2}$ and $\mathbf{s_2^2}$ be the variances of independent random samples of size n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2, respectively.
- **(4)** The random variable $\dfrac{\mathbf{s_1^2}}{\mathbf{s_2^2}} \sim \mathrm{F}_{n_1-1,n_2-1}$ is a test statistic for the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > (<)\sigma_2^2$.
- Let $H_1 : \sigma_1^2 > \sigma_2^2$. If the level of significance is α, then the acceptance interval is $\left(0, F_{\alpha,n_1-1,n_2-1}\right)$ and the critical interval is $\left[F_{\alpha,n_1-1,n_2-1}, \infty\right)$.

- Let $H_1 : \sigma_1^2 < \sigma_2^2$. If the level of significance is α, then the acceptance interval is

$$(F_{1-\alpha,n_1-1,n_2-1}, \infty) = \left(\frac{1}{F_{\alpha,n_2-1,n_1-1}}, \infty \right).$$

and the critical interval is

$$(0, F_{1-\alpha,n_1-1,n_2-1}] = \left(0, \frac{1}{F_{\alpha,n_2-1,n_1-1}} \right].$$

9.10. The p-value

- Assume again that the parameter value characterizing the distribution of the random vector (or sample) $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is $\theta \in \Theta \subset \mathbb{R}$.
- Let $\tilde{t} = g\left(\tilde{X}\right)$ be a test statistic whose distribution under the null hypothesis is $P_{\tilde{t}}\left(\cdot; \theta_0\right)$.
- Consider the test of the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$.
- The p-value is the probability under the null hypothesis of obtaining a value for the test statistic at least as extreme as that given by the sample value, $t = g(X)$.
- In other words, is the minimal size of the critical region for which we would reject the null hypothesis if the value of the statistic $\tilde{t}$ were t.

- If $t = g(X)$ is such that $P\{\tilde{t} \leq t; \theta_0\} \leq 1/2$, then

$$p\text{-value} = 2P\{\tilde{t} \leq t; \theta_0\} = 2P_{\tilde{t}}\{(-\infty, t]; \theta_0\},$$

whereas, if $P\{\tilde{t} \leq t; \theta_0\} > 1/2$, then

$$p\text{-value} = 2P\{\tilde{t} \geq t; \theta_0\} = 2P_{\tilde{t}}\{[t, \infty); \theta_0\}.$$

- Equivalently,

$$p\text{-value} = 2 \cdot \min\{P\{\tilde{t} \leq t; \theta_0\}, P\{\tilde{t} \geq t; \theta_0\}\}.$$

- If we fix the level of significance at α, then we reject H_0 if p-value $\leq \alpha$ and we do not reject H_0 if p-value $> \alpha$.
- In the one-tailed test of the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > (<)\theta_0$, the p-value is either

$$p\text{-value} = P\{\tilde{t} \geq t; \theta_0\} \quad \text{or} \quad p\text{-value} = P\{\tilde{t} \leq t; \theta_0\}.$$

- We use stars to flag levels of significance for three of the most commonly used levels.
- If a p-value is less than 0.05, it is flagged with one star ($*$). If a p-value is less than 0.01, it is flagged with two stars ($**$). If a p-value is less than 0.001, it is flagged with three stars ($***$).
- Note that the p-value gives us directly the strength of the evidence against the null hypothesis. This is an advantage of the p-value over checking whether the value of the test statistic falls in the critical region.
- This is a very interesting article about the use, misuse or abuse of p-values in Economics:

  Imbens, Guido W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives* 95, 157-174.

9.11. Tests concerning means, differences between means, variances, proportions, and differences among several proportions

- See the handout.

9.12. Contingency tables - 9.13. Goodness of fit

- See the handout.

Monotone Likelihood Ratio Property

Proposition 1. If the family of density (or probability) functions $f(x;\theta)$, with $\theta \in \Theta \subset \mathbb{R}$, has a MLR then, for every $\theta_2 > \theta_1$, the following holds:

(a) First-order stochastic dominance:

$$F(x;\theta_2) \leq F(x;\theta_1) \text{ for all } x \in \mathbb{R}. \tag{1}$$

(b)

$$P\{\tilde{x} > x;\theta_2\} \geq P\{\tilde{x} > x;\theta_1\} \text{ for all } x \in \mathbb{R}. \tag{2}$$

(c) Non-increasing hazard rate:

$$\frac{f(x;\theta_1)}{1-F(x;\theta_1)} \geq \frac{f(x;\theta_2)}{1-F(x;\theta_2)} \text{ for all } x \in \mathbb{R}. \tag{3}$$

Proof. Let us prove it for the case where $f(x;\theta)$ is a pdf. The case where $f(x;\theta)$ is a pmf has an analogous proof by using sums instead of integrals.

Since $f(x;\theta)$ has a MLR, if $\theta_2 > \theta_1$ then

$$\frac{f(x_1;\theta_2)}{f(x_1;\theta_1)} \leq \frac{f(x_2;\theta_2)}{f(x_2;\theta_1)}, \text{ for } x_2 \geq x_1,$$

or

$$f(x_1;\theta_2)\cdot f(x_2;\theta_1) \leq f(x_1;\theta_1)\cdot f(x_2;\theta_2), \text{ for } x_2 \geq x_1,$$

or

$$f(x_1;\theta_2)\cdot f(x;\theta_1) \leq f(x_1;\theta_1)\cdot f(x;\theta_2) \text{ for } x \geq x_1 \tag{4}$$

or

$$f(x;\theta_2)\cdot f(x_2;\theta_1) \leq f(x;\theta_1)\cdot f(x_2;\theta_2) \text{ for } x_2 \geq x. \tag{5}$$

Let us fix x and integrate both sides of the inequality (4) with respect to x_1 on $(-\infty, x]$,

$$\int_{(-\infty,x]} f(x_1;\theta_2)\cdot f(x;\theta_1)dx_1 \leq \int_{(-\infty,x]} f(x_1;\theta_1)\cdot f(x;\theta_2)dx_1$$

$$\Longrightarrow F(x;\theta_2)\cdot f(x;\theta_1) \leq F(x;\theta_1)\cdot f(x;\theta_2)$$

or

$$\frac{f(x;\theta_2)}{f(x;\theta_1)} \geq \frac{F(x;\theta_2)}{F(x;\theta_1)}. \tag{6}$$

Similarly, let us fix x and integrate both sides of the inequality (5) with respect to x_2 on (x, ∞),

$$\int_{(x,\infty)} f(x;\theta_2)\cdot f(x_2;\theta_1)dx_2 \leq \int_{(x,\infty)} f(x;\theta_1)\cdot f(x_2;\theta_2)dx_2$$

$$\Longrightarrow f(x;\theta_2)\cdot[1-F(x;\theta_1)] \leq f(x;\theta_1)\cdot[1-F(x;\theta_2)]$$

or

$$\frac{1-F(x;\theta_2)}{1-F(x;\theta_1)} \geq \frac{f(x;\theta_2)}{f(x;\theta_1)}. \tag{7}$$

Combining (6) and (7), we get

$$\frac{1-F(x;\theta_2)}{1-F(x;\theta_1)} \geq \frac{F(x;\theta_2)}{F(x;\theta_1)},$$

which, after rearranging, simplifies to (1)

Expression (1) implies that

$$P\{\tilde{x} > x;\theta_2\} = 1-F(x;\theta_2) \geq 1-F(x;\theta_1) = P\{\tilde{x} > x;\theta_1\}, \text{ for all } x \in \mathbb{R},$$

which proves (2).

Finally, rearranging (7), we get (3). *Q.E.D.*

First-Order Stochastic Dominance

Proposition 2 (Rothschild and Stiglitz). Consider two random variables $\tilde{x}$ and $\tilde{y}$ having distributions whose supports are subsets of the bounded interval $[a, b]$ and their associated distribution functions are $F_{\tilde{x}}$ and $F_{\tilde{y}}$, respectively. Then, $F_{\tilde{x}}(z) \leq F_{\tilde{y}}(z)$ for all $z \in \mathbb{R}$ if and only if

$$\mathrm{E}\left[u(\tilde{x})\right] \equiv \int_{\mathbb{R}} u(z) dF_{\tilde{x}}(z) \geq \mathrm{E}\left[u(\tilde{y})\right] \equiv \int_{\mathbb{R}} u(z) dF_{\tilde{y}}(z),$$

for all Bernoulli utility functions u that are non-decreasing and continuous on $[a, b]$.

Proof.
(a) $F_{\tilde{x}}(z) \leq F_{\tilde{y}}(z)$ for all $z \in \mathbb{R} \Longrightarrow \mathrm{E}\left[u(\tilde{x})\right] \geq \mathrm{E}\left[u(\tilde{y})\right]$ for all u non-decreasing and continuous on $[a, b]$.

From the integration by parts formulae in Section 4.4 of Chapter 4, and since the function u can be viewed as a continuous distribution function on $[a, b]$, we have

$$\mathrm{E}\left[u(\tilde{x})\right] = \int_{[a,b]} u\left(z\right) dF_{\tilde{x}}(z) = u(a)\left[F_{\tilde{x}}(a) - F_{\tilde{x}}(a^{-})\right] + \int_{(a,b]} u\left(z\right) dF_{\tilde{x}}(z)$$

$$= F_{\tilde{x}}(a)u(a) - F_{\tilde{x}}(a^{-})u(a) + \overbrace{F_{\tilde{x}}(b)u(b) - \underbrace{F(a^{+})}_{\substack{F(a) \\ \text{by right-continuity}}} u(a) - \int_{(a,b]} F_{\tilde{x}}(z)du(z)}^{\text{integrating by parts}}$$

$$= \underbrace{F_{\tilde{x}}(b)}_{1}u(b) - \underbrace{F_{\tilde{x}}(a^{-})}_{P\{\tilde{x}<a;\theta\}=0} u(a) - \int_{(a,b]} F_{\tilde{x}}(z)du(z) = u(b) - \int_{(a,b]} F_{\tilde{x}}(z)du(z)$$

$$= u(b) - \int_{[a,b]} F_{\tilde{x}}(z)du(z),$$

where the last equality follows from the continuity of the "distribution function" u on $[a, b]$.

Similarly,

$$\mathrm{E}\left[u(\tilde{y})\right] = \int_{[a,b]} u\left(z\right) dF_{\tilde{y}}(z) = u(b) - \int_{[a,b]} F_{\tilde{y}}(z)du(z).$$

From the previous two expressions,

$$\mathrm{E}\left[u(\tilde{x})\right] - \mathrm{E}\left[u(\tilde{y})\right] = -\overbrace{\int_{[a,b]} \underbrace{\left[F_{\tilde{x}}(z) - F_{\tilde{y}}(z)\right]}_{\substack{\leq 0 \\ \text{Stochastic dominance}}} du\left(x\right)}^{\leq 0} \geq 0. \tag{8}$$

(b) $\mathrm{E}\left[u(\tilde{x})\right] \geq \mathrm{E}\left[u(\tilde{y})\right]$ for all u non-decreasing and continuous on $[a,b] \Longrightarrow F_{\tilde{x}}(z) \leq F_{\tilde{y}}(z)$ for all $z \in \mathbb{R}$.

By contradiction. Assume that there exists a $z' \in [a,b)$ such that $F_{\tilde{x}}(z') > F_{\tilde{y}}(z')$. The right-continuity of the distribution functions implies that there exists a $c \in (z',b)$ such that $F_{\tilde{x}}(z) > F_{\tilde{y}}(z)$ for all $z \in [z',c) \subset [a,b]$.

Consider the following Bernoulli utility function u:

$$u(z) = \begin{cases} \bar{u}_0 & \text{for } z < z' \\ u_0 + \left(\dfrac{\bar{u}_1 - \bar{u}_0}{c - z'}\right)(z - z') & \text{for } z \in [z', c) \\ \bar{u}_1 & \text{for } z > c, \end{cases}$$

where the two constants $\bar{u}_0$ and $\bar{u}_1$ satisfy $\bar{u}_1 > \bar{u}_0$. Note that this Bernoulli utility function u is non-decreasing and continuous for all $z \in \mathbb{R}$, and differentiable a.e. (it is non-differentiable only at z' and c).

Then, from (8),

$$\mathrm{E}\left[u(\tilde{x})\right] - \mathrm{E}\left[u(\tilde{y})\right] = -\int_{[a,b]} \left[F_{\tilde{x}}(z) - F_{\tilde{y}}(z)\right] du\left(z\right)$$

$$= -\int_{[z',c)} \underbrace{\left[F_{\tilde{x}}(z) - F_{\tilde{y}}(z)\right]}_{>0} \overbrace{\underbrace{\left(\frac{\bar{u}_1 - \bar{u}_0}{c - z'}\right)}_{>0} dz}^{du(z)} < 0$$

since $du\left(z\right) = u'(z)dz = 0$ for $z \notin [z',c)$. Thus, we arrive at a contradiction. *Q.E.D.*

The Karlin-Rubin Theorem

Theorem. Assume that (i) $\tilde{t} = t\left(\tilde{X}\right)$ is a test statistic that is a real-valued (or univariate) sufficient estimator for the parameter $\theta \in \Theta \subset \mathbb{R}$ characterizing the distribution of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ and (ii) the family $g(t;\theta)$ of density (or probability) functions for $\tilde{t}$ has a MLR. Then, in testing the null hypothesis $H_0 : \theta \leq \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$, the test

$$\widehat{\phi}\left(\tilde{t}\right) = \begin{cases} 1 & \text{if } \tilde{t} > t_0 \\ \\ 0 & \text{if } \tilde{t} \leq t_0, \end{cases}$$

is a UMP test with size $\alpha = P\left\{\tilde{t} > t_0; \theta_0\right\} = \mathrm{E}_{\theta_0}\left(\widehat{\phi}\left(\tilde{t}\right)\right)$.

Proof. Fix a parameter value $\theta' \in \Theta$ such that $\theta' > \theta_0$ and consider testing $H_0' : \theta = \theta_0$ versus $H_1' : \theta = \theta'$ by using the test $\widehat{\phi}$. The MLR property implies that

$$t > t_0 \text{ if and only if } \frac{g(t;\theta')}{g(t;\theta_0)} > k',$$

with

$$k' = \inf_{t \in \mathcal{T}} \frac{g(t;\theta')}{g(t;\theta_0)},$$

where $\mathcal{T}$ is the set of real numbers t such that $t > t_0$ and $g(t;\theta')/g(t;\theta_0)$ is well defined in $\overline{\mathbb{R}}$ (i.e., $g(t;\theta') > 0$ or $g(t;\theta_0) > 0$ or both).

Since $\tilde{t}$ is a sufficient estimator for θ then, according to the factorization theorem and its proof in Chapter 8, the pdf/pmf of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ can be written as

$$f\left(X;\theta\right) = h\left(X\right) \cdot g(t;\theta),$$

where the function $g(t;\theta)$ is the pdf/pmf of the estimator $\tilde{t}$. Thus,

$$\frac{g(t;\theta')}{g(t;\theta_0)} = \frac{h\left(X\right) \cdot g(t;\theta')}{h\left(X\right) \cdot g(t;\theta_0)} = \frac{f\left(X;\theta'\right)}{f\left(X;\theta_0\right)}.$$

Therefore, under the test $\widehat{\phi}$ we reject H_0' if and only if

$$\frac{f\left(X;\theta'\right)}{f\left(X;\theta_0\right)} > k' \iff \frac{f\left(X;\theta_0\right)}{f\left(X;\theta'\right)} < \frac{1}{k'} \equiv k.$$

Thus, according to the Neyman-Pearson lemma, the test $\phi(X) = \widehat{\phi}\left(t\left(\tilde{X}\right)\right)$ is a most powerful test with significance level no larger than $\alpha = P\left\{\tilde{t} > t_0; \theta_0\right\}$ for testing $H_0' : \theta = \theta_0$ versus $H_1' : \theta = \theta'$. Since θ' was arbitrary, the test based

on the test function $\widehat{\phi}\left(\tilde{t}\right)$ is simultaneously UMP with significance level α for every $\theta' > \theta_0$. Therefore, the test $\widehat{\phi}$ is UMP with significance level α for testing $H_0' : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Let

$$\pi_{\widehat{\phi}}(\theta) = P\left\{\tilde{t} > t_0; \theta\right\} = \mathrm{E}_\theta\left(\widehat{\phi}\left(\tilde{t}\right)\right)$$

be the power function of the test $\widehat{\phi}$ for testing the null hypothesis $H_0 : \theta \leq \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$. From the MLR assumption and equation (2) in Proposition 1, the power function $\pi_{\widehat{\phi}}(\theta)$ is non-decreasing since, for any $\theta_2 > \theta_1$,

$$\pi_{\widehat{\phi}}(\theta_2) = P\left\{\tilde{t} > t_0; \theta_2\right\} \geq P\left\{\tilde{t} > t_0; \theta_1\right\} = \pi_{\widehat{\phi}}(\theta_1).$$

Therefore,

$$\sup_{\theta \leq \theta_0} \pi_{\widehat{\phi}}(\theta) = \pi_{\widehat{\phi}}(\theta_0) = \alpha,$$

which means that the significance level of the test $\widehat{\phi}$ when the null hypothesis is $H_0 : \theta \leq \theta_0$ is the same as when the null hypothesis is $H_0' : \theta = \theta_0$.

Hence, the test based on the test function $\widehat{\phi}\left(\tilde{t}\right)$ it is a UMP test with significance level $\alpha = P\left\{\tilde{t} > t_0; \theta_0\right\}$ for testing the composite null hypothesis $H_0 : \theta \leq \theta_0$ against the composite alternative hypothesis $H_1 : \theta > \theta_0$. *Q.E.D.*

Hypothesis Testing

The test procedure for simple null hypotheses consists of the following four steps:

(1) State the null hypothesis H_0 and an appropriate alternative hypothesis H_1.

(2) Using the sampling distribution of an appropriate test statistic, determine a critical region of size α, where α is specified.

(3) Compute the value of the test statistic from sample data.

(4) Decide whether to reject the null hypothesis or whether to accept it.

Tests concerning means $(H_0 : \mu = \mu_0)$

Use the value of the test statistic

$$\frac{\overline{\boldsymbol{x}} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1),$$

where $\overline{\boldsymbol{x}}$ is the mean of the random sample from a normal population.

If $n \geq 30$, we can replace σ by s. Moreover, for large n we can use the central limit theorem, even if the population is not normal.

If $n < 30$, σ^2 is unknown, and the random sample comes from a normal population, use the statistic

$$\frac{\overline{\boldsymbol{x}} - \mu_0}{\boldsymbol{s}/\sqrt{n}} \sim \mathrm{t}_{n-1}.$$

Notes (arising from the Karlin-Rubin theorem): (a) The previous procedure can be used to test the simple null hypothesis $\mu = \mu_0$ against the alternative hypothesis $\mu \neq \mu_0$ (composite alternative), $\mu = \mu_1 > \mu_0$ (simple alternative), $\mu = \mu_1 < \mu_0$ (simple alternative), $\mu > \mu_0$ (composite alternative), or $\mu < \mu_0$ (composite alternative).

(b) The same procedure can also be used to test the composite null hypothesis $\mu \geq \mu_0$ (or $\mu \leq \mu_0$) against the composite alternative hypothesis $\mu < \mu_0$ (or $\mu > \mu_0$). In this case α is the the maximum probability of committing a type I error for any value of μ assumed under the null hypothesis.

(c) The same logic applies to the next tests concerning differences between means, variances, proportions, and differences between proportions.

Tests concerning differences between means $(H_0 : \mu_1 - \mu_2 = \delta)$

Use the test statistic

$$\frac{\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \delta}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \overset{a}{\sim} \mathrm{N}(0,1)$$

where $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ are the means of two independent random samples.

If $n_1 \geq 30$ and $n_2 \geq 30$, we can replace σ_1^2 and σ_2^2 by s_1^2 and s_2^2, respectively.

If $n_1 < 30$ or $n_2 < 30$, σ_1^2 and σ_2^2 are equal but unknown, and the two random samples are independent and come from a normal population, use the test statistic

$$\frac{\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 - \delta}{\boldsymbol{s}_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim \mathrm{t}_{n_1+n_2-2}, \quad \text{where} \quad \boldsymbol{s}_p^{\boldsymbol{2}} = \frac{(n_1-1)\boldsymbol{s}_1^{\boldsymbol{2}} + (n_2-1)\boldsymbol{s}_2^{\boldsymbol{2}}}{n_1+n_2-2}.$$

Tests concerning variances

Assume that you have random samples from normal populations.

a) $H_0 : \sigma^2 = \sigma_0$. Use the test statistic

$$\frac{(n-1)\boldsymbol{s}^{\boldsymbol{2}}}{\sigma_0^2} \sim \chi^2_{n-1}.$$

b) $H_0 : \sigma_1^2 = \sigma_2^2$. Use the test statistic

$$\frac{\boldsymbol{s}_1^{\boldsymbol{2}}}{\boldsymbol{s}_2^{\boldsymbol{2}}} \sim \mathrm{F}_{n_1-1,n_2-1}.$$

Tests concerning proportions $(H_0 : \theta = \theta_0)$

Let $\tilde{x} \sim \mathrm{B}(n;\theta)$. Then, we use the test statistic

$$\frac{\tilde{x} - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \overset{a}{\sim} \mathrm{N}(0,1).$$

Tests concerning differences among several proportions

Suppose that x_1, $x_2, \ldots,$ x_k are observed values of a set of random variables $\tilde{x}_1$, $\tilde{x}_2, \ldots,$ $\tilde{x}_k$ having binomial distributions with the respective parameters n_1 and θ_1, n_2 and θ_2, …, n_k and θ_k. If the n's are sufficiently large we use the test statistic

$$\frac{\tilde{x}_i - n_i\theta_i}{\sqrt{n_i\theta_i(1-\theta_i)}} \overset{a}{\sim} \mathrm{N}(0,1) \quad \text{for } i = 1,\, 2,\, \ldots,\, k.$$

Therefore,

$$\sum_{i=1}^{k} \frac{(\tilde{x}_i - n_i\theta_i)^2}{n_i\theta_i(1-\theta_i)} \overset{a}{\sim} \chi^2_k.$$

To test the null hypothesis $H_0 : \theta_1 = \theta_2 = \ldots = \theta_k = \theta_0$, against the alternative that at least one of the θ's does not equal θ_0, we can use the test statistic

$$\sum_{i=1}^{k} \frac{(\tilde{x}_i - n_i\theta_0)^2}{n_i\theta_0(1-\theta_0)} \overset{a}{\sim} \chi^2_k.$$

When θ_0 is not specified, that is, when we are interested only in the null hypothesis $\theta_1 = \theta_2 = \ldots = \theta_k$, we use the pooled estimator

$$\boldsymbol{\theta} = \frac{\tilde{x}_1 + \tilde{x}_2 + \ldots + \tilde{x}_k}{n_1 + n_2 + \ldots + n_k}.$$

Then

$$\sum_{i=1}^{k} \frac{(\tilde{x}_i - n_i\boldsymbol{\theta})^2}{n_i\boldsymbol{\theta}(1-\boldsymbol{\theta})} \overset{a}{\sim} \chi^2_{k-1}. \tag{1}$$

The loss of one degree of freedom is due to the fact that we use an estimate for the unknown parameter θ (see the discussion in the next section).

An alternative form of the chi-square statistic for this kind of test is the following:

	successes	failures
Sample 1	x_1	$n_1 - x_1$
Sample 2	x_2	$n_2 - x_2$
...	...	...
Sample k	x_k	$n_k - x_k$

Let us refer to its entries as the observed cell frequencies, n_{ij}, $i = 1, 2, \ldots, k$ and $j = 1, 2$. Under the null hypothesis $\theta_1 = \theta_2 = \ldots = \theta_k = \theta_0$ the expected cell frequencies for the first column are $e_{i1} = n_i\theta_0$ for $i = 1, 2, \ldots, k$ and those for the second column are $e_{i2} = n_i(1 - \theta_0)$. When θ_0 is unknown, we substitute for it, as before, the value θ of the pooled estimator $\boldsymbol{\theta}$ and estimate the cell frequencies as

$$e_{i1} = n_i\theta \quad \text{and} \quad e_{i2} = n_i(1-\theta), \quad \text{for } i = 1, 2, \ldots, k.$$

The value of the test statistic (1) can be written as

$$\sum_{i=1}^{k}\sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}. \quad \leftarrow \text{Check it!!} \tag{2}$$

Contingency tables (Generalization of the previous section)

A $r \times c$ contingency table has c columns representing the different categories $A_1, A_2, \ldots, A_c$ of one variable, r rows representing the different categories $B_1, B_2, \ldots, A_r$ of another variable, and n_{ij} is the cell frequency for $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$. In the previous section we had $r = k$ and $c = 2$.

Schematically,

	A_1	A_2	...	A_c	
B_1	n_{11}	n_{12}	...	n_{1c}	$n_{1\cdot}$
B_2	n_{21}	n_{22}	...	n_{2c}	$n_{2\cdot}$
...	...	...	...	...	...
B_r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot c}$	n

,

where the row and the column totals $n_{i\cdot} = \sum_{j=1}^{c} n_{ij}$ and $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$ are called the marginal frequencies and n, the sum of all cell frequencies, is called the grand total.

The null hypothesis that we want to test is that the two variables are independent. If θ_{ij} is the probability that an item will fall into the cell (i, j), $\theta_{i\cdot}$

is the probability that an item will fall into the i^{th} row and $\theta_{\cdot j}$ is the probability that an item will fall into the j^{th} column. The null hypothesis we want to test is that

$$\theta_{ij} = (\theta_{i\cdot})(\theta_{\cdot j}) \quad \text{for } i = 1, 2, \ldots, r \text{ and } j = 1,\ 2, \ldots, c.$$

To test the null hypothesis against the alternative, that $\theta_{ij} \neq (\theta_{i\cdot})(\theta_{\cdot j})$ for at least a pair of values of i and j, we estimate the probabilities $\theta_{i\cdot}$ and $\theta_{\cdot j}$ as

$$\hat{\theta}_{i\cdot} = \frac{n_{i\cdot}}{n} \quad \text{and} \quad \hat{\theta}_{\cdot j} = \frac{n_{\cdot j}}{n},$$

and, hence, the expected cell frequencies as

$$e_{ij} = \left(\hat{\theta}_{i\cdot}\right)\left(\hat{\theta}_{\cdot j}\right) n = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} \cdot n = \frac{(n_{i\cdot})(n_{\cdot j})}{n}.$$

Then, we base our decision on

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

and reject the null hypothesis if the value we obtain exceeds $\chi^2_{\alpha,(r-1)(c-1)}$. The degrees of freedom come from the fact that, if we know the frequencies of the $(r-1)(c-1)$ cells corresponding to the first $(r-1)$ rows and to the first $(c-1)$ columns, then we know the frequencies of the remaining cells since the total frequencies of rows and columns are fixed (think about it!).

Notes: (a) Since the previous test statistic has only approximately a chi-square distribution, we should use this test only when none of the e_{ij}'s is less than five. This sometimes requires that we combine some of the cells with a corresponding loss of degrees of freedom.

(b) *A general method to find the number of degrees of freedom:* Whenever expected cell frequencies in chi-square formulas are estimated on the basis of sample data, the number of degrees of freedom is $s - t - 1$, where s is the number of terms in the sum and t is the number of independent parameters replaced by estimates. The -1 in the previous formula is due to the additional restriction that the sum of all the cell frequencies must be equal to the fixed grand total n.

(c) When testing for differences among k proportions (previous section), we had $s = 2k$ and $t = k$, since we had to estimate the k parameters θ_1, $\theta_2, \ldots,$ θ_k and the number of degrees of freedom was $2k - k - 1 = k - 1$.

(d) When testing for independence with an $r \times c$ contingency table, we have $s = r \cdot c$ and $t = r + c - 2$ since the $r + c$ parameters $\theta_{i\cdot}$ and $\theta_{\cdot j}$ are subject to the two restrictions that their respective sum must equal one. Hence, $s - t - 1 = rc - (r + c - 2) - 1 = (r - 1)(c - 1)$.

Goodness of fit (for discrete random variables)

This test applies to the situation in which we want to determine whenever a set of data may be looked as a random sample from a population having a given distribution. Let n_i be the observed absolute frequencies and e_i the expected

absolute frequencies. To test the null hypothesis that a set of observed data comes from a population having a specified distribution against the alternative, that the population has some other distribution, we compute the value of the Pearson's statistic,

$$\sum_{i=1}^{m} \frac{(n_i - e_i)^2}{e_i},$$

and reject H_0 at the level of significance α, if the computed value exceeds $\chi^2_{\alpha,m-t-1}$, where m is the number of terms in the summation and t is the number of independent parameters estimated on the basis of sample data (see previous discussion).

Notes: (a) The set of values of the random variable is partitioned in adjacent classes so that each expected absolute frequency is at least five.

(b) The statistic considered in this section has an asymptotic distribution that is chi-square regardless of the distribution being tested.

Goodness of fit (for all types of random variables)

Consider a random sample of size n and let us order the values of the random sample as $x_1 < x_2 < \ldots < x_n$. The "sample distribution function" is constructed as follows:

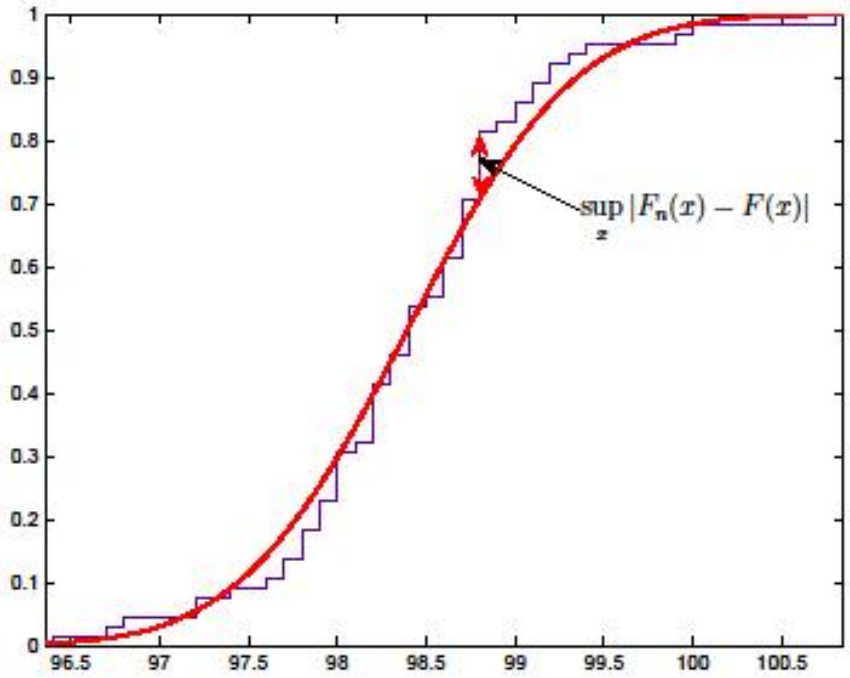
$$F_n(x) = \frac{j}{n} \quad \text{for} \quad x_j \leq x < x_{j+1}, \; j = 0, 1, \ldots, n,$$

with $x_0 = -\infty$ and $x_{n+1} = \infty$.

We compute the value of the Kolmogorov-Smirnov test statistic as follows:

$$D_n = \sup_x |F_n(x) - F(x)|,$$

where $F(x)$ is the distribution function of the population according to the null hypothesis H_0. If H_0 is true, D_n should be small.



Sample and true distribution functions.

The distribution of D_n is independent of F, but not of n. The following table gives the acceptance limits for such a test:

SAMPLE SIZE (n)	LEVEL OF SIGNIFICANCE $\propto$ FOR $D_n = \sup_x \lvert F_n(x) - F(x) \rvert$				
	.20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.210	.220	.240	.270	.320
30	.190	.200	.220	.240	.290
35	.180	.190	.210	.230	.270
OVER 35	$\frac{1.07}{\sqrt{N}}$	$\frac{1.14}{\sqrt{N}}$	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

The most serious limitation of this test is that the distribution must be fully specified. That is, if the parameters of $F(x)$ are estimated from the data, the critical region of the Kolmogorov-Smirnov test is no longer valid. It typically must be determined by simulation.

Comparison between two distributions

Let $F_{n_1}(x)$ be the sample distribution function of a random sample of size n_1 and $G_{n_2}(x)$ be the sample distribution function of a random sample of size n_2

(see previous discussion). The two random samples are independent.

Consider the following value of the two sample Kolmogorov-Smirnov test statistic:

$$D_{n_1,n_2} = \sup_x |F_{n_1}(x) - G_{n_2}(x)|.$$

The null hypothesis is that the two random samples come from identical populations. If H_0 is true, D should be small. The following table gives the acceptance limits for such a test:

Table gives critical D-values for $\alpha = 0.05$ (upper value) and $\alpha = 0.01$ (lower value) for various sample sizes. * means you cannot reject H_0 regardless of observed D.

$n_2 \backslash n_1$	3	4	5	6	7	8	9	10	11	12
1	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	16/16	18/18	20/20	22/22	24/24
	*	*	*	*	*	*	*	*	*	*
3	*	*	15/15	18/18	21/21	21/24	24/27	27/30	30/33	30/36
	*	*	*	*	*	24/24	27/27	30/30	33/33	36/36
4		16/16	20/20	20/24	24/28	28/32	28/36	30/40	33/44	36/48
		*	*	24/24	28/28	32/32	32/36	36/40	40/44	44/48
5			*	24/30	30/35	30/40	35/45	40/50	39/55	43/60
			*	30/30	35/35	35/40	40/45	45/50	45/55	50/60
6				30/36	30/42	34/48	39/54	40/60	43/66	48/72
				36/36	36/42	40/48	45/54	48/60	54/66	60/72
7					42/49	40/56	42/63	46/70	48/77	53/84
					42/49	48/56	49/63	53/70	59/77	60/84
8						48/64	46/72	48/80	53/88	60/96
						56/64	55/72	60/80	64/88	68/96
9							54/81	53/90	59/99	63/108
							63/81	70/90	70/99	75/108
10								70/100	60/110	66/120
								80/100	77/110	80/120
11									77/121	72/132
									88/121	86/132
12										96/144
										84/144

For larger sample sizes, the approximate critical value D_α is given by the equation

$$D_\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

where the coefficient is given by the table below.

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

Examples: (1) At $\alpha = 0.05$ and samples sizes 5 and 8, $D_\alpha = 30/40 = 0.75$.

(2) At $\alpha = 0.01$ and samples sizes 15 and 28, $D_\alpha = 1.63\sqrt{\frac{15+28}{15 \cdot 28}} = 0.522$.

Exercises. Probability and Statistics. IDEA.
9. Hypothesis Testing

1. Suppose we want to test the null hypothesis that the mean μ of a normal population with the variance $\sigma^2 = 1$ is μ_0 against the alternative hypothesis that it is μ_1, where $\mu_1 > \mu_0$. Let $\bar{x}$ be the value of the mean of a random sample of size n. Find the value of k (as a function of n) such that $\bar{x} > k$ provides a critical region of size $\alpha = 0.05$. Also determine the minimum sample size n needed for testing the null hypothesis $\mu = 10$ against the alternative hypothesis $\mu = 11$ by this procedure so that the probability β of committing a type II error is smaller than 0.05, $\beta \leq 0.05$.

2. A random sample of size n from a normal population with known variance σ^2 is to be used to test the null hypothesis $\mu = \mu_0$ against the alternative hypothesis $\mu = \mu_1$, where $\mu_1 > \mu_0$. Use the Neyman-Pearson lemma to find the uniformly most powerful test of size α.

3. Consider a coin having the probability θ of landing on heads and the following composite hypotheses:

$$H_0 \ : \quad 0 \leq \theta \leq 0.5 \qquad \text{(null hypothesis)}$$

$$H_1 \ : \quad 0.5 < \theta \leq 1 \qquad \text{(alternative hypothesis)}$$

   (a) Find the power functions of the following two tests:

   - *Test 1*: observe three independent flips of the coin and reject H_0 if three heads come up;

   - *Test 2*: observe three independent flips of the coin and reject H_0 if either two or three heads come up.

   (b) Discuss the properties of the previous two tests in terms of the associated probabilities of committing type I and type II errors.

4. Let $\tilde{x}_i$ have a Bernoulli distribution with unknown parameter θ, and let

$$\tilde{y} = \tilde{x}_1 + \tilde{x}_2 + \tilde{x}_3 + \tilde{x}_4 + \tilde{x}_5$$

   denote the number of successes in five independent trials. To test $\theta = 0.5$ against $\theta \neq 0.5$ we make a test with critical region $\tilde{y} = 0, 1, 4$ or 5. Plot the power function and find the significance level of the test.

5. (a) Find the critical region of the likelihood ratio test with a level of significance α for testing the null hypothesis

$$H_0 : \ \mu = \mu_0$$

   against the composite alternative

$$H_1 : \ \mu \neq \mu_0$$

on the basis of a random sample of size n from a normal population with the known variance σ^2.

(b) Do the same with both the Wald test and the score test.

6. Suppose it is known from experience that the standard deviation of the weight of 8-ounce packages of cookies made by a certain bakery is 0.16 ounce. Let us assume that the weight of those packages of cookies is normally distributed. To check whether its production is under control on a given day, namely, to check whether the true average weight of the packages is 8 ounces, they select a random sample of 25 packages and find that their mean weight is $\bar{x} = 8.112$ ounces. Since the bakery stands to lose money when $\mu > 8$ and the customer loses out when $\mu < 8$, test the null hypothesis $\mu = 8$ against the alternative $\mu \neq 8$ using $\alpha = 0.01$. Find the corresponding p-value.

7. Suppose that 100 tires of a certain brand lasted on the average 21431 miles with standard deviation of 1295 miles. Using $\alpha = 0.05$, test the null hypothesis $\mu = 22000$ miles against the alternative hypothesis $\mu < 22000$ miles. Find the corresponding p-value.

8. Suppose that the specifications for a certain kind of ribbon call for a mean breaking strength of 185 pounds, and that five pieces randomly selected from different rolls have a mean breaking strength of 183.1 pounds with a sample standard deviation of 8.2 pounds. Assuming that we can look upon the data as a random sample from a normal population, test the null hypothesis $\mu = 185$ against the alternative $\mu < 185$ at $\alpha = 0.05$. Find the corresponding p-value.

9. Suppose that the nicotine content of two brands of cigarettes are being measured. If in an experiment fifty cigarettes of the first brand had an average nicotine content of $\bar{x}_1 = 2.61$ milligrams with a sample standard deviation of $s_1 = 0.12$ milligram, while forty cigarettes of the second brand had an average nicotine content of $\bar{x}_2 = 2.38$ milligrams with a sample standard deviation of $s_2 = 0.14$ milligram, test the null hypothesis $\mu_1 - \mu_2 = 0.2$ against the alternative $\mu_1 - \mu_2 \neq 0.2$, using $\alpha = 0.05$. Find the corresponding p-value.

10. In the comparison of two kind of paint, a consumer testing service finds that four one-gallon cans of one brand cover on average 512 square feet with a standard deviation of 31 square feet, while four one-gallon cans of another brand cover on the average 492 square feet with a sample standard deviation of 26 square feet. Test the null hypothesis $\mu_1 - \mu_2 = 0$ against the alternative $\mu_1 - \mu_2 \neq 0$ at the significance level $\alpha = 0.05$. Assume that the two populations are normal and have equal variances. Find the corresponding p-value.

11. Suppose that the thickness of a part used in a semiconductor is its critical dimension and that measurements of the thickness of a random sample of 18 such parts have the variance $s^2 = 0.68$, where the measurements are in the thousandths of an inch. The process is considered to be under control if the variation of the thickness is given by a variance not greater that 0.36. Assuming

that the measurements constitute a random sample from a normal population, test the null hypothesis $\sigma^2 = 0.36$ against the alternative hypothesis $\sigma^2 > 0.36$ at $\alpha = 0.05$. Find the corresponding p-value.

12. In comparing the variability of the tensile strength of two kind of structural steel, an experiment yielded the following results: $n_1 = 13$, $s_1^2 = 19.2$, $n_2 = 16$ and $s_2^2 = 3.5$, where the units of measurement are 1000 pounds per square inch. Assuming that the measurements constitute independent random samples from two normal populations, test the null hypothesis $\sigma_1^2 = \sigma_2^2$ against the alternative hypothesis $\sigma_1^2 \neq \sigma_2^2$ at the $\alpha = 0.02$ level of significance. Find the corresponding p-value.

13. An oil company claims that at most 20 percent of all automobile market owners buy brand A gasoline. Test this claim at $\alpha = 0.01$, if a random check indicates that 58 of 200 automobile owners buy brand A gasoline. Find the corresponding p-value.

14. Determine, on the basis of the sample data shown in the following table, whether the true proportion of shoppers favoring detergent A over detergent B is the same in all three cities:

	Number favoring A	Number favoring B	
Los Angeles	232	168	$n_1 = 400$
San Diego	260	240	$n_2 = 500$
Fresno	197	203	$n_3 = 400$

Use the significance level $\alpha = 0.05$. Find the corresponding p-value.

15. For the data shown in the table below, test for independence between a person's ability in mathematics and his or her interest in statistics. Use the 0.01 significance level. Moreover, find the corresponding p-value.

		Ability in Math		
		Low	Average	High
Interest in Statistics	Low	63	42	15
	Average	58	61	31
	High	14	47	29

16. For the data in the table below, test whether the number of errors the compositor makes in setting a galley of type is a random variable having a Poisson distribution. Use the significance level $\alpha = 0.05$. Moreover, find the corresponding p-value.

Number of errors per galley	Observed frequencies n_i
0	18
1	53
2	103
3	107
4	82
5	46
6	18
7	10
8	2
9	1
	440

17. Test at the 5% significance level the hypothesis that a distribution is normal with mean 32 and variance 3.24, using the following observed values of a random sample of size 10: 31.0, 31.4, 33.3, 33.4, 33.5, 33.7, 34.4, 34.9, 36.2, and 37.0.

18. The random variable $\tilde{x}$ has a lognormal density function,

$$f(x;\mu,\sigma) = \frac{1}{\left(\sqrt{2\pi}\right)\sigma \cdot x} e^{-\frac{1}{2}\left(\frac{(\ln x)-\mu}{\sigma}\right)^2}, \text{ for } 0 < x < \infty.$$

Note that the random variable $\ln \tilde{x}$ is normal with the mean μ and the variance σ^2.

Consider the null and alternative hypotheses:

$$H_0: \quad \mu = 80;\ \sigma = 20$$

$$H_1: \quad \mu = 110;\ \sigma = 20$$

Given a random sample of size n from a population having the density $f(x;\mu,\sigma)$, use the Neyman-Pearson Lemma to devise a statistical test to discriminate between H_0 and H_1 with the level of significance $\alpha = 0.05$. That is, apply the Lemma to find the appropriate test statistic and the precise critical region with size $\alpha = 0.05$ yielding the most powerful test. The critical region you must construct should be a function of the sample size n only.

19. On the basis of **a single** observation, we want to test the simple null hypothesis that the probability function of $\tilde{x}$ is

x	1	2	3	4	5	6	7
$f(x)$	1/12	1/12	1/12	1/4	1/6	1/6	1/6

against the composite alternative that the probability function is

x	1	2	3	4	5	6	7
$g(x)$	$a/3$	$b/3$	$c/3$	$2/3$	0	0	0

where $a + b + c = 1$.

(a) Using the likelihood ratio technique, find the critical region with size $\alpha = 0.25$ for which the null hypothesis is rejected. Which is the corresponding probability of a type II error?

(b) Consider the critical region for which the null hypothesis is rejected only when $x = 4$. Which is the size of such a critical region? Which is the corresponding probability of a type II error?

(c) Compare the test obtained in (a) with the one in (b). Which one do you prefer? Does the comparison contradict the Neyman-Pearson lemma?

20. R.D. Clarke divided the south of London into 576 squares of 0.1 square miles each and counted the number of flying bombs hitting those squares during the WWII. He obtained the following results:

Number of bombs per square	0	1	2	3	4	5	6 or more
Number of squares (frequency)	229	211	93	35	7	1	0

Test the null hypothesis that the bombs were hitting randomly those squares against the alternative that the bombs were directed towards a specific target. Use a significance level of 0.01. Find the corresponding p-value.

21. Let $\tilde{x} : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B})$ be a random variable whose range is $\tilde{x}(\Omega) = \{0, 1, 2\}$. We observe the value $\{x_1, x_2\}$ of a random sample $\{\tilde{x}_1, \tilde{x}_2\}$ of size 2 from the population $\tilde{x}$.

We want to test the null hypothesis that the probability function of the random variable $\tilde{x}$ is given by the following uniform probability function:

$$f(x) = \frac{1}{3}\,, \quad \text{for } x = 0, 1, 2; \tag{$*$}$$

against the alternative that the probability function of $\tilde{x}$ is

$$g(x) = \begin{cases} 0 & \text{for } x = 0 \\ \dfrac{2}{3} & \text{for } x = 1 \\ \dfrac{1}{3} & \text{for } x = 2 \end{cases}$$

(a) Devise the most powerful test with a level of significance $\alpha = 1/9$. What is the power of this test?

(b) Consider a test where the sample mean $\bar{\boldsymbol{x}} = \dfrac{\tilde{x}_1 + \tilde{x}_2}{2}$ is the test statistic. Find the value of k such that $\bar{x} \geq k$ provides a critical region of size $\alpha = 1/9$, where $\bar{x}$ is the value of the sample mean $\bar{\boldsymbol{x}}$. What is the power of this test?

Suppose now that we want to test the null hypothesis that the probability function of the discrete random variable $\tilde{x}$ is the uniform probability function given in $(*)$ against the alternative COMPOSITE hypothesis that the probability function of $\tilde{x}$ IS NOT the one given in $(*)$.

(c) Using again the value $\{x_1, x_2\}$ of a random sample $\{\tilde{x}_1, \tilde{x}_2\}$ of size 2 from the population $\tilde{x}$, devise a likelihood ratio test with a level of significance $\alpha = 1/3$. You should provide the test statistic and the critical region for this test.

22. Prove that the two test statistics given in (1) and (2) in the handout "Hypothesis Testing" are equivalent.

23. Assume that you have obtained x "successes" in three identical and independent trials. Let θ be the probability of success in each trial. Devise a likelihood ratio test with a level of significance $\alpha = 1/4$ to test the null hypothesis that $\theta = 1/2$ against the alternative hypothesis that $\theta \neq 1/2$. You should provide the test statistic and the critical region for this test.

    Hint: Note that $\lim_{x \to 0} x^x = 1$ or, equivalently, $\lim_{x \to 0} (x \cdot \ln x) = 0$, which follows from L'Hôpital's rule.

24. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample of size n from a normal population with unknown mean μ and unknown variance σ^2. Consider the following statistic:

$$\hat{\boldsymbol{s}}^2 = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \overline{\boldsymbol{x}})^2}{n},$$

    where $\overline{\boldsymbol{x}}$ is the sample mean.

    Assume that the sample size is $n = 20$. Use a level of significance $\alpha = 0.05$ to construct an acceptance interval (a, b) for the statistic $\hat{\boldsymbol{s}}^2$ to test the null hypothesis $H_0 : \sigma^2 = 5$ against the alternative $H_1 : \sigma^2 \neq 5$. The interval should be such that you do not reject H_0 if $a < \hat{s}^2 < b$ and you reject it if $\hat{s}^2 \notin (a, b)$. Moreover, $P\left\{\hat{\boldsymbol{s}}^2 \leq a\right\} = P\left\{\hat{\boldsymbol{s}}^2 \geq b\right\} = \alpha/2$ under the null hypothesis.

25. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample of size n from a normal population with unknown mean μ and known variance σ^2. Consider testing $H_0 : \mu \geq \mu_0$ against $H_1 : \mu < \mu_0$. Use the Karlin-Rubin theorem to construct a uniformly most powerful (UMP) test with significance level α based on the test statistic $\bar{\boldsymbol{x}}$ (the sample mean).

26. You know that $\tilde{x}$ is a normally distributed random variable with a standard deviation of 200 and that its mean is either 700 or 800. You wish to make a judgement about the value of the mean of $\tilde{x}$. You may choose among the following two pieces of information (or statistics) for your judgement:

(a) The mean $\overline{\boldsymbol{x}}$ of a random sample of size 25 from the population of $\tilde{x}$.

(b) The proportion $\hat{\boldsymbol{\theta}}$ of observations with a value smaller than 800 in a random sample of size 49 from the population of $\tilde{x}$.

We impose that the probability of the mistake of saying $\mu = 700$ if in fact $\mu = 800$ must be the same as the probability of the mistake of saying $\mu = 800$ if in fact $\mu = 700$.

Which of the previous two statistics do you prefer for judging μ?

Note: For the statistic in (b) use the corresponding approximation by the normal distribution (as dictated by the central limit theorem) since such an approximation is excellent in this case.

27. Let $\{\tilde{x}_i\}_{i=1}^{n}$ be a random sample of size $n = 70$ from a population $\tilde{x}$. Assume that the population distribution is Poisson with the parameter $\lambda > 0$. Consider the null hypothesis $H_0 : \lambda = 5$ and the alternative hypothesis $H_1 : \lambda = 6$.

(a) Use the Neyman-Pearson Lemma to test H_0 against H_1 with the level of significance $\alpha = 0.05$. That is, apply the Lemma to find the appropriate test statistic and the precise critical region with size $\alpha = 0.05$ yielding the most powerful test. *Hint:* Use the Central Limit Theorem to find a good approximation for the standardization of the sample mean.

(b) What is the power of the test you have devised in part (a)?

28. Check that the two formulae to compute the expected utility shown in the handout "First-Order Stochastic Dominance",

$$\mathrm{E}\left[u(\tilde{x})\right] = \int_{[a,b]} u\left(x\right) dF_{\tilde{x}}(x) = u(b) - \int_{[a,b]} F_{\tilde{x}}(x) du(x),$$

yield exactly the same value for the following example. The random variable $\tilde{x}$, which represents consumption, is discrete and its support is a subset of the closed interval $[0, 1]$. In particular, $\tilde{x}$ only takes on three values: 0 with probability 1/2, 1/2 with probability 1/4, and 1 with probability 1/4. The Bernoulli utility function is $u(x) = -x^2 + 2x + 1$, which is continuous and increasing for $x \leq 1$.

29. Consider a normal population with the known variance $\sigma^2 = 3$ and a random sample $\{\tilde{x}_1, ..., \tilde{x}_n\}$ with size $n = 12$.

(a) We want to test the null hypothesis that the mean μ of this population is equal to 7 ($H_0 : \mu = 7$) against the alternative simple hypothesis that it is 9 ($H_1 : \mu = 9$). Let $\bar{x}$ be the value of the mean of the random sample. Find the value of k such that $\bar{x} > k$ provides a critical region of size $\alpha = 0.05$. Compute the power of this test.

(b) Construct a likelihood ratio test with a level of significance $\alpha = 0.05$ for testing the null hypothesis $H_0 : \mu = 7$ against the composite alternative

$H_1 : \mu \neq 7$. You should provide the appropriate test statistic and the precise critical region. *Hint:* for this part, you should prove that

$$\sum_{i=1}^{12} \left[(x_i - 7)^2 - (x_i - \bar{x})^2 \right] = 12(\bar{x} - 7)^2.$$

(c) Plot the power function of the test found in part (b).