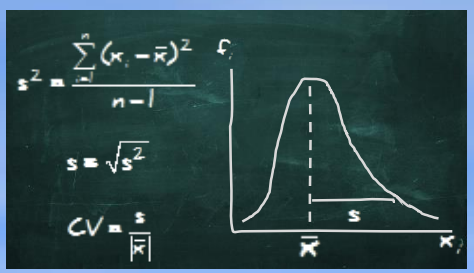
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.4
Fundamentos de estadística inferencial

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2016). Fundamentos de estadística inferencial. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra. (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. Capítulo III.4. http://ddd.uab.cat/record/163560

Capítulo acabado de redactar en agosto de 2016

Contenido

PARTE III
Capítulo 4

Fundamentos de estadística inferencial

El análisis de datos estadísticos, como destacamos en el capítulo anterior, comprende dos aspectos importantes, diferenciables, pero que se tratan al unísono en un proceso de tratamiento, análisis e interpretación de datos obtenidos por muestreo estadístico: la vertiente descriptiva y la vertiente inferencial. Se trata de responder a una cuestión fundamental: ¿en qué medida las conclusiones que se extraen del análisis descriptivo tienen validez[1] y pueden extrapolarse en consecuencia al conjunto de la población?

En el capítulo II.4 introdujimos diversos conceptos de muestreo estadístico al hablar del diseño muestral de un estudio. Recuperaremos esos conceptos, muchos de los cuales tienen un fundamento de conceptualización estadística que abordaremos con algo más de detenimiento.

En este capítulo buscamos sentar las bases de una primera aproximación al razonamiento inferencial en el análisis de los datos cuantitativos. Se abordarán diversas cuestiones que suponen el primer paso en el tratamiento inferencial considerando la información de una sola variable. No obstante, la vertiente inferencial estará presente en todo ejercicio de análisis en todas las técnicas estadísticas que trabajen con muestras de la población. En los capítulos siguientes de este manual se irán introduciendo las extensiones de este tipo de razonamiento a medida que sean necesarias para el uso completo y adecuado de cada instrumento analítico.

Inicialmente, los conceptos que se derivan del razonamiento inferencial suelen ser los más complejos y abstractos de comprensión. Por ello intentaremos abordarlos de forma didáctica con el objetivo de alcanzar un conocimiento de sus fundamentos que nos facilite la compresión básica de su sentido. Una vez asimilado el aprendizaje de esos fundamentos veremos que se trata de un aspecto muy mecánico que se reproduce sistemáticamente en cada técnica de análisis que trata con datos muestrales y que suele

[1] Se trata de un tipo de validez específica, la denominada validez externa, la que me permite establecer que mis conclusiones son generalizables al conjunto de la población sobre la base de la información que se dispone de una muestra representativa estadísticamente de dicha población.

expresarse de forma sencilla en afirmaciones como "(no) es estadísticamente significativo", o bien "considerando un nivel de significación del 5% (0,05) la prueba de hipótesis estadística (no) es significativa". En este tipo de conclusiones se traduce el aspecto inferencial, establece la relevancia de una afirmación descriptiva. Por ejemplo, valida si las diferencias de ingresos salariales entre varones y mujeres que en un momento se pueden observar en una muestra determinada, afectadas por un determinado error, son significativas estadísticamente, o bien si esas diferencias no permiten afirmar que sean suficientemente importantes como para ser extrapoladas al conjunto de la población, a pesar de que existan, pero resultan ser pequeñas, pues están por debajo del error estadístico que podemos admitir con nuestros datos y que en cada caso se establece. Veremos estos razonamientos estadísticos considerando diversas fórmulas que se aplican y calculan en cada caso, con el objetivo de comprender que hay detrás de cada posible conclusión inferencial. En el momento de la aplicación de las técnicas veremos también que de esta tarea se encargará el software estadístico y nosotros/as nos dedicaremos tan sólo a interpretar sencillamente si se da o no significación de un resultado concreto.

El capítulo se estructura en cuatro apartados principales. En primer lugar se introduce la temática de la probabilidad. Las bases del comportamiento estadístico inferencial se fundamentan en el concepto del azar, de la aleatoriedad, y en la idea de la probabilidad de que se dé un determinado suceso. Cada suceso tendrá una probabilidad mayor o menor, y el conjunto de sucesos da lugar a una distribución de probabilidades. Las distribuciones de probabilidades pueden corresponder a nuestros datos o pueden determinarse teóricamente desde un punto de vista estadístico. Aquí veremos diversas distribuciones muestrales estadísticas que nos ayudarán a establecer la significación de los resultados del análisis de los datos: la distribución normal, la distribución de chi-cuadrado, de Fisher, la t de student, entre otras. Veremos cómo interpretar básicamente esta información que se utilizará de forma recurrente en todas las técnicas estadísticas.

En segundo lugar veremos que la idea de probabilidad y de distribución muestral son la base de nuestras estimaciones. Todo ejercicio de análisis pretender afirmar alguna característica de la población a partir de una parte de ella, la muestra. Al hacerlo comete, necesariamente, por tener una parte del todo, un determinado error que se puede calcular. El resultado que observamos es una estimación muestral, llamado estadístico muestral, que estima el verdadero valor poblacional, siendo el error el que establece el margen de desviación del valor muestral respecto del valor poblacional, llamado parámetro poblacional. Los estadísticos y los parámetros son el mismo cálculo, efectuados en la muestra y en la población, respectivamente, y pueden ser cálculos tan sencillos como, por ejemplo, un porcentaje o una media aritmética. Veremos por tanto cómo un determinado resultado (un porcentaje o una media) se ve afectado por el error estadístico y cómo nos conduce a hablar de intervalos de confianza, es decir, el conjunto de valores (entre un mínimo y un máximo) donde probablemente se situará el valor poblacional. Es decir, el error que comentemos con una muestra nos obligará a realizar afirmaciones descriptivas con un cierto grado de incertidumbre, de variabilidad.

Una primera aplicación del razonamiento inferencial de los estudios empíricos por muestreo es el de seleccionar una muestra representativa de la población y determinar,

en particular, cuál "debe ser" el tamaño de la muestra, es decir, la magnitud suficiente del número de unidades (entrevistados/as, hogares, secciones censales, etc.) que garantice la relevancia de nuestras conclusiones analíticas. Veremos cómo se calcula el tamaño de la muestra y la relación que se establece con el tamaño de la población y el error muestral.

Finalmente, el razonamiento estadístico inferencial persigue un objetivo fundamental: establecer la validez de nuestras hipótesis de investigación. En este caso se trata de expresarlas en forma de hipótesis estadísticas y realizar un contraste de hipótesis: entre una hipótesis que afirma un estado de cosas, por ejemplo, existen diferencias de ingresos salariales entre varones y mujeres, y la anulación (falsación) de la misma, es decir, que no existen diferencias significativas. Veremos a lo largo del manual que cada técnica de análisis introducirá su particular prueba de hipótesis estadística, por lo que será una cuestión recurrente en los próximos capítulos. Aquí realizaremos una primera presentación y veremos dos tipos de pruebas sencillas para el cálculo de una media y una proporción. Así pues, los dos primeros contenidos del capítulo constituyen un bagaje conceptual instrumental destinado a resolver dos aspectos de la investigación cuantitativa: la necesidad de todo diseño de investigación por muestreo de determinar tamaño de la muestra, y establecer las bases del contraste de hipótesis, objetivo fundamental de la investigación con datos estadísticos.

1. Probabilidad y distribuciones estadísticas

1.1. Conceptos y definiciones fundamentales

Del tratamiento de la información estadística a través de sus diversos métodos y técnicas se derivan afirmaciones y conclusiones basadas en criterios probabilísticos. La estadística constituye un instrumento de la investigación científica que nos permite inferir resultados y conclusiones, con el fin de describir fenómenos y validar hipótesis que establecen generalizaciones respecto de una población que es objeto de estudio, que por estar basadas en muestras estarán afectadas de un cierto grado de incertidumbre. Habitualmente estos resultados y conclusiones no son más que estimaciones de la realidad que comportan un determinado nivel de error que nos llevan a enunciar que el fenómeno en cuestión analizado es más o menos probable. Y ello es así dadas las dificultades de aseverar con total seguridad el comportamiento social en un espacio y tiempo determinados y, en particular, dadas las dificultades de predecir el comportamiento futuro por su imprevisibilidad.

En este sentido, la probabilidad se tiene que entender como una medida de incertidumbre. En los estudios sociales empíricos es habitual recoger información de naturaleza estadística a partir de una muestra de la población con la intención de extrapolar aquello que caracteriza a la muestra, al conjunto de la población. El denominado muestreo aleatorio o al azar es un concepto que se relaciona estrechamente con el de probabilidad. Tal como hemos visto previamente, la inferencia es el proceso de razonar a partir de una parte (de lo particular), que constituye la muestra con respecto a un todo (lo general), que representa la población. La validez de este razonamiento dependerá de cuán representativa sea esa parte, y por tanto, también

su nivel de error y la incertidumbre de las afirmaciones que se deriven[2]. El error siempre existirá, e igual que la validez, es una cuestión de grado, muchas veces difícil de precisar. Nos referimos al error en general pues el error estadístico en particular es más fácil de determinar al basarse en fórmulas que calculan unos valores que ponen en relación muestra y población. Pero recordemos que junto al error estadístico está el error no estadístico o sistemático que tiene que ver con cualquier aspecto de producción de la información y, por tanto, de difícil determinación y precisión. Sin embargo, el error estadístico es posible calcularlo con precisión siempre que nuestros datos se obtengan de una muestra representativa de la población. ¿Cómo se obtiene una muestra representativa? Uno de los procedimientos básicos es el denominado muestreo aleatorio simple, donde los elementos de la muestra se eligen al azar de forma que se garantiza que todo elemento de la población tenga igual probabilidad de ser elegido en la muestra (sea equiprobable) y que toda combinación de elementos también tenga la misma probabilidad de ser elegida.

En esa elección al azar siempre existirá un grado de imprevisibilidad, una probabilidad de acertar y, en consecuencia, de cometer un error, simplemente porque se dispone de una parte del todo y porque cada posible muestra que pudiéramos obtener, al contener elementos distintos, generaría un resultado distinto. Un claro ejemplo de imprevisibilidad que suele aparecer en los manuales de estadística son los juegos de azar. De hecho históricamente los juegos de azar han motivado el interés por el estudio de la probabilidad. Pero que algo sea imprevisible no es incompatible con conocer cómo se comporta y bajo qué condiciones se da cierto comportamiento regular. Este comportamiento precisamente se puede expresar en términos de probabilidad. Por ejemplo podemos afirmar que "lo más probable es que el partido socialista gane las próximas elecciones" o que "es poco probable que alguien nacido en un barrio marginal con una familia sin estudios y con problemas de desempleo tenga movilidad social ascendente y alcance una ocupación de director de una gran empresa". En cada caso, con los datos estadísticos, será posible establecer esa probabilidad para una población en un espacio y tiempo fijados.

El concepto de probabilidad requiere una definición para convertirlo en un instrumento objetivable y aplicable en el trabajo científico. No es una tarea sencilla y a lo largo de la historia de la estadística se han propuesto diversas alternativas que brevemente destacaremos. Primero, es conveniente precisar algunos conceptos básicos de probabilidad.

En primer lugar cabe destacar que el concepto de probabilidad se emplea no para referirse a sucesos particulares sino a un gran número de acontecimientos repetidos y, por tanto, se trata de una probabilidad que ocurre a la larga y respecto a una estabilidad o regularidad que caracteriza a un fenómeno de una población. La obtención de una probabilidad se concreta en el contexto de una operación determinada que, en idénticas condiciones, nos permite observar un comportamiento. En este contexto y

[2] Se trata de un doble proceso deductivo e inductivo. Deductivo en la medida en que la teoría del muestreo nos permite razonar sobre el comportamiento general de una población cuando se expresa en términos de una muestra representativa y porque en cualquier proceso de investigación, especialmente con metodología cuantitativa, se precisa definir primero conceptualmente la información que se analiza. Inductivo porque obtenemos de una realidad empírica concreta, situada espacial y temporalmente, resultados que son extrapolados al conjunto de la población y que permiten generalizar, validar y teorizar un comportamiento social.

condiciones es importante que se dé la imposibilidad de predecir o controlar cuál será el resultado exacto del comportamiento.

La teoría de la probabilidad es la modelización del azar. Mediante un muestreo aleatorio, se trata de que actúe el azar y que no exista la posibilidad de condicionar un resultado. De esta forma razonar teóricamente sobre el comportamiento muestral y aplicar el potente instrumental de la teoría de probabilidades para evaluar el margen de error y el riesgo de nuestras estimaciones.

En el lenguaje estadístico, a la operación de obtención de la muestra aleatoria se denomina experimento aleatorio. Un suceso aleatorio es cada uno de los resultados, elementos o hechos observables a los que puede dar lugar un experimento aleatorio. Los sucesos se suelen identificar con letras mayúsculas (A, B, C,...). Por ejemplo, si consideramos a la población española a 1 de enero de 2016 tenemos la distribución por sexo que aparece en la Tabla III.4.1 donde V sería el suceso Varón y M el suceso Mujer. La noción de probabilidad, que denotaremos con Pr, la podemos expresar a partir de esta distribución conocida de frecuencias de la población.

Tabla III.4.1. Distribución de la población española según sexo. 1/1/2016

	Frecuencia	Frecuencia relativa	Porcentaje
V Varón	22.805.060	0,491	49,1
M Mujer	23.633.362	0,509	50,9
Total	46.438.422	1,000	100,0

Fuente: Instituto Nacional de Estadística. Cifras de población.

Si queremos elegir una muestra al azar de esta población podemos responder a las preguntas siguientes:

- ¿Cuál es la probabilidad de que una persona elegida al azar sea varón?, es:

$$Pr(V) = \frac{22.805.060}{46.438.422} = 0,491$$

- ¿Cuál es la probabilidad de que una persona elegida al azar sea mujer?, es:

$$Pr(M) = \frac{22.805.060}{23.633.362} = 0,509$$

Las probabilidades se expresan como frecuencias relativas o tanto por 1 (proporciones). Sabiendo que el porcentaje de mujeres en la población es del 50,9 la probabilidad de ser mujer al elegir aleatoriamente una persona de la población es de 0,509.

Otros sucesos aleatorios podrían ser: que salga un 6 en el lanzamiento de un dado, que un encuestado elegido al azar nos diga que va a votar al partido comunista, o tener un nivel de ingresos por debajo del umbral de pobreza, que se tenga conexión a internet, o que el nivel de aceptación de determinada política municipal sea superior a 5 en una escala de 0 a 10.

Se suele emplear la expresión éxito aplicado a un suceso cuando éste ocurre, y fracaso cuando éste no ocurre. Así se habla de probabilidad de éxito y de probabilidad de fracaso.

Un suceso es elemental si no es posible descomponerlo en otros más elementales (tener un hijo/a), y es compuesto cuando se puede descomponer en otros sucesos más elementales (tener 3 o más hijos).

Se denomina espacio muestral, identificado por Ω, al conjunto de todos los resultados elementales posibles a los que puede dar lugar un experimento aleatorio[3]. Estos resultados pueden ser finitos, infinitos numerables o infinitos no numerables. Cualquier suceso aleatorio es un subconjunto del espacio muestral.

Un suceso se denomina universal o suceso seguro si siempre se da, si contiene por tanto todos los resultados posibles. En este caso coincide con el espacio muestral Ω y se cumple que su probabilidad es uno:

$$Pr(\Omega) = 1 \qquad \textit{Ecuación 1}$$

En el ejemplo anterior del sexo, el suceso "ser varón o ser mujer", es un suceso seguro.

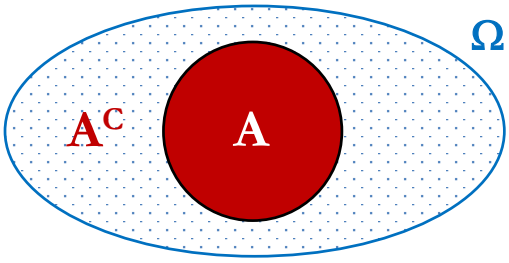
De forma inversa, un suceso que nunca se da se denomina imposible, se identifica con $\varnothing$ (conjunto vacío) y se cumple que su probabilidad es cero:

$$Pr(\varnothing) = 0 \qquad \textit{Ecuación 2}$$

Si consideramos el suceso "ser varón y ser mujer" es evidente que se trata de un suceso imposible.

Estas ideas de espacio muestral y de suceso se puede representar de diversas formas. En particular como puntos en el espacio y como conjuntos. En el Gráfico III.4.1 se representa el espacio muestral por un conjunto Ω, y un suceso A particular, que constituye un subconjunto de Ω. Se denomina suceso complementario o contrario de un suceso A, con notación A^C, a otro suceso que contiene todos los elementos que no contiene el suceso A, el suceso de que no ocurra A.

Gráfico III.4.1. Representación de un espacio muestral, de un suceso y del suceso complementario



[3] Concepto desarrollado a partir de los trabajos de Von Mises (1931).

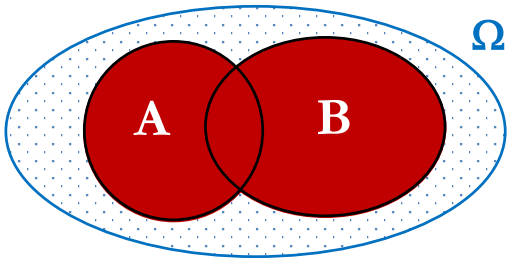
En el ejemplo anterior del sexo de la población española podemos expresar el espacio muestral así: $\Omega=\{V,M\}$, es decir, un conjunto formado por dos elementos, "ser varón" y "ser mujer". Su probabilidad es uno menos la probabilidad de A:

$$Pr(A^C)=1-Pr(A)$$ *Ecuación 3*

El suceso complementario de ser varón es ser mujer: $V^C=M$, e inversamente, el de ser mujer es ser varón: $M^C=C$. El conjunto Ω podría ser el de todas las edades de las personas y un suceso A particular podría ser tener entre 18 y 30 años. El suceso complementario sería entonces el de todas las edades de 0 a 17 y de 31 hasta la más alta.

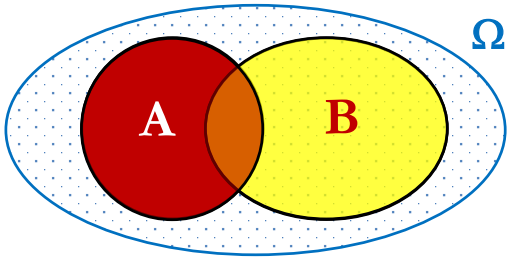
La unión de dos sucesos A y B, es el suceso $A \cup B$ que contiene todos los sucesos elementales en A o en B, e implica que o se da A o se da B. Gráficamente se representa como en el Gráfico III.4.2. Siguiendo con el ejemplo anterior, A podría ser el suceso tener menos de 18 años (ser menor de edad) y B tener de 16 a 64 años (edad laboral habitual en España). La unión de A y B sería el espacio muestral de todas las edades hasta 64 años.

Gráfico III.4.2. Representación de la unión de sucesos



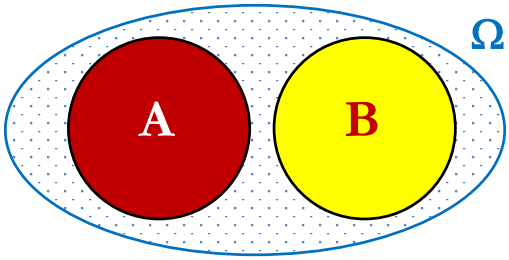
La intersección de dos sucesos A y B, es el suceso $A \cap B$ que contiene sólo los elementos comunes a A y B, e implica que se da A y se da B. Con el ejemplo anterior la intersección entre los que tienen menos de 18 años y los que tienen entre 16 a 64 años da lugar al suceso tener 16, 17 o 18 años.

Gráfico III.4.3. Representación de la intersección de sucesos



Dos sucesos son incompatibles cuando son mutuamente excluyentes, es decir, no tienen ningún elemento en común. Por tanto, la intersección entre ambos sucesos es el conjunto vacío: $A \cap B=\varnothing$. Es el caso de los sucesos tener menos de 18 años y tener más de 65.

Gráfico III.4.4. Representación de dos sucesos incompatibles



Cuando se asocia la ocurrencia de un suceso con un espacio muestral Ω se está planteando un enunciado de probabilidad. La probabilidad se expresa como *Pr(A)*. Para definir la probabilidad de un suceso *A*, es decir *Pr(A)*, se pueden adoptar diferentes puntos de vista:

1) Definición clásica. Es la definición más antigua y más intuitiva. Establece que la probabilidad de un suceso aleatorio *A* es el número de casos favorables dividido por el número de casos posibles:

$$Pr(A) = \frac{casos\ favorables}{casos\ posibles} \qquad \textit{Ecuación 4}$$

Es una definición de la probabilidad de ocurrencia que especifica básicamente su forma de cálculo. El problema que puede presentar es que no podría aplicarse en el caso de desconocer los casos posibles. Por otro lado, es una definición que presupone que todos los elementos son igualmente probables, si no se cumpliera esta condición tampoco sería aplicable.

2) Definición frecuencialista. Se trata de una formulación expresada en términos empíricos. La probabilidad formal de un suceso se corresponde con su frecuencia relativa de ocurrencia. Esta ocurrencia hace alusión a ciertos fenómenos cuya frecuencia relativa en situaciones similares se aproximan a un valor fijo constante al aumentar el número de observaciones. Así, la probabilidad se define como el límite de la frecuencia relativa cuando el número de observaciones crece indefinidamente:

$$Pr(A) = \lim_{n\to\infty} f(A) \qquad \textit{Ecuación 5}$$

Es una definición vinculada a la ley de los grandes números: a medida que aumenta el número de casos, el valor de probabilidad converge hacia el verdadero valor. Esta concepción de la probabilidad no está exenta de problemas matemáticos, en particular cuando no existe este límite. Por ello se ha propuesto un concepto desarrollado axiomáticamente.

3) Definición axiomática. Kolmogorov (1933) estableció una definición axiomática que debe cumplir toda medida de probabilidad a partir de una función de probabilidad. Una función de probabilidad es una relación de correspondencia que para cualquier suceso *A* asocia un número *P(A)* del espacio muestral Ω tal que:

- $Pr(A) \geq 0$, la probabilidad es cero o positiva.
- $Pr(\Omega) = 1$, la suma de todos los sucesos posibles es 1.

– $A \cap B = \varnothing \Rightarrow Pr(A \cup B) = Pr(A) + Pr(B)$, si la probabilidad de la intersección es el conjunto vacío, la probabilidad de la unión es la suma de probabilidades de los sucesos.

De esta definición se derivan las propiedades siguientes:

– $0 \leq Pr(A) \leq 1$, la probabilidad de un suceso siempre se encuentra entre 0 y 1.
– $Pr(\varnothing) = 0$, la probabilidad del suceso imposible es cero.
– $Pr(A^C) = 1 - Pr(A)$, uno menos la probabilidad de A es la probabilidad del suceso complementario.

4) Definición subjetivista. Se trata de medidas de incertidumbre basadas en la creencia o en la valoración que atribuye un individuo a un suceso dependiendo de su grado de información y de la experiencia sobre determinados fenómenos difícilmente cuantificables.

Trabajaremos con la primera definición, por lo que consideraremos un espacio muestral finito definido por los sucesos o posibles valores tomando muestras de tamaño 1 de una población donde todos los individuos tienen igual probabilidad de ser escogidos en una muestra aleatoria. En ese caso las probabilidades son exactamente las frecuencias relativas o proporciones poblacionales, como vimos anteriormente en el caso del sexo de la población (Tabla III.4.1).

Una vez establecida la definición de probabilidad podemos formalizar la definición de probabilidad de la unión y la probabilidad conjunta de dos sucesos.

La probabilidad de la unión de dos sucesos cualesquiera A y B es la suma de probabilidades de cada suceso menos la probabilidad de la intersección.

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B) \qquad \textit{Ecuación 6}$$

La probabilidad conjunta de dos sucesos A y B es la probabilidad de su intersección $A \cap B$, es decir, la probabilidad de que se den los dos sucesos simultáneamente. En general la propiedad de la intersección satisface:

$$Pr(A \cap B) = Pr(A) \cdot Pr(B) \text{ si } A \text{ y } B \text{ son independientes} \qquad \textit{Ecuación 7}$$

$$Pr(A \cap B) \neq Pr(A) \cdot Pr(B) \text{ si } A \text{ y } B \text{ no son independientes} \qquad \textit{Ecuación 8}$$

Cuando la probabilidad de un suceso es independiente del resultado del otro, estas probabilidades se denominan marginales. En estas condiciones se cumple que:

– $Pr(A \cap B) \leq Pr(A)$.
– $Pr(A \cap B) \leq Pr(B)$.

El concepto de independencia es de crucial importancia en Estadística y en la investigación pues, cuando se relacionan fenómenos a través de variables, se establece la existencia o no de relaciones entre ellas. Por ejemplo, si los datos de ingresos son semejantes entre varones y mujeres cabe concluir la independencia entre las dos

variables, ingresos y sexo. En otras palabras, independientemente de si se es varón o mujer los ingresos son iguales.

Junto a la idea de independencia otro concepto fundamental es el de dependencia o de probabilidad condicionada. La probabilidad condicional de un suceso A condicionado por un suceso B, expresada como $Pr(A|B)$, es el cociente entre la probabilidad conjunta y la probabilidad marginal del suceso que condiciona:

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)} \qquad \textit{Ecuación 9}$$

En la probabilidad condicional se considera el espacio muestral Ω definido por el suceso B, y la probabilidad que se considera se refiere a la parte del suceso B que contiene el suceso A. En el ejemplo citado, la relación entre ingresos y sexo es de dependencia cuando la probabilidad de ser rico es mayor entre los varones que entre la mujeres, es decir, a condición de ser varón la probabilidad aumenta y a condición de ser mujer la probabilidad disminuye.

De esta definición de la probabilidad condicional se deriva la definición general de probabilidad conjunta o de la intersección:

$$Pr(A \cap B) = Pr(A \mid B) \cdot Pr(B) \qquad \textit{Ecuación 10}$$

Si consideramos además $Pr(B|A)$, se verifica que:

$$Pr(A \cap B) = Pr(A \mid B) \cdot Pr(B) = Pr(B \mid A) \cdot Pr(A) \qquad \textit{Ecuación 11}$$

Cuando los sucesos A y B son independientes, no se condicionan, no existe intersección, es decir, $Pr(A \cap B)=\varnothing$, entonces se verifica que:

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{Pr(A) \cdot Pr(B)}{Pr(B)} = Pr(A) \qquad \textit{Ecuación 12}$$

Y de la misma forma $Pr(B \mid A) = Pr(B)$.

Pero cuando no se da la independencia se concluye que:

$$Pr(A \mid B) = \frac{Pr(B \mid A) \cdot Pr(B)}{Pr(B)} \qquad \textit{Ecuación 13}$$

Y de forma equivalente para $Pr(B \mid A)$.

A partir de estas relaciones se formula el Teorema de Bayes, de gran utilidad para entender cómo se modifican las probabilidades cuando se introduce información suplementaria. Es un teorema que ha desarrollado la llamada Estadística Bayesiana y es fundamental en Teoría de la Decisión.

Si $Pr(A)$ es la probabilidad a priori de un suceso, cuando introducimos la información del suceso B, el primero se ve modificado de forma que $Pr(A|B)$ es la probabilidad a

posteriori del suceso. Con el teorema obtenemos la probabilidad final (a posteriori) a partir de la probabilidad inicial (a priori).

1.2. Distribuciones de probabilidad de una variable aleatoria

Cuando se consideran los valores de una variable como resultados obtenidos a partir de una elección de individuos al azar, decimos que se trata de una variable aleatoria[4]. Las variables aleatorias no son más que la traducción de los sucesos aleatorios en términos numéricos. Cuando se asigna un valor numérico a cada uno de los sucesos elementales de un espacio muestral obtenemos la variable aleatoria.

Si consideramos la variable X definida como el voto en las elecciones, por ejemplo las últimas Elecciones Generales al Congreso de los Diputados en España, de 27 de junio de 2016, cada suceso se define como el voto a cada una de las opciones políticas que se presentan en las elecciones (Tabla III.4.2).

Tabla III.4.2. Resultados de las Elecciones Generales al Congreso de los Diputados en España, 27 de junio de 2016.

Partido	Frecuencia	Frecuencia relativa	Porcentaje	Escaños
PP	7.906.185	0,330	33,0	137
PSOE	5.424.709	0,227	22,7	85
UNIDOS PODEMOS	5.049.734	0,211	21,1	71
Ciudadanos	3.123.769	0,131	13,1	32
ERC-CATSI	629.294	0,026	2,6	9
CDC	481.839	0,020	2,0	8
EAJ-PNV	286.215	0,012	1,2	5
EH Bildu	184.092	0,008	0,8	2
CCa-PNC	78.080	0,003	0,3	1
Resto	336.393	0,024	2,4	0
Total	23.500.310	1	100,0	350

Fuente: Ministerio del Interior. Gobierno de España.

Podemos asignar de forma arbitraria un valor a cada uno de los partidos:

1 PP
2 PSOE
3 UNIDOS PODEMOS
4 Ciudadanos
5 ERC-CATSI
6 CDC
7 EAJ-PNV
8 EH Bildu
9 CCa-PNC
10 Resto

En este caso hemos efectuado una asignación numérica arbitraria puesto que se trata de una variable discreta de naturaleza cualitativa, en concreto se corresponde con un nivel de medición nominal. De hecho, la definición de las variables aleatorias se relaciona directamente con los tipos de escalas de medición. Las variables aleatorias

[4] También denominada variable estocástica o variable de probabilidad.

más sencillas son las variables discretas, aquellas que sólo pueden tomar un número finito de valores (o infinito numerable).

Ahora la cuestión que se plantea es la probabilidad asociada a cada suceso, es decir, la probabilidad que tiene cada partido de ganar las elecciones.

Si X es una variable aleatoria de valores discretos $x_1, x_2, \ldots, x_k$ de la cual conocemos las probabilidades de sus valores $\pi_1, \pi_2, \ldots, \pi_k$, decimos que tenemos una distribución de probabilidad de la variable X:

x_i	π_i
x_1	π_1
x_2	π_2
⋮	⋮
x_k	π_k
Total	1

donde se verifica que:

– La suma de probabilidades es 1: $\sum_{i=1}^{k} \pi_i = 1$

– Cada probabilidad es un valor entre 0 y 1: $0 \leq \pi_i \leq 1$

La función $\pi(x)$ que toma valores $\pi_1, \pi_2, \ldots, \pi_k$ para $x_1, x_2, \ldots, x_k$ se llama función de probabilidad o función de frecuencia de X, es una regla que asigna las probabilidades a los valores de la variables aleatoria. Más estrictamente se dice que es una aplicación del espacio muestral Ω sobre el conjunto de números reales $\Re$.

En el ejemplo sobre el voto en las Elecciones Generales, la distribución de probabilidades sería la que se ha presentado en la Tabla III.4.2:

x_i	Partido	π_i
1	PP	0,330
2	PSOE	0,227
3	UNIDOS PODEMOS	0,211
4	Ciudadanos	0,131
5	ERC-CATSI	0,026
6	CDC	0,020
7	EAJ-PNV	0,012
8	EH Bildu	0,008
9	CCa-PNC	0,003
10	Resto	0,024
	Total	1,000

La distribución de probabilidad es análoga a la distribución de frecuencias. Pero se considera a las distribuciones de probabilidad que son teóricamente las distribuciones de frecuencias relativas llevadas al límite, es decir, cuando el número de observaciones

es muy grande. Las distribuciones de probabilidad se pueden ver como distribuciones de poblaciones a partir de las cuales se obtienen, en muestras extraídas aleatoriamente, distribuciones de frecuencias relativas.

En relación con la probabilidad del voto en la Elecciones Generales de 2016 podemos disponer de los resultados y de la distribución de frecuencias de un sondeo electoral realizado días antes de la convocatoria a las urnas. El estudio 3141 del Centro de Investigaciones Sociológicas de mayo de 2016 arrojaba estos resultados de intención de voto (Tabla III.4.3)[5]:

Tabla III.4.3. Resultados de las Elecciones Generales y sondeo del CIS.

x_i	Partido	Elecciones	CIS
1	PP	33,0	26,4
2	PSOE	22,7	23,0
3	UNIDOS PODEMOS	21,1	28,8
4	Ciudadanos	13,1	13,4
5	ERC-CATSI	2,6	2,8
6	CDC	2,0	1,6
7	EAJ-PNV	1,2	1,3
8	EH Bildu	0,8	0,8
9	CCa-PNC	0,3	0,2
10	Resto	2,4	1,9
	Total	100,0	100,0

Fuente: Ministerio del Interior del Gobierno de España y Centro de Investigaciones Sociológicas (CIS).

Como se aprecia, un cierto grado de discrepancia se pone de manifiesto en relación a los resultados del primer y tercer partido[6].

Las distribuciones de probabilidad pueden representarse gráficamente mediante un diagrama de barras o un histograma de la misma forma que las distribuciones de frecuencias relativas. En términos de probabilidad esta representación de datos poblacionales se denomina función de probabilidades y nos indica cuál es la probabilidad de que la variable aleatoria sea igual a un valor determinado:

$$f(x_i) = Pr(x = x_i) \qquad \textit{Ecuación 14}$$

donde, para todo i el valor de probabilidad es mayor o igual que cero, es decir, $f(x_i) \geq 0$ y la suma de las probabilidades es la unidad, esto es, se cumple que $\sum_{\min}^{\max} f(x) = 1$. Si la variable aleatoria X es continua, es decir, puede tomar un número infinito de valores (infinito no numerable[7]), se puede considerar como un caso límite

[5] Los datos se calculan a partir de la pregunta de intención directa de voto eliminando la falta de información: no sabe, no contesta, no vota, vota nulo y vota en blanco.

[6] Fueron unas elecciones donde se registraron generalizadas discordancias con los resultados finales, todas las empresas demoscópicas compartieron esos errores de estimación. Diversas razones podrían explicarlo pero finalmente se produjo un cambio de último momento en la definición del voto.

[7] En Teoría de Conjuntos, a partir de las contribuciones de George Cantor (1874), se establece la distinción entre conjuntos finitos e infinitos. En los conjuntos finitos existe un número (natural) que establece cuántos elementos contiene y esos elementos son numerables (por ejemplo, el número de partidos que se presentan a las elecciones).

de una población donde el polígono de frecuencias llega a ser una curva continua cuya ecuación $Y=\pi(x)$ se denomina función de densidad de probabilidades.

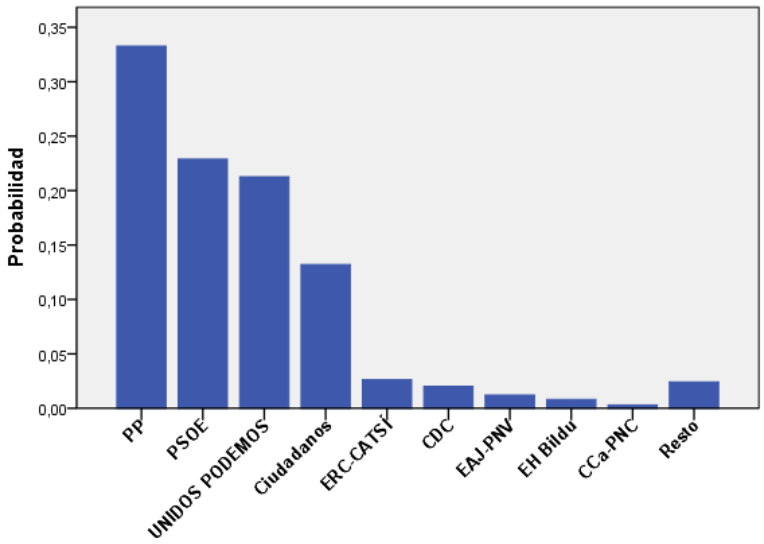
Además se pueden construir las distribuciones de frecuencias acumuladas. La función asociada a esta distribución es la llamada función de distribución, y nos indica cuál es la probabilidad de que la variable aleatoria sea menor o igual de un valor determinado:

$$F(x_i) = Pr(x \leq x_i) \qquad \textit{Ecuación 15}$$

Siendo una función no negativa, para todo i, $F(x_i) \geq 0$, no decreciente y acotada entre 0 y 1, $0 \leq F(x_i) \leq 1$.

Con el ejemplo del voto podemos representar el diagrama de barras con las probabilidades asociadas a cada partido político (Gráfico III.4.5), a partir de las frecuencias de los resultados electorales obtenemos la representación de la función de probabilidades. Con esta información, si quisiéramos realizar un estudio postelectoral sobre el comportamiento político, tanto este gráfico como la tabla de frecuencias nos indican la probabilidad de encontrar una persona de cada opción política en la extracción de una muestra aleatoria.

Gráfico III.4.5. Diagrama de barras de los resultados de las Elecciones Generales de 27 de junio de 2016



Fuente: Ministerio del Interior. Gobierno de España.

Como la variable carece de la propiedad de la distancia y de la continuidad de sus valores no nos es posible obtener un histograma ni un histograma de frecuencias o porcentajes acumulados. Pero si consideramos otras variables cuantitativas como el número de miembros del hogar o la edad de la población, según el Censo de Población y Vivienda de 2011, podemos obtener estas representaciones. La variable número de

En los conjuntos infinitos, que se definen por la negación de los finitos, implican un proceso de numeración que da lugar a una secuencia que nunca se detiene. En los conjuntos infinitos se establece además la distinción entre numerables y no numerables. Por ejemplo, el conjunto de los números naturales (1, 2, 3, 4, 5,…) es infinito y se puede numerar, pero el de los números reales (los que se expresan con decimales con una secuencia infinita de dígitos) es no numerable.

miembros (Tabla III.4.4 y Gráfico III.4.6) es discreta y disponemos de un número reducido de valores.

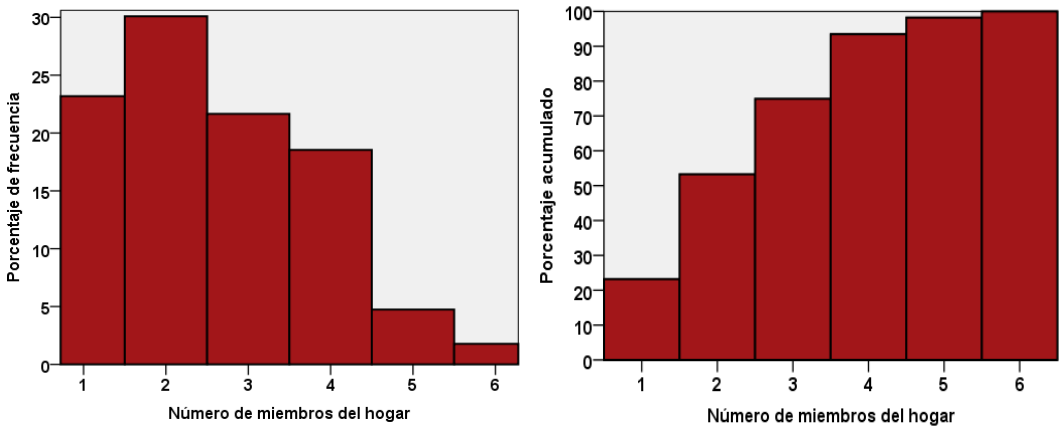
Tabla III.4.4. Tabla de frecuencias del número de miembros del hogar. España. 2011.

Número de miembros	Frecuencia	Porcentaje	Porcentaje acumulado
1	4.193.319	23,2	23,2
2	5.441.840	30,1	53,3
3	3.916.574	21,7	74,9
4	3.353.076	18,5	93,5
5	857.861	4,7	98,2
6 o más	321.022	1,8	100,0
Total	18.083.692	100,0	

Suceso (valor)	Probabilidad	Probabilidad acumulada
x_i	$\pi(x_i)$	$F(x_i)$
1	0,232	0,232
2	0,301	0,533
3	0,217	0,749
4	0,185	0,935
5	0,047	0,982
6 o más	0,018	1,000
Total	1,000	

Fuente: Instituto Nacional de Estadística. Censo de Población y Viviendas 2011

Gráfico III.4.6. Histograma e histograma de porcentaje acumulado del número de miembros del hogar. España. 2011



Fuente: Instituto Nacional de Estadística. Censo de Población y Viviendas 2011

Con estos datos nos podemos formular diversas preguntas:

- ¿Cuál es la probabilidad de que los hogares españoles tengan 2 miembros?

$Pr(x=2)=\pi(x_i=2)=0,301$

- ¿Cuál es la probabilidad de que los hogares españoles tengan 5 miembros?

$Pr(x=5)=\pi(x_i=5)=0,047$

- ¿Cuál es la probabilidad de que los hogares españoles tengan hasta 2 miembros?

$Pr(x\leq 2)=F(x_i=2)=0,533$

- ¿Cuál es la probabilidad de que los hogares españoles tengan hasta 5 miembros?

$Pr(x\leq 5)=F(x_i=5)=0,982$

- ¿Cuál es la probabilidad de que los hogares españoles tengan más de 2 miembros?

$Pr(x>2)=1-Pr(x\leq 2)=1-F(x_i=2)=1-0,533=0,467$

- ¿Cuál es la probabilidad de que los hogares españoles tengan más de 5 miembros?

$Pr(x>5)=1-Pr(x\leq 5)=1-F(x_i=5)=1-0,982=0,018$

- ¿Cuál es la probabilidad de que los hogares españoles tengan entre 3 y 5 miembros?

$$Pr(3 \le x_i \le 5) = F(x_i = 5) - F(x_i = 2) = 0,982 - 0,533 = 0,449$$
$$Pr(3 \le x_i \le 5) = \pi(x_i = 3) + \pi(x_i = 4) + \pi(x_i = 5) = 0,217 + 0,185 + 0,047 = 0,449$$

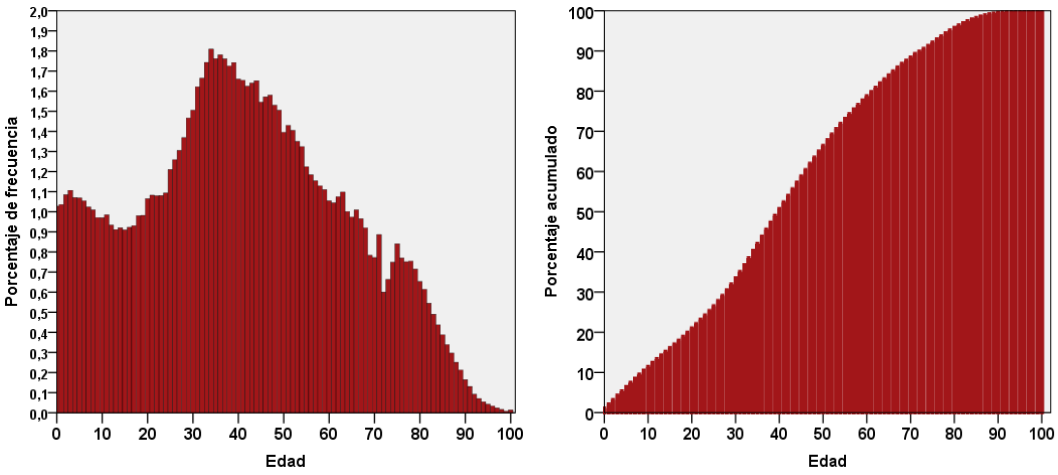
En el caso de la variable edad de la población (Tabla III.4.5 y Gráfico III.4.7) disponemos de un mayor número de valores y se puede dibujar una forma más continua de la distribución de la variable.

Tabla III.4.5. Tabla de frecuencias de la edad de la población. España. 2011.

	Frecuencia	Porcentaje	Porcentaje acumulado		Frecuencia	Porcentaje	Porcentaje acumulado
0	478.660	1,0	1,0	51	665.900	1,4	67,8
1	482.635	1,0	2,1	52	654.340	1,4	69,2
2	505.265	1,1	3,1	53	628.680	1,3	70,6
3	515.480	1,1	4,3	54	616.640	1,3	71,9
4	498.480	1,1	5,3	55	570.065	1,2	73,1
5	497.885	1,1	6,4	56	551.320	1,2	74,3
6	491.085	1,1	7,4	57	537.275	1,2	75,5
7	476.840	1,0	8,5	58	525.735	1,1	76,6
8	469.935	1,0	9,5	59	517.065	1,1	77,7
9	452.275	1,0	10,5	60	490.905	1,1	78,7
10	451.880	1,0	11,4	61	486.250	1,0	79,8
11	458.795	1,0	12,4	62	500.620	1,1	80,9
12	434.640	0,9	13,3	63	511.400	1,1	82,0
13	424.605	0,9	14,3	64	466.255	1,0	83,0
14	428.745	0,9	15,2	65	453.400	1,0	83,9
15	423.775	0,9	16,1	66	469.965	1,0	84,9
16	429.740	0,9	17,0	67	449.530	1,0	85,9
17	433.640	0,9	17,9	68	428.405	0,9	86,8
18	456.525	1,0	18,9	69	364.885	0,8	87,6
19	457.450	1,0	19,9	70	359.645	0,8	88,4
20	496.130	1,1	21,0	71	412.995	0,9	89,3
21	504.220	1,1	22,0	72	279.800	0,6	89,9
22	502.775	1,1	23,1	73	308.780	0,7	90,5
23	503.400	1,1	24,2	74	348.725	0,7	91,3
24	509.165	1,1	25,3	75	391.535	0,8	92,1
25	563.680	1,2	26,5	76	358.815	0,8	92,9
26	586.300	1,3	27,8	77	350.015	0,8	93,7
27	608.010	1,3	29,1	78	351.415	0,8	94,4
28	637.490	1,4	30,4	79	333.175	0,7	95,1
29	683.100	1,5	31,9	80	304.030	0,7	95,8
30	700.780	1,5	33,4	81	286.100	0,6	96,4
31	755.465	1,6	35,0	82	253.855	0,5	96,9
32	775.135	1,7	36,7	83	228.490	0,5	97,4
33	811.995	1,7	38,4	84	203.670	0,4	97,9
34	842.640	1,8	40,3	85	180.635	0,4	98,2
35	820.660	1,8	42,0	86	157.885	0,3	98,6
36	829.345	1,8	43,8	87	138.250	0,3	98,9
37	820.255	1,8	45,6	88	116.955	0,3	99,1
38	803.555	1,7	47,3	89	99.020	0,2	99,3
39	811.625	1,7	49,0	90	76.910	0,2	99,5
40	773.085	1,7	50,7	91	61.040	0,1	99,6
41	770.260	1,7	52,3	92	42.840	0,1	99,7
42	756.695	1,6	54,0	93	33.135	0,1	99,8
43	763.800	1,6	55,6	94	25.545	0,1	99,9
44	769.305	1,7	57,3	95	19.760	0,0	99,9
45	719.850	1,5	58,8	96	15.580	0,0	99,9
46	732.195	1,6	60,4	97	11.140	0,0	100,0
47	735.945	1,6	62,0	98	7.620	0,0	100,0
48	712.670	1,5	63,5	99	3.720	0,0	100,0
49	700.790	1,5	65,0	100	6.505	0,0	100,0
50	649.845	1,4	66,4	Total	46.574.720	100,0	

Fuente: Instituto Nacional de Estadística. Censo de Población y Viviendas 2011.

Gráfico III.4.7. Histograma e histograma de porcentaje acumulado de la edad de la población. España. 2011



Fuente: Instituto Nacional de Estadística. Censo de Población y Viviendas 2011.

De igual forma que en estadística descriptiva calculábamos una serie de medidas estadísticas de resumen de la distribución de una variable, con las distribuciones de probabilidad de una variable aleatoria se pueden calcular las mismas medidas para caracterizarlas. Nos centraremos en dos medidas, la media y la varianza, por tanto aplicables tan sólo a variables medidas a nivel de intervalo o de razón, y que se obtienen a través del concepto de esperanza matemática.

La esperanza matemática de una variable aleatoria es el valor promedio o valor esperado $E(X)$, denominado también momento de primer orden, que se calcula mediante la expresión:

$$E(X) = \sum_{i=1}^{k} \pi_i \cdot x_i = \mu \qquad \textit{Ecuación 16}$$

y se interpreta como la media aritmética aplicada a una variable aleatoria, expresada en términos poblacionales. Representa de hecho el valor promedio que se obtendría si realizáramos de forma aleatoria un gran número de veces una determinada observación. En esta fórmula si se sustituyen las probabilidades π_i por las frecuencias obtenidas en una muestra aleatoria de tamaño n, se obtiene la media aritmética $\overline{x}$. Cuando el tamaño de la muestra n aumenta, las frecuencias muestrales tienden a aproximarse a los valores de probabilidad, e igualmente la media muestral tiende a aproximarse a la media poblacional y conduce a interpretar que $E(X)$ representa la media poblacional.

Si se elige al azar una muestra de tamaño n de una población suponiendo que todas las muestras son igualmente probables, entonces se demuestra que el valor esperado de la media muestral $\overline{x}$ es la media poblacional μ.

La varianza de una variable aleatoria, denominada también momento de segundo orden, se define como el promedio o la esperanza matemática de las desviaciones cuadráticas respecto a la esperanza matemática y se calcula mediante la expresión:

$$Var(X) = E\left[x - E(x)\right]^2 = \sum_{i=1}^{k} \pi_i \cdot (x_i - \mu)^2 = \sigma^2 \qquad \textit{Ecuación 17}$$

Se entiende de la misma forma como la varianza poblacional. A diferencia de lo que sucede con la media, el valor esperado de la varianza muestral no es la varianza poblacional, sino este valor corregido por un factor: $\frac{n-1}{n} \cdot \sigma^2$.

Sobre los ejemplos que hemos visto del número de miembros y la edad, se presentan en la tabla siguiente los cálculos de la media (μ) y la varianza (σ^2) poblacionales:

	Número de miembros del hogar	Edad de la población
Media	2,57	40,82
Varianza	1,578	508,48

En el caso de la variable número de miembros el cálculo se realiza de la forma siguiente:

$$E(X) = \sum_{i=1}^{k} \pi_i \cdot x_i = 1 \times 0,232 + 2 \times 0,301 + 3 \times 0,217 + 4 \times 0,185 + 5 \times 0,047 + 6 \times 0,018 = 2,57$$

$$Var(X) = \sum_{i=1}^{k} \pi_i \cdot (x_i - \mu)^2 = 0,232 \times (1 - 2,57)^2 + 0,301 \times (2 - 2,57)^2 + 0,217 \times (3 - 2,57)^2 +$$
$$+ 0,185 \times (4 - 2,57)^2 + 0,047 \times (5 - 2,57)^2 + 0,018 \times (6 - 2,57)^2 =$$
$$= 1,58$$

En promedio, pues, el número de miembros de los hogares es de 2,57 personas, y, en promedio, la dispersión expresada a través de la varianza (la desviación cuadrática) es de 1,58.

1.3. La distribución normal

1.3.1. *Características de la distribución normal*

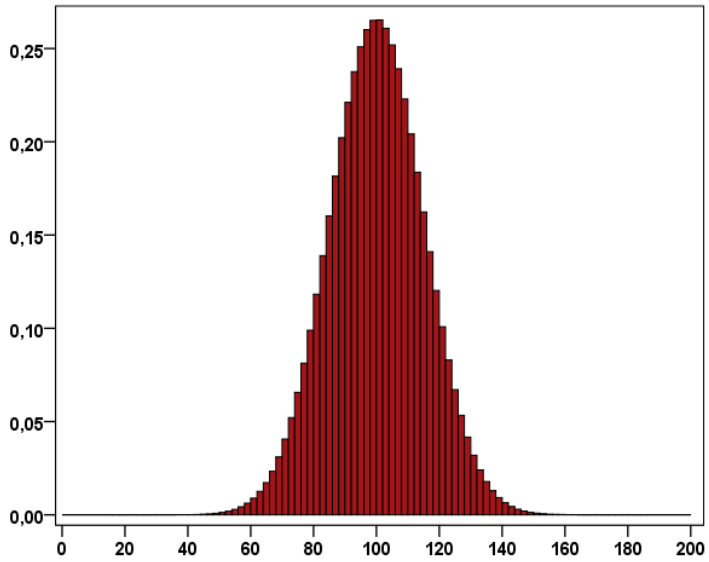
La distribución de una variable, como vimos en el capítulo anterior, puede caracterizarse por una medida de tendencia central y por una de dispersión, y también por su forma. Existen diferentes formas de una distribución, una de las más importantes utilizadas en estadística es la llamada distribución normal, caracterizada en particular por su simetría.

La distribución normal es una distribución teórica que marca un modelo ideal de comportamiento con el que se comparan las formas de las distribuciones empíricas obtenidas en los estudios de la realidad. Cuando se realizan observaciones a una

muestra de una población, en algunos casos, se encuentra que las distribuciones de las variables observadas se aproximan a esta forma de distribución teórica. En ciencias sociales no suele ser habitual, lo más "normal" es que no se observen distribuciones normales. Su utilidad se justifica por la importancia que adquiere en el terreno de la estadística inferencial, pues la distribución de los estadísticos, como la media por ejemplo, siguen una distribución normal. Además, como veremos, otras distribuciones teóricas, se derivan de ella.

Algunos ejemplos de comportamientos normales de variables empíricas observables son el coeficiente (o cociente) de inteligencia, los test estandarizados de conocimiento, la altura, el peso de las personas, el consumo de ciertos productos, entre otros. Los errores de medición siguen también ese comportamiento. La distribución de todas las medias que se pueden obtener de todas las posibles muestras de una población, con un mismo tamaño dado, siguen igualmente una distribución normal.

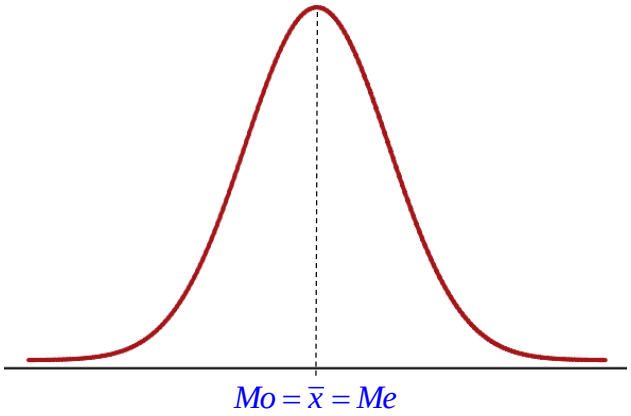
Gráfico III.4.8. Distribución del Coeficiente de Inteligencia con media 100 y desviación típica 15



La distribución normal fue reconocida por primera vez por el francés Abraham de Moivre (1667-1754). Posteriormente, Carl Friedrich Gauss (1777-1855) elaboró desarrollos más profundos y formuló la ecuación de la curva, de ahí que también se la conozca, más comúnmente, como la campana de Gauss.

La distribución normal es una distribución teórica de valores estandarizados (valores centrados y reducidos) que da lugar a una curva de probabilidad estandarizada con una forma perfectamente simétrica respecto del centro de la distribución y en forma de campana (Gráfico III.4.9).

Gráfico III.4.9. Forma de la curva normal (Campana de Gauss)



El área bajo la curva normal se interpreta en términos de probabilidades, proporciones o porcentajes de casos. La curva normal se caracteriza por tres aspectos:

- Es unimodal.
- Es simétrica, el área que queda por debajo de la curva se puede dividir en dos mitades exactamente iguales, cada una de las cuales contiene el 50% de los casos. Siendo una distribución unimodal y simétrica coinciden la moda, la mediana y la media en el centro de la distribución.
- Es asintótica con respecto al eje de abscisas, a medida que aumenta el valor en la escala de las abscisas la curva se aproxima al eje pero nunca llega a cortarlo. Por ello, cualquier valor entre $-\infty$ y $+\infty$ es teóricamente posible. El área total bajo la curva es, por tanto, igual a 1.

La distribución normal caracteriza teóricamente a las variables continuas, a aquellas que pueden tener un número infinito de valores y donde cualquier valor fraccionario es posible. No obstante, no todas las distribuciones empíricas tienen un rango infinito de valores, y, por otro lado, la distribución normal también puede emplearse como modelo para representar medidas discretas, con un error mínimo, siempre que el número de valores de la variable sea suficientemente alto.

La curva normal se corresponde con una expresión matemática[8], su función de densidad, donde intervienen la media (μ) y la desviación típica (σ) de la siguiente forma:[9]

$$Y = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$ *Ecuación 18*

En la ecuación, Y es la altura de la curva para un valor dado de X.

Cada variable que consideremos tendrá su propia media y desviación típica, por tanto, existen muchas curvas normales, una para cada combinación de media y desviación

[8] La ecuación matemática fue dada por Laplace en 1874.
[9] Esta ecuación se corresponde con una función, llamada función de densidad de la normal.

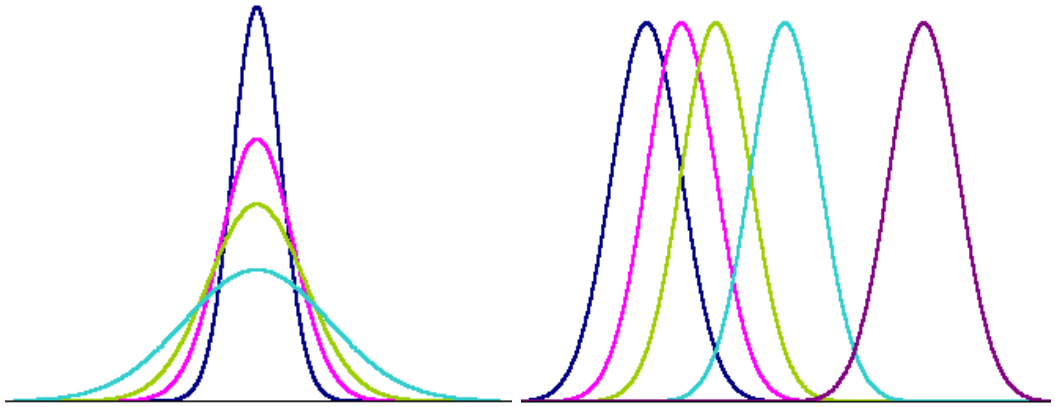
típica. Se dice que una variable X sigue una distribución normal de media μ y varianza σ^2, y se denota como:

$$X \sim N(\mu, \sigma^2)$$ *Ecuación 19*

con formas diversas que guardan las propiedades de ser unimodales, simétricas y asintóticas.

La media indica la posición de la curva, de modo que para diferentes valores de μ el gráfico se desplaza a lo largo del eje horizontal. Por otra parte, la desviación típica determina el grado de apuntamiento de la curva. Cuanto mayor sea el valor de σ, más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución. Un valor alto aumenta las probabilidades de encontrar datos más extremos.

Gráfico III.4.10. Distribuciones normales según diferentes medias y desviaciones



(a) Igual media y distintas desviaciones (b) Igual varianza y distintas medias

Por tanto, no existe una única distribución normal, sino una familia de distribuciones con una forma común, diferenciadas por los valores de su media y su varianza. De entre todas ellas, la más utilizada es la distribución normal tipificada, que corresponde a una distribución de media 0 y varianza 1. Si nos fijamos en la ecuación de la normal, el valor de X está estandarizado mediante la expresión $\frac{X-\mu}{\sigma}$, por tanto, podemos substituirla por la puntuación típica Z:

$$Z = \frac{X - \mu}{\sigma}$$ *Ecuación 20*

Así obtenemos la expresión tipificada de la curva normal que queda como:

$$Y = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot Z^2}$$ *Ecuación 21*

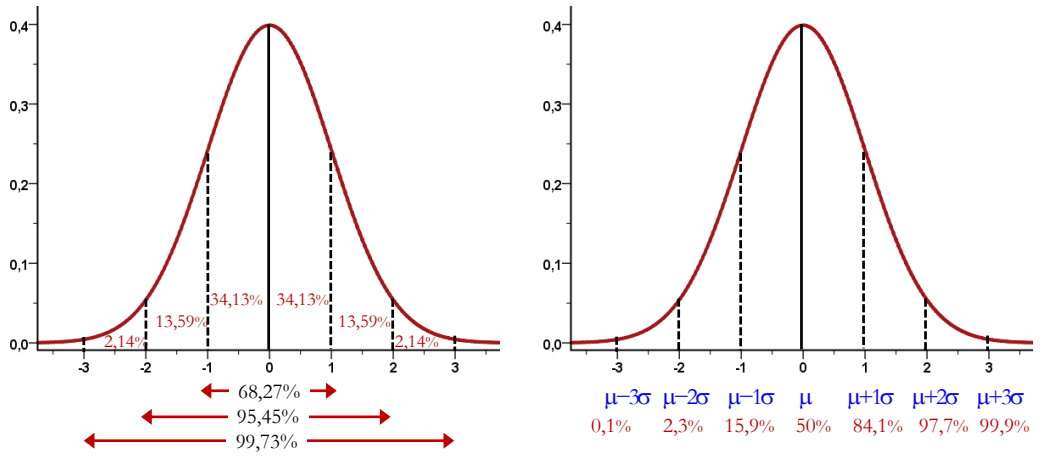
En este caso se dice que la variable X se distribuye normalmente con media 0 y varianza 1, o bien, $Z \sim N(0,1)$. Así pues, a partir de cualquier variable X que siga una

distribución normal, $X \sim N(\mu, \sigma^2)$, se puede obtener otra característica Z con una distribución normal tipificada, sin más que efectuar la transformación $Z = \frac{X - \mu}{\sigma}$.

Cuando tipificamos en valores Z, conseguimos expresar las variables en unidades de desviación, donde la media de todas las Z es 0 y su desviación típica es 1. Si la variable se distribuye normalmente (o es aproximadamente normal) esta característica se traduce en una propiedad de interés: el área que queda por debajo de la curva en el intervalo de valores definido entre la media menos un valor dado de unidades de desviación típica $(\mu - \sigma)$, y, la media más un valor dado de unidades de desviación típica $(\mu + \sigma)$, es decir, en el intervalo $(\mu - \sigma, \mu + \sigma)$, incluye siempre el mismo porcentaje de casos, ya que el área por debajo de la curva es constante e igual 1.

Así, por ejemplo, como se puede observar en el gráfico siguiente, entre la media y más/menos una unidad de desviación, es decir, entre $Z=-1$ y $Z=1$, se encuentra el 68,27% de los casos o el área por debajo de la curva, entre la media y más/menos dos unidades de desviación, entre $Z=-2$ y $Z=2$, se encuentra el 95,45%, y entre la media y más/menos tres unidades de desviación, entre $Z=-3$ y $Z=3$, se encuentra el 99,73% de los casos.

Gráfico III.4.11. Área por debajo de la curva de la normal

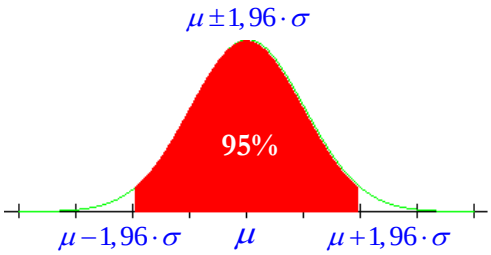


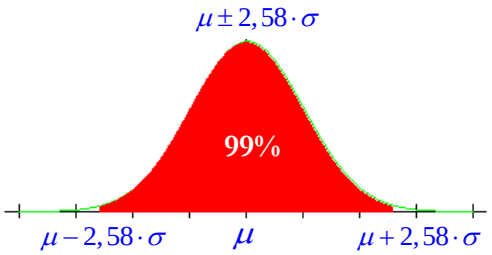
(a) Área central (b) Área acumulada desde la izquierda

Más adelante veremos el papel que juegan los valores Z de la distribución y el porcentaje del área asociado. En particular, valores entre −1,96 y 1,96 determinan el 95% central de la distribución, es decir, que el 95% de la población se encuentra en ese intervalo, mientras que entre −2 y 2 se encuentra el 95,45%. Si los valores son −2,58 y 2,58, entre ellos se se sitúa el 99% de los casos, mientras que entre −3 y 3 el porcentaje es del 99,73%.

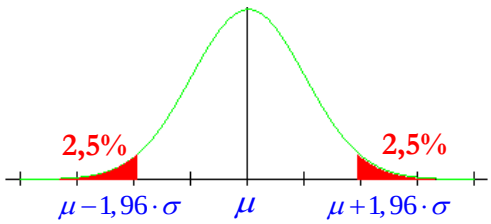
Los intervalos, que denominaremos intervalos de probabilidad o intervalos de confianza, se construyen a partir de la media μ, sumando y restando el número de

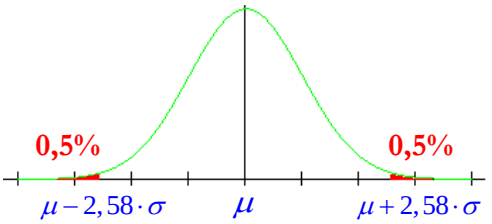
desviaciones σ. En los gráficos siguientes se representan esos intervalos considerando el 95% y el 99% del área central:





El área complementaria o restante de los extremos, también denominadas colas, acumula respectivamente el 5% y el 1%:





Veremos cómo estos porcentajes serán importantes para determinar la significación de un resultado estadístico, y determinarán en particular la región de aceptación o rechazo de una determinada hipótesis estadística, afirmando esa conclusión con un 95% o un 99% de confianza, o lo que es lo mismo, con un 5% o un 1% de nivel de significación. El nivel de significación se expresará con α y en tanto por uno: 0,05 o 0,01.

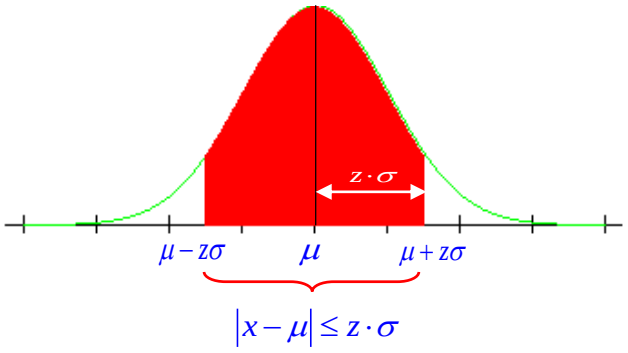
Por tanto, considerando una variable aleatoria X que se distribuye normalmente encontraremos que la mayoría de los casos están a una distancia de la media $|x-\mu|$ inferior o igual a un determinado número de desviaciones:

$$|x-\mu| \leq z \cdot \sigma \qquad \textit{Ecuación 22}$$

Se consideran, como lo hace la literatura estadística habitualmente, estos valores de unidades de desviación z:

Si z=1,28 entonces la probabilidad o el porcentaje de casos es del 90%.
Si z=1,96 entonces la probabilidad o el porcentaje de casos es del 95%.
Si z=2 entonces la probabilidad o el porcentaje de casos es del 95,5%.
Si z=2,58 entonces la probabilidad o el porcentaje de casos es del 99%.
Si z=3 entonces la probabilidad o el porcentaje de casos es casi del 100%.

Gráficamente, y de forma general para cualquier posible valor de z, se puede representar de la forma siguiente:



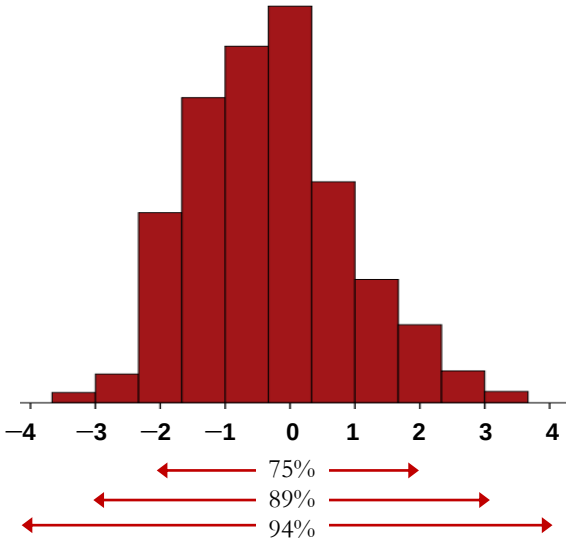
1.3.2. *La distribución normal y la desigualdad de Chebychev*

La desigualdad de Chebychev es un teorema estadístico que muestra unos resultados de interés en relación al comportamiento de las distribuciones estadísticas. De forma sencilla establece que en una distribución cualquiera de una variable aleatoria, como mínimo:

- El 75% del casos están a ±2 desviaciones típicas, entre $\mu - 2\sigma$ y $\mu + 2\sigma$.
- El 89% del casos están a ±3 desviaciones típicas, entre $\mu - 3\sigma$ y $\mu + 3\sigma$.
- El 94% del casos están a ±4 desviaciones típicas, entre $\mu - 4\sigma$ y $\mu + 4\sigma$.

Como se muestra en el Gráfico III.4.12.

Gráfico III.4.12. Desigualdad de Chebychev



Tal como hemos visto, en el caso en que la variable aleatoria siga una distribución normal estos valores se amplían y nos permiten ser más precisos en el conocimiento del comportamiento de la distribución, cumpliéndose, como hemos visto, que:

- El 95,5% del casos están a ±2 desviaciones típicas, entre $\mu - 2\sigma$ y $\mu + 2\sigma$.
- El 99,7% del casos están a ±3 desviaciones típicas, entre $\mu - 3\sigma$ y $\mu + 3\sigma$.
- Casi el 100% de los casos está a ±4 desviaciones típicas, entre $\mu - 4\sigma$ y $\mu + 4\sigma$.

1.3.3. *Cálculos con la distribución normal*

Estas características y propiedades de la distribución normal permitirán resolver diversas cuestiones de probabilidad acerca del comportamiento de variables de las que se sabe o se asume que siguen una distribución aproximadamente normal. En particular nos permitirá resolver dos problemas prácticos complementarios: determinar el porcentaje de casos que se corresponden con una valor determinado de Z, es decir, la probabilidad de que ocurra una observación dentro de los límites de área, y, de forma inversa, determinar qué valor de Z se corresponde a un porcentaje o intervalo dado de casos.

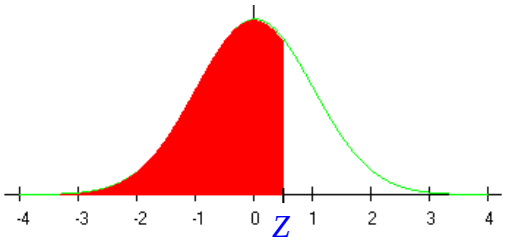
Si Φ es la función de distribución (función de probabilidades acumuladas) de una variable que se distribuye como una normal, con media 0 y varianza 1, es decir, $Z \sim N(0,1)$, entonces:

- la probabilidad que Z tenga valores más pequeños que z es: $P\{Z \leq z\} = \Phi(z)$

De forma similar podríamos considerar[10]:

- la probabilidad que Z tenga valores más grandes que z es: $P\{Z > z\} = 1 - \Phi(z)$
- la probabilidad que Z tenga valores más grandes o iguales que z es: $P\{Z \geq z\} = 1 - \Phi(z)$
- la probabilidad que Z tenga valores definidos por un intervalo entre a y b es: $P\{a \leq Z \leq b\} = \Phi(b) - \Phi(a)$

Para facilitar esta tarea no hará falta emplear la ecuación de la normal para efectuar los cálculos, se utiliza la tabla de la distribución normal (Tabla III.4.6), o bien utilizamos cualquier programa informático (SPSS, R, Excel) o alguna aplicación destinada específicamente a ello[11], que da la correspondencia entre las puntuaciones Z y las áreas o proporción de casos por debajo de la curva. Los elementos de esta tabla o el cálculo que se obtiene son las probabilidades de que una variable aleatoria con una distribución normal tipificada tenga un valor entre $-\infty$ y Z. Esta probabilidad representa el área sombreada bajo la curva:



En los márgenes de la Tabla III.4.6 se encuentran los valores de Z: en las filas se incluyen las unidades y el primer decimal, y en las columnas el segundo decimal.

[10] La probabilidad de un valor concreto es cero, sólo tienen probabilidad no nula los intervalos.

[11] Por ejemplo, se puede emplear la aplicación "gauss.exe" desarrollada por Jordi Lagares, http://www.xtec.cat/~jlagares/matemati.htm#ESTADISTICA, que utilizaremos a continuación para obtener también una representación.

La utilización de la tabla para determinar la proporción de casos que están por debajo de la curva a partir de una variable X, distribuida normalmente, requiere los siguientes pasos:

- Determinar la media (μ) y la desviación típica (σ) de la variable X.
- Estandarizar la variable mediante: $Z = \frac{X - \mu}{\sigma}$
- Localizar el valor de Z en los márgenes de la tabla de la normal, el resultado de cruzar la fila y la columna correspondiente nos dará el valor de la proporción.

La tabla Tabla III.4.6 contiene la probabilidad acumulada desde $-\infty$ hasta el valor crítico z, y considerando como primer valor crítico el que corresponden a la mitad de la distribución, es decir, al 0. Como se trata de una distribución simétrica, a partir de estos valores, los positivos de la distribución normal, se pueden derivar todos los demás.

Veamos un caso numérico concreto para ejemplificar el uso de la tabla de la normal. Suponemos la distribución normal de una variable X, por ejemplo los resultados de un ejercicio de evaluación, con notas entre 0 y 10, que arrojan un resultado de una media de 5 y una desviación típica de 2, es decir, una normal $N(5,4)$. A partir de estos datos nos planteamos las siguientes cuestiones:

1) ¿Qué proporción de casos tienen una evaluación de 7 o superior? O lo que lo mismo ¿qué probabilidad existe de obtener una nota 7 o más?
   - A partir de los valores de la media y la desviación procedemos a calcular en primer lugar la puntuación típica Z. En este caso se trata de considerar un área en el intervalo entre el 7 e infinito. En primer lugar tipificamos el valor:

     $$Z = \frac{X - \mu}{\sigma} = \frac{7 - 5}{2} = 1$$

     Es decir, el valor 7 se encuentra a 1 unidad de desviación típica con respecto a la media.
   - A continuación buscamos en la tabla de la normal el valor de probabilidad que corresponden a $Z=1$. La tabla proporciona la información del área acumulada entre $-\infty$ y un valor dado a partir de la media. En nuestro caso, hasta 1 unidad de desviación es el valor 0,84134, es decir, se acumula el 84,13% del área que queda a la izquierda, son los casos que obtienen hasta un 7. Como el total del área es 1, la proporción que corresponde a obtener 7 o más es la diferencia entre 1 y 0,84134, es decir, 0,15866. Por tanto, el porcentaje del área a partir del 7 es del 15,87% de la curva. Existe una probabilidad del 0,15866 de encontrar una evaluación de 7 o superior. Gráficamente corresponde a la zona sombreada:

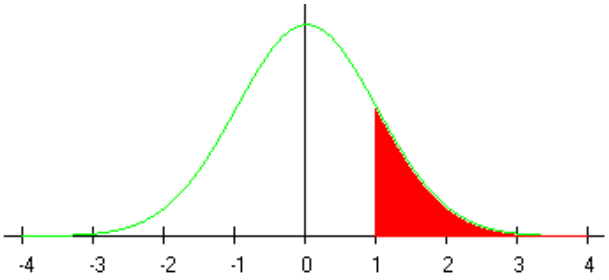


Tabla III.4.6. Tabla de la distribución normal tipificada, N(0,1)

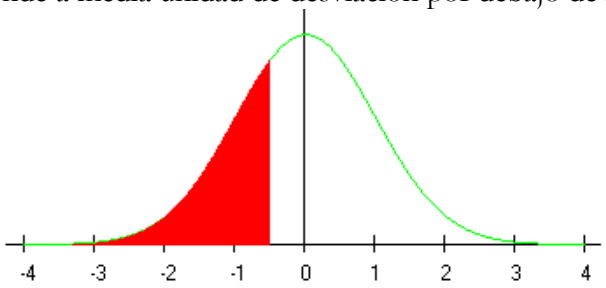
Probabilidad acumulada desde $-\infty$ hasta el valor crítico z

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4,0	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998

2) ¿Qué proporción de alumnos han obtenido una nota inferior a 4?
   - Calculamos la puntuación típica Z correspondiente al 4:

$$Z = \frac{X - \mu}{\sigma} = \frac{4-5}{2} = -0,5$$

Un 4 corresponde a media unidad de desviación por debajo de la media.



   - Para determinar el valor de área de la normal que se acumula hasta −0,5 unidades de desviación buscamos el valor de la tabla que corresponde al valor tipificado positivo de 0,5, este valor es 0,69416. Como se trata de una distribución simétrica, el área por encima de 0,5 es la misma que corresponde a por debajo de −0,5. Este valor se obtiene restando el total del área por debajo de la normal, 1, menos el valor 0,69146, es decir, un área de 0,30854. En consecuencia, un 30,85% del alumnado obtiene una nota inferior a 4.

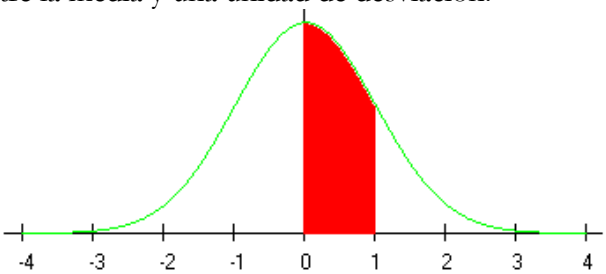
3) ¿Qué proporción de alumnos han obtenido un resultado entre 5 y 7?
   - A partir de los valores de la media y la desviación procedemos a calcular la puntuación típica Z. En este caso se trata de saber cuántos casos hay entre la media 5 y el valor 7 de la distribución, por tanto, la tipificación se expresa como:

En el caso de 7 $Z = \frac{X - \mu}{\sigma} = \frac{7-5}{2} = 1$

En el caso de 5: $Z = \frac{X - \mu}{\sigma} = \frac{5-5}{2} = 0$

Es decir, como antes, el valor 7 se encuentra a 1 unidad de desviación típica con respecto de la media, y 5 corresponde al valor de la media por lo que está a 0 unidades de desviación.

   - A continuación buscamos en la tabla de la normal los valores de probabilidad que corresponden a $Z=0$ y $Z=1$ para saber la proporción de casos que se encuentran entre la media y una unidad de desviación.



En el cruce de la fila correspondiente al 1 y la columna del 0,00, aparece el valor 0,84134, que es el área acumulada desde $-\infty$ hasta 1. La proporción de área que se acumula hasta $Z=0$ en la normal tipificada es la mitad del área por debajo de la curva, es decir, 0,5, como se puede ver en la primera casilla de la tabla. Por tanto, el área entre 0 y 1 es la diferencia entre 0,84134 y 0,5, el valor 0,34134, es decir, el 34,13% de la curva. Esto es, el 34,13% de los alumnos obtienen una

nota entre 5 y 7, o lo que es lo mismo, existe una probabilidad de 0,34134 de obtener una nota en ese intervalo.

Como hemos señalado, también se puede tratar el problema inverso de determinar un valor concreto que corresponda a una proporción dada de casos que están por debajo de la curva. Si α es un número entre 0 y 1, el percentil $\alpha \times 100$ de la distribución $N(0,1)$ es el valor:

$$z_\alpha \text{ tal que } P\{Z \leq z_\alpha\} = \alpha$$

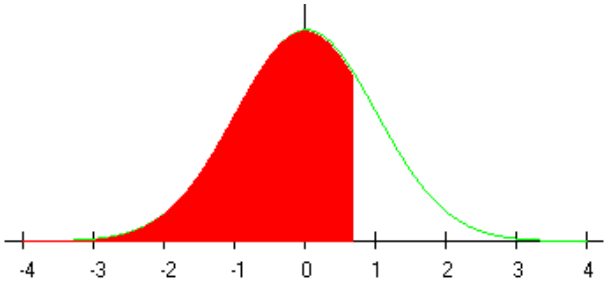
Si consideramos una distribución $N(\mu,\sigma^2)$, el percentil $\alpha \times 100$ es el valor:

$$x_\alpha \text{ tal que } P\{X \leq x_\alpha\} = \alpha$$

En este caso, a partir de una variable X, distribuida normalmente, se requieren los siguientes pasos:

- Localizar la proporción del área por debajo de la curva e identificar el valor de Z correspondiente en los márgenes de la tabla de la normal.
- Determinar el valor de X, conociendo los valores de la media (μ) y la desviación típica (σ) de la variable X, a partir de la expresión de los valores tipificados: $Z = \frac{X - \mu}{\sigma}$, despejando X resulta que: $X = \mu + Z \cdot \sigma$.

Siguiendo el ejemplo anterior, podríamos plantear ¿a qué valor corresponde el 75% de la distribución? Un área del 75% corresponde aproximadamente a un valor de $Z=0,67$[12]. El valor de X será pues: $X = \mu + Z \cdot \sigma = 5 + 0,67 \cdot 2 = 6,34$.



Estos mismos cálculos pueden realizarse con la ayuda del ordenador. Veamos cómo sería con el software SPSS. Los cálculos se realizan con el comando **COMPUTE** (Calcular en el menú Datos) y se utilizan palabras clave que identifican diversas funciones tanto para la distribución normal como para otras distribuciones. Las funciones se especifican en un prefijo de la palabra clave y la distribución se recoge en un sufijo. Como prefijo utilizaremos principalmente la función de distribución (Cumulative Distribution Function) CDF.d_spec(x,a,...) que devuelve una probabilidad p de que una variable aleatoria con la distribución especificada (d_spec) caiga por debajo de x para funciones continuas y en o por debajo de x para funciones discretas. Empleando la distribución normal el comando de SPSS es:

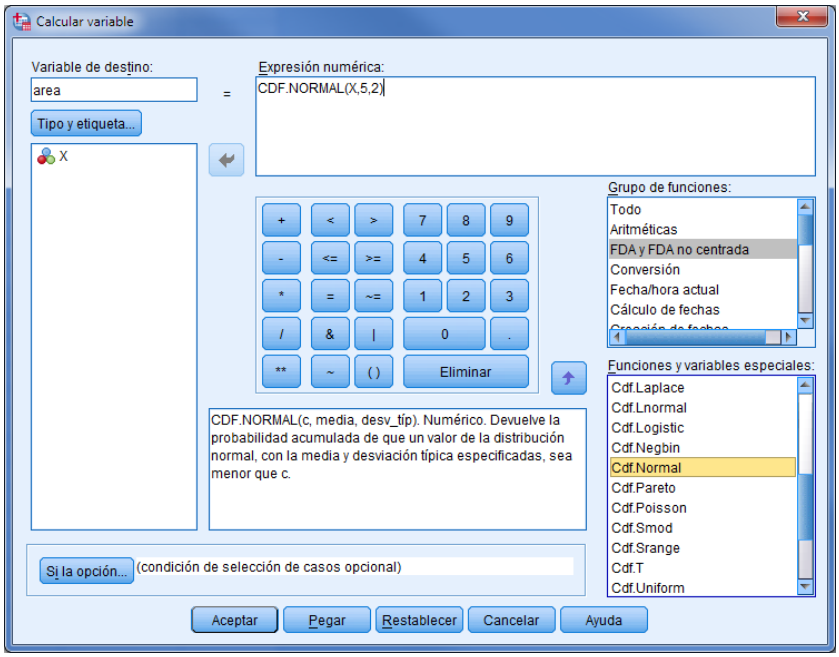
[12] Con la ayuda del ordenador se podrá calcular de forma exacta.

CDF.NORMAL(cant, media, desv_típ) y devuelve la probabilidad acumulada de que un valor de la distribución normal, con la media y la desviación típica especificadas, sea menor que la cantidad cant[13]. Si se considera **CDFNORM(valorz)** es el caso de una variable aleatoria con media 0 y desviación típica 1.

Para realizar los cálculos del ejemplo anterior es necesario crear primero una matriz de datos con una columna con los valores de la variable *X* (con los valores 4, 5 y 7) y ejecutar la siguiente instrucción a través del editor de sintaxis:

```
COMPUTE area=CDF.NORMAL(X,5,2).
```

o bien a través del menú Transformar / Calcular variable, eligiendo en el grupo de funciones FDA y FDA no centrada la función de la probabilidad acumulada de la normal:



donde se crea la variable area que calcula la proporción del área o probabilidad acumulada hasta el valor de *X*. El resultado se puede ver en la imagen de la matriz de datos que se obtiene:

	X	area
1	4	,30854
2	5	,50000
3	7	,84134

[13] Otras funciones son IDF, función de distribución inversa; PDF, la función de densidad de probabilidad; RV, la función aleatoria de generación de números; NCDF, función de distribución acumulada no central; NPDF, función de densidad de probabilidad no central; SIG, función de probabilidad de la cola. Otras distribuciones son: UNIFORM, distribución uniforme; WEIBULL, distribución de Weibull. BERNOULLI, distribución de Bernoulli; BINOM, distribución binomial; GEOM, distribución geométrica; HIPER, distribución hipergeométrica; NEGBIN, distribución binomial negativa; POISSON, distribución de Poisson; T, t de student; F, distribución de Fisher-Snedecor; CHISQ, distribución de chi-cuadrado; LOGISTIC, distribución logística, PARETO, distribución de Pareto, entre otras más.

Con estos datos reproduciríamos los cálculos e interpretaciones que comentamos anteriormente. Para obtener el valor que corresponde el 75% de la distribución, ejecutaríamos la instrucción de la función inversa:

```
COMPUTE z=IDF.NORMAL(0.75,5,2).
```

que da como resultado 6,35.

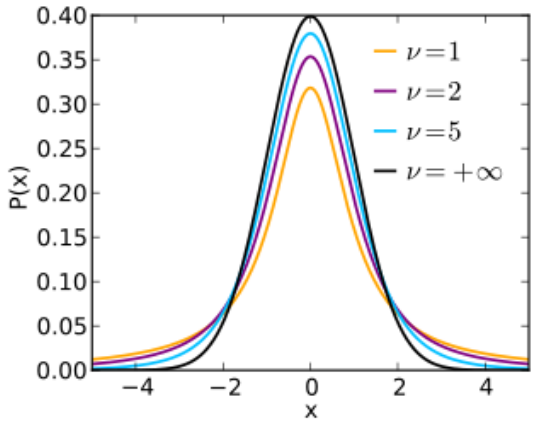
1.4. La distribuciones t, F y χ^2

La distribución normal es la principal función de distribución estadística, aunque existen otras más que se utilizan frecuentemente en diferentes técnicas de análisis de datos. Presentamos de forma breve tres de las más habituales: la t de student, la F de Fisher-Snedecor y la de Chi-cuadrado (χ^2).

1.4.1. *La distribución t de student*

La distribución recibe ese nombre por el apodo que utilizó William Sealy Gosset (1876–1937), estadístico y químico que trabajó en la Cervecería Guinness, para ocultar su identidad en sus publicaciones. Se trata de una distribución semejante a la normal pero con más colas, es decir, se hace más plana. Se emplea en estadística cuando no se conoce la varianza poblacional y para estimar la media poblacional con muestras pequeñas. Existen diferentes curvas posibles que varían según un parámetro denominado grados de libertad (ν)que expresa o depende del tamaño de la muestra n (se calcula como $\nu = n-1$)[14].

Gráfico III.4.13. Distribución t de student



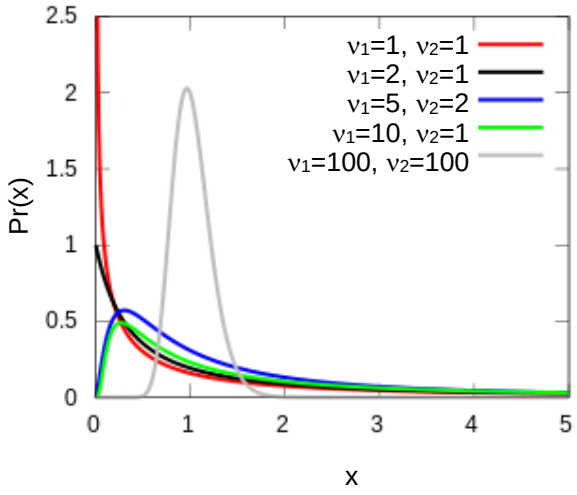
[14] Los grados de libertad (*degrees of freedom* en inglés) es un concepto matemático que podemos asociar a la idea del espacio de libertad en que una medida de resumen puede moverse y tomar diferentes valores (De La Cruz-Oré, 2013). Por ejemplo, podemos elegir al azar 3 personas jóvenes, a condición que en promedio tengan 20 años, por ejemplo 18, 20 y 21. En cualquier muestra, las dos primeras que elijamos puede tener las edades que queramos, por ejemplo 18 y 22 o 21 y 22, pero la tercera no puede tener cualquier edad para que la media sea 20; en el primer caso la persona debe tener 20 y en el otro caso 17. Los grados de libertad son el número de casos menos uno, en este ejemplo 2. Otro ejemplo que nos aproxima al concepto de grados de libertad es el conocido juego japonés del Sudoku, donde se trata de ir colocando los números del 1 al 9 en sus casillas según la información del resto de casillas, los valores se ponen solamente cuando se reducen los grados de libertad (como mucho 9 grados de libertad) a uno. En estadística los grados de libertad se determinan por el número de casos de la muestra y también por el número de variables y valores, que no son más que restricciones que tienen nuestros datos, siendo por tanto los grados de libertad el número de observaciones independientes.

En un apartado posterior de este capítulo y en los capítulos III.8 de análisis de varianza y III.9 de análisis de regresión veremos su aplicación. En el Tabla de distribución teórica de la t de StudentAnexo I se pueden consultar los valores de la tabla de la t de student.

1.4.2. *La distribución F de Fisher-Snedecor*

A partir de la contribución de R. Fisher (1890–1962) y G. W. Snedecor (1881–1974) contemplamos otra distribución teórica utilizada en los análisis de varianza para realizar comparaciones entre más de dos medias. Se trata de una distribución que no es simétrica, a diferencia de las anteriores, y depende de dos parámetros que definen los grados de libertad ν_1 y ν_2, configurando así diversas representaciones de la curva. Como veremos se trata de una distribución similar a la de Chi-cuadrado.

Gráfico III.4.14. Distribución F de Fisher-Snedecor

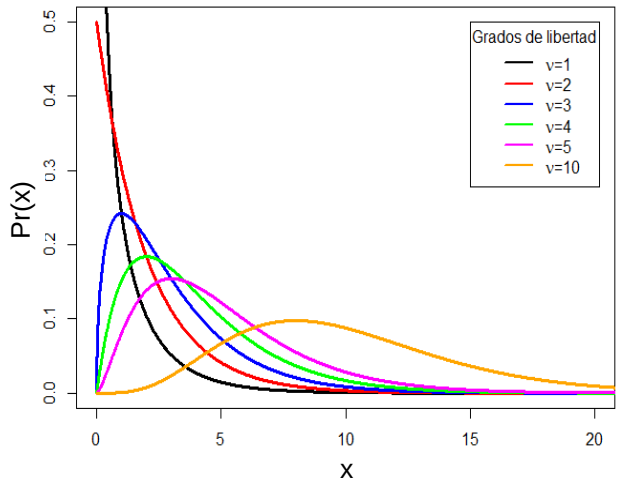


En los capítulos III.8 de análisis de varianza y III.9 de análisis de regresión veremos su aplicación.

1.4.3. *La distribución de Chi-cuadrado*

Se conoce como la distribución de Chi-cuadrado de Pearson (1857–1936) y como distribución teórica permite la comparación o contraste entre dos distribuciones diferentes, y en particular para establecer la asociación entre dos variables cualitativas en una tabla de contingencia. Es una distribución no simétrica que también se corresponde con una familia de distribuciones dependiendo de los grados de libertad ν de cada caso.

Gráfico III.4.15. Distribución de Chi-cuadrado



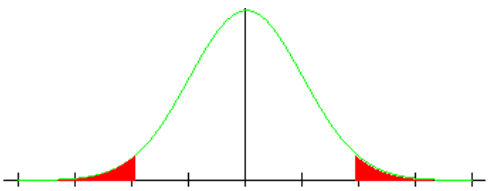
En el capítulo III.7 de análisis de tablas de contingencia y III.7 de análisis log-lineal veremos su aplicación.

1.5. La significación de un valor de la distribución

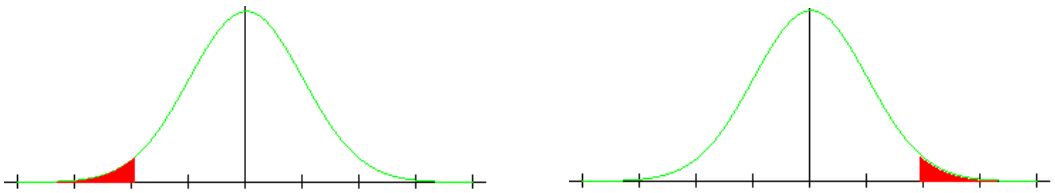
Las distribuciones estadísticas teóricas que hemos visto nos proporcionan la probabilidad de que una variable que sigue una de esas distribuciones tenga valores más extremos que uno dado *x*. Un valor de una distribución se valora en términos probabilísticos como un valor muy extremo, es decir, poco probable, o como no tan extremo y más probable. El valor extremo del 5% se considera habitualmente como el umbral a partir del cual cabe entender que el valor es significativo (menor o igual de ese 5%) o no significativo (mayor al 5%). Ese valor del 5%, o del 0,05 en términos de probabilidad, marca la separación entre la región de aceptación (nivel de confianza del 95%) y la región de rechazo (nivel de significación del 5%). Cuando hablemos de los contrastes de hipótesis veremos cómo se aplican estos conceptos.

En este tipo de razonamiento se pueden diferenciar dos situaciones en función de si la distribución es simétrica, como la normal o la t de student, o no es simétrica, como la F de Fisher-Snedecor o la Chi-cuadrado.

En el primer caso la simetría permite realizar valoraciones de la significatividad a dos colas (significación bilateral o a dos colas), valorando los dos extremos de la distribución:



o a una cola (significación unilateral o a una cola), valorando el extremo negativo o el extremo positivo según el signo del valor del que se trate cuando se establece la significación:

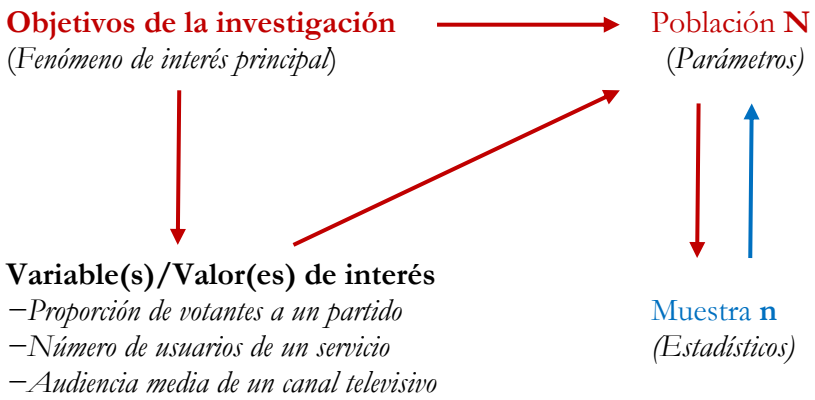


o porque la distribución, como la F de Fisher-Snedecor o la Chi-cuadrado solamente tienen valores positivos.

2. Estimación e intervalos de confianza

En la investigación social siempre se plantea de manera genérica la necesidad de indagar en el conocimiento de determinadas características de una sociedad. En el contexto de un estudio con datos estadísticos obtenidos por muestreo cabe plantear una reflexión general que esquematizamos en el **¡Error! No se encuentra el origen de la referencia.**. Inicialmente las características que quieren ser estudiadas están definidas por los objetivos de la investigación y por una reflexión teórica que ayudan a precisar y definir el fenómeno de interés de estudio. Así, por ejemplo, podríamos tener interés en conocer el comportamiento electoral de una población en las próximas elecciones, o bien saber las características de los potenciales usuarios de un servicio social del ayuntamiento, o bien determinar la audiencia potencial de un canal de televisión.

Gráfico III.4.16. Planteamiento de los estudios por muestreo estadístico



Si la información que disponemos hace referencia a toda la población, como es el caso de los datos censales o, por ejemplo, en el caso de disponer del escrutinio de una consulta electoral o de la información de todos los asociados/as de una organización, entonces tenemos un conocimiento exhaustivo y cierto de las características que queremos conocer: el porcentaje de la población ocupada o desocupada, la distribución de la población según su condición socioeconómica, el número de jóvenes emancipados, el porcentaje de votos que obtiene cada partido político en las elecciones, etc.

Esta información de toda la población constituye un valor cierto, es el resultado de contabilizar una característica de todos y cada uno de los individuos de esta población. Sin embargo, no es lo más habitual disponer de información exhaustiva de cada una de las unidades poblacionales, al contrario, la investigación social se nutre de estudios y datos que son el resultado de la selección de una parte de la población, una muestra estadística, fundamentalmente por el elevado coste económico, de tiempo y recursos materiales que supone llegar a tener información de cada unidad de la población. El objetivo final es el mismo, obtener información de determinadas características de la población, pero se opta por la extracción de una muestra aleatoria de un conjunto de unidades de esta población con las que se estima aquella característica poblacional.

Cuando procedemos de esta forma, decimos que hacemos estimaciones estadísticas (calculamos estadísticos muestrales, como una media o una proporción) de las características poblacionales (los mismo cálculos referidos a toda la población que se denominan parámetros poblacionales). Al ser estimaciones, los resultados o las medidas que obtenemos están afectadas por un determinado margen de error, por el hecho de disponer tan sólo de una parte de la población (la muestra) con la que se quiere conocer un rasgo para todo el conjunto poblacional. Las estimaciones más el margen de error constituyen el intervalo de confianza. Sobre estas cuestiones tratemos seguidamente.

2.1. Conceptos: parámetros, estimadores y estimaciones

Consideremos una población con N individuos donde el tamaño N de esta población es "muy grande" comparativamente con el tamaño de la muestra de tamaño n, que suele ser la situación habitual en los sondeos o en los estudios sociales por muestreo. En términos matemáticos se dice que esa población es infinita. Desde el punto de vista del análisis estadístico, identificamos la población con una variable X de interés: tenemos una población de valores de la variable. Definamos a continuación los conceptos de parámetros, referidos a la población, y de estimadores y estimaciones, referidos a la muestra estadística.

2.1.1. *Los parámetros poblacionales*

Las características que queremos conocer o estimar de la población se denominan parámetros poblacionales: son los valores verdaderos que caracterizan a la población y que se suelen expresar en términos de algún cálculo aritmético que afecta a toda la población.

Los parámetros poblacionales son las características de posición, de dispersión, etc., más relevantes de la variable, por ejemplo:

- La media poblacional de una variable numérica, un parámetro de posición central. Por ejemplo, la media de visitas diarias a un centro comercial, la audiencia media de un medio de comunicación, el salario o los ingresos medios de un grupo social, etc. La media poblacional de una variable X también se llama valor esperado o esperanza y, habitualmente, se denota por μ. Con un tamaño poblacional de N individuos, la media es:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

- La varianza poblacional de una variable numérica, un parámetro que mide la dispersión. Habitualmente, denotamos la varianza poblacional por σ^2 y por σ a la desviación típica poblacional, sus expresiones son:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \quad \text{y} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

  La expresión de la varianza cuando se divide por N−1: $S^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N-1}$, se denomina varianza ajustada.

- La proporción (o el porcentaje) de presencia de cierta característica de la población. Por ejemplo, la proporción de los que ven cada día la televisión, el porcentaje favorable a la despenalización del aborto, el porcentaje de parados, etc. La denotamos con P:

$$P = \frac{C}{N} \quad \text{o bien} \quad P = \frac{C}{N} \times 100$$

  La proporción se puede considerar un caso particular de variable dicotómica y utilizarla como una variable cuantitativa a través de una codificación dicotómica o binaria, que transforma los valores de una variable con tan sólo dos valores de presencia-ausencia de cierta característica. Por ejemplo, la variable cualitativa sobre el nivel educativo (diferenciando tres valores: primario, secundario y terciario) se puede considerar como una (o más) variable(s) con solo dos valores, siempre en relación a cada nivel educativo: "Tiene nivel primario" vs. "No tiene nivel primario", "Tiene nivel secundario" vs. "No tiene nivel secundario" y "Tiene nivel terciario" vs. "No tiene nivel terciario". En los tres casos estos valores se codifican como 0 (para el valor que niega una característica A, es decir, la ausencia de A) y 1 (para el valor que refleja la característica A, su presencia):

$$X = \begin{cases} 0, & \textit{ausencia de la característica A} \\ 1, & \textit{presencia de la característica A} \end{cases}$$

  Si P es la proporción de casos que presentan la característica A en la población, tendremos que:

$$P = \frac{\textit{Número de individuos de la población con la característica A}}{\textit{Número total de individuos en la población}} = \frac{C}{N}$$

  siendo C el número de casos con 1, los individuos que poseen la característica A. Con variables dicotómicas o binarias codificadas de esta forma es posible calcular la media y darle un sentido. Se demuestra fácilmente que, para una variable X con valores 0 y 1, su media, μ, es la proporción P de casos que presentan la característica A. En efecto, como los valores X_i son solo 0 y 1, tenemos que:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{1+1+\cdots+1+0+0+\cdots+0}{N} = \frac{C}{N} = P$$

También se demuestra que la varianza de una variable dicotómica es el producto de P por Q, donde Q es el complementario de P (los que no presentan la característica A) e igual a $1-P$, por tanto, $P+Q=1$. Operando la fórmula de la varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N} X_i^2 - \mu^2}{N} = \frac{1^2 - P^2 + \cdots + 1^2 - P^2 + 0^2 - P^2 + \cdots + 0^2 - P^2}{N} =$$

$$= \frac{A - NP^2}{N} = P - P^2 = P \cdot (1-P) = P \cdot Q$$

con $Q=1-P$ (o bien $Q=100-P$ si se consideran porcentajes).

De todo lo anterior deducimos que la proporción es un caso particular de media, e igual a P, con una varianza igual a $P \times Q$.

Veamos un ejemplo a partir de los datos de la Tabla III.4.7 sobre el nivel de estudios de la población española según el Censo de Población de 2011.

Tabla III.4.7. Nivel de estudios de la población. España. 2011.

Grado	Frecuencia	Porcentaje
1 Analfabetos	729.860	1,6
2 Sin estudios	3.538.180	7,6
3 Primer grado	5.819.905	12,5
4 Segundo grado	21.508.105	46,2
5 Tercer grado	7.487.685	16,1
6 No es aplicable	7.490.990	16,1
Total	46.574.725	100,0

Fuente: Instituto Nacional de Estadística.
Censo de Población y Viviendas 2011

Sobre un total poblacional N de 46.574.725 personas, la proporción de personas que alcanzan el grado terciario es de $P=0,161$, el 16,1%. Por tanto, la proporción complementaria será $Q=1-P=0,839$, es decir, la población que no tiene estudios superiores. Centrando la atención en este nivel educativo podemos considerar la variable dicotómica X:

$$X = \begin{cases} 0, \text{ ausencia de la característica, no tiene el tercer grado} \\ 1, \text{ presencia de la característica, tiene el tercer grado} \end{cases}$$

La distribución de la nueva variable es:

X Tercer Grado	Frecuencia	Porcentaje
0 No lo tiene	39.087.040	83,9
1 Lo tiene	7.487.685	16,1
Total	46.574.725	100,0

Si calculamos la media de la variables dicotómica X obtendríamos que $\mu = P = 0{,}161$ y la varianza sería $\sigma^2 = PQ = 0{,}161 \times 0{,}839 = 0{,}135$. Lo mismo podríamos hacer con el resto de categorías educativas al convertirlas en variables dicotómicas.

2.1.2. *Los estadísticos muestrales*

Cuando no se pueden tener las observaciones de toda la población respecto de la característica de interés que se quiere estudiar, no conocemos la distribución de la variable poblacional X, y no podemos calcular los valores de los parámetros poblacionales, nos encontramos en la necesidad de escoger una muestra aleatoria de observaciones independientes para efectuar una estimación de los valores de esos parámetros.

Denotamos por $x_1, \ldots, x_n$ a los posibles valores de una muestra aleatoria de tamaño n. Con una formulación análoga a la de los parámetros poblacionales, se definen los mismos indicadores para la muestra, llamados estadísticos muestrales: la media muestral $\overline{x}$, la varianza muestral σ^2 , la desviación típica muestral σ y la proporción muestral p .

Tabla III.4.8. Principales parámetros poblacionales y estadísticos muestrales

	Parámetros poblacionales	**Estadísticos muestrales**
Media	$\mu = \dfrac{\sum_{i=1}^{N} X_i}{N}$	$\hat{\mu} = \overline{x} = \dfrac{\sum_{i=1}^{n} X_i}{n}$
Varianza	$\sigma^2 = \dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$	$\hat{\sigma}^2 = s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$
Varianza ajustada	$S^2 = \dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{N-1}$	$\hat{S}^2 = s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$
Proporción	$P = \dfrac{C}{N}$	$\hat{P} = p = \dfrac{c}{n}$
Porcentaje	$P = \dfrac{C}{N} \times 100$	$\hat{P} = p = \dfrac{c}{n} \times 100$

En la Tabla III.4.8 se presenta un cuadro de resumen con la formulación de las medidas más importantes. Obsérvese la diferente notación empleada: los parámetros se expresan en letras griegas y en mayúsculas mientras que los estadísticos contienen letras latinas y minúsculas. Para simbolizar que un estadístico muestral estima un parámetro poblacional, se suele escribir el símbolo poblacional con un sombrero; por ejemplo, $\hat{\mu} = \bar{x}$ que significa que la media muestral estima la media poblacional.

Para expresar que utilizamos un estadístico para estimar un parámetro decimos que el estadístico es un estimador del parámetro. Los valores concretos que se obtienen de una muestra reciben el nombre de estimaciones o bien estimaciones puntuales.

Por ejemplo, supongamos desconocida la media de ingresos μ de la población de un municipio. Una estimación de este valor se puede obtener tomando una muestra aleatoria de sus habitantes. Pongamos que obtenemos una muestra de $n=1.000$ personas a las que les preguntamos por sus ingresos, y con las 1.000 respuestas u observaciones calculamos la media de los ingresos. Supongamos que los ingresos medios obtenidos fueran $\bar{x} = 1.500€$, la estimación puntual de la media poblacional es $\hat{\mu} = 1.500€$. Tengamos presente que en otra posible muestra de 1.000 personas hubiera dado unos ingresos medios diferentes, por ejemplo de 1.600€. De hecho, cada posible muestra generaría puntualmente un resultado distinto, lo que nos está indicado la existencia de una variabilidad y de que estamos cometiendo un error de estimación. Este es un aspecto crucial que retomaremos y nos conducirá a hablar de estimaciones por intervalos teniendo en cuenta el error que se comete.

Otro ejemplo se puede plantear con el estadístico muestral de la proporción. Imaginemos que hay elecciones al rectorado de la universidad y se presentan dos candidaturas, A y B. Desconocemos el porcentaje de votos que obtendrá cada candidatura. Para estimar la intención de voto de la comunidad universitaria realizamos un sondeo encuestando a 800 personas del censo del profesorado, estudiantado y personal de administración y servicios. Si los resultados muestran que un 25% se manifiesta como partidario de la candidatura A y un 75% partidarios de la candidatura B, en proporciones, tendremos que $p_A = 0,25$ como estimación puntual de la proporción poblacional, P_A, desconocida, de la candidatura A. De forma similar podríamos comentar el resultado de la candidatura B con la proporción estimada en este caso de $p_B = 0,75$. Como en el ejemplo anterior, otra muestra habría producido estimaciones diferentes, si bien en este caso las diferencias son tan importantes que es fácil concluir la victoria de la candidatura B. Cosa distinta sería si los porcentajes de cada candidatura fueran próximos, por ejemplo del 49 y de 51 por ciento, en ese caso las diferencias son tan pequeñas que estaríamos dentro del margen de error de muestras estimaciones y otras muestras podrían arrojar resultados distintos, incluso que los porcentajes se invirtieran entre las candidaturas, habría que ser cautos en nuestras conclusiones. Veremos cómo el margen de error nos ayudará a establecer cuándo una diferencia es relevante, significativa estadísticamente y cuando no.

2.2. Distribuciones muestrales, margen de error e intervalo de confianza

Como hemos remarcado en los dos ejemplos anteriores, cada posible muestra de tamaño n que podemos extraer de una misma población da como resultado una estimación particular del parámetro. Por el hecho de disponer de una parte de la población, aunque sea representativa e incluso numerosa, cada estimación del valor verdadero (y desconocido) del parámetro a partir de los estadísticos en la muestra, tiene asociado un margen de error, error muestral o error simplemente, respecto del valor real y desconocido del parámetro. Por ejemplo, si estimamos el porcentaje de voto que cierto partido obtendrá las próximas elecciones a partir de una muestra del censo electoral, obtendremos un resultado que puede estar más cercano (si tenemos buena suerte con la muestra!) o menos cercano (si tenemos mala suerte!), pero que difícilmente coincidirá con el verdadero porcentaje de votos, que de momento, y hasta que no se hagan las elecciones, es desconocido. El margen de error e es la diferencia máxima entre la estimación puntual obtenida en la muestra y el valor verdadero del parámetro poblacional.

El error muestral o margen de error de las estimaciones nos mide el grado de exactitud o de precisión con el que inferimos de la muestra a la población. Este valor vendrá determinado por la variabilidad del estadístico, es decir, que el error se cuantifica mediante las varianzas del estadístico considerado en cada caso. Como veremos esta variabilidad se denomina error típico del estadístico, y se corresponde con un cálculo que depende de la varianza y del tamaño de la muestra, además de otra característica como el nivel de confianza. Pero el error muestral se define simplemente como la divergencia entre los valores de los estadísticos obtenidos en la muestra de una variable y los existentes en la población.

Si consideráramos por ejemplo el estadístico de la media, el error muestral sería la diferencia entre el valor muestral de la media, $\overline{x}$, y el valor poblacional de la media, μ, es decir, $\overline{x} - \mu$. Esta es una definición de lo real pero no conocido, ya que el parámetro poblacional es desconocido y es lo que se quiere saber. Para solucionar este interrogante, razonaremos en términos probabilísticos a partir del conocimiento de la forma de la distribución del estadístico. En este sentido, cualquier afirmación para dar cuenta del valor de un parámetro poblacional desconocido a partir del valor del estadístico que lo estima, tendrá un grado de desviación, de variación al alza o de variación a la baja, que es el que mide el error de muestreo. De aquí se deriva una relación básica del muestreo y de las estimaciones estadísticas:

Valor poblacional = **Valor estimado en la muestra** ± **Error de muestreo**
(**Parámetro**) (**Estadístico**)

De esta forma, cada estadístico ($\overline{x}$, p, etc.) debemos entenderlo como si fuera una variable, en el sentido de que puede tomar varios valores según la muestra que consideramos (de todas las muestras posibles que podemos tomar de la población). Así, todos los posibles valores que estiman el único valor verdadero del parámetro forman lo que se llama una distribución muestral del estadístico. Si bien jamás nos dedicaremos a obtener todas las posibles muestras podemos llegar a conocerlas a partir del conocimiento general de la "forma" de la distribución del estadístico, pudiendo

determinar, con una certeza suficiente, el margen de error que cometemos al hacer estimaciones. Este es un aspecto fundamental que tenemos que considerar siempre en cualquier estudio inferencial, ya que nos indica el nivel de precisión de nuestra estimación, cuanto menor sea el error, más precisa será nuestra estimación.

Cuando a una estimación puntual le añadimos un margen de error, sumando y restando el error de muestreo, estamos realizando una estimación por intervalos de confianza. Más adelante veremos cómo calcular o aproximar el margen de error asociado a cada estimación. Pero veamos un ejemplo para clarificar qué significa asumir un margen de error y determinar un intervalo de confianza. Situémonos en el contexto del ejemplo de las elecciones al rectorado con dos candidaturas A y B. Recordemos que la candidatura A tenía una estimación puntual del 25% de los votos. Supongamos ahora que el margen de error asociado a la estimación del porcentaje de las candidaturas fuera del 3%. Esto significaría que el valor poblacional se situaría en un intervalo que resulta de sumar y restar este margen error a la estimación del porcentaje. Así, afirmaríamos que la candidatura A obtendrá unos resultados que se situarán en una horquilla definida por:

$$25\% \pm 3\%$$

es decir, en un intervalo entre el

$$22\% \text{ y el } 28\%.$$

y que escribimos así $(22\%,28\%)$.

Cualquier valor puntual estimado es una aproximación al valor poblacional desconocido afectado por el error muestral, por lo tanto, esto no significa que el porcentaje sea estrictamente del 25% en nuestro ejemplo, sino que sólo podemos establecer un intervalo de valores, entre un mínimo y un máximo, en cuyo interior se situará, probablemente el parámetro poblacional. Y decimos probablemente porque, como veremos, el afirmar sobre el intervalo se establece con un grado de probabilidad o nivel de confianza, habitualmente del 95%, lo que significa que en el 5% de los casos nos podemos equivocar, si bien es un nivel suficientemente bajo de incertidumbre que garantiza la bondad de nuestras estimaciones. En consecuencia, tenemos dos fuentes de incertidumbre, la del error de la muestra que hemos tomado (el candidato puede obtener entre el 22% y el 28% de los votos) y la probabilidad de de equivocarnos en esa estimación, valorada en un 5%, de que el valor del parámetro no se encuentre en el intervalo que hemos definido.

El error muestral es un aspecto clave de los estudios por muestreo que debemos considerar en dos momentos de la investigación: antes de proceder a obtener la muestra y de proceder a la recogida de información, en el que hay que determinar a priori qué nivel de error estadístico estamos dispuestos a asumir junto a la decisión del tamaño de la muestra, y después de obtener la muestra, para determinar el error asociado a cada estimación o prueba estadística de hipótesis que podamos plantear.

En este sentido, y como situación frecuente de particular interés en la investigación, conviene tener presente la necesidad, que hay que prever como objetivo del diseño muestral para garantizar un error muestral suficiente, si se busca realizar análisis específicos y autónomos de colectivos o zonas geográficas. Considerar estas subpoblaciones de forma separada implica trabajar con submuestras de la muestra

original que aumentan rápidamente los errores muestrales y la validez de los resultados. Por lo tanto, esto será posible siempre que se haya previsto un tamaño muestral suficiente.

2.3. Distribución de la media muestral: el teorema central del límite

Como hemos comentado, los estadísticos (los estimadores de los parámetros) son variables: varían en cada muestra de tamaño n que podamos extraer de una misma población. También hemos comentado que la configuración de los posibles valores constituye la distribución muestral del estadístico. En particular, la media muestral $\bar{x}$ es una variable: si consideramos las medias de todas las posibles muestras de determinado tamaño de una misma población (por ejemplo, la media de horas trabajadas semanalmente en un territorio) y extraemos todas las muestras de tamaño 1.000 calculando las medias de todas estas muestras, entonces, la distribución de todos los posibles valores de estas medias es la distribución muestral del estadístico $\bar{x}$.

Esta distribución muestral, para muestras de tamaño suficientemente grande, siempre tiene una misma forma conocida que se deriva de un razonamiento central de la estadística que nos fundamenta la distribución muestral del estadístico de la media. A partir de este resultado podremos determinar el margen de error que podemos cometer en cada caso.

Se trata del Teorema Central del Límite que afirma: si se extraen repetidamente muestras aleatorias de tamaño n de una población cualquiera, con independencia de la distribución de esta población, a medida que aumenta el tamaño de la muestra, las medias muestrales que se obtienen siguen aproximadamente una distribución normal de media μ (la media poblacional) y varianza σ^2/n (la varianza poblacional de la variable dividida por el tamaño de la muestra).

Por tanto, si tenemos una variable X de una población con media poblacional μ y varianza poblacional σ^2 y escogemos muestras de tamaño n suficientemente grande (por la ley de los grandes números cuando $n \geq 30$), las medias muestrales (el estimador) se distribuyen aproximadamente siguiendo una ley normal $\bar{x} \sim N(\mu, \sigma^2/n)$ con:

- Media del estimador $\bar{x}$: μ
- Varianza del estimador $\bar{x}$: σ^2/n , que es menor a la varianza de la variable σ^2 y que decrece a medida que aumenta el tamaño de la muestra n.
- Desviación del estadístico $\bar{x}$, que se denomina error típico de la media: $\sigma/\sqrt{n}$, y se denota por $\sigma_{\bar{x}}$.

El mismo razonamiento podríamos aplicar al caso de una proporción. Como vimos, la proporción p es un caso particular de media para variables dicotómicas, y como tal

tendrá el mismo tipo de distribución. Para muestras suficientemente grandes (cuando $nPQ>20$) entonces: $p \sim N(P, {}^{PQ}\!/_{n})$

Estos resultados nos permitirán calcular o aproximar (con un grado aceptable de certeza o confianza) el error que cometemos cuando hacemos estimaciones de medias o porcentajes, y deducir el intervalo de confianza de la estimación: el conjunto de valores que esperamos (confiamos) que contenga el parámetro poblacional. Veamos seguidamente la noción de confianza.

2.4. Conceptos de margen de error, nivel de confianza e intervalo de confianza

Las estimaciones puntuales que se efectúan en una muestra deben matizarse o completar con dos elementos adicionales:

- El error muestral o margen de error, e, y el intervalo. Toda estimación está afectada de un margen de error asociado a cada estadístico, que nos indica el grado de imprecisión de nuestra estimación, es decir, en qué medida el valor que hemos obtenido en la muestra puede diferir del valor poblacional. Escribiremos que el parámetro pertenece al intervalo: $estimación \pm e$.
- El nivel de confianza. La certeza absoluta y el margen de error nulo solo se tienen cuando se dispone de toda la población, nunca con una muestra. Si quisiéramos asegurar con certeza absoluta del 100% que el valor del parámetro se encuentra en el intervalo, tendríamos que ampliar muchísimo la muestra o bien considerar un intervalo muy grande, lo que significa tomar un margen de error importante e ir en detrimento de la precisión. Así pues, un determinado margen de error tendrá asociada una certeza, llamado nivel de confianza, siempre inferior al 100%. Se considera un buen compromiso aceptar una certeza cercana e inferior al cien por ciento: 95%, 95,5% o 99% que son, por este orden, los niveles de confianza más utilizados. Precisamos más detalladamente el significado del nivel de confianza. Un nivel de confianza del 95% significa que, en el 95% de todas las posibles muestras de tamaño n, la diferencia entre el valor obtenido con la muestra y el valor poblacional será inferior al margen de error. De manera equivalente: en un 95% de todas las muestras aleatorias de tamaño n de la población, el valor del parámetro estará dentro del intervalo determinado a partir de la muestra. Al valor de probabilidad complementario del nivel de confianza se denomina nivel de significación, o también nivel de riesgo. Habitualmente el nivel de significación se expresa con la letra α y en tanto por uno, mientras que el nivel de confianza, también en tanto por uno, sería $1-\alpha$. Si consideramos un 95% de confianza $1-\alpha$ es 0,95, y el nivel de significación α, 0,05. Es decir, un 5% de las posibles muestras dan estimaciones malas en el sentido de que están más alejadas del valor poblacional que el margen de error. Equivalentemente, en el 5% de muestras restantes el parámetro no estará dentro del intervalo que ellas mismas determinen.

A continuación veremos, para las medias y las proporciones, cómo realizar los cálculos o utilizar los programas informáticos para calcular, o aproximar, el margen de error y obtener el intervalo.

2.5. Estimación de la media poblacional

Para realizar una estimación de la media poblacional y obtener el intervalo distinguiremos dos situaciones experimentales. Una primera situación, muy poco frecuente, supone conocer la varianza poblacional σ^2 de la variable desconociendo la media. Podríamos disponer de la información de la varianza por estimaciones de un estudio anterior. Una segunda situación, la habitual, es desconocer el valor de la varianza poblacional por lo que precisamos estimarla a través de los resultados obtenidos en la muestra.

2.5.1. *Estimación de la media conocida la varianza*

Hemos visto por el teorema central del límite que las medias muestrales tienen una distribución aproximadamente normal: $\bar{x} \sim N(\mu, \sigma^2/n)$. Conocida la forma de la distribución podemos aplicar lo que vimos en el apartado 1.3 de este capítulo. Allí razonábamos sobre la base de una variable aleatoria (como los ingresos, la edad, etc.). Aquí hablaremos de una variable particular: la de las medias que se obtienen en cada muestra posible. Los valores x son ahora medias $\bar{x}$ muestrales y la desviación típica σ ahora es el error típico $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Siguiendo la *Ecuación 22* de la distribución normal:

$$|x-\mu| \leq z \cdot \sigma_{\bar{x}} = |x-\mu| \leq z \cdot \sigma/\sqrt{n} \qquad \textit{Ecuación 23}$$

Concluimos que: en particular, con un nivel de confianza del 95%, cuando $z=1{,}96$, en el 95% de las muestras la diferencia entre la media poblacional y la media muestral es menor o igual que $1{,}96 \cdot \sigma/\sqrt{n}$ que es el error muestral e. La diferencia máxima o margen de error para estimar la media poblacional, conocida la desviación poblacional será:

$$e = z \cdot \sigma_{\bar{x}} = z \cdot \sigma/\sqrt{n} \qquad \textit{Ecuación 24}$$

En términos de intervalo de confianza diremos que la media poblacional μ se encuentra en un intervalo de valores definidos por la media muestral $\bar{x}$ más, menos, el error:

$$\bar{x} \pm e = \bar{x} \pm z \cdot \sigma/\sqrt{n} \qquad \textit{Ecuación 25}$$

el intervalo es:

$$\left[\bar{x} - z \cdot \sigma/\sqrt{n}, \bar{x} + z \cdot \sigma/\sqrt{n}\right] \qquad \textit{Ecuación 26}$$

Habitualmente construiremos los intervalos asumiendo un nivel de confianza del 95%, $z=1{,}96$, del 95,5%, $z=2$.

Veamos un ejemplo para ilustrar estos resultados. Supongamos que queremos estimar la media poblacional desconocida del salario medio mensual de un subsector productivo, el hostelero pongamos por caso. De un estudio realizado con anterioridad sabemos que la desviación típica del salario es de aproximadamente 600 euros. Tomamos una muestra aleatoria de 900 trabajadores/as del subsector y les preguntamos a través de un cuestionario de encuesta, entre otras cosas, su salario actual. Calculamos la media salarial y obtenemos que es de 1.500 euros. Si consideramos un nivel de confianza del 95,5%, el error de nuestra estimación es:

$$e = z \cdot \sigma_{\bar{x}} = z \cdot \sigma / \sqrt{n} = 2 \cdot 600 / \sqrt{900} = 40$$

El intervalo que resulta es $\bar{x} \pm e = 1500 \pm 40$, por tanto, [1460 , 1540]. Es decir, con un nivel de confianza del 95,5% el salario medio de los trabajadores/as del subsector se sitúa entre 1460 y 1540 euros mensuales, puede ser cualquier valor de ellos, esta es nuestra estimación por intervalo.

2.5.2. *Estimación de la media cuando la varianza es desconocida*

Cuando la varianza poblacional no es conocida, como sucede habitualmente, se utiliza la varianza (desviación típica) muestral s^2 y se sustituye la distribución normal estándar por la t de Student. La formulación es semejante, tan sólo hay que sustituir el valor de la z por el valor de la t, y la σ por la s. El intervalo viene ahora determinado por:

$$\bar{x} \pm e = \bar{x} \pm t \cdot s / \sqrt{n} \qquad \textit{Ecuación 27}$$

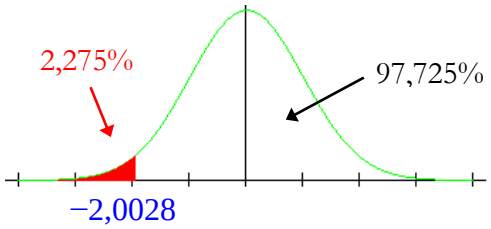
Todos los cálculos son semejantes, y necesitamos saber el valor correspondiente a la t de Student. No obstante, mientras que el valor z de la distribución normal los podemos fijar, en el caso del valor de la t depende del tamaño de la muestra, y se necesita una tabla estadística de la distribución teórica o utilizar un programa informático que nos proporcione el dato. Se trata de una distribución que a medida que el tamaño de la muestra crece el valor de la t se aproxima al valor de la z; y con tamaño muestrales de 120 casos o más las diferencias son mínimas, a partir del tercer decimal. Por tanto en nuestro caso podríamos considerar, para un nivel de confianza del 95,5%, el mismo valor que en la normal, $t=2$.

Si quisiéramos obtener el valor preciso de la t con la ayuda del SPSS podríamos ejecutar la siguiente instrucción:

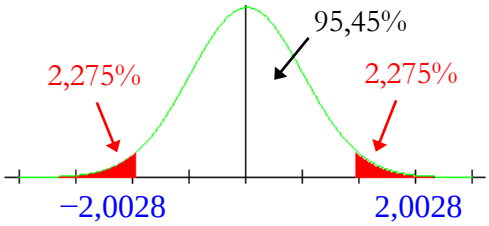
```
COMPUTE valort=IDF.T(0.02275,899).
```

El resultado es −2,0028, redondeando a dos decimales, el valor −2. Veamos cómo interpretarlo. En primer lugar, utilizamos la función inversa de la distribución de probabilidad (IDF) en relación a la t de student (T) que devuelve el valor de la distribución t, dada una probabilidad acumulada p y unos grados de libertad gl. En

nuestro caso consideramos "estratégicamente" una probabilidad de 0,02275 que corresponde a un nivel de confianza del 95,45% (y un nivel de significación del 4,55%) por la razón siguiente: el ordenador nos da el valor acumulado de la distribución, hasta el 2,275% del área que es el valor de la $t=-2{,}0028$:



Por tanto, este valor es considerando una cola. Nosotros queremos obtener el 95,45% central de la distribución a dos colas:



Los grados de libertad de la instrucción son 899, el resultado de $gl=n-1=900-1$. Como vemos, por tanto, si la muestra supera los 120 casos podemos asumir un valor de la t de 2.

2.5.3. *Estimación de la proporción*

El mismo razonamiento visto de cómo determinar el error y los intervalos de confianza de una media se puede aplicar en el caso de la estimación de una proporción o de un porcentaje p, sabiendo que siguen igualmente una distribución normal $p \sim N(P, PQ/n)$. En este caso el error muestral será:

$$e = z \cdot \sigma_p = z \cdot \sqrt{PQ/n} \qquad \textit{Ecuación 28}$$

Donde P es la proporción poblacional desconocida y Q su complementaria e igual a $1-P$. El intervalo de confianza de la proporción se encuentra en un intervalo de valores definidos por la proporción muestral p más, menos, el error:

$$p \pm e = p \pm z \cdot PQ/\sqrt{n} \qquad \textit{Ecuación 29}$$

y el intervalo es:

$$\left[p - z \cdot PQ/\sqrt{n},\ p + z \cdot PQ/\sqrt{n} \right] \qquad \textit{Ecuación 30}$$

Para realizar una estimación de una proporción debemos resolver el problema de determinar el valor poblacional de P y Q. Podríamos optar por buscar información

que nos diera un valor aproximado, pero carece de sentido dado que justamente nuestro objetivo es saber cuánto vale P. Alternativamente existen dos soluciones sencillas que podemos aplicar.

La primera opción consiste en sustituir el producto PQ por la estimación de esta cantidad a partir de la muestra, pq; es decir, PQ se aproxima por pq, con $q=1-p$. Como hemos hecho una aproximación, el intervalo de confianza y el margen de error que obtenemos no son exactos sino aproximados y decimos que encontramos el intervalo aproximando (o estimando) el error. En este caso las fórmulas son las siguientes. Del error:

$$e = z \cdot s_p = z \cdot \sqrt{pq/n} \qquad \textit{Ecuación 31}$$

y del intervalo

$$p \pm e = p \pm z \cdot \sqrt{pq/n} \qquad \textit{Ecuación 32}$$

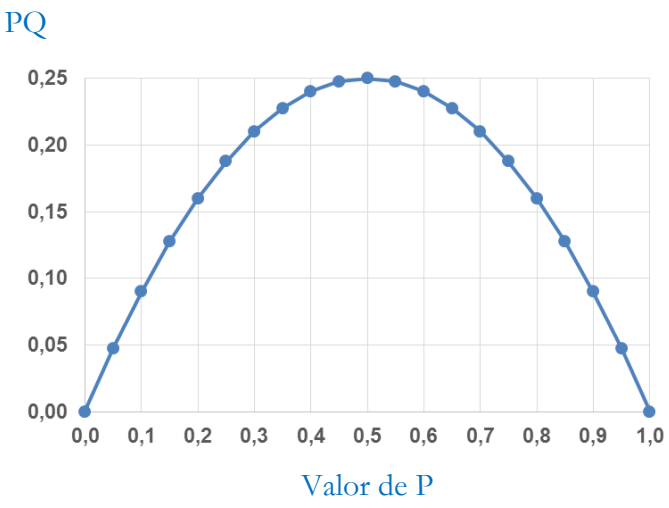
Una segunda opción consiste en considerar el error máximo, sabiendo que siempre, para cualquier par P y Q, se cumple que:

$$PQ \leq 1/4 \qquad \textit{Ecuación 33}$$

Entonces, se sustituye PQ por este valor máximo de ¼, es decir, PQ se aproxima por ¼. Este resultado se puede ver facilmente calculando y representando todos los posibles valores del producto de P por Q (Gráfico III.4.17). El máximo valor de todas las combinaciones posibles de P y Q se alcanza cuando $P=0,5$ y $Q=0,5$, cuyo producto $PQ=0,25$, es decir, ¼.

Gráfico III.4.17. Determinación del valor máximo de PQ

P	Q	PQ
0,00	1,00	0,0000
0,05	0,95	0,0475
0,10	0,90	0,0900
0,15	0,85	0,1275
0,20	0,80	0,1600
0,25	0,75	0,1875
0,30	0,70	0,2100
0,35	0,65	0,2275
0,40	0,60	0,2400
0,45	0,55	0,2475
0,50	0,50	0,2500
0,55	0,45	0,2475
0,60	0,40	0,2400
0,65	0,35	0,2275
0,70	0,30	0,2100
0,75	0,25	0,1875
0,80	0,20	0,1600
0,85	0,15	0,1275
0,90	0,10	0,0900
0,95	0,05	0,0475
1,00	0,00	0,0000



Decimos que calculamos el intervalo de máxima incertidumbre o de máxima indeterminación. Este supuesto es el que se asume también habitualmente en los sondeos electorales y encuestas sociológicas en el momento de determinar el tamaño de la muestra donde, como veremos a continuación, también se plantea el conocimiento del parámetro poblacional de la varianza.

En este caso las fórmulas son las siguientes. Del error:

$$e = z \cdot s_p = z \cdot \sqrt{1/4n} \qquad \textit{Ecuación 34}$$

y del intervalo

$$p \pm e = p \pm z \cdot \sqrt{1/4n} \qquad \textit{Ecuación 35}$$

Apliquemos en un ejemplo lo que acabamos de comentar. Retomamos el caso de las candidaturas donde con una muestra de $n=900$ personas habíamos encontrado una proporción muestral de partidarios de la candidatura A del 25%, $p=0,25$, siendo $q=1-p=0,75$. Para determinar el intervalo calculamos el margen de error aproximado considerando un nivel de confianza del 95,5%, por tanto, $z=2$:

$$e = z \cdot s_p = 2 \cdot \sqrt{0,25 \cdot 0,75/900} = 0,0144$$

Es decir, un error del 1,44%. El intervalo de votos favorables a la candidatura A será, a partir de:

$$p \pm e = 0,75 \pm 0,0144$$

en proporciones: [0,7356 , 0,7644], en porcentaje [73,56% , 76,44%]. Cualquier valor de ese intervalo será el resultado previsto dado el tamaño de la muestra, el error asumido y el nivel de confianza. Esto es, a partir de las respuestas de 900 personas, con un nivel de confianza del 95,5% y asumiendo un error máximo del 1,44%, la candidatura A obtendrá un porcentaje de votos que se sitúa en una horquilla entre el 73,6 y el 76,4%.

Para determinar este intervalo, pero ahora asumiendo el supuesto de máxima indeterminación, calculamos el error de esta forma:

$$e = z \cdot s_p = z \cdot \sqrt{1/4n} = 2 \cdot \sqrt{1/4 \cdot 900} = 0,0333$$

dando un error mayor del 3,33%, siendo el intervalo: $p \pm e = 0,75 \pm 0,0333$, en proporciones: [0,7167 , 0,7833] y en porcentaje [71,67% , 78,33%].

¿Qué opción es mejor, el intervalo aproximando el error o el intervalo de máxima indeterminación? Nos inclinamos por el primero, el intervalo estimando el error es menos amplio, por tanto, es más preciso, y por consiguiente es preferible en este sentido. No obstante, el otro intervalo me proporciona "más confianza". Con todo, si disponemos de muestras grandes y para valores de proporciones ni muy pequeños ni muy grandes, los cálculos de intervalos son similares.

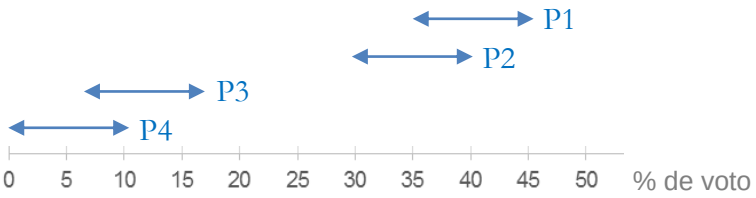
3. Error muestral y tamaño de la muestra

En este capítulo nos hemos planteado el problema de la estimación de un parámetro de la población a partir de la extracción de una muestra aleatoria. Al efectuar la estimación a través del estadístico sabemos que asumimos un riesgo y que tenemos un margen de error. También hemos sugerido que el margen de error tiene una relación inversa con el tamaño de la muestra: cuanto mayor es la muestra n menor es el error típico $\sigma_{\bar{x}} = \sigma / \sqrt{n}$. Hasta ahora hemos asumido tamaños de muestras sin valorar si eran suficientes para obtener estimaciones precisas, con poco margen de error. La cuestión que nos planteamos ahora es ¿qué tamaño de muestra es necesario para tener una muestra representativa?

La pregunta no tiene una respuesta automática, pero se podría resumir diciendo que depende del margen de error que queramos asumir, es decir, de la calidad o de la precisión de nuestras estimaciones. Para ser más precisos y tener un menor margen de error, necesitamos más observaciones muestrales.

Márgenes de error del 2% o del 3% son, por ejemplo, aceptables para estimar porcentajes en las encuestas. Un error superior implica a menudo que el estudio asuma excesivos márgenes de error. Imaginemos que damos unas predicciones del resultado electoral con un margen de error del 5%. Supongamos que las estimaciones de voto para el partido P1 son de un 40%, para el partido P2 de un 35%, para el partido P3 del 12%, y para el partido P4 un 5% (el resto hasta el 100% puede corresponder a otros partidos minoritarios que no consideramos).

En estas condiciones, el margen de error del ±5% significa que nos moveremos en un intervalo de 10 puntos porcentuales de variación; estaríamos en la situación siguiente: el partido P1 obtendría entre un 35 y un 45% de los votos, P2 entre un 30 y un 40%, el partido P3 entre un 7% y un 17%, el partido P4 entre un 0 y un 10%, gráficamente podríamos representarlo de esta forma:



Es evidente que estos intervalos no son satisfactorios para ninguno de los partidos y no satisfacen las exigencias de precisión que se deben esperar de un sondeo: no se sabe qué partido ganará, tanto P1 como P2 pueden alcanzar la primera posición los dos mayoritarios podrían ganar (P1 podría obtener un 37% y P2 un 38%, por ejemplo); al partido minoritario con esperanza de voto escasa se le pronostica un intervalo que seguro conocía antes del sondeo, pero además se indica que lo mismo obtiene un 0% que un 10%. En el mejor de los casos incluso adelantaría al tercer partido en disputa, como sus intervalos se cruzan, tanto uno como otro podría estar por delante. Se trata

pues de resultados muy imprecisos que invalidan cualquier ejercicio válido de estimación y predicción de resultados.
Para que no sea así debemos reducir el error muestral, lo que significa aumentar el tamaño de la muestra, y también los costes del trabajo de campo. Pero mantener un nivel de error bajo o aceptable implica considerar tamaños de muestra suficientemente elevados a partir de 800, 1.000 o 2.000 individuos, cuando no más dependiendo de los los objetivos de la operación estadística. Por tanto, la decisión final de qué tamaño es necesario dependerá, sobre todo, de los recursos económicos de la investigación. Por más que queramos ser cuidadosos en nuestras estimaciones, si no disponemos de estos recursos, tendremos que ajustarnos a muestras más pequeñas y errores más elevados. Sin embargo, debemos procurar que fuera dentro de unos límites tolerables.

Determinar el tamaño de la muestra es una decisión que debe tomarse antes de extraerla, cuando se diseña el estudio. Para determinar el tamaño de la muestra se debe definir el tipo de estimación que se desea realizar. Distinguiremos aquí el caso de la media y el de la proporción, pero podría ser la estimación de un número total de casos o de una razón o cociente.

3.1. Tamaño de la muestra para estimar una media

Analicemos primeramente de qué depende el error y cómo es esta dependencia. Hemos visto que la expresión del margen de error tenía la fórmula siguiente en el caso de una media:

$$e = z \cdot \sigma / \sqrt{n}$$

Vemos que el error e depende inversamente del tamaño de la muestra n: cuanto mayor sea la muestra (mayor es el denominador), menor será el error (menor será el cociente). En cambio, depende directamente del nivel de confianza: al aumentar el nivel de confianza, mayor es z, que es un factor multiplicador, crece el error. Por ello, se suele considerar una confianza del 95% (o del 95.5%) y no del 99%. Finalmente, el error también es mayor en función de la varianza poblacional. La varianza es un parámetro que está dado y no lo podemos controlar; por el contrario, sí que podemos disminuir el error si aumentamos el tamaño de la muestra.

Para determinar el tamaño de la muestra, podemos aislar n en función de e y del resto de parámetros de la expresión anterior. De este modo, si fijamos un nivel de confianza z y un error máximo e (nivel de precisión), dada una varianza, podemos determinar el tamaño de la muestra necesaria para tener aquella precisión en la estimación mediante:

$$n = \frac{z^2 \cdot \sigma^2}{e^2} \qquad \textit{Ecuación 36}$$

El resultado de esta operación debe ser siempre un número entero, ya que para la muestra debemos extraer unidades enteras de la población. Por ello siempre aproximamos el valor resultante al entero inmediatamente superior.

Veamos un ejemplo de determinación del tamaño de la muestra. Consideremos que se busca determinar el consumo medio de unidades de cierto producto. Se sabe, según las estimaciones de estudios anteriores, que la varianza poblacional del número de unidades de consumo se sitúa en un valor de 100 unidades. Lo que queremos es determinar el tamaño de la muestra necesaria para realizar la estimación de la media de unidades consumidas. Asumimos un nivel de confianza del 95,5% ($z=2$) y fijamos un margen de error máximo de 2 unidades del producto. El tamaño de muestra que se obtiene es de 1.000 individuos:

$$n = \frac{z^2 \cdot \sigma^2}{e^2} = \frac{2^2 \cdot 1000}{2^2} = 1000$$

3.2. Tamaño de la muestra para estimar una proporción

Ahora consideraremos el caso de determinar el tamaño de la muestra cuando se trata estimar una proporción. La expresión del margen de error cuando estimamos una proporción poblacional, para el caso de máxima incertidumbre o indeterminación (cuando $P=Q=0,5$), es:

$$e = z \cdot \sqrt{1/4n}$$

Despejando n en función del resto de parámetros, resulta que el tamaño de la muestra necesaria, para tener una precisión fijada en el estimación, se calcula mediante:

$$n = \frac{z^2}{4e^2} \qquad \textit{Ecuación 37}$$

Si además considerásemos un nivel de confianza del 95,5%, es decir, $z=2$, la expresión se simplificaría aún más: $n = 1/e^2$, expresión que evidencia la relación inversa entre el tamaño de la muestra y el error, si bien esta relación inversa es más que proporcional, ya que el error está elevado al cuadrado. Esto significa que aumentos progresivos de tamaño de muestra disminuyen cada vez menos el margen de error, hasta un punto tal, que una disminución del error no vale la pena si no es a costa de añadir grandes cantidades de individuos en la muestra.

La *Ecuación 37* se generaliza para el caso de considerar cualquier valor P *y* Q con la expresión:

$$n = \frac{z^2 \cdot PQ}{e^2} \qquad \textit{Ecuación 38}$$

Ejemplificaremos estos cálculos con la información de un sondeo elaborado por el Instituto Opina con motivo de las elecciones nacionales al Parlamento de Cataluña de 2003, publicado en el diario El País, donde se adjuntaba la ficha técnica siguiente:

ELECCIONES EN CATALUÑA *Sondeo de Opina para EL PAÍS*

FICHA TÉCNICA

La encuesta ha sido realizada por el INSTITUTO OPINA. **Realización del trabajo de campo:** 31 de octubre, 1 y 2 de noviembre de 2003. **Ámbito geográfico:** Cataluña. **Recogida de la información:** mediante entrevista telefónica. **Universo de análisis:** población mayor de 18 años residente en hogares con teléfono. **Error muestral:** El margen de error para el total de la muestra es de ±2,14% para un margen de confianza del 95% y bajo el supuesto de máxima indeterminación (p=q=50%). **Procedimiento de muestreo:** selección polietápica del entrevistado: unidades primarias de muestreo (MUNICIPIOS) seleccionadas de forma aleatoria proporcional para cada provincia. unidades secundarias (HOGARES) mediante la selección aleatoria de números de teléfono. Unidades últimas (INDIVIDUOS) según cuotas cruzadas de SEXO, EDAD y RECUERDO DE VOTO GENERALES 2000.

Tamaño de la muestra: 2.100 entrevistas. 900 en la provincia de Barcelona (290 en Barcelona ciudad, 256 en el área metropolitana, 354 en el resto de la provincia). **400 en la provincia de Girona** (53 en Girona ciudad, 347 en el resto de la provincia). **400 en la provincia de Lleida** (126 en Lleida ciudad, 274 en el resto de la provincia). **400 en la provincia de Tarragona** (78 en Tarragona ciudad, 322 en el resto de la provincia).

Ponderación: el total de la muestra se ha ponderado para otorgar a cada una de las provincias su peso real dentro del conjunto de la población de Cataluña. **De 900 en la provincia de Barcelona a 1.594** (514 en Barcelona ciudad, 453 en el área metropolitana, 627 en el resto de la provincia). **De 400 en la provincia de Girona a 185** (25 en Girona ciudad, 160 en el resto de la provincia). **De 400 en la provincia de Lleida a 122** (38 en Lleida ciudad, 84 en el resto de la provincia). **De 400 en la provincia de Tarragona a 199** (39 en Tarragona ciudad, 160 en el resto de la provincia).

Como se puede leer en la ficha, se ha realizado una encuesta con una muestra de 2.100 personas mayores de 18 años en Cataluña. El margen de error ha sido del 2,14% para un nivel de confianza del 95% ($z=1,96$) y bajo el supuesto de máxima indeterminación (P=Q=50%). Estos datos son el resultado de aplicar la fórmula del cálculo del tamaño muestral:

$$n = \frac{z^2 \cdot P \cdot Q}{e^2} = \frac{1,96^2 \cdot 0,5 \cdot 0,5}{0,0214^2} \cong 2100$$

El mismo resultado se obtiene si hacemos los cálculos en porcentajes:

$$n = \frac{z^2 \cdot P \cdot Q}{e^2} = \frac{1,96^2 \cdot 50 \cdot 50}{2,14^2} \cong 2100$$

3.3. Ajuste por finitud en las fórmulas del tamaño de la muestra

Tanto en el caso de la media como de la proporción hemos asumido implícitamente que no era relevante otro parámetro: el tamaño de la población. De hecho hemos procedido asumiendo que la población era infinita (100.000 y más unidades) y las fórmulas anteriores no precisan introducir el valor de la población. No obstante, cuando la población se considera finita es posible introducir un factor de corrección por finitud.

Al hablar de poblaciones (o universos) se establece la distinción entre una población finita y una infinita. Desde el punto de vista del muestreo, la distinción se basa en la importancia relativa que tiene el tamaño de la muestra en relación al tamaño de población. Si el tamaño de la muestra es muy pequeño respecto al tamaño de la población (habitualmente se admite que represente menos del 5%) se suele considerar infinita la población. En cambio, si la muestra necesaria es considerable en relación a la población (por encima del 10% se suele considerar necesario, y entre un 5% y un 10% recomendable) entonces la población se considera finita y se deben utilizar factores de corrección de población finita. Igualmente, se considera que una población es finita a toda población formada por menos de 100.000 unidades, e infinita a la que tiene 100.000 o más.

El factor de corrección por finitud tiene en cuenta la relación entre el tamaño de la población y el tamaño de la muestra con la expresión: $\sqrt{\frac{N-n}{N-1}}$. Al introducir este elemento en las ecuaciones anteriores (*Ecuación 36* y *Ecuación 38*) se obtiene la fórmula para determinar el tamaño de la muestra con poblaciones finitas en el caso de la media:

$$n = \frac{z^2 \cdot \sigma^2 \cdot N}{(N-1) \cdot e^2 + z^2 \cdot \sigma^2}$$ *Ecuación 39*

y en el caso de la proporción:

$$n = \frac{z^2 \cdot PQ \cdot N}{(N-1) \cdot e^2 + z^2 \cdot PQ}$$ *Ecuación 40*

3.4. Determinación del tamaño de la muestra: resumen

En la determinación del tamaño de la muestra se conjugan cuatro elementos:

1. La amplitud de la población o universo, en función de si es finita (menor de 100.000 unidades) o infinita (mayor o igual a 100.000).
2. El nivel de confianza adoptado z.
3. El error muestral de la estimación e.
4. La varianza o desviación típica de la población σ^2.

En función de estos aspectos, y dependiendo de si estimamos una media o una proporción, se concretan las fórmulas de la Tabla III.4.9[15]. Primero aparecen las cuatro fórmulas que hemos visto y después esas mismas despejando el error en función del tamaño de la muestra, expresión práctica pues muchas veces, de lo que se trata, es de fijar el tamaño según el presupuesto disponible y se mira el error que implica.

Tabla III.4.9. Fórmulas del tamaño de la muestra

(a) Tamaño de la muestra en función del margen de error

Tamaño en función del error		**Población**	
		Finita	Infinita
Parámetro	Media	$n = \frac{z^2 \cdot \sigma^2 \cdot N}{(N-1) \cdot e^2 + z^2 \cdot \sigma^2}$	$n = \frac{z^2 \cdot \sigma^2}{e^2}$
	Proporción	$n = \frac{z^2 \cdot PQ \cdot N}{(N-1) \cdot e^2 + z^2 \cdot PQ}$	$n = \frac{z^2 \cdot PQ}{e^2}$

[15] Son expresiones en condiciones del llamado muestreo aleatorio simple (MAS).

(b) Margen de error en función del tamaño de la muestra

Error en función del tamaño		**Población**	
		Finita	Infinita
Parámetro	Media	$e = z \cdot \sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}}$	$e = z \cdot \sqrt{\frac{\sigma^2}{n}}$
	Proporción	$e = z \cdot \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}$	$e = z \cdot \sqrt{\frac{PQ}{n}}$

3.5. Tablas para la determinación del tamaño de la muestra

A continuación aparecen varias tablas destinadas a facilitar la determinación del tamaño de la muestra sin tener que realizar manualmente los cálculos a través de las expresiones que hemos visto. Se presentan las tablas de tamaño de la muestra en función del error muestral, tanto para poblaciones finitas como infinitas, y también del error en función del tamaño de la muestra.

Estas tablas de cálculo del tamaño y del error se adjuntan en un archivo Excel que se puede encontrar en la página web del manual con el nombre FórmulasMAS.xlsx. Igualmente se incluye una primera pestaña con todas las fórmulas de la Tabla III.4.9 que facilitan el cálculo automático del tamaño y del error introduciendo los valores de los diferentes parámetros en el caso de la estimación de proporciones:

Muestreo aleatorio simple (MAS)
Fórmulas para determinar el tamaño de la muestra
Estimación de proporciones (porcentajes)

	Poblaciones finitas $n=\frac{z^2 \cdot PQ \cdot N}{(N-1) \cdot e^2 + z^2 \cdot PQ}$ $e = z \cdot \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}$			**Poblaciones infinitas** $n=\frac{z^2 \cdot PQ}{e^2}$ $e = z \cdot \sqrt{\frac{PQ}{n}}$	
Nivel de confianza	z	2		z	2
Error muestral (%)	e	2,00		e	3,00
Tamaño de la población	N	10.000		n	1.111
Tamaño de la muestra	n	2.000		P	50
Porcentaje	P	50		$Q=100-P$	50
Porcentaje complementario	$Q=100-P$	50			
Si el error es del: 2,00	**n?**	**2.000**	3,00	**n?**	**1.111**
Si el tamaño de muestra es de: 2.000	**e?**	**2,00**	1.111	**e?**	**3,00**

Número de unidades de la muestra (n), con una población finita (N<100.000), para un nivel de confianza del 95,5% (z=2) y bajo el supuesto de máxima indeterminación (P=Q=50%)

$$n = \frac{z^2 \cdot PQ \cdot N}{(N-1) \cdot e^2 + z^2 \cdot PQ}$$

	Margen de error (e)										
Población (N)	**1,0**	**1,5**	**2,0**	**2,5**	**3,0**	**3,5**	**4,0**	**4,5**	**5,0**	**7,5**	**10,0**
250	244	237	227	216	204	191	179	166	154	104	71
500	476	449	417	381	345	310	278	248	222	131	83
1.000	909	816	714	615	526	449	385	331	286	151	91
1.500	1.304	1.121	938	774	638	529	441	372	316	159	94
2.000	1.667	1.379	1.111	889	714	580	476	396	333	163	95
2.500	2.000	1.600	1.250	976	769	615	500	412	345	166	96
3.000	2.308	1.791	1.364	1.043	811	642	517	424	353	168	97
3.500	2.593	1.958	1.458	1.098	843	662	530	433	359	169	97
4.000	2.857	2.105	1.538	1.143	870	678	541	440	364	170	98
4.500	3.103	2.236	1.607	1.180	891	691	549	445	367	171	98
5.000	3.333	2.353	1.667	1.212	909	702	556	449	370	172	98
6.000	3.750	2.553	1.765	1.263	938	719	566	456	375	173	98
7.000	4.118	2.718	1.842	1.302	959	731	574	461	378	173	99
8.000	4.444	2.857	1.905	1.333	976	741	580	465	381	174	99
9.000	4.737	2.975	1.957	1.358	989	748	584	468	383	174	99
10.000	5.000	3.077	2.000	1.379	1.000	755	588	471	385	175	99
15.000	6.000	3.429	2.143	1.446	1.034	774	600	478	390	176	99
20.000	6.667	3.636	2.222	1.481	1.053	784	606	482	392	176	100
25.000	7.143	3.774	2.273	1.504	1.064	791	610	484	394	177	100
30.000	7.500	3.871	2.308	1.519	1.071	795	612	486	395	177	100
35.000	7.778	3.944	2.333	1.530	1.077	798	614	487	395	177	100
40.000	8.000	4.000	2.353	1.538	1.081	800	615	488	396	177	100
45.000	8.182	4.045	2.368	1.545	1.084	802	616	488	396	177	100
50.000	8.333	4.082	2.381	1.550	1.087	803	617	489	397	177	100
100.000	9.091	4.255	2.439	1.575	1.099	810	621	491	398	177	100

Número de unidades de la muestra (n), con una población finita (N<100.000), para un nivel de confianza del 99,7% (z=3) y bajo el supuesto de máxima indeterminación (P=Q=50%)

$$n = \frac{z^2 \cdot PQ \cdot N}{(N-1) \cdot e^2 + z^2 \cdot PQ}$$

	Margen de error (e)										
Población (N)	**1,0**	**1,5**	**2,0**	**2,5**	**3,0**	**3,5**	**4,0**	**4,5**	**5,0**	**7,5**	**10,0**
250	247	244	239	234	227	220	212	204	196	154	118
500	489	476	459	439	417	393	369	345	321	222	155
1.000	957	909	849	783	714	647	584	526	474	286	184
1.500	1.406	1.304	1.184	1.059	938	826	726	638	563	316	196
2.000	1.837	1.667	1.475	1.286	1.111	957	826	714	621	333	202
2.500	2.250	2.000	1.731	1.475	1.250	1.059	900	769	662	345	206
3.000	2.647	2.308	1.957	1.636	1.364	1.139	957	811	692	353	209
3.500	3.029	2.593	2.158	1.775	1.458	1.205	1.003	843	716	359	211
4.000	3.396	2.857	2.338	1.895	1.538	1.259	1.040	870	735	364	213
4.500	3.750	3.103	2.500	2.000	1.607	1.304	1.071	891	750	367	214
5.000	4.091	3.333	2.647	2.093	1.667	1.343	1.098	909	763	370	215
6.000	4.737	3.750	2.903	2.250	1.765	1.406	1.139	938	783	375	217
7.000	5.339	4.118	3.119	2.377	1.842	1.455	1.171	959	797	378	218
8.000	5.902	4.444	3.303	2.483	1.905	1.494	1.196	976	809	381	219
9.000	6.429	4.737	3.462	2.571	1.957	1.525	1.216	989	818	383	220
10.000	6.923	5.000	3.600	2.647	2.000	1.552	1.233	1.000	826	385	220
15.000	9.000	6.000	4.091	2.903	2.143	1.636	1.286	1.034	849	390	222
20.000	10.588	6.667	4.390	3.051	2.222	1.682	1.314	1.053	861	392	222
25.000	11.842	7.143	4.592	3.147	2.273	1.711	1.331	1.064	869	394	223
30.000	12.857	7.500	4.737	3.214	2.308	1.731	1.343	1.071	874	395	223
35.000	13.696	7.778	4.846	3.264	2.333	1.745	1.352	1.077	877	395	224
40.000	14.400	8.000	4.932	3.303	2.353	1.756	1.358	1.081	880	396	224
45.000	15.000	8.182	5.000	3.333	2.368	1.765	1.364	1.084	882	396	224
50.000	15.517	8.333	5.056	3.358	2.381	1.772	1.368	1.087	884	397	224
100.000	18.367	9.091	5.325	3.475	2.439	1.804	1.387	1.099	892	398	224

Número de unidades de la muestra (n), con una población infinita (N≥100.000), para un nivel de confianza del 95,5% (z=2) y diferentes valores de P y Q.

$$n = \frac{z^2 \cdot PQ}{e^2}$$

	Valors de P i Q en % (P + Q = 100)													
Margen de error (e)	P= 1	2	3	4	5	10	15	20	25	30	35	40	45	50
	Q= 99	98	97	96	95	90	85	80	75	70	65	60	55	50
0,1	39.600	78.400	116.400	153.600	190.000	360.000	510.000	640.000	750.000	840.000	910.000	960.000	990.000	1.000.000
0,2	9.900	19.600	29.100	38.400	47.500	90.000	127.500	160.000	187.500	210.000	227.500	240.000	247.500	250.000
0,3	4.400	8.711	12.933	17.067	21.111	40.000	56.667	71.111	83.333	93.333	101.111	106.667	110.000	111.111
0,4	2.475	4.900	7.275	9.600	11.875	22.500	31.875	40.000	46.875	52.500	56.875	60.000	61.875	62.500
0,5	1.584	3.136	4.656	6.144	7.600	14.400	20.400	25.600	30.000	33.600	36.400	38.400	39.600	40.000
0,6	1.100	2.178	3.233	4.267	5.278	10.000	14.167	17.778	20.833	23.333	25.278	26.667	27.500	27.778
0,7	808	1.600	2.376	3.135	3.878	7.347	10.408	13.061	15.306	17.143	18.571	19.592	20.204	20.408
0,8	619	1.225	1.819	2.400	2.969	5.625	7.969	10.000	11.719	13.125	14.219	15.000	15.469	15.625
0,9	489	968	1.437	1.896	2.346	4.444	6.296	7.901	9.259	10.370	11.235	11.852	12.222	12.346
1,0	396	784	1.164	1.536	1.900	3.600	5.100	6.400	7.500	8.400	9.100	9.600	9.900	10.000
1,5	176	348	517	683	844	1.600	2.267	2.844	3.333	3.733	4.044	4.267	4.400	4.444
2,0	99	196	291	384	475	900	1.275	1.600	1.875	2.100	2.275	2.400	2.475	2.500
2,5	63	125	186	246	304	576	816	1.024	1.200	1.344	1.456	1.536	1.584	1.600
3,0	44	87	129	171	211	400	567	711	833	933	1.011	1.067	1.100	1.111
3,5	32	64	95	125	155	294	416	522	612	686	743	784	808	816
4,0	25	49	73	96	119	225	319	400	469	525	569	600	619	625
4,5	20	39	57	76	94	178	252	316	370	415	449	474	489	494
5,0	16	31	47	61	76	144	204	256	300	336	364	384	396	400
6,0	11	22	32	43	53	100	142	178	208	233	253	267	275	278
7,0	8	16	24	31	39	73	104	131	153	171	186	196	202	204
8,0	6	12	18	24	30	56	80	100	117	131	142	150	155	156
9,0	5	10	14	19	23	44	63	79	93	104	112	119	122	123
10,0	4	8	12	15	19	36	51	64	75	84	91	96	99	100
15,0	2	3	5	7	8	16	23	28	33	37	40	43	44	44
20,0	1	2	3	4	5	9	13	16	19	21	23	24	25	25
25,0	1	1	2	2	3	6	8	10	12	13	15	15	16	16
30,0	0	1	1	2	2	4	6	7	8	9	10	11	11	11
35,0	0	1	1	1	2	3	4	5	6	7	7	8	8	8
40,0	0	0	1	1	1	2	3	4	5	5	6	6	6	6

Número de unidades de la muestra (n), con una población infinita (N≥100.000), para un nivel de confianza del 99,7% (z=3) y diferentes valores de P y Q.

$$n = \frac{z^2 \cdot PQ}{e^2}$$

	Valors de P i Q en % (P + Q = 100)													
Margen de error (e)	P= 1	2	3	4	5	10	15	20	25	30	35	40	45	50
	Q= 99	98	97	96	95	90	85	80	75	70	65	60	55	50
0,1	89.100	176.400	261.900	345.600	427.500	810.000	1.147.500	1.440.000	1.687.500	1.890.000	2.047.500	2.160.000	2.227.500	2.250.000
0,2	22.275	44.100	65.475	86.400	106.875	202.500	286.875	360.000	421.875	472.500	511.875	540.000	556.875	562.500
0,3	9.900	19.600	29.100	38.400	47.500	90.000	127.500	160.000	187.500	210.000	227.500	240.000	247.500	250.000
0,4	5.569	11.025	16.369	21.600	26.719	50.625	71.719	90.000	105.469	118.125	127.969	135.000	139.219	140.625
0,5	3.564	7.056	10.476	13.824	17.100	32.400	45.900	57.600	67.500	75.600	81.900	86.400	89.100	90.000
0,6	2.475	4.900	7.275	9.600	11.875	22.500	31.875	40.000	46.875	52.500	56.875	60.000	61.875	62.500
0,7	1.818	3.600	5.345	7.053	8.724	16.531	23.418	29.388	34.439	38.571	41.786	44.082	45.459	45.918
0,8	1.392	2.756	4.092	5.400	6.680	12.656	17.930	22.500	26.367	29.531	31.992	33.750	34.805	35.156
0,9	1.100	2.178	3.233	4.267	5.278	10.000	14.167	17.778	20.833	23.333	25.278	26.667	27.500	27.778
1,0	891	1.764	2.619	3.456	4.275	8.100	11.475	14.400	16.875	18.900	20.475	21.600	22.275	22.500
1,5	396	784	1.164	1.536	1.900	3.600	5.100	6.400	7.500	8.400	9.100	9.600	9.900	10.000
2,0	223	441	655	864	1.069	2.025	2.869	3.600	4.219	4.725	5.119	5.400	5.569	5.625
2,5	143	282	419	553	684	1.296	1.836	2.304	2.700	3.024	3.276	3.456	3.564	3.600
3,0	99	196	291	384	475	900	1.275	1.600	1.875	2.100	2.275	2.400	2.475	2.500
3,5	73	144	214	282	349	661	937	1.176	1.378	1.543	1.671	1.763	1.818	1.837
4,0	56	110	164	216	267	506	717	900	1.055	1.181	1.280	1.350	1.392	1.406
4,5	44	87	129	171	211	400	567	711	833	933	1.011	1.067	1.100	1.111
5,0	36	71	105	138	171	324	459	576	675	756	819	864	891	900
6,0	25	49	73	96	119	225	319	400	469	525	569	600	619	625
7,0	18	36	53	71	87	165	234	294	344	386	418	441	455	459
8,0	14	28	41	54	67	127	179	225	264	295	320	338	348	352
9,0	11	22	32	43	53	100	142	178	208	233	253	267	275	278
10,0	9	18	26	35	43	81	115	144	169	189	205	216	223	225
15,0	4	8	12	15	19	36	51	64	75	84	91	96	99	100
20,0	2	4	7	9	11	20	29	36	42	47	51	54	56	56
25,0	1	3	4	6	7	13	18	23	27	30	33	35	36	36
30,0	1	2	3	4	5	9	13	16	19	21	23	24	25	25
35,0	1	1	2	3	3	7	9	12	14	15	17	18	18	18
40,0	1	1	2	2	3	5	7	9	11	12	13	14	14	14

Determinación del margen de error (e), según el tamaño de la muestra (n) y diferentes valores de P y Q , para un nivel de confianza del 95,5% (z=2) con una población infinita (N≥100.000)

$$e = z \cdot \sqrt{\frac{PQ}{n}}$$

	Valors de P i Q en % (P + Q = 100)													
Tamaño de la	P=	1	2	3	4	5	10	15	20	25	30	35	40	45	50
muestra (n)	Q=	99	98	97	96	95	90	85	80	75	70	65	60	55	50
25		4,0	5,6	6,8	7,8	8,7	12,0	14,3	16,0	17,3	18,3	19,1	19,6	19,9	20,0
50		2,8	4,0	4,8	5,5	6,2	8,5	10,1	11,3	12,2	13,0	13,5	13,9	14,1	14,1
75		2,3	3,2	3,9	4,5	5,0	6,9	8,2	9,2	10,0	10,6	11,0	11,3	11,5	11,5
100		2,0	2,8	3,4	3,9	4,4	6,0	7,1	8,0	8,7	9,2	9,5	9,8	9,9	10,0
150		1,6	2,3	2,8	3,2	3,6	4,9	5,8	6,5	7,1	7,5	7,8	8,0	8,1	8,2
200		1,4	2,0	2,4	2,8	3,1	4,2	5,0	5,7	6,1	6,5	6,7	6,9	7,0	7,1
250		1,3	1,8	2,2	2,5	2,8	3,8	4,5	5,1	5,5	5,8	6,0	6,2	6,3	6,3
300		1,1	1,6	2,0	2,3	2,5	3,5	4,1	4,6	5,0	5,3	5,5	5,7	5,7	5,8
400		1,0	1,4	1,7	2,0	2,2	3,0	3,6	4,0	4,3	4,6	4,8	4,9	5,0	5,0
500		0,9	1,3	1,5	1,8	1,9	2,7	3,2	3,6	3,9	4,1	4,3	4,4	4,4	4,5
600		0,8	1,1	1,4	1,6	1,8	2,4	2,9	3,3	3,5	3,7	3,9	4,0	4,1	4,1
800		0,7	1,0	1,2	1,4	1,5	2,1	2,5	2,8	3,1	3,2	3,4	3,5	3,5	3,5
1.000		0,6	0,9	1,1	1,2	1,4	1,9	2,3	2,5	2,7	2,9	3,0	3,1	3,1	3,2
1.200		0,6	0,8	1,0	1,1	1,3	1,7	2,1	2,3	2,5	2,6	2,8	2,8	2,9	2,9
1.500		0,5	0,7	0,9	1,0	1,1	1,5	1,8	2,1	2,2	2,4	2,5	2,5	2,6	2,6
2.000		0,4	0,6	0,8	0,9	1,0	1,3	1,6	1,8	1,9	2,0	2,1	2,2	2,2	2,2
2.500		0,4	0,6	0,7	0,8	0,9	1,2	1,4	1,6	1,7	1,8	1,9	2,0	2,0	2,0
3.000		0,4	0,5	0,6	0,7	0,8	1,1	1,3	1,5	1,6	1,7	1,7	1,8	1,8	1,8
4.000		0,3	0,4	0,5	0,6	0,7	0,9	1,1	1,3	1,4	1,4	1,5	1,5	1,6	1,6
5.000		0,3	0,4	0,5	0,6	0,6	0,8	1,0	1,1	1,2	1,3	1,3	1,4	1,4	1,4
7.500		0,2	0,3	0,4	0,5	0,5	0,7	0,8	0,9	1,0	1,1	1,1	1,1	1,1	1,2
10.000		0,2	0,3	0,3	0,4	0,4	0,6	0,7	0,8	0,9	0,9	1,0	1,0	1,0	1,0
15.000		0,2	0,2	0,3	0,3	0,4	0,5	0,6	0,7	0,7	0,7	0,8	0,8	0,8	0,8
25.000		0,1	0,2	0,2	0,2	0,3	0,4	0,5	0,5	0,5	0,6	0,6	0,6	0,6	0,6
50.000		0,1	0,1	0,2	0,2	0,2	0,3	0,3	0,4	0,4	0,4	0,4	0,4	0,4	0,4

Determinación del margen de error (e), según el tamaño de la muestra (n) y diferentes valores de P y Q , para un nivel de confianza del 99,7% (z=3) con una población infinita (N≥100.000)

$$e = z \cdot \sqrt{\frac{PQ}{n}}$$

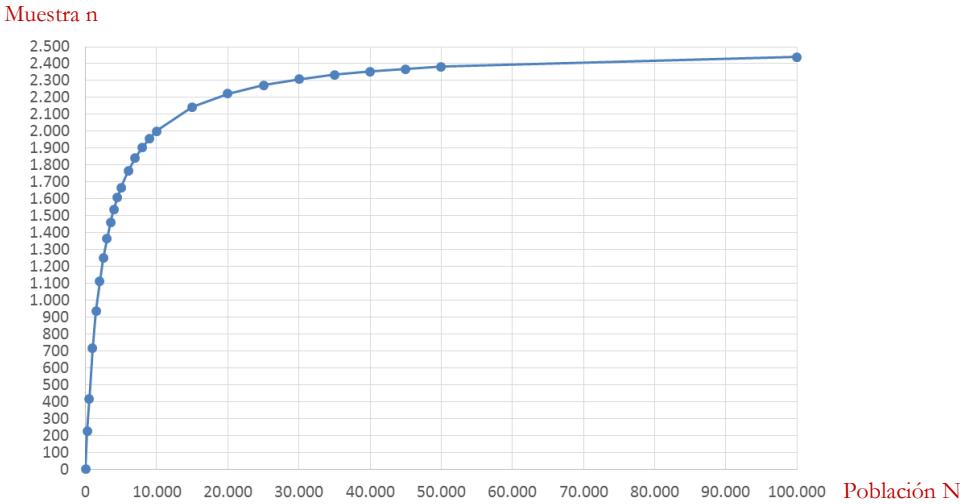
	Valors de P i Q en % (P + Q = 100)													
Tamaño de la	P=	1	2	3	4	5	10	15	20	25	30	35	40	45	50
muestra (n)	Q=	99	98	97	96	95	90	85	80	75	70	65	60	55	50
25		6,0	8,4	10,2	11,8	13,1	18,0	21,4	24,0	26,0	27,5	28,6	29,4	29,8	30,0
50		4,2	5,9	7,2	8,3	9,2	12,7	15,1	17,0	18,4	19,4	20,2	20,8	21,1	21,2
75		3,4	4,8	5,9	6,8	7,5	10,4	12,4	13,9	15,0	15,9	16,5	17,0	17,2	17,3
100		3,0	4,2	5,1	5,9	6,5	9,0	10,7	12,0	13,0	13,7	14,3	14,7	14,9	15,0
150		2,4	3,4	4,2	4,8	5,3	7,3	8,7	9,8	10,6	11,2	11,7	12,0	12,2	12,2
200		2,1	3,0	3,6	4,2	4,6	6,4	7,6	8,5	9,2	9,7	10,1	10,4	10,6	10,6
250		1,9	2,7	3,2	3,7	4,1	5,7	6,8	7,6	8,2	8,7	9,0	9,3	9,4	9,5
300		1,7	2,4	3,0	3,4	3,8	5,2	6,2	6,9	7,5	7,9	8,3	8,5	8,6	8,7
400		1,5	2,1	2,6	2,9	3,3	4,5	5,4	6,0	6,5	6,9	7,2	7,3	7,5	7,5
500		1,3	1,9	2,3	2,6	2,9	4,0	4,8	5,4	5,8	6,1	6,4	6,6	6,7	6,7
600		1,2	1,7	2,1	2,4	2,7	3,7	4,4	4,9	5,3	5,6	5,8	6,0	6,1	6,1
800		1,1	1,5	1,8	2,1	2,3	3,2	3,8	4,2	4,6	4,9	5,1	5,2	5,3	5,3
1.000		0,9	1,3	1,6	1,9	2,1	2,8	3,4	3,8	4,1	4,3	4,5	4,6	4,7	4,7
1.200		0,9	1,2	1,5	1,7	1,9	2,6	3,1	3,5	3,8	4,0	4,1	4,2	4,3	4,3
1.500		0,8	1,1	1,3	1,5	1,7	2,3	2,8	3,1	3,4	3,5	3,7	3,8	3,9	3,9
2.000		0,7	0,9	1,1	1,3	1,5	2,0	2,4	2,7	2,9	3,1	3,2	3,3	3,3	3,4
2.500		0,6	0,8	1,0	1,2	1,3	1,8	2,1	2,4	2,6	2,7	2,9	2,9	3,0	3,0
3.000		0,5	0,8	0,9	1,1	1,2	1,6	2,0	2,2	2,4	2,5	2,6	2,7	2,7	2,7
4.000		0,5	0,7	0,8	0,9	1,0	1,4	1,7	1,9	2,1	2,2	2,3	2,3	2,4	2,4
5.000		0,4	0,6	0,7	0,8	0,9	1,3	1,5	1,7	1,8	1,9	2,0	2,1	2,1	2,1
7.500		0,3	0,5	0,6	0,7	0,8	1,0	1,2	1,4	1,5	1,6	1,7	1,7	1,7	1,7
10.000		0,3	0,4	0,5	0,6	0,7	0,9	1,1	1,2	1,3	1,4	1,4	1,5	1,5	1,5
15.000		0,2	0,3	0,4	0,5	0,5	0,7	0,9	1,0	1,1	1,1	1,2	1,2	1,2	1,2
25.000		0,2	0,3	0,3	0,4	0,4	0,6	0,7	0,8	0,8	0,9	0,9	0,9	0,9	0,9
50.000		0,1	0,2	0,2	0,3	0,3	0,4	0,5	0,5	0,6	0,6	0,6	0,7	0,7	0,7

3.6. Relación entre el tamaño de la muestra, de la población y del error

Finalmente incluimos un apartado con un análisis ilustrativo de las relaciones que dan entre el tamaño de la muestra y el tamaño de la población así como entre el tamaño de la muestra y el error muestral.

En el primer caso una representación gráfica del tamaño de la muestra en función del tamaño muestral (Gráfico III.4.18) dibuja una curva que pasa por el origen de coordenadas con un comportamiento asintótico que nos pone de manifiesto una relación directa y un comportamiento creciente (a medida que aumenta el tamaño de la población necesitamos más muestra para un determinado nivel de error muestral) pero no es un comportamiento de proporcionalidad, de crecimiento lineal. Al inicio, con poblaciones pequeñas, se requiere un número importante de elementos de la muestra, y a medida que crece el tamaño poblacional los elementos muestrales requeridos crecen notablemente, pero a partir de un determinado *N* los incrementos son cada vez más reducidos, hasta llegar a un momento en que los aumentos de *N* no producen aumentos en el tamaño de la muestra. A partir de entonces la población es infinita.

Gráfico III.4.18. Relación entre el tamaño de la muestra y el tamaño de la población
Poblaciones finitas, e=2%, z=2, P=Q=50%

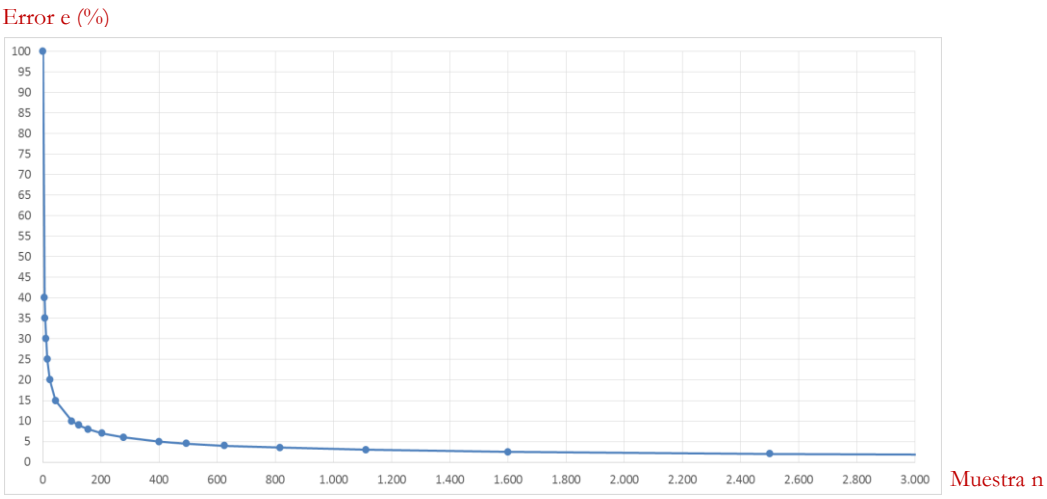


En particular, al pasar de una población 1 a 1000 la muestra necesaria pasa de 1 a 714, mientras que de 1000 a 2000 tan sólo necesitamos aumentar la muestra en 397, y de 2000 a 3000, el aumento es de 253, progresivamente la necesidad de nuevos elementos en la muestra disminuye hasta estabilizarse a partir de 100.000 individuos.

Un comportamiento similar se da entre el error muestral y el tamaño de la muestra en el sentido de que tampoco se trata de una relación proporcional lineal. Error y tamaño tienen una relación inversa, a medida que aumenta el tamaño muestral se reduce el error. Pero el tamaño *n* es inversamente proporcional al cuadrado del error de muestreo, lo que implica que cualquier pequeño incremento de unidades en la muestra, cuando éstas son pequeñas, producen un reducción muy importante del error cometido, pero llega un momento en que las reducciones son cada vez más pequeñas

de modo que reducciones mínimas del error suponen incrementos muy elevados del tamaño de la muestra.

Gráfico III.4.19. Relación entre el tamaño de la muestra y el margen de error. Población infinita, z=2, P=Q=50%



En particular, al pasar de un tamaño de 1 a 100 se reduce el error del 100% al 10%, una reducción de 90 puntos porcentuales, pero al pasar de 100 a 200, la reducción es tan solo del 3%.

4. Contraste de hipótesis

La inferencia estadística es el ejercicio de extrapolar un resultado obtenido en una muestra a toda la población. Hasta ahora hemos visto un tipo de problema de la inferencia estadística: la estimación puntual de parámetros mediante intervalos de confianza. Se trataba de determinar el error estadístico asociado a la estimación de una media o de una proporción y, vinculado a ello, también la cuestión de la determinación del tamaño de la muestra para efectuar estas estimaciones. En todos estos casos hemos razonado desde el estadística inferencial para dar significación y precisión a los resultados de estos cálculos. Pero existen otros tipos de cuestiones y resultados que requieren asimismo un razonamiento inferencial y que son fundamentales en la investigación científica. Nos referimos a las pruebas estadísticas destinadas a contrastar o validar hipótesis.

Una hipótesis no es más que una afirmación o enunciado fundamentado que se establece provisionalmente como respuesta a una problemática de investigación. Las hipótesis están formadas por conceptos que se relacionan entre sí, y están destinadas a orientar la investigación. Las hipótesis pueden ser desde generales y teóricas hasta concretas y operativas. Aquí nos referiremos a las hipótesis estadísticas, que son la expresión concreta de hipótesis operativas de investigación.

El contraste de hipótesis estadísticas se plantea cuando queremos inferir, a partir de una muestra, un enunciado poblacional. Se trata, por tanto, de determinar si, en términos de probabilidad, podemos aceptar o no una determinada hipótesis referida a

una característica de la población con los datos obtenidos en una muestra. El contraste mide la probabilidad de que un determinado resultado de un cálculo realizado en la muestra (el valor del estadístico muestral) sea fruto del azar y no significativo para extrapolarlo al conjunto de la población, o, por el contrario, el resultado sea significativo y generalizable a la población a la que se refiere.

Las hipótesis estadísticas se formulan en términos de una doble distinción: la hipótesis nula y la hipótesis alternativa. La hipótesis nula (H_0) afirma una determinada característica, comportamiento o valor de la población, mientras que la hipótesis alternativa (H_A) niega la anterior, es decir, no se da esta característica, comportamiento o el valor es diferente, mayor o menor.

Todo contraste de hipótesis implica tomar la decisión sobre si hay evidencias significativas para rechazar la hipótesis nula y, por tanto, considerar como buena la hipótesis alternativa, o bien no hay evidencias significativas para rechazarla, no podemos falsarla, y nos tenemos que quedar con la hipótesis nula. La decisión se toma siempre en términos de probabilidad, es lo que llamaremos como nivel de significación del contraste. La significación mide la probabilidad de obtener un valor del estadístico más extremo o igual que el obtenido en la muestra si la hipótesis nula es cierta: si supera determinado nivel, por ejemplo, como es habitual, considerando el nivel 0,05, entonces aceptaremos la hipótesis nula; si no supera este nivel, entonces, rechazaremos la hipótesis nula y aceptaremos la alternativa, puesto que consideramos que la probabilidad de obtener un valor como aquel es muy pequeña.

Ahora bien, cuando tomamos una decisión en un contraste, podemos correr el riesgo de equivocarnos. En estadística se definen formalmente dos tipos de errores en el contraste de hipótesis:

- El error de tipo I, con la probabilidad asociada α, que se comete cuando se rechaza la hipótesis nula y era cierta. Se concluye que la prueba estadística era significativa pero en realidad no lo era.
- El error de tipo II, con la probabilidad asociada β, el que se comete cuando no se rechaza la hipótesis nula y era falsa. Se concluye que la prueba estadística no era significativa pero en realidad sí lo era. asociada $1-\beta$ se conoce como el poder o potencia del test.

Estas son las diferentes situaciones que nos podemos encontrar en un contraste:

		Decisión estadística	
		Se acepta H_0	*No se acepta* H_0
Realidad	*Es cierta* H_0	**Correcta**	**Error Tipo I**
	Es falsa H_0	**Error Tipo II**	**Correcta**

En la práctica, solo controlamos que la probabilidad α del error de tipo I sea pequeña. Esto crea una asimetría entre las dos hipótesis: cuando hacemos un test, miramos si hay evidencias significativas para rechazar la hipótesis nula. Si no hay evidencias, no la podemos rechazar y aceptamos como cierta la hipótesis nula, aunque eso no quiere decir que se haya demostrado que sea cierta; simplemente no hay evidencias de que sea falsa.

El contraste de hipótesis da lugar a la realización de un test o prueba estadística cuyo resultado determina el rechazo o no de la hipótesis nula. Hay una amplia variedad de pruebas que nos permiten contrastar hipótesis, de hecho, cualquier resultado de un análisis de datos estadísticos tiene implicado algún contraste de hipótesis con la prueba correspondiente que da significación al resultado.

Los contrastes de hipótesis pueden plantearse en relación a un solo parámetro de una población, con parámetros de dos poblaciones o de más de dos. A partir del próximo capítulo veremos cómo van apareciendo cuando se plantean diversas hipótesis relacionales entre dos o más variables: cuando se comparan dos medias poblacionales o más de de dos, cuando se determina la asociación entre dos variables cualitativas, cuando se establece la significación de un coeficiente en un análisis de regresión, etc.

Todos los contrastes de hipótesis se basan en las distribuciones de probabilidades de los estadísticos. En los contrastes veremos que estas distribuciones se basan en la normal, en la t de Student, en la F de Fisher-Snedecor o en la distribución de Chi cuadrado.

Para realizar el contraste de una hipótesis estadística siempre se sigue el procedimiento siguiente:

1. Formulación de la hipótesis.
2. Cálculo del valor del estadístico muestral.
3. Determinación de la significación.
4. Decisión sobre la significación del estadístico, aceptando o rechazando la hipótesis nula.

De hecho se trata de un procedimiento cuya aplicación es muy mecánica y habitualmente se limita a ver si el valor de la probabilidad es menor o igual a 0,05 (es significativo –existe diferencia entre los valores testeados–) o por encima (no es significativo –no hay diferencia entre valores testeados–), a partir del dato que nos proporciona el software estadístico en una tabla.

Cuando en una prueba estadística se utiliza la distribución de probabilidad de la población considerada, se dice que la inferencia o la prueba es paramétrica, cuando se hace uso exclusivamente de la muestra, la prueba es no paramétrica.

Para acabar este capítulo veremos dos pruebas estadísticas sencillas destinadas a establecer la significación de una media y de una proporción.

4.1. Prueba estadística de una media

El propósito de la prueba estadística es evaluar, a partir de una muestra escogida al azar, si la media muestral $\overline{x}$ es igual o no a un cierto valor poblacional μ que se considera como un comportamiento teórico dado. Para dar cuenta de la prueba sigamos los cuatro pasos que hemos comentado.

1. Formulación de la hipótesis.

El test de una sola media establece como hipótesis nula que la media muestral es igual a la poblacional, lo que implica que la diferencia entre $\bar{x}$ y μ es pequeña y se debe a un efecto aleatorio por el hecho de haber escogido esta muestra al azar, y como alternativa que ambas medias difieren suficientemente como para que se deba a una fluctuación aleatoria. Las hipótesis nula y alternativa se expresan formalmente en los términos siguientes:

$$H_0: \mu=\mu_0$$
$$H_A: \mu\neq\mu_0$$

Es decir, si la media poblacional es un determinado valor μ_0 (hipótesis nula H_0) o bien es un valor distinto (hipótesis alternativa H_A). Si es un valor distinto no establecemos si es mayor o menor, por ello el contraste se denomina a dos colas pues no afirmamos, en el caso de que no sea igual, si el valor de la media es superior (positivo) o inferior (negativo).

Alternativamente podríamos formular un contraste a una sola cola afirmando bien que el valor de la media poblacional será mayor o menor a uno dado:

$$H_0: \mu\leq\mu_0 \qquad H_0: \mu\geq\mu_0$$
$$H_A: \mu>\mu_0 \qquad H_A: \mu<\mu_0$$

Consideremos como ejemplo ilustrativo el caso de la asistencia media al cine durante el año, tomando como indicador el número de veces que se asistió a una sala de cine. Según los datos más recientes disponibles del sector audiovisual la media de asistencia es de 2,5 veces al año. Realizamos una encuesta sobre hábitos y prácticas culturales y obtenemos que la media obtenida en una muestra de 2.000 personas es de 2,8. Nos preguntamos si este valor muestral difiere significativamente del dato poblacional, y establecemos la hipótesis de que difieren y que se está produciendo un aumento de la asistencia como resultado de la recuperación de la crisis.

2. Cálculo del estadístico muestral.

Para realizar el contraste de la hipótesis planteada debemos calcular el estadístico estableciendo la forma de la distribución del mismo. En este caso el estadístico compara dos valores, la media muestral $\bar{x}$ y la media poblacional μ, evaluando si son iguales, o lo que es lo mismo, si su diferencia $\bar{x}-\mu$ es cero. La forma de la distribución de este estadístico es la distribución normal o la distribución t de student, que depende de la situación:

- Con muestras pequeñas y conociendo la varianza poblacional así como con muestras grandes (a partir de 30 casos) se considera que la distribución es normal:

$$z=\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \qquad \textit{Ecuación 41}$$

  Donde $\sigma_{\bar{x}}$ es el error típico del estadístico de la media, e igual a σ/n si la desviación típica poblacional es conocida, o bien igual a s/n si la muestra es grande y se usa la desviación típica muestral.

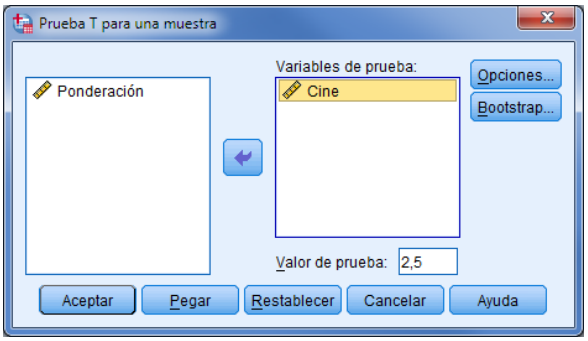
– Si la muestra es pequeña y la varianza poblacional es desconocida entonces se considera: la distribución de la t de student y el error típico $s_{\bar{x}}$ toma en cuenta la desviación típica muestral s/n.

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \textit{Ecuación 42}$$

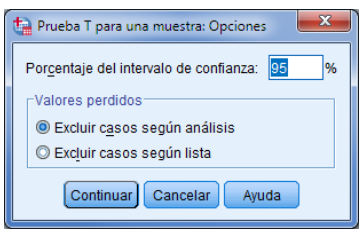
En nuestro ejemplo se trata de valorar si la diferencia entre $\bar{x} = 2{,}8$ y $\mu = 2{,}5$, $\bar{x} - \mu = 0{,}3$, es estadísticamente significativa, es una diferencia que se puede extrapolar a la población. Como el tamaño de nuestra muestra es muy grande, 2.000 individuos, podríamos considerar que el estadístico sigue una distribución normal, si bien dado el tamaño muestral el valor no diferirá del cálculo con un estadístico t [16]. Del estudio sabemos que la desviación típica muestral de la variable, del número de veces que se ha ido al cine, es de 0,8. Con estos datos construimos el valor del estadístico t:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2{,}8 - 2{,}5}{0{,}8/\sqrt{2000}} = 16{,}766$$

A este mismo resultado podríamos llegar con el software SPSS si dispusiéramos de la variable en una matriz de datos, aplicando el procedimiento de contraste de una media implementado en el software estadístico. Para ello se ejecuta el procedimiento del menú Analizar / Medias / Prueba t para una muestra, donde se trata de elegir la variable de la que se calcula la media y la desviación, pasándola al recuadro de la derecha, y que se comparará con el Valor de prueba 2,5:



En opciones no es necesario especificar nada, consideraremos el valor por defecto del 95% de nivel de confianza.



[16] Como vimos, a medida que aumenta el tamaño de la muestra la distribución t se va aproximando a la normal.

Tras ejecutar el procedimiento se obtiene este resultado:

Estadísticas de muestra única

	N	Media	Desviación estándar	Media de error estándar
Cine	15000	2,8000	,80000	,00653

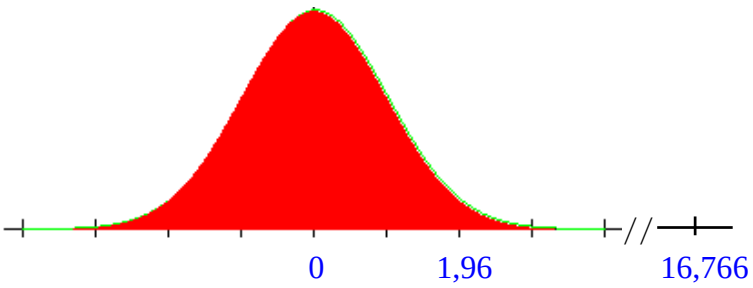
Prueba de muestra única

	Valor de prueba = 2.5					
					95% de intervalo de confianza de la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
Cine	16,766	1999	,000	,30000	,2649	,3351

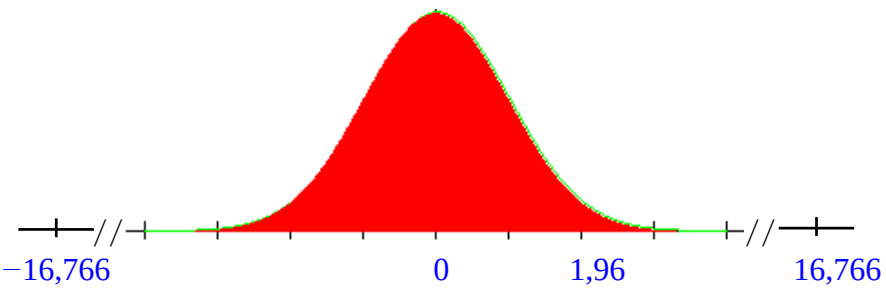
3. Determinación de la significación.

Establecida la hipótesis y calculado el estadístico se trata, en tercer lugar, de valorarlo en términos de probabilidad para determinar posteriormente si es significativo: si la diferencia observada es relevante o no. La probabilidad asociada al valor del estadístico *t* se puede consultar en una tabla teórica (que se puede obtener también con el software estadístico) o mediante los resultados que se obtienen con el software estadístico donde además del valor del estadístico que vimos antes nos proporciona la probabilidad asociada y otros datos de interés.

En el primer caso el valor $t=16,766$, con $v=n-1=2000-1=1999$ grados de libertad, determina una probabilidad acumulada del 100% prácticamente, dejando un 0% a la derecha[17]:



Como realizamos un contraste a dos colas esa probabilidad debe aplicarse al extremo derecho y el izquierdo:



[17] Este valor se puede obtener ejecutando la instrucción de SPSS: **COMPUTE** p=CDF.T (16.766,1999).

pero como es prácticamente cero se concluye igualmente que la probabilidad asociada a la hipótesis nula es inexistente.

Este mismo resultado se puede observar en los resultados obtenidos con el SPSS, donde la probabilidad de 0,000, es decir, de 0 a 3 decimales, es la significación en un contraste bilateral o a dos colas.

4. Decisión sobre la significación del estadístico.

Con este resultado nos queda finalmente extraer la conclusión de la prueba aceptando o rechazando la hipótesis nula. Bien sencillo:

- Si la probabilidad es mayor a 0,05 (sig. ≥ 0,05) se acepta la hipótesis nula, nuestra media muestral, 2,8, no difiere de la media poblacional, 2,5.
- Si la probabilidad es menor o igual a 0,05 (sig. < 0,05) se rechaza la hipótesis nula, y se concluye que nuestra media muestral, 2,8, difiere significativamente de la media poblacional, 2,5.

Como en nuestro caso la probabilidad observada es de 0,000, menor a 0,05, concluimos que podemos rechazar la hipótesis nula y afirmamos que las medias difieren, o afirmamos que las diferencias son significativas, por tanto, que se produce un aumento en la asistencia al cine en la actualidad.

En términos formales la prueba estadística considera el valor de significación α, habitualmente con $\alpha=0,05$, y se plantea en estos términos[18]:

Si $Pr(t_0) \geq \alpha$ aceptamos la hipótesis nula, es igual a la media poblacional.
Si $Pr(t_0) < \alpha$ rechazamos la hipótesis nula, no es igual a la media poblacional.

También se podría plantear en términos de los valores del estadístico:

Si $t_0 \leq t$ aceptamos la hipótesis nula, es igual a la media poblacional.
Si $t_0 > t$ rechazamos la hipótesis nula, no es igual a la media poblacional.

Donde t es un valor teórico que se fija, considerando una probabilidad o nivel de significación de $\alpha=0,05$ y considerando los grados de libertad $\nu=n-1$ de cada caso, si bien cuando el tamaño de la muestra es muy grande, a partir de un valor superior a 120, este valor se corresponde con el de la normal, 1,96. En este caso, por tanto, todo valor observado del estadístico superior a 1,96 implica rechazar la hipótesis nula, conclusión a la que llegamos en nuestro ejemplo pues: 16,766>1,96.

Nuestra estimación muestral difiere de la poblacional en tres décimas porcentuales: $\bar{x}-\mu=2,8-2,5=0,3$. Como se trata de una estimación puntual podemos construir su intervalo de confianza a partir de la diferencia entre las medias, el valor $0,3$, y sumando y restando el margen de error, donde consideramos una $t=1,96$: que corresponde a un nivel de confianza del 95% a dos colas:

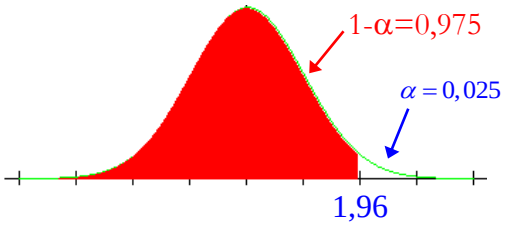
$$\bar{x} \pm t \cdot s_{\bar{x}} = 0,3 \pm 1,96 \cdot 0,8 / \sqrt{2000} = 0,3 \pm 0,0351$$

[18] t_0 es la expresión del estadístico observado con los datos de la muestra.

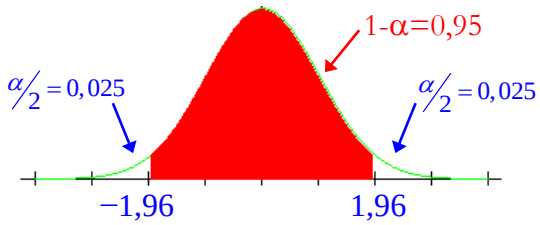
Por tanto, el intervalo [0,2649, 0,3351]. Estos mismos resultados se obtienen con el software SPSS.

El intervalo no contiene el valor cero indicando que la diferencia entre la media de la muestra y la de la población no es cero, que es una diferencia significativamente distinta de cero. Para el cálculo del intervalo, el valor de la *t* que aparece es el que corresponde a la distribución teórica de t de student considerando un nivel de confianza del 95% y 1999 grados de libertad en un contraste a dos colas (donde hay que buscar el valor correspondiente a una probabilidad de α=0,025, ver Anexo I), es decir, el valor 1,96, que coincide con el valor *z* de la normal.[19]

La tabla teórica proporciona el valor acumulado, que para el 97,5% del área se representa de esta forma donde solo aparece una cola:



En nuestro caso consideramos dos colas y el 95% del área (que implica un 5% de significación, es decir, 0,05, que repartido entre las dos colas es 0,025) el intervalo se situará entre el −1,96 y el 1,96:



A pesar de que hemos dedicado cuatro páginas a dar cuenta de este contraste de hipótesis, en resumidas cuentas, se trata de cuatro pasos sencillos:

1. Formulación de la hipótesis: la media muestral 2,8 difiere de la poblacional 2,5.
2. Cálculo del valor del estadístico muestral: una t de student con t_0=16,77.
3. Determinación de la significación: la probabilidad asociada es 0,000.
4. Decisión sobre la significación, aceptando o rechazando la hipótesis nula: como 0,000 ≤ 0,05 rechazamos la hipótesis nula y concluimos que la media muestral difiere del valor teórico poblacional.

[19] Este valor se puede obtener ejecutando la instrucción de SPSS: **COMPUTE** t=IDF.T(0.975,1999).

4.2. Prueba estadística de una proporción

Aquí de nuevo se plantea una prueba estadística similar, se trata de saber si la proporción muestral p se corresponde con la proporción P de la población que se considera como un comportamiento teórico dado. La proporción es un caso particular de media como vimos y la estimación de una proporción poblacional también se comporta como una normal (con muestras suficientemente grandes, siempre que $n \times p \geq 5$) que nos ayudará a determinar si la proporción muestral, a pesar de diferir de la teórica, se debe a las fluctuaciones del azar, generando diferencias no significativas, o bien obedece a una diferencia cierta. Aplicamos el contraste en las cuatro etapas que hemos visto.

1. Formulación de la hipótesis.

En general se establece como hipótesis nula que la proporción muestral es igual a la poblacional, lo que implica que la diferencia entre p y P es pequeña y se debe al azar de la muestra elegida, y como alternativa que ambas proporciones difieren suficientemente como para que se deba al azar:

$$H_0: P = P_0$$
$$H_A: P \neq P_0$$

Para ilustrar la prueba estadística podemos tomar como ejemplo la estimación del nivel de abstención de las próximas elecciones, después de que en las celebradas unos pocos meses antes no condujeron a la formación de un gobierno y tuvieron que convocarse de nuevo. En estas condiciones nos preguntamos si el nivel de abstencionismo será similar al de las últimas elecciones o bien será distinto. También podríamos matizar nuestra hipótesis afirmando que cabe esperar un mayor nivel de abstención. Sabemos que el porcentaje de personas que se abstuvieron en las pasadas elecciones fue del 30%. Realizamos un sondeo preelectoral a una muestra representativa de 1.000 personas con derecho a voto y obtenemos un porcentaje de abstencionistas del 31%. Cabe plantearse pues si nuestro resultado nos indica que aumenta el nivel de abstención o por el contrario el cambio no es significativo y se deben a fluctuaciones aleatorias. Por tanto, se plantean en este caso las hipótesis nula y alternativa siguientes:

$$H_0: P = 30\%$$
$$H_A: P \neq 30\%$$

2. Cálculo del estadístico muestral.

Si la proporción muestral p pertenece al intervalo de probabilidad teórico se considera que la diferencia entre p y P no difieren significativamente, más allá de una variación aleatoria. En ese caso no se rechaza la hipótesis nula. En caso contrario, si la proporción observada en la muestra p se sitúa fuera del intervalo, se considera que la diferencia entre p y P es significativa y se acepta la hipótesis alternativa.

Para efectuar esta comparación calculamos el estadístico z:

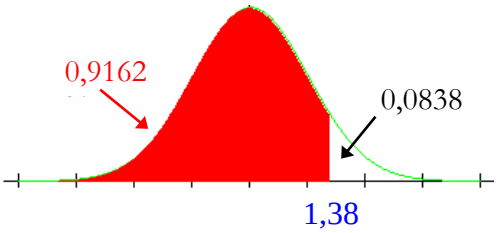
$$z = \frac{p - P}{s_p} = \frac{p - P}{\sqrt{pq/n}} \qquad \textit{Ecuación 43}$$

En nuestro ejemplo:

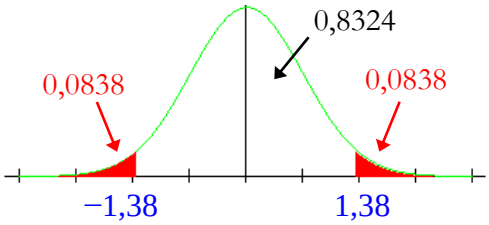
$$z = \frac{p - P}{\sqrt{pq/n}} = \frac{0,32 - 0,30}{\sqrt{0,3 \cdot 0,7/1000}} = 1,38$$

3. Determinar la significación.

La probabilidad asociada al valor observado de z=1,38 de la normal, la $N(0,1)$ es 0,91621 según la tabla de la normal (Tabla III.4.6). Es decir, el valor 1,38 deja a su derecha un área del 8,38%, esto es, la probabilidad de obtener un valor del estadístico más extremo o igual que el obtenido en la muestra si la hipótesis nula es cierta es de 0,0838:



Como estamos realizando un contraste a dos colas, este valor de probabilidad debemos multiplicarlo por dos, es decir, 0,1676 y plantear el test en estos términos:



Por tanto, la probabilidad asociada al estadístico es 0,1676.

4. Decisión sobre la significación del estadístico.

El contraste de la hipótesis se establece en los siguientes términos:

Si $Pr(z) \geq \alpha$ aceptamos la hipótesis nula, es igual a la proporción poblacional.
Si $Pr(z) < \alpha$ rechazamos la hipótesis nula, no es igual a la proporción poblacional.

Considerando el nivel de significación habitual α=0,05, como hemos obtenido una probabilidad superior: 0,1676>0,05, concluimos que debemos aceptar la hipótesis nula[20]. En otras palabras, la probabilidad de que se dé la hipótesis nula es 0,1676, un valor que se considera muy alto y nos conduce a no poder rechazarla, por lo que el 31% de abstencionismo obtenido en nuestra muestra es equivalente al 30% y, en consecuencia, cabe esperar un nivel de abstencionismo similar en las próxima convocatoria electoral.

[20] Lo mismo se podría concluir afirmando que el valor 1,38 es inferior a 1,96.

4.3. Razonamiento estadístico de una prueba de hipótesis

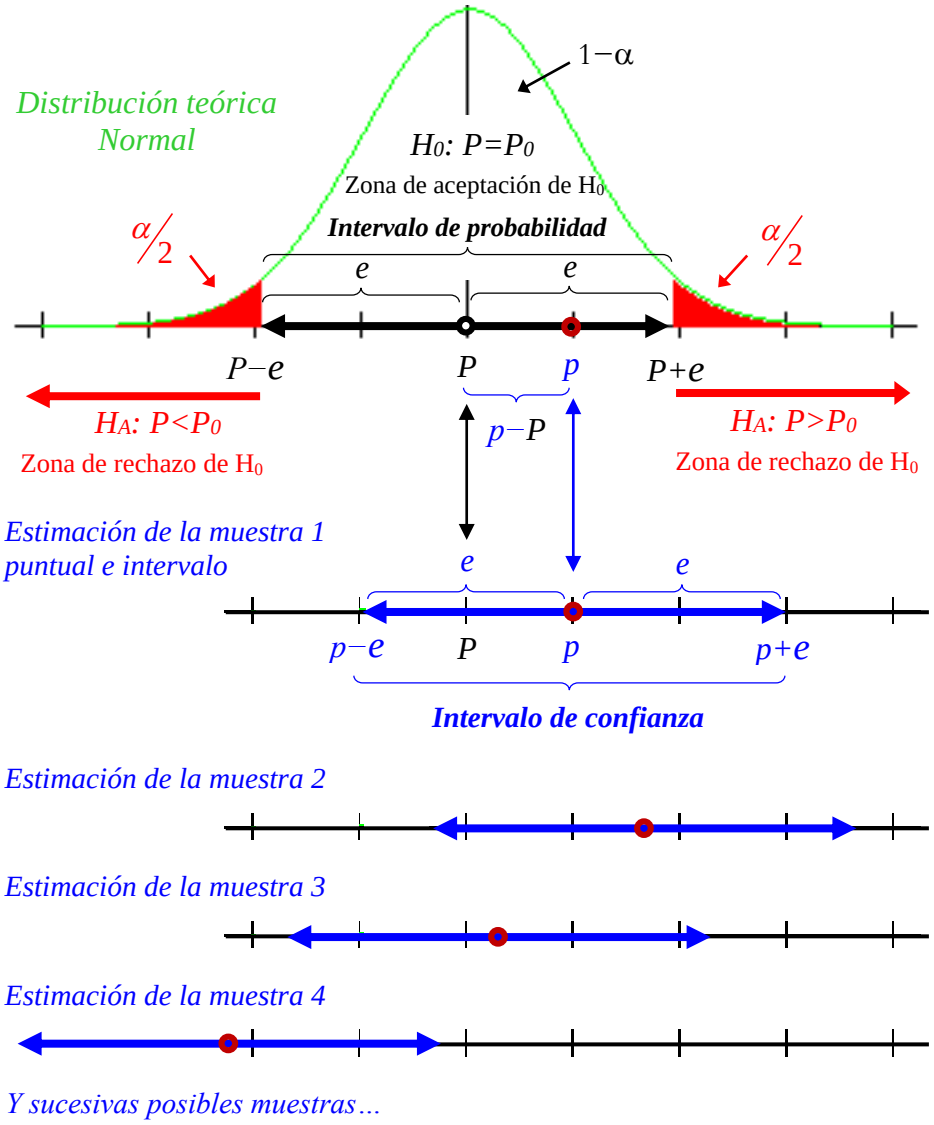
Cualquier prueba de hipótesis, como lo acabamos de ver, se plantea en términos de dos afirmaciones: la hipótesis nula y la hipótesis alternativa. Sabiendo la distribución teórica del estadístico que permite realizar el contraste de hipótesis obtenemos un valor al cual se le asocia una probabilidad. Cuando esta probabilidad es inferior a 0,05 concluimos que la prueba es significativa se acepta la hipótesis alternativa.

El Gráfico III.4.20 ilustra el razonamiento estadístico que se realiza. Tomamos el ejemplo de la estimación de la proporción que vimos en el apartado anterior. Sabiendo que el estadístico de la proporción se distribuye como una normal se considera el intervalo de probabilidad (flecha doble de color negro) que se construye a partir de la proporción poblacional P (marcada con una redonda o círculo negro) y los distintos valores muestrales que dan lugar, en cada muestra, a un valor distinto estimado p (se presentan en el gráfico 4 muestras con diferentes valores de p marcados con una redonda o círculo granate). Esta variabilidad de posibles valores se expresa en términos de unidades de desviación, de desvíos hacia la derecha y la izquierda que en el gráfico se expresan por e. En relación a la proporción poblacional se establece así el intervalo de probabilidad, en particular, sabiendo que en el 95% de las muestras estos valores se sitúan entre −1,96 y 1,96 unidades de desviación, lo que en el gráfico está expresado de forma genérica como $P-e$ y $P+e$, marcando los límites del intervalo.

Cuando estimamos una proporción en la muestra y este valor cae dentro del intervalo de probabilidad (como en el casp de p de la primera muestra que aparece en el gráfico) significa que las diferencias entre el valor muestral y la proporción poblacional, $p-P$, es pequeña, esto es, se considera que no es estadísticamente significativa. Todo valor de p que se obtenga en cada posible muestra (redondas de color granate en el gráfico) es un posible valor que puede caer dentro del intervalo de probabilidad y se valorará como igual al poblacional y se acepta la hipótesis nula. En cambio si cae fuera de ese intervalo querrá decir que es diferente y se rechazaría. Las muestras 1, 2 y 3 dan lugar a tres estimaciones que caen dentro del intervalo de probabilidad, coinciden con la proporción poblacional, son valores distintos porque se obtienen en muestras distintas donde las diferencias se deben a las fluctuaciones del azar, pero no a verdaderas diferencias. En cambio, la muestra 4 da lugar a un valor que se sitúa fuera del intervalo mostrando que no es igual a la proporción poblacional.

Por otro lado, en cada posible muestra, obtenemos un valor concreto de p, y, sabiendo que se distribuye normalmente, podemos construir un intervalo de confianza (flecha doble de color azul) que se obtiene sumando y restando el margen de error e. Este intervalo nos permite garantizar que en el 95% de los casos nuestra muestra contendrá la proporción poblacional.

Gráfico III.4.20. Representación de una prueba de hipótesis



5. Bibliografía

Agresti, A.; Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Upper Saddle RIver (New Jersey): Pearson Prentice Hall. 4ª edición.

Alabert, A. (2015). *The Modelling of Random Phenomena*. Materials Matemàtics, 3. Departament de Matemàtiques, Universitat Autònoma de Barcelona. http://ww.mat.uab.cat/matmat

Badiella, Ll. (2015). *Com fer-se ric amb la lotería (i aprendre Estadística en l'intent)*. Materials Matemàtics, 5. Departament de Matemàtiques, Universitat Autònoma de Barcelona. http://ww.mat.uab.cat/matmat

Bardina, X.; Farré, M.; López-Roldán, P. (2005). *Estadística: un curs introductori per a estudiants de ciències socials i humanes. Volum 2: Descriptiva i exploratòria bivariant*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona. Col·lecció Materials, 166.

Blalock, H. M. (1978). *Estadística Social*. México: Fondo de Cultura Económica.

Bryman, A., Duncan, C. (1990). *Quantitative data analysis for social scientists*. London: Routledge.

Castillo, J. del (2011). *La Campana de Gauss*. Materials Matemàtics, 4. Departament de Matemàtiques, Universitat Autònoma de Barcelona. http://ww.mat.uab.cat/matmat

Dalgaard, P. (2008). *Introductory Statistics with R*. New York: Springer.

De La Cruz-Oré, J. L. (2013). ¿Qué significan los grados de libertad? *Revista Peruana de Epidemiología*, 17, 2, agosto, 1-6. www.redalyc.org/pdf/2031/203129458002.pdf

Domènech, J. M. (1982). *Bioestadística. Métodos estadísticos para investigadores*. Barcelona: Herder. 4ª. Edición.

Everitt, B. S.; Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. New York: Cambridge University Press. Fourth Edition.

Finkelstein, M. O.; Levin, B. (2001). *Statistics for Social Science and Public Policy*. New York: Springer. 2ª edición.

García Ferrando, M. (1994) *Socioestadística. Introducción a la estadística en sociología*. 2a edició rev. i amp. Madrid: Alianza. Alianza Universidad Textos, 96.

Guisan de González, C.; Vaamonde Liste, A,; Barreiro Felpeto, A. (2011). *Tratamiento de datos con R, STATISTICA y SPSS*. Madrid: Díaz Santos.

Hopkins, K. D.; Hopkins, B. R.; Glass, G. V. (1997). *Estadística Básica para las ciencias sociales y del comportamiento*. 3ª. edición. Naucalpan de Juárez: Prentice-Hall Hispanoamericana.

Jolis, M. (2013). *Probabilitat i Anàlisi, uns grans amics*. Materials Matemàtics, 2. Departament de Matemàtiques, Universitat Autònoma de Barcelona. http://ww.mat.uab.cat/matmat

Landau, S.; Everitt, B. S. (2004). *A Handbook of Statistical Analyses using SPSS*. Boca Raton: Chapman & Hall/CRC.

Losilla, J. M.; Navarro, B.; Palmer, A.; Rodrigo, M. F.; Ato, M. (2005). *Análisis de Datos. Del contraste de hipótesis al modelado estadístico*. Girona: Documenta Universitaria.

Marqués, J. P. (2007). *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*. Berlin: Springer. 2ª edición.

Moncho, J. (2015). *Estadística aplicada a las ciencias de la salud*. Barcelona: Elsevier.

Peró, M. et al. (2012). *Estadística aplicada a las ciencias sociales mediante R y R-Commander*. Madrid: Ibergarceta Publicaciones.

Sánchez Carrión, J. J. (1999). *Manual de análisis estadístico de los datos*. Madrid: Alianza. Manuales 055.

Suñer, J. del (2010). *Històries de la probabilitat*. Materials Matemàtics, 6. Departament de Matemàtiques, Universitat Autònoma de Barcelona. http://ww.mat.uab.cat/matmat

Visauta Vinacua, B. (2002). *Análisis estadístico con SPSS 11.0 para Windows. Volumen I. Estadística básica*. 2a. Edició. Madrid: McGraw-Hill.

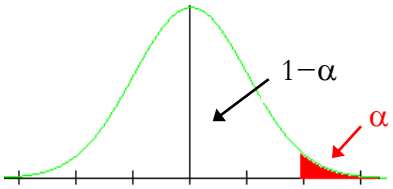
Von Mises, R. (1931). Wahrscheinlichkeitsrechnung und IhreAnwnedung in der Statistik und Theoretischen Physik. Leipzig: F. Deuticke.

Wonacott, Th. H.; Wonacott, R. J. (1979). *Fundamentos de Estadística para Administración y Economía*. México: Limusa.

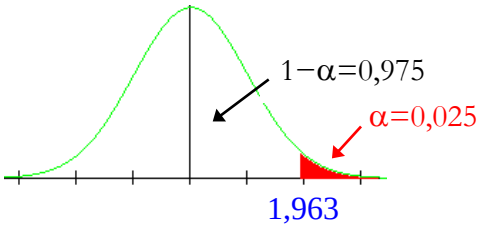
Anexo I. Tabla de distribución teórica de la t de Student

Grados de libertad ν	Probabilidades α (asociadas al valor crítico en un contraste a una cola)						
	0,10	0,075	0,05	0,025	0,01	0,005	0,001
1	3,078	4,165	6,314	**12,706**	31,821	63,657	318,309
2	1,886	2,282	2,920	**4,303**	6,965	9,925	22,327
3	1,638	1,924	2,353	**3,182**	4,541	5,841	10,215
4	1,533	1,778	2,132	**2,776**	3,747	4,604	7,173
5	1,476	1,699	2,015	**2,571**	3,365	4,032	5,893
6	1,440	1,650	1,943	**2,447**	3,143	3,707	5,208
7	1,415	1,617	1,895	**2,365**	2,998	3,499	4,785
8	1,397	1,592	1,860	**2,306**	2,896	3,355	4,501
9	1,383	1,574	1,833	**2,262**	2,821	3,250	4,297
10	1,372	1,559	1,812	**2,228**	2,764	3,169	4,144
11	1,363	1,548	1,796	**2,201**	2,718	3,106	4,025
12	1,356	1,538	1,782	**2,179**	2,681	3,055	3,930
13	1,350	1,530	1,771	**2,160**	2,650	3,012	3,852
14	1,345	1,523	1,761	**2,145**	2,624	2,977	3,787
15	1,341	1,517	1,753	**2,131**	2,602	2,947	3,733
16	1,337	1,512	1,746	**2,120**	2,583	2,921	3,686
17	1,333	1,508	1,740	**2,110**	2,567	2,898	3,646
18	1,330	1,504	1,734	**2,101**	2,552	2,878	3,610
19	1,328	1,500	1,729	**2,093**	2,539	2,861	3,579
20	1,325	1,497	1,725	**2,086**	2,528	2,845	3,552
21	1,323	1,494	1,721	**2,080**	2,518	2,831	3,527
22	1,321	1,492	1,717	**2,074**	2,508	2,819	3,505
23	1,319	1,489	1,714	**2,069**	2,500	2,807	3,485
24	1,318	1,487	1,711	**2,064**	2,492	2,797	3,467
25	1,316	1,485	1,708	**2,060**	2,485	2,787	3,450
26	1,315	1,483	1,706	**2,056**	2,479	2,779	3,435
27	1,314	1,482	1,703	**2,052**	2,473	2,771	3,421
28	1,313	1,480	1,701	**2,048**	2,467	2,763	3,408
29	1,311	1,479	1,699	**2,045**	2,462	2,756	3,396
30	1,310	1,477	1,697	**2,042**	2,457	2,750	3,385
32	1,309	1,475	1,694	**2,037**	2,449	2,738	3,365
34	1,307	1,473	1,691	**2,032**	2,441	2,728	3,348
36	1,306	1,471	1,688	**2,028**	2,434	2,719	3,333
38	1,304	1,469	1,686	**2,024**	2,429	2,712	3,319
40	1,303	1,468	1,684	**2,021**	2,423	2,704	3,307
42	1,302	1,466	1,682	**2,018**	2,418	2,698	3,296
44	1,301	1,465	1,680	**2,015**	2,414	2,692	3,286
46	1,300	1,464	1,679	**2,013**	2,410	2,687	3,277
48	1,299	1,463	1,677	**2,011**	2,407	2,682	3,269
50	1,299	1,462	1,676	**2,009**	2,403	2,678	3,261
60	1,296	1,458	1,671	**2,000**	2,390	2,660	3,232
70	1,294	1,456	1,667	**1,994**	2,381	2,648	3,211
80	1,292	1,453	1,664	**1,990**	2,374	2,639	3,195
90	1,291	1,452	1,662	**1,987**	2,368	2,632	3,183
100	1,290	1,451	1,660	**1,984**	2,364	2,626	3,174
110	1,289	1,450	1,659	**1,982**	2,361	2,621	3,166
120	1,289	1,449	1,658	**1,980**	2,358	2,617	3,160
150	1,287	1,447	1,655	**1,976**	2,351	2,609	3,145
200	1,286	1,445	1,653	**1,972**	2,345	2,601	3,131
250	1,285	1,444	1,651	**1,969**	2,341	2,596	3,123
500	1,283	1,442	1,648	**1,965**	2,334	2,586	3,107
750	1,283	1,441	1,647	**1,963**	2,331	2,582	3,101
1000	1,282	1,441	1,646	**1,962**	2,330	2,581	3,098
2000	1,282	1,440	1,646	**1,961**	2,328	2,578	3,094
3000	1,282	1,440	1,645	**1,961**	2,328	2,577	3,093
4000	1,282	1,440	1,645	**1,961**	2,327	2,577	3,092
5000	1,282	1,440	1,645	**1,960**	2,327	2,577	3,092
10000	1,282	1,440	1,645	**1,960**	2,327	2,576	3,091
∞	1,282	1,440	1,645	**1,960**	2,326	2,576	3,090

Cada valor de la tabla hace referencia al valor *t* que deja un área (probabilidad) a la derecha del mismo, igual al valor de la columna:



Por ejemplo, el valor 1,963 se corresponde con la línea ν=750 grados de libertad y la columna de probabilidad o nivel de significación α=0,025.



Este valor es el que corresponde a un contraste a dos colas considerando una significación del 0,05, dejando a cada lado el 0,025: