*Curso de Formación de Personal Investigador Usuario de Animales para Experimentación*

# STATISTICAL INFERENCE

**Jesús Piedrafita Arilla**
jesus.piedrafita@uab.cat
*Departament de Ciència Animal i dels Aliments*

UAB

# Learning objectives

- State the concept of Statistical Inference.
- Distinguish between parameters and statistics.
- State the difference between point and interval estimation.
- Define the standard normal distribution and its applications.
- Define the concept of standard error, in particular of a mean.
- Define confidence intervals, both when the variances are known or unknown (estimated), as well as the *t*-distribution.
- Establish the four steps of a Test of Hypothesis:
    1. Define the Null hypothesis ($H_0$) and the Alternative hypothesis ($H_1$)
    2. Find an appropriate Test statistic
    3. Calculate the *p*-value
    4. Establish a decision rule
- Distinguish Type I and Type II errors, Significance level and Power of the test.
- Make some calculations with R Commander and R.

# Statistical inference

## Drawing conclusions based on data taking into account the inherent random variation.

1. Second step after the description of the data.
2. **We want to extrapolate to the population which we observe in a sample**.
3. Need to assume a particular **data distribution**:
   - **Normal**: adult weight, loin muscle area, average daily gain, …
   - **T**: distribution of some statistics.
   - Bernoulli: ill *vs.* not ill.
   - Poisson: number of microorganisms in a microscope field.
4. **In inferential statistics we estimate** –obtain an approximate value of- **the true value of the parameter** (a mean for example) **through an adequate statistic** (sample mean, for example)**.**
5. There are a many contexts in which inference is desirable, and there are many approaches to performing inference.
6. Some methods do not need to assume a distribution: **non parametric methods**.

# Parameters and statistics

Usually the **parameters** of the distribution are designed with **Greek** characters, whereas the corresponding **statistics** are designed with **Latin** characters. The next table includes some examples:

| | **Parameter (population)** | **Statistic (sample)** |
|---|---|---|
| **Mean** | $\mu$ | $\bar{y}$ |
| **Variance** | $\sigma^2$ | $s^2$ |
| **Standard deviation** | $\sigma$ | $s$ |
| **Proportion** | $\pi$ | $p$ |

**Estimator:** Some equation that allows us to estimate some parameter.

**Estimate:** The value obtained.

# Estimation of parameters

1. **Point estimation**: a value is obtained as an estimate of the parameter.
2. **Interval estimation**: we calculate an interval in which we affirm that with a certain probability we can find the true value of the parameters.

So far we have presented some point estimators of several parameters.

In practice, when we work with the unknown parameter of the population, in addition to this point estimate we are usually interested in an interval (**confidence interval, CI**) that gives an idea of the **uncertainty** of the estimate.

We will present the way to construct **intervals** through some classical examples. The procedure is **based upon the distribution of the statistic**.

# One application of the standard normal

$z_i = \dfrac{y_i - \bar{y}}{s}$ , where $z_i$ follows a standard normal distribution.

A male has *BODYwt* of 3.3 kg in our cat dataset. Assuming that the distribution is normal, we can compute the probability of having a value lower or equal than this one (CDF) using R syntaxis:

```
> z <- (3.3-2.9)/0.4675; z
[1] 0.855615
> pnorm(z)
[1] 0.8038946
```

The complement of `pnorm(z)`, i.e. `1-pnorm(z)`, will be the probability of having a value greater than 3.3, in this case ≅ 0.2.

An easier way to compute this probability and the quantile is:

```
> pnorm(3.3,2.9,0.4675)
[1] 0.8038946
> qnorm(0.8038946,2.9,0.4675)
[1] 3.3
```

# Another application of the standard normal

Following with the previous example, imagine we want to know the **central interval that includes 95% of the values**. Then we have 2.5% of the values to the left and 2.5% to the right, outside this interval. Then:

$$z_i = \frac{y_i - \bar{y}}{s}; \quad y_i = \bar{y} + z_i s$$

We can compute $z_{.025}$ and $z_{.975}$ in R:

```
> qnorm(.025)
[1] -1.959964
> qnorm(.975)
[1] 1.959964
```

`qnorm` gives the *z*-value for a given quantile of the cdf of a standard normal

Then, the lower and upper limits of the interval are:

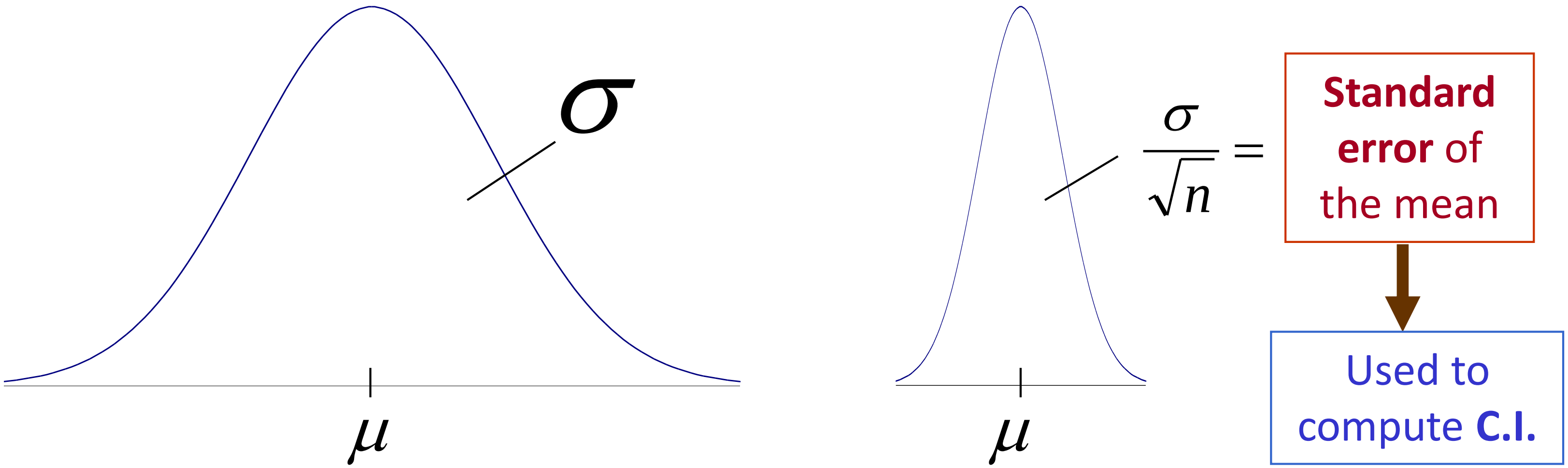
$$y_{lower} = \bar{y} + z_{.025} s = 2.90 + (-1.96) \times 0.4675 = 1.984$$

$$y_{upper} = \bar{y} + z_{.975} s = 2.90 + 1.96 \times 0.4675 = 3.816$$

# Standard error of the mean

Suppose we want to know the mean for *BODYwt* of male cats. Usually we do not have the entire population, but a sample. Let assume that we take repeated samples with replacement of size $n$ (in our case $n$ = 97) from that entire population, that is normally distributed.

For each sample we will have a different, but close mean, for example 2.90, 2.81, 2.93, 2.76, 2.83, … and so on. It can be shown that:



$$\frac{\sigma}{\sqrt{n}} =$$

**Standard error** of the mean

Used to compute **C.I.**

Distribution of the entire population

Distribution of the estimated means

# CI: mean of a normal, variance known

If we fix some confidence level $\gamma$ (for example 95%), with $\alpha = 1 - \gamma$, the true mean $\mu$ is found in the interval:

$$\bar{y} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

1.96 for $\gamma$ = 95%

Confidence limits

$$\left( \bar{y} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{y} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Interval length

$$\ell = 2 z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

In this formula we have used the same principle as in slide 7, but applied to the distribution of the means.

Note that $\frac{\sigma}{\sqrt{n}}$ is the **standard error**, i.e., the standard deviation of the distribution of means under a repeated sampling (infinite) of size *n*.

## CI: mean of a normal, variance known (example)

Assuming that the estimated variance for *BODYwt* is the true variance, and that *BODYwt* is normally distributed, the 95% confidence interval of the mean is:

$$\left(2.90 - 1.96 \times 0.0475, \quad 2.90 + 1.96 \times 0.0475\right)$$

$$\left(2.807, \quad 2.993\right)$$

With interval length:

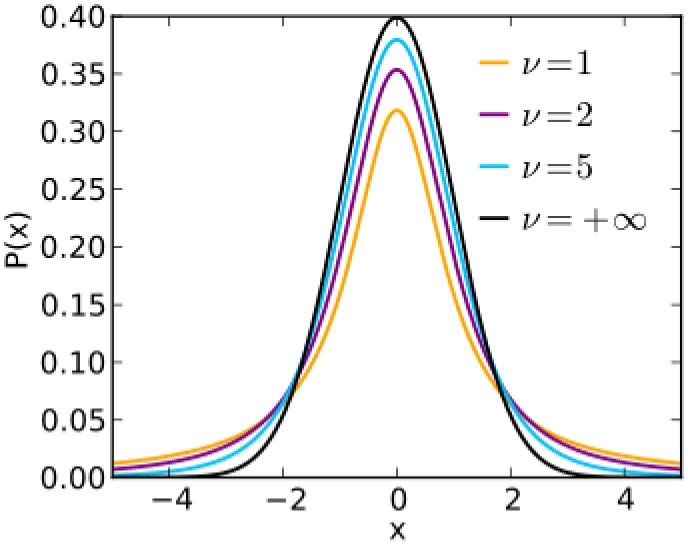
$$\ell = 2 \times 1.96 \times 0.0475 = 0.186$$

# *t* distribution (1)

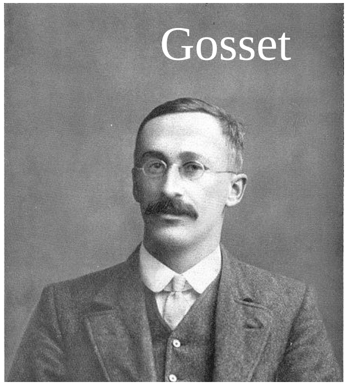
$T \sim t_{n-1}$ if its p.d.f. is

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

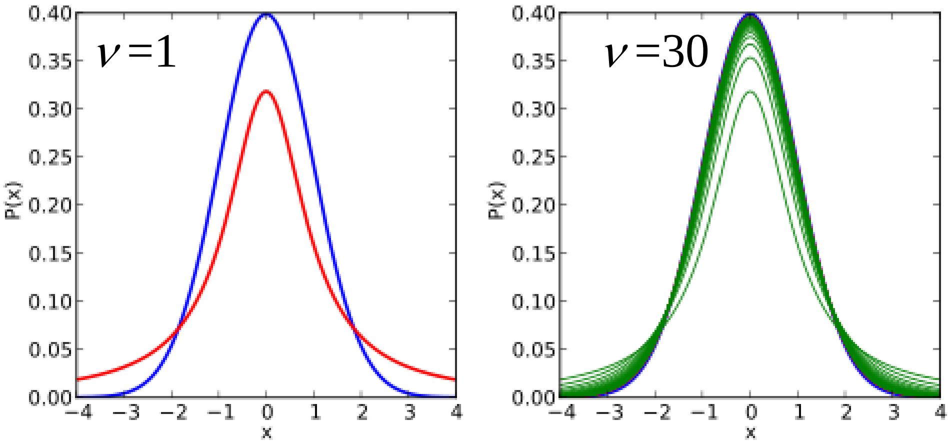
with $\nu = n - 1$



http://en.wikipedia.org/wiki/Student%27s_t-distribution



http://en.wikipedia.org/wiki/William_Sealy_Gosset



Note that the *t*-distribution (red or green line) approaches the normal distribution (blue line) as $\nu$ increases

# CI: mean of a normal, variance unknown (1)

This is the **common case**, because we have an estimate of the variance or the standard deviation instead of the true value.

It is similar to the previous case but using the *t*-distribution instead of the Normal distribution.

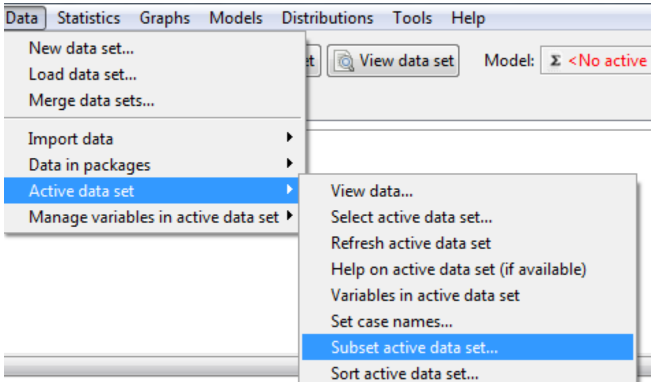
$$\left( \bar{y} - t_{1-\frac{\alpha}{2}}^{n-1} \frac{s}{\sqrt{n}}, \bar{y} + t_{1-\frac{\alpha}{2}}^{n-1} \frac{s}{\sqrt{n}} \right)$$

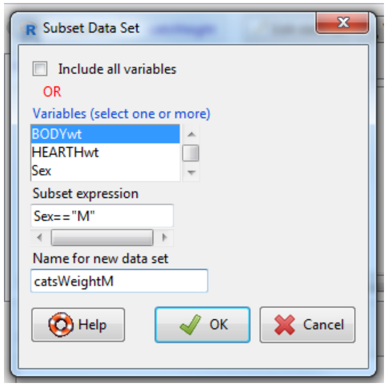
Or in a synthetic expression:

$$\bar{y} \pm t_{1-\frac{\alpha}{2}}^{n-1} \frac{s}{\sqrt{n}}$$

# CI: mean of a normal, variance unknown (2)

First we need a **subset** with male weights only: catsWeightM





□ Include all variables

Subset expression: Sex=="M"

Name for new data set: catsWeightM

Data: catsWeightM

Statistics > Means > Single-sample t-test

Variable: BODYwt; Alternative Hypothesis: Population mean !=mu0

```
        One Sample t-test
data:  BODYwt
t = 61.097, df = 96, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.805781 2.994219
sample estimates:
mean of x
      2.9
```

## CI: Interpretation

If we would take all possible samples of size 97 of body weight in the cat males, and for each of them we would made the above calculations, 95% of the intervals found, approximately, would contain $\mu$.

**We do not know whether the interval we have found contains or not $\mu$**, because this parameter is unknown (in fact it is what we are looking for), but we are 95% confident in that it be so.

# Test of Hypothesis (Neyman-Pearson approach)

**Hypothesis testing** is the use of Statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of **four steps**.

1. Formulate the **null hypothesis** $H_0$ (commonly, that the **observations are the result of pure chance**) and the **alternative hypothesis** $H_1$ (commonly, that the observations are the result of a **real effect** **combined** with a component of **chance** variation).
2. Identify a **test statistic** that can be used to assess the truth of the null hypothesis.
3. Compute the ***p*-value**, which is the probability of finding a value of the test statistic bigger (or lower, depending on the tail) as the one observed, assuming that the null hypothesis were true. **The smaller the *p*-value, the stronger the evidence against the null hypothesis**.
4. Construct a **decision rule**: Compare the $p$-value to an acceptable **significance level** $\alpha$ . If $p \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is rejected, and the alternative hypothesis is accepted.

# Test of Hypothesis (cont.)

| **Null hypothesis** | **Accepted** | **Rejected** |
|---|---|---|
| **True** | Correct decision | Type I error ($\alpha$) or significance level<br>**(False positives)** |
| **False** | Type II error ($\beta$)<br>**(False negatives)** | Correct decision<br>**($1-\beta$ = Power of test)** |

The decision rule is based upon Type I error:

- If $p$-value > 0.05 ➔ accept $H_0$
- If $p$-value < 0.05 ➔ reject $H_0$

# Contrasting a mean

To contrast that the mean of weight of male cats is different from 0, we can use a previous procedure and result:

Data: catsWeightM

Statistics > Means > Single-sample t-test

Variable: BODYwt; Alternative Hypothesis: Population mean !=mu0

```
       One Sample t-test
data:  BODYwt
t = 61.097, df = 96, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.805781 2.994219
sample estimates:
mean of x
      2.9
```

The null hypothesis ($H_0$: $\mu = 0$) is rejected, as the *p*-value < 0.05

# Contrasting normality: Shapiro-Wilk test

Statistics > Summaries > Test of normality

Variable: BODYwt; Normality test: Shapiro-Wilk

```
Shapiro-Wilk normality test
data:  BODYwt
W = 0.95188, p-value = 0.00006731
```

The null hypothesis ($H_0$) is that the distribution is normal

Test by groups: Sex

```
Sex = F
data:  BODYwt
W = 0.89096, p-value = 0.0003754

Sex = M
data:  BODYwt
W = 0.97883, p-value = 0.119
```

Male’s weight is normally distributed, but Females and the mixture distribution are not