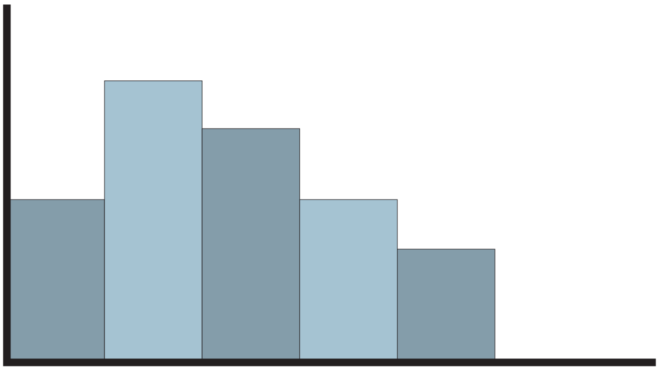
www.cisolog.com



Fórmulas y apuntes de Estadística aplicada a las ciencias sociales

Mayo 2013

Rubén Crespo
cisolog@cisolog.com

Nota Previa

Este documento no tiene ninguna finalidad pedagógica. Se trata de una compilación de las principales fórmulas estadísticas y apuntes teóricos recogidos del los libros de Luis Camarero Rioja et al. 2010. *Estadística para la investigación social*. Ibergarceta; y Alejandro Almazán et al. 2011. *Análisis estadístico para la investigación social*. Ibergarceta, que constituye respectivamente los manuales básicos de las asignaturas *Estadística social aplicada a las ciencias sociales I* y *II* de los planes de estudio de Grado de Sociología en la UNED.

La finalidad de este documento es ofrecer una herramienta de apoyo tanto para el profesional como para el estudiante que ha asimilado los conocimientos básicos de estadística aplicada a las ciencias sociales, de manera que cuando realice prácticas de investigación social donde se requiere el uso de la estadística, pueda disponer de un material de consulta que concentra las principales fórmulas y elementos teóricos útiles para resolver la mayoría de los problemas que se le planteen. Por tanto, el contenido de este documento, al no disponer de explicaciones precisas y exposición de ejemplos, no es conveniente para el aprendizaje de la estadística enfocada a la investigación social. Para ello se recomienda la lectura y comprensión del citado manual. Si bien hay otros muchos manuales cuyo aprendizaje servirían para entender las formulas y apuntes teóricos que se compilan aquí, en este documento los contenidos se han estructurado en el mismo orden que los manuales señalados más arriba.

La mayoría de las tablas y gráficos son de elaboración propia, excepto un par de ellos que se han extraído del primer manual, así como las tablas de los anexos que se pueden encontrar al final del documento.

Rubén Crespo
Mayo, 2013

Índice

Fórmulas y apuntes de Estadística aplicada a las Ciencias Sociales

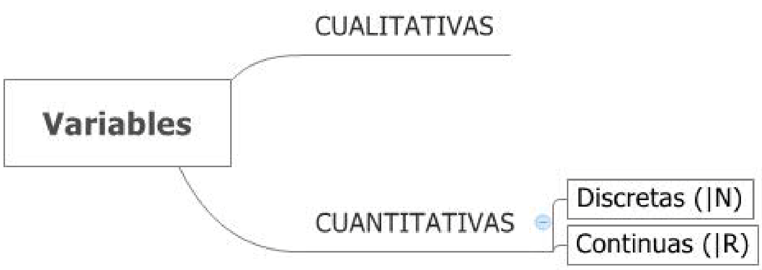
Tipología de los datos

	Ventajas	Inconvenientes
CENSOS	• Incluyen a toda la población. • Posibilidad de desagregación en subpoblaciones.	• Coste económico elevado. • Periodicidad dilatada.
REGISTROS	• Actualización frecuente.	• Disponible sólo para poblaciones específicas. • Incluyen pocas variables • Consulta restringida.
ENCUESTAS	• Costes menores. • Información instantánea. • Pueden incluir mayor número de variables. • Mayor posibilidad de verificación de los datos.	• Están sujetas a errores de muestreo. • Dificultades de desagregación de la información de forma detallada por el tamaño limitado de la muestra. • Necesidad de previa existencia de censos para seleccionar la muestra.

Tipos de escala de las variables

	Propiedades	Ejemplos
NOMINAL (N-1)	Clasifica	Lugar de nacimiento, sexo, estado civil, lugar de residencia, etc.
ORDINAL (N-2)	Clasifica y ordena	Nivel de estudios, grado de satisfacción, jerarquía de mando, etc.
INTERVALO (N-3)	Clasifica, ordena y posee unidad de medida	Tamaño del hogar, fecha, temperatura, etc.
RAZÓN (N-4)	Clasifica, ordena, posee unidad de medida, y origen = “0 absoluto” “0” es la ausencia de característica	Número de hijos, renta familiar, peso, distancia, etc.

Otra clasificación de las variables



Distribuciones de Frecuencias

n_i → frecuencias absoluta

$N = n_1 + n_2 + n_3 + \ldots + n_i$ N → Número de casos de la población

Frecuencia Relativa

$$\boxed{fr_i = \frac{n_i}{N}}$$

Frecuencia Acumulada

$$\boxed{Na_i = Na_{i-1} + n_i}$$

Límites reales para creación de categorías según variables de intervalo

Los intervalos reales de las categorías dependen de la naturaleza de la variable

- Variable:
  - Discreta → Límites Clase = Límites Reales
  - Continua → Depende de la variable

    Ejemplos:

    Distancia → de 3 hasta 4 → de 2,5 hasta 4,4999… [2,5 – 4,5[

    Edad → de 3 hasta 4 → de 3 hasta 4,4999… [3 – 5[

Tipos Representación Gráfica distribución de frecuencias según "apuntamientos"

1. **PLATICÚRTICA** → Aplanada Frecuencias similares a lo largo del recorrido.
2. **LEPTOCÚRTICA** → Puntiaguda Frecuencias altas en pocos valores.
3. **MESOCÚRTICA** → Intermedia Intermedia entre las dos anteriores.

Medidas de Tendencia Central

Media Aritmética

$$\boxed{\bar{x} = \frac{\sum x_i}{N}}$$

$$\bar{x} = \frac{\sum x_i}{N} = \frac{x_1 + x_2 + x_3 + \ldots + x_i}{N}$$

Media Aritmética datos agrupados

$$\boxed{\bar{x} = \frac{\sum x_i n_i}{N}}$$

x_i ⟶ Marca de clase

Media estadística Ponderada

$$\boxed{\bar{x} = \frac{\sum x_i p_i}{\sum p_i}}$$

$$\bar{x} = \frac{x_1p_1 + x_2p_2 + x_3p_3 + \ldots + x_np_n}{p_1 + p_2 + p_3 + \ldots + p_n}$$

Cálculo de las Marcas de Clase

$$\boxed{x_i = \frac{L_{inf} + L_{sup}}{2}}$$

A partir de los límites reales[1]

Mediana

$$\boxed{Me = L_i + \frac{\left(\frac{N}{2} - Na_{i-1}\right) C_i}{n_i}}$$

$$Me = L_i + \left(\frac{N}{2} - Na_{i-1}\right) \cdot \frac{C_i}{n_i}$$

L_i → Límite Inferior

C_i → Amplitud del Intervalo $C_i = L_{sup} - L_{inf}$

Cuartiles

$$\boxed{Q_1 = L_i + \frac{\left(\frac{N}{4} - Na_{i-1}\right) C_i}{n_i}} \qquad \boxed{Q_3 = L_i + \frac{\left(\frac{3N}{4} - Na_{i-1}\right) C_i}{n_i}}$$

$Q_2 = Me$

$Q_4 = L_{sup}$ = Valor Máximo de la Población/Muestra

Deciles

$$\boxed{D_1 = L_i + \frac{\left(\frac{N}{10} - Na_{i-1}\right) C_i}{n_i}} \qquad \boxed{D_9 = L_i + \frac{\left(\frac{9N}{10} - Na_{i-1}\right) C_i}{n_i}}$$

Centiles

$$\boxed{C_1 = L_i + \frac{\left(\frac{N}{100} - Na_{i-1}\right) C_i}{n_i}} \qquad \boxed{C_{199} = L_i + \frac{\left(\frac{99N}{100} - Na_{i-1}\right) C_i}{n_i}}$$

Moda

$$\boxed{Mo = L_i + \frac{n_{i+1}}{n_{i+1} + n_{i-1}} C_i} \qquad \boxed{Mo = L_i + \frac{n_{i+1}C_i}{n_{i+1}C_{i-1} + n_{i-1}C_{i+1}} C_i}$$

Cálculo de Percentiles

$$\boxed{i = \left(\frac{p}{100}\right) \cdot N}$$

Rango Intercuántico: $\boxed{\boldsymbol{I = Q_3 - Q_1}}$ **Rango Semi-Intercuantílico:** $\boxed{\boldsymbol{I = \frac{Q_3 - Q_1}{2}}}$

---

[1] Ejemplo: si para la variable "edad" de una muestra quisiéramos agregar los datos en diferentes grupos de edades, en el caso del grupo "de 25 a 29 años" el límite inferior es 25, y el límite superior es 30 [25 – 30[. Por tanto, la marca de clase para ese grupo de edad es 27,5 años.

Medidas de Dispersión

Varianza

$$V_x = S_x^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

$$S_x^2 = \frac{\sum x_i{}^2 n_i}{N} - \bar{x}^2$$ ← Fórmula Alternativa V_x

Desviación Estándar (o Típica)

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

$$S_x = \sqrt{\frac{\sum x_i{}^2 n_i}{N} - \bar{x}^2}$$ ← Fórmula Alternativa S_x

Varianza para datos agrupados

$$V_x = S_x^2 = \frac{\sum(x_i - \bar{x})^2 \cdot n_i}{N}$$

Desviación Estándar (o Típica) para datos agrupados

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2 \cdot n_i}{N}}$$

Coeficiente de Variación (para comparar el grado de dispersión entre varias distribuciones)

$$V = \frac{S_x}{\bar{x}}$$

$$V = \frac{S_x}{\bar{x}} \cdot 100$$ Expresado en porcentaje (%)

Medidas de Localización Relativa

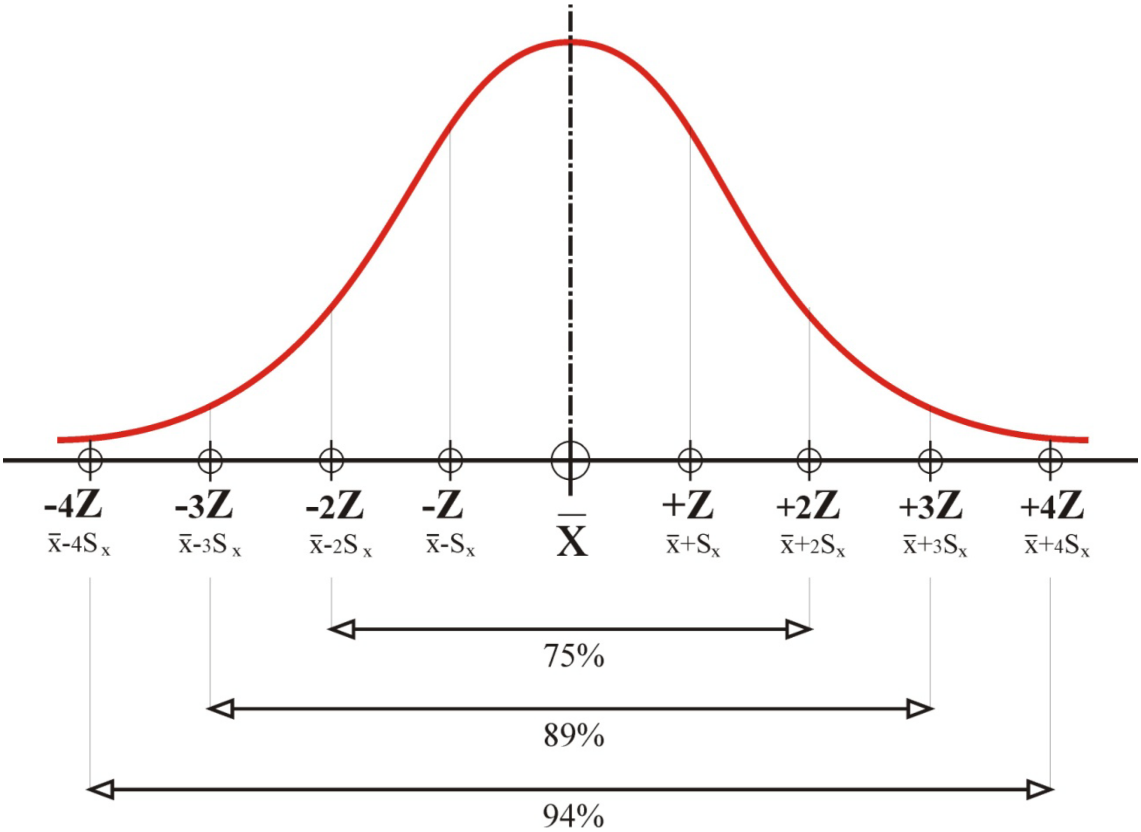
Valores de Z

$$Z = \frac{x_i - \bar{x}}{S_x}$$

Nos indica a cuántas desviaciones estándar está cualquier valor (x_i) de la Media ($\bar{x}$).

Los valores de Z los utilizamos para localizar de manera relativa cada valor en la distribución en la que se encuentra, permitiendo la comparación entre valores de distribuciones distintas. Se establece en cada distribución la distancia de cualquier valor x a la media ($\bar{x}$) y se mide esta distancia en unidades de desviación típica o estándar (S_x).

El Teorema de Chebyshev



*Porcentaje de casos (%) independientemente de que la distribución sea normal o no.

Para cualquier otro valor se calcula:

$$1 - \frac{1}{Z^2}$$

Otras formas de calcular la Media

Media Armónica

$$MH = \frac{N}{\sum\left(\frac{1}{x_i}\right)}$$

Media Cuadrática

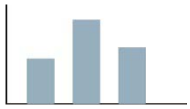
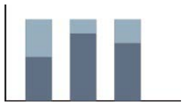
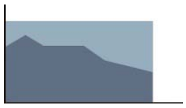
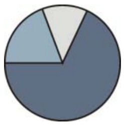
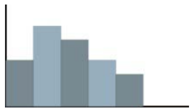
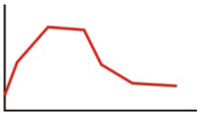
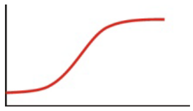
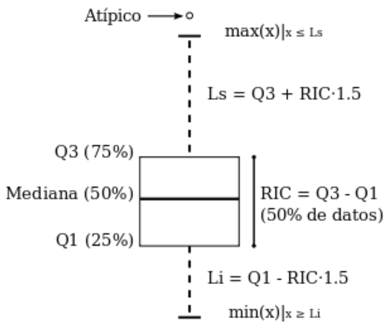
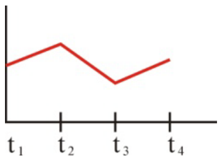
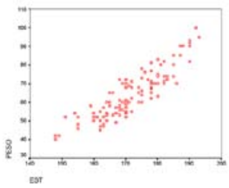
$$MC = \sqrt{\frac{\sum x_i^2}{N}}$$

Media Geométrica

$$MG = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

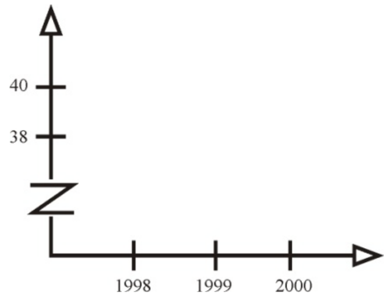
$$MG = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$$

Tipos de Gráfico

Problema de Investigación	Tipo de Gráfico	Representación Gráfica
Descripción y Comparación de Categorías	DIAGRAMA DE BARRAS	
	DIAGRAMA DE BARRAS APILADAS	
	DIAGRAMA DE ÁREAS APILADAS	
	DIAGRAMA DE SECTORES	
Análisis y Comparación de Distribuciones	HISTOGRAMA S = b x h → h = S/b	
	POLÍGONO DE FRECUENCIAS	
	OJIVA	
	DIAGRAMA DE CAJAS	
Análisis Temporales	DIAGRAMA DE LÍNEAS	
Distribución conjunta de dos variables	DIAGRAMA DE DISPERSIÓN	

Elementos básicos para la construcción de gráficos

1. Fuente
2. Título y subtítulos
3. Año o periodo
4. Etiquetado de los Ejes
5. Diferenciación de las categorías
6. Base (el "0" se debe representar si forma base de la escala)
7. Discontinuidad en el Eje Vertical (si es necesario)



Escala del Gráfico:

- **Aritmética**
- **Logarítmica** → Gráfico semilogarítmico.

Teoría y cálculo de Probabilidades

Probabilidad *a priori* y frecuencia relativa

$$\boxed{P(S) = \frac{n}{N}}$$

$$0 \leq P(S) \leq 1$$

Suceso Elemental → **Unidad Muestral**

$\mathscr{E}$ = Espacio de Sucesos → Conjunto de todos los sucesos posibles → **Espacio Muestral**

Suceso Complementario

$$\boxed{P(S) + P(\bar{S}) = 1}$$

Técnicas de Conteo (Agrupaciones/Ordenaciones)

Permutación

$$\boxed{P_n = n!}$$

Variaciones

$$\boxed{V_{n,r} = n(n-1)(n-2)\ldots(n-r+2)(n-r+1)}$$

Combinaciones

$$\boxed{C_{n,r} = \binom{n}{r} = \frac{n!}{r! \cdot (n-r)!}}$$

$$\boldsymbol{C_{n,r} = \frac{n(n-1)(n-2)\ldots(n-r+2)(n-r+1)}{r!}}$$

Tipos de Sucesos

- **Simultáneos** → son, o no, mutuamente excluyentes.
- **Sucesivos/Secuenciales** → Observar si el anterior condiciona al posterior

  - **Elemental**
  - **Compuesto**
    - ∪ "unión" → (V) / "o" [Simultáneo]
    - ∩ "intersección → (∧) / "y" [Secuencial]

Regla de la Adicción: Unión de Sucesos (simultáneos)

Si son mutuamente excluyentes: $P(S_1 \cap S_2) = P(\emptyset) = 0$

$$\boxed{P(S_1 \cup S_2) = P(S_1) + P(S_2)}$$

Sin tener en cuenta la mutua exclusión:

$$\boxed{P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2)}$$

Regla de la Multiplicación: Sucesos Condicionados o Intersección de Sucesos (secuenciales)

Si están condicionados

$$\boxed{P(S_1 \cap S_2) = P(S_1) \cdot P\left(\frac{S_2}{S_1}\right)}$$

Si $P(S_1)$ y $P(S_2)$ no están condicionados: $P\left(\frac{S_2}{S_1}\right) = P(S_2)$

$$\boxed{P(S_1 \cap S_2) = P(S_1) \cdot P(S_2)}$$

<u>Regla de Laplace</u>

«La probabilidad de un suceso es igual a su cardinal entre el cardinal del espacio muestral en el que se inscribe».

$$P(S) = \sum_{i=1}^{n} P(c_i) = P(c_1) + P(c_2) + \cdots + P(c_n) = \frac{1}{N} + \frac{1}{N} + \ \ldots n \ldots \ + \frac{1}{N} = \frac{n}{N}$$

$$\boxed{P(S) = \frac{n}{N}}$$

Propiedades básicas de la Probabilidad

- Suceso Seguro → $P(U) = 1$
- $P(S_1 \cup \bar{S_1}) = 1$
- $P(S_1) + P(\bar{S_1}) = 1$

$\boxed{p + q = 1}$ $\boxed{p = 1 - q}$ $\boxed{q = 1 - p}$

- $P(\emptyset) = 0$
- $P(S_1 \cap \bar{S_1}) = 0$
- $0 \leq P(S) \leq 1$

Distribuciones Teóricas de Probabilidad

Según el tipo de variable, las distribuciones de probabilidad pueden ser:

- **Discretas**
- **Continuas**

Función de la Distribución Normal

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{u}}{\sigma}\right)^2}; \ \forall x \in \mathbb{R}}$$

Cuando trabajamos con muestras distinguimos entre la distribución de la muestra y la de la población:

Parámetros:

$\bar{\mu}$ = Media de la Población

σ = Desviación Típica de la Población

Estadísticos:

$\bar{x}$ = Media de la Muestra

s_x = Desviación Típica de la Muestra

- **Parámetro:** es el valor que toma la población respecto a un indicador utilizado para resumir la información que nos interesa de dicha población.
- **Estadístico:** es el valor que obtenemos a partir de una muestra y que representa al parámetro.

Valor Esperado

$$\boxed{E(x) = \mu}$$

$$E(x) = \sum x_i \cdot p(x_i) = \sum x_i \cdot \frac{n_i}{N} = \mu$$

Distribución Real, Teórica y Empírica

- **Distribución Real** → distribución de la **Población** a investigar
- **Distribución Teórica** → ***distribución muestral*** (media de todas las muestras)
  INTERMEDIA entre la Real (Población) y la Empírica (Muestra)
- **Distribución Empírica** → distribución de una **muestra** seleccionada

Una **distribución de probabilidad** es una distribución de frecuencias relativas de una variable que denominamos aleatoria, por estar asociada a un experimento de carácter aleatorio; no presenta resultados ciertos, sólo podemos conocer la probabilidad de sus resultados posibles. En el campo de la investigación social está asociada al **muestreo**: a la posibilidad de extraer de una población una serie de individuos que presentan una determinada característica.

Igual que toda distribución, la de probabilidad puede ser **discreta** o **continua**, según sea el **espacio muestral** del experimento que la define, es decir, según sea el conjunto de valores que puede tomar la variable: el conjunto de resultados posibles del experimento.

En el campo de la **investigación social** podemos trabajar con **distribuciones empíricas**, viendo cómo se distribuyen los datos en una **población** o en una **muestra**. Pero cuando trabajamos con una muestra lo que pretendemos es saber hasta qué punto la distribución de una variable que obtenemos en la muestra se corresponde con la **distribución real** de la variable en la población. Para ello necesitamos hacer uso de una **distribución intermedia** entre la de la **muestra** y la de la **población**: la ***distribución muestral***, que es una distribución teórica. Al calcular el valor esperado en una muestra que hemos supuesto, al trabajar con variables numéricas (no nominales), todas las medidas de todas las muestras posibles (de un determinado tamaño), la distribución de estas medias constituye la ***distribución muestral***, que no hay que confundir con la **distribución empírica** de los datos de una muestra, ni con la **distribución real** de los datos de la población.

Cuando realizamos **el muestreo**, automáticamente las distribuciones de frecuencias relativas de los datos de una población o de una muestra se convierten en una **distribución de probabilidad**.

Por tanto, hay que subrayar la diferencia entre la **distribución empírica** (de los datos de la muestra o de la población) y la **distribución teórica** que generan las medias de todas las muestras que podemos extraer de la población investigada en el caso de variables de tipo cuantitativo.

Las **distribuciones teóricas** son referencia obligada para constrastar **distribuciones empíricas** observadas en muestras y poder conocer la validez de los datos observados en éstas, ya que nos indicarán hasta qué punto los datos se corresponden con los de la población de donde extraemos las muestras.

Distribución Uniforme

Es la más sencilla de las distribuciones de probabilidad: aquella en que la probabilidad se distribuye por igual en todos los casos o en todos los grupos de la población. No existen probabilidades diferentes parea casos o grupos diferentes.

Distribución Uniforme Discreta

Cuando contamos con n grupos y la probabilidad de cada grupo (x) es:

$$p(x) = \frac{1}{n}$$

Función de Densidad

La función de densidad se aplica para calcular la probabilidad que toma un intervalo dentro de distribuciones continuas. Se utilizan áreas para conocer la probabilidad de un intervalo de valores.

Distribución Binomial

$$p + q = 1$$

$$p = 1 - q$$

$$q = 1 - p$$

p → probabilidad de éxito

q → probabilidad de fracaso

Experimento Binomial

$$\binom{n}{x} \cdot p^x \cdot q^{n-x}$$

Para cada número de sucesos que fijemos (n) y para cada probabilidad de éxito (p) en cada uno de estos sucesos tendremos una distribución binomial distinta. Cada distribución binomial es determinada por estos dos parámetros: n y p.

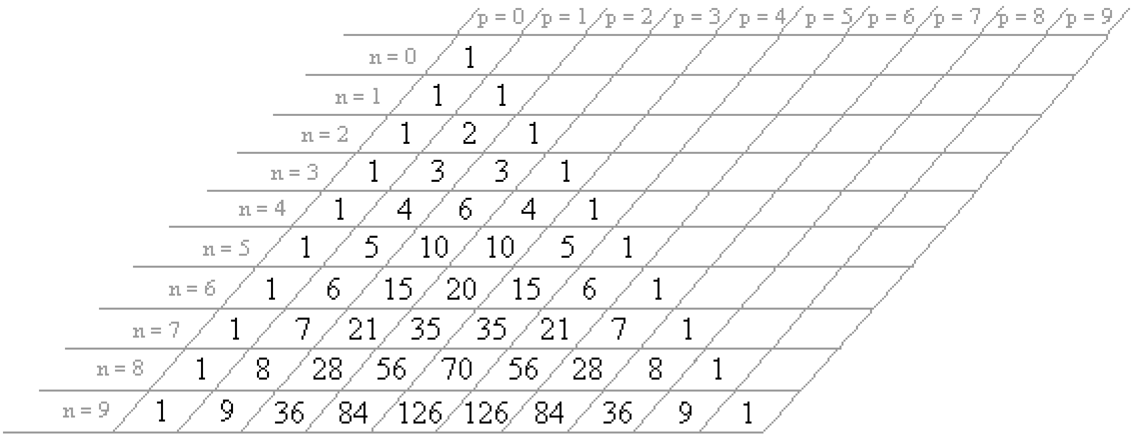
Forma General de la Distribución Binomial

Nº de éxitos *(x)*	Probabilidad *p(x)*
0	$\binom{n}{0} \cdot p^0 \cdot q^n$
1	$\binom{n}{1} \cdot p^1 \cdot q^{n-1}$
2	$\binom{n}{2} \cdot p^2 \cdot q^{n-2}$
3	$\binom{n}{3} \cdot p^3 \cdot q^{n-3}$

…	…
X	$\binom{n}{x} \cdot p^x \cdot q^{n-x}$
…	…
n-1	$\binom{n}{n-1} \cdot p^{n-1} \cdot q^1$
n	$\binom{n}{n} \cdot p^n \cdot q^0$

Esta distribución se describe por los términos del desarrollo del binomio de Newton $(p+q)^n$. Los coeficientes de cada término $\left[\binom{n}{0}, \binom{n}{1}, \binom{n}{2}, \binom{n}{3}, \dots, \binom{n}{x}, \dots, \binom{n}{n-1}, \binom{n}{n}\right]$ corresponden con una de las líneas del triángulo de Tartaglia o de Pascal.

Triángulo de Tartaglia (o de Pascal)



Para cada n (número de elementos de la muestra) y cada x (número de éxitos de ésta) obtenemos los coeficientes a partir del Triángulo de Tartaglia (o de Pascal).

Cada línea representa los coeficientes de una distribución binomial, coeficientes que corresponden con estos números combinatorios:

$$\binom{0}{0}$$

$$\binom{1}{0} \quad \binom{1}{1}$$

$$\binom{2}{0} \quad \binom{2}{1} \quad \binom{2}{3}$$

$$\binom{3}{0} \quad \binom{3}{1} \quad \binom{3}{2} \quad \binom{3}{3}$$

$$\binom{4}{0} \quad \binom{4}{1} \quad \binom{4}{2} \quad \binom{4}{3} \quad \binom{4}{4}$$

Media de una distribución binomial	Varianza de una distribución binomial	Desviación Estándar (o Típica) de una distribución binomial
$\mu = n \cdot p$	$\sigma^2 = n \cdot p \cdot q$	$\sigma = \sqrt{n \cdot p \cdot q}$

Para muestras de variable nominal → biniomializar / dicotomizar

Valor esperado en la Distribución Binomial

$$E(x) = \sum x_i \cdot p_i = n \cdot p$$

Distribución Normal

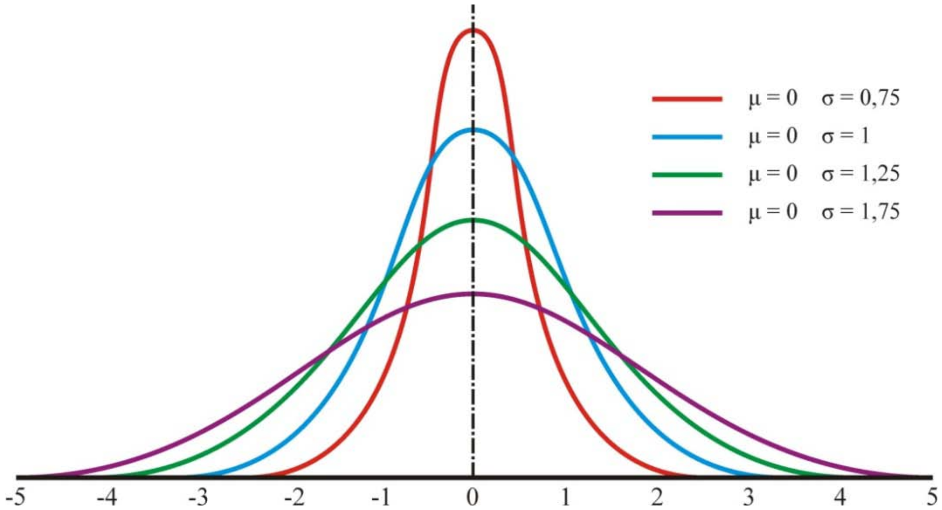
La **distribución normal** es la distribución teórica más usada en estadística. Aparte de que múltiples conjuntos de datos pueden ajustarse a ella, es la clave de **la estadística inferencial**. Toda **distribución muestral** de medias se aproxima a la **distribución normal**.

Función de la Distribución Normal

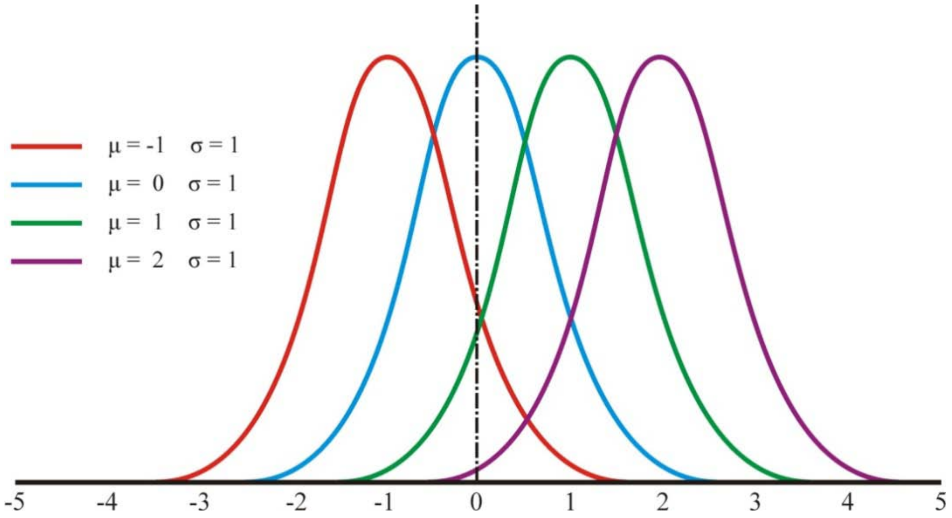
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{u}}{\sigma}\right)^2}; \quad \forall x \in \mathbb{R}$$

Como e y π son dos constantes matemáticas transcendentes ($e = 2{,}7181 \ldots$; $\pi = 3{,}1415 \ldots$), $f(x)$, aparte del valor variable de x, depende de la media de todos los valores $x(\mu)$ y de su desviación típica o estándar (σ). Cada distribución normal depende de estos dos parámetros (su media y su desviación típica); fijados estos, se puede conocer el valor de la función $f(x)$ para cada valor de x.

Distribuciones normales con distintas desviaciones típicas



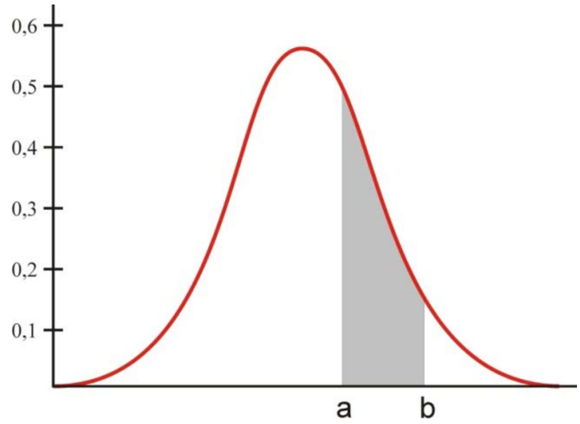
Distribuciones normales con distintas medias



Función de Densidad Normal

$$p(a \leq x \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{u}}{\sigma}\right)^2} dx$$

La función de densidad normal nos define la probabilidad de cualquier intervalo.



Distribución Normal Estandarizada

Transformación valores de x en unidades de Z

$$Z_x = \frac{x - \mu}{\sigma}$$

Los valores de Z los utilizamos para localizar de manera relativa cada valor en la distribución en la que se encuentra, permitiendo la comparación entre valores de distribuciones distintas. Se establece en cada distribución la distancia de cualquier valor x a la media (μ) y se mide esta distancia en unidades de desviación típica o estándar (σ).

Si una distribución continua la convertimos en valores de Z decimos que la estandarizamos. Al estandarizarla, su media se convierte en "0" y desviación típica en 1.

Las transformaciones que convierten unidades de Z a x (y viceversa) permiten generalizar esto a todas las distribuciones normales, tengan la media y la desviación típica que tengan.

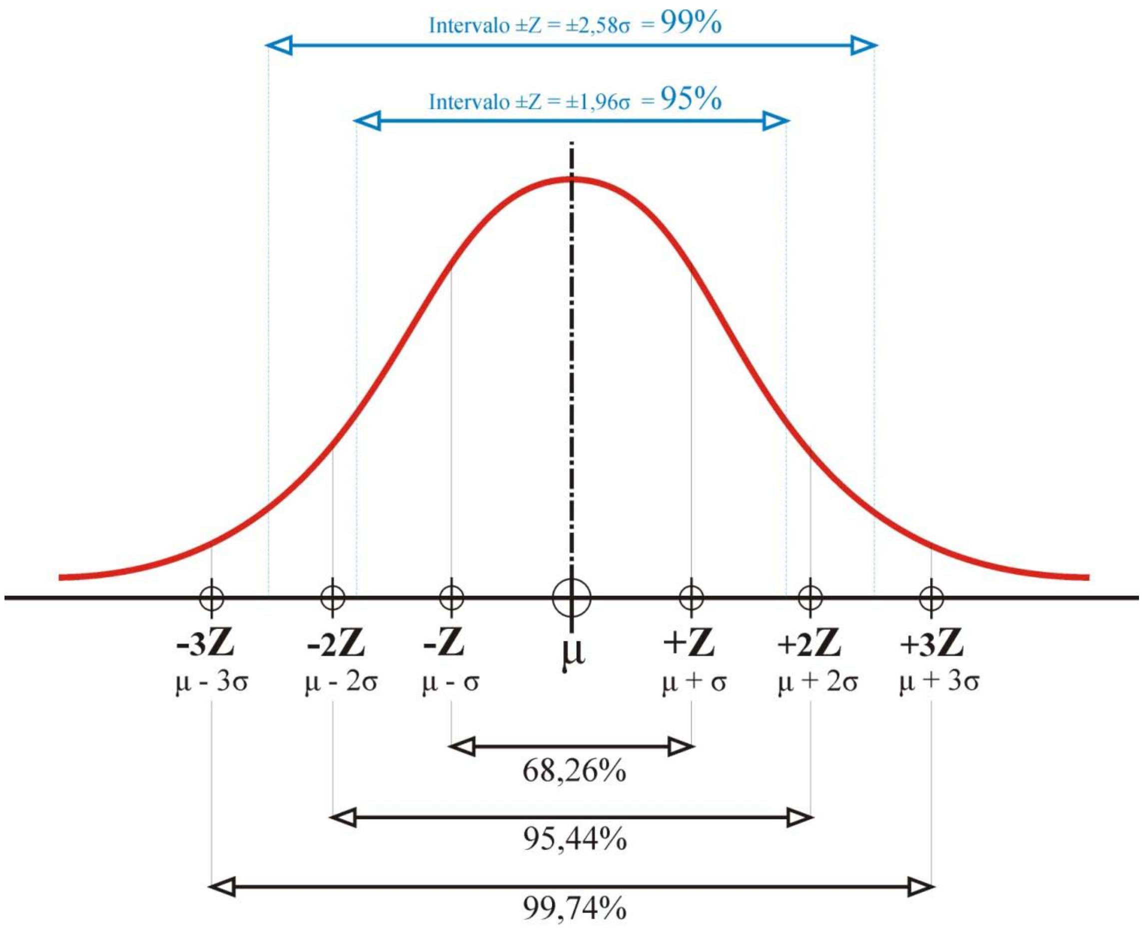
Función Normal Estandarizada

$$f(Z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

Función de Densidad Estandarizada

$$p(a \leq Z \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dx$$

Cálculo de probabilidades en Distribuciones Normales



Aproximación de la Distribución Binomial a la Normal

Si p no es una probabilidad extrema (próxima al "0" o al "1"), entonces podemos sustituir la binomial por una normal cuando $n \cdot p \geq 5$ o $n \geq 30$.

Distribución *t* de Student

La distribución t de Stundent nos sirve para hacer inferencias sobre la media poblacional a partir de la media de la muestra cuando se desconoce la desviación típica de la población. La distribución t de Student es una distribución asociada a la normal. Se puede decir que es una distribución normal corregida.

En muchos manuales se sostiene de forma arbitraria que si $n > 30$ entonces la distribución normal puede sustituir a la distribución t de Student. Se puede dar este límite por bueno.

Transformación valores de x en unidades de *t*

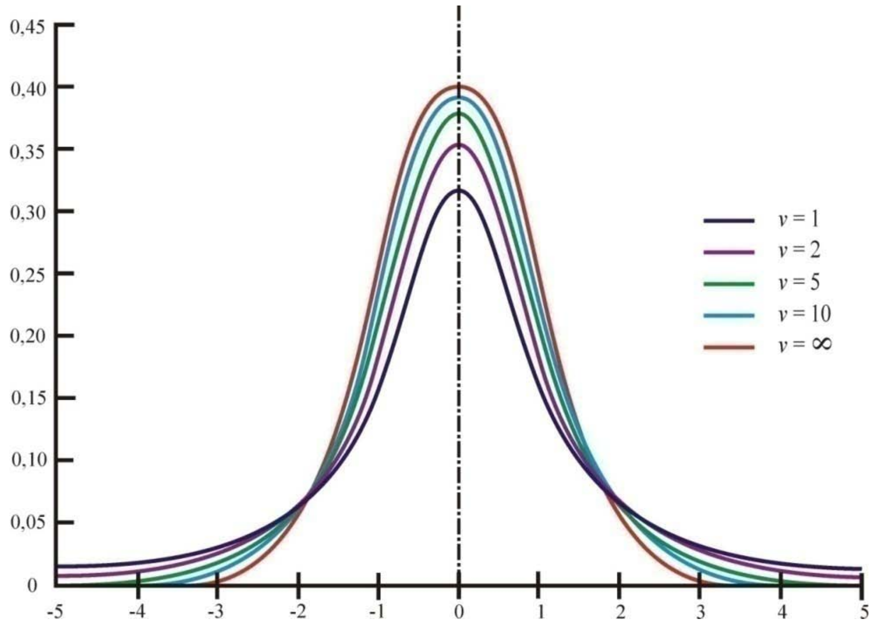
$$t = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n-1}}}$$

Hay una distribución t distinta para cada tamaño muestral "n". En este caso se habla de los **Grados de Libertad**, que serán siempre **"$\boldsymbol{n-1}$"**, y que están asociados a l cálculo de la desviación típica de la muestra.

Grados de Libertad de la Distribución *t* de Student:

$$\boldsymbol{n-1}$$

Distribución *t* de Student para varios grados de libertad (v)



Desviación Típica de la Distribución *t* de Student

$$S_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

Cuadro Resumen Medias y Varianzas de las diferentes distribuciones de probabilidad

Tipo de Distribución	Media	Varianza
Uniforme b = máximo a = mínimo	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Binominal n = número de casos p = probabilidad de éxito $q = (1-p)$	$n \cdot p$	$n \cdot p \cdot q$
Normal	μ	σ^2
Normal Estándar	0	1
***t* de Student** v = grados de libertad	$0\ para\ v > 1$	$\frac{\boldsymbol{v}}{\boldsymbol{v}-2}\ para\ v > 2$

Las Muestras Estadísticas

Investigación mediante Muestras. **Condiciones:**

1. Definir el **Universo** (Conjunto o Población de referencia)

2. Fijar las condiciones para la **Estimación**:

   - **Error**
   - **Confianza**

   → Representatividad de la Muestra

3. Método de **Selección de la Muestra**:

   a. Muestreo Aleatorio

   - **Estratificado**

     Afijación
     - **Uniforme**
     - **Proporcional**
     - **Neyman** (proporcional a la Varianza)

   - **Por Conglomerados**

   b. Muestreo por Cuotas

4. **Inferirencia** de los resultados para el Universo.

Parámetros y Estadísticos

- **Parámetro → Constante Fija** (μ , σ)
  Es el valor que toma la población respecto a un indicador utilizado para resumir la información que nos interesa de dicha población.
- **Estadístico → Constante Variable** ($\bar{x}$, s_x , ${s_x}^2$)
  Es el valor que obtenemos a partir de una muestra y que representa al parámetro.

Parámetro ≠ Estadístico

Con el estadístico vamos buscando el parámetro

Error y Sesgo

- **Error Estadístico**

  Es aquel que procede del hecho de utilizar observaciones muestrales, es decir, de no observar la totalidad de los datos. Este error puede ser conocido e incorporado en los resultados a través de un intervalo.

Error Estadístico ≠ Error Típico (Desviación Típica o Estándar)

- **Errores No Estadísticos**

  Procede de defectos en los instrumentos de medida y de las condiciones en las que se establece la medida, así como de otro conjunto de errores en la transmisión de datos. Este tipo de errores se pueden (y deben) ser evitados. Entre los errores no estadísticos se encuentra el sesgo:

  - **Sesgo**
    Desviación sistemática de nuestras observaciones respecto a lo que estamos midiendo. En investigación mediante muestras el sesgo se produce al obtener muestra que no se adecúan a la población, es decir, por el desfase que existe entre la población objetivo y la población de la que obtenemos la muestra.

Muestreo Aleatorio. Uso de tres tipos de distribuciones distintas:

1. **Distribución de la Población:** el conjunto de todos los datos de la población y a cuyos indicadores de resumen denominamos ***parámetros***. Generalmente esta distribución es desconocida, por ello recurrimos a muestras.
2. **Distribución de la Muestra:** el conjunto de datos pertenecientes a la muestra seleccionada. Generalmente sólo obtenemos una muestra de la población. Sus indicadores de resumen se denominan ***estadísticos***.
3. **Distribución Muestral:** la distribución de los estadísticos (las medias, por ejemplo) de todas las muestras posibles de tamaño n que proceden de una población. Es una distribución teórica. No se desarrolla en forma de tabla porque se necesitaría conocer todos los elementos de la población, pero, aunque no se conozcan, se puede construir a través de sus ***parámetros***.

Error Estadístico

Es la medida de la distancia entre el valor del ***estadístico*** obtenido en la **muestra** y el valor del ***parámetro*** en la **población.**

Nivel de Confianza

Es la probabilidad que existe de que esa distancia, o error estadístico, no sea mayor que la preestablecida.

Si en una distribución muestral se aumenta el valor de n, la varianza de la distribución disminuirá y los resultados de las distintas muestras se concentrarán más en el valor del parámetro. Por consiguiente, para un mismo nivel de error, el nivel de confianza mejora con tamaños muestrales mayores.

Los tres pilares de la Teoría Muestral

1. **TAMAÑO MUESTRAL:** es el número de elementos que extraemos de una población para su observación y estudio.
2. **ERROR ESTADÍSTICO:** es la diferencia máxima (en valor absoluto) que admitimos entre el valor del estadístico y el del parámetro.
3. **NIVEL DE CONFIANZA:** es la probabilidad de que la muestra seleccionada no supere el error preestablecido.

Estos tres términos están absolutamente relacionados y la variación en uno produce variaciones en los otros dos. Así, por ejemplo, si aumentamos el tamaño muestral y mantenemos fijo el error, el nivel de confianza aumenta. Si aumentamos el error sin cambiar el tamaño muestral, el nivel de confianza aumenta.

Relaciones:

- Tamaño Muestral – Nivel de Confianza → $\mathcal{R}$ directa
- Error Estadístico – Tamaño Muestral → $\mathcal{R}$ inversa
- Error Estadístico – Nivel de Confianza → $\mathcal{R}$ inversa

Error Típico (o error estándar)

Es la medida de dispersión (varianza) de la distribución muestral.

Error Estadístico ≠ Error Típico (Desviación Típica o Estándar)

Teorema del Límite Central

La distribución muestral de las medias de una población grande con media μ y desviación típica σ se aproxima, según aumenta n, a una distribución normal con media μ y desviación típica $\frac{\sigma}{\sqrt{n}}$

Con poblaciones grandes $(N > 30)$: $\bar{x}_{muestra} = \mu$ y:

$$\boxed{S_x = \sigma_x = \frac{\sigma}{\sqrt{n}}}$$

Distribución Normal en Z

$$\boxed{Z = \frac{x - \mu}{\sigma_x}}$$

Error Estadístico

$$\boxed{e = |\bar{x} - \mu|}$$

Teorema Límite Central

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

Relación entre Error, Nivel de Confianza y Tamaño de la Muestra

$$\boxed{Z = \frac{e}{\frac{\sigma}{\sqrt{n}}}}$$

Cálculo del Tamaño Muestral

$$\boxed{n = \frac{Z^2 \sigma^2}{e^2}}$$

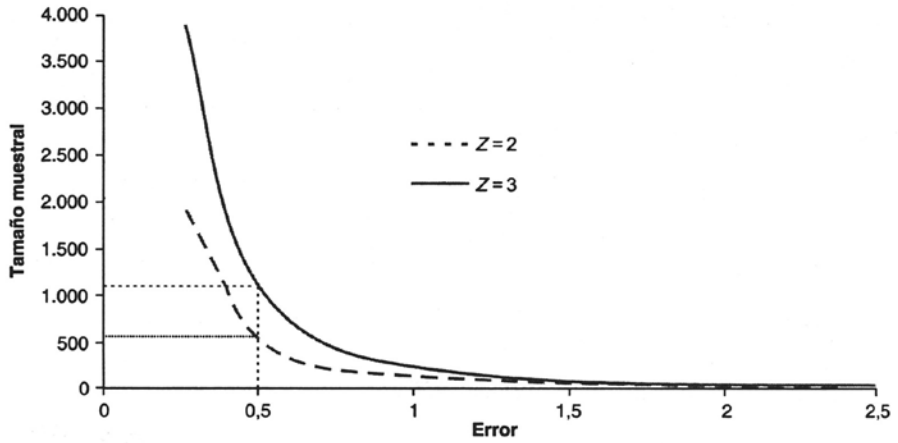
Cálculo del Error Estadístico

$$\boxed{e = \frac{Z\sigma}{\sqrt{n}}}$$

Valores de Z para Niveles de Confianza más usuales

Nivel de Confianza	Z
95%	$\pm 1{,}96$
95,45%	± 2
99%	$\pm 2{,}58$
99,7%	± 3

Tamaños muestrales para distintos errores y niveles de confianza



Fuente: Camarero Rioja, Luis et al. 2010. *Estadística para la investigación social*. Ibergarceta. p. 220.

Nomenclatura Distribuciones Población/Muestra

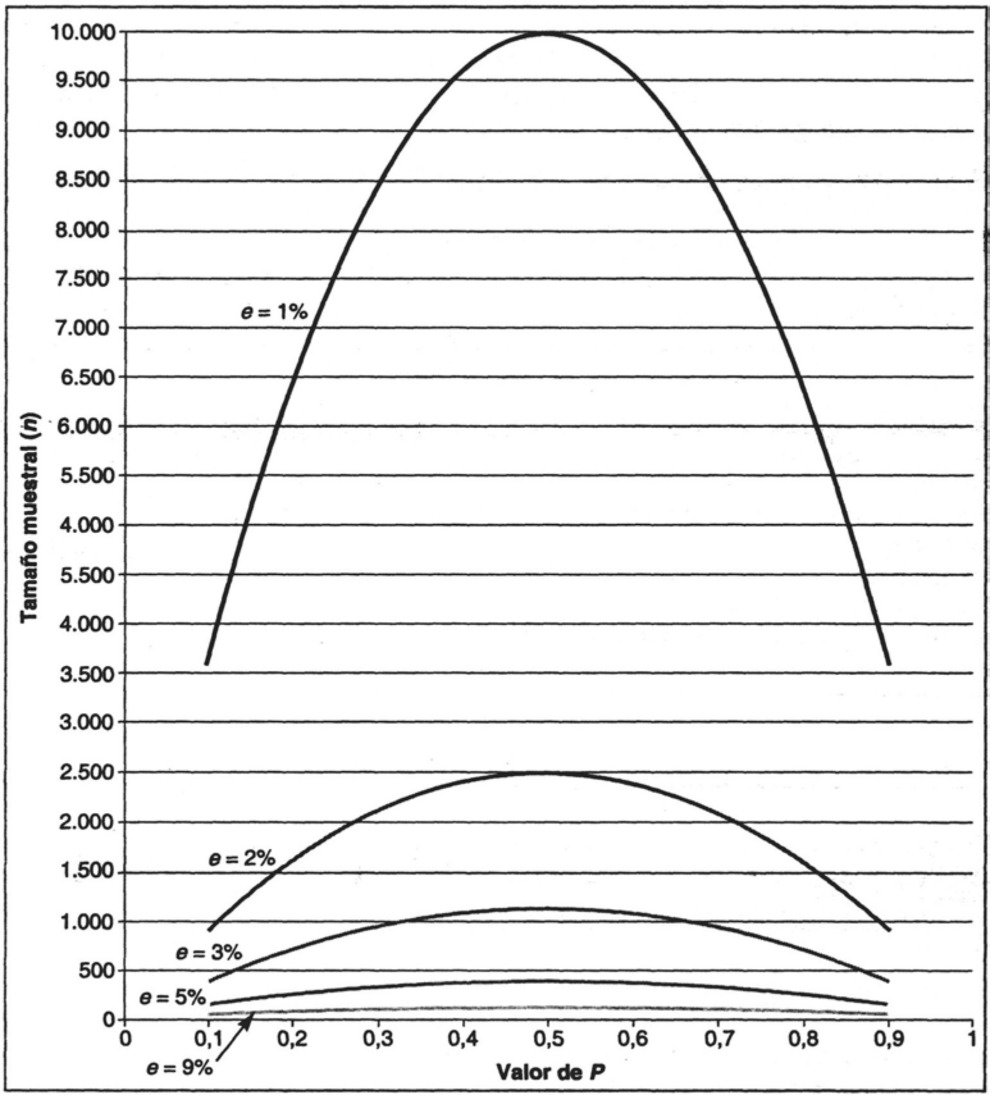
	Parámetro	Estadístico
MEDIA	Media: μ	Media: $\bar{x}$
	Varianza: σ^2	Varianza: S_x^2
PROPORCIÓN	Proporción/Probabilidad P	Proporción/Probabilidad p
	Complementario $Q = (1 - P)$	Complementario $q = (1 - p)$

Errores Típicos y Fórmulas del Tamaños Muestral para los estadísticos de la Media y de la Proporción (Poblaciones Infinitas)

	Media	Proporción*
Parámetro	μ	P
Error Típico Error Estándar	$\sigma_\mu = \frac{\sigma}{\sqrt{n}}$	$\sigma_p = \sqrt{\frac{PQ}{n}}$
La distribución se considera como la Normal cuando:	$n > 30$	$n > 30$
Tamaño Muestral n	$n = \frac{Z^2\sigma^2}{e^2}$	$n = \frac{Z^2 PQ}{e^2}$

*En el caso de la Proporción, cuando desconocemos la varianza PQ, elegimos el caso más desfavorable, varianza máxima, cuando $P = Q = 0{,}5$.

Tamaños muestrales para el estadístico de la proporción con distintos valores de P.
Nivel de Confianza 95,45%



Fuente: Camarero Rioja, Luis et al. 2010. *Estadística para la investigación social*. Ibergarceta. p. 222.

Poblaciones

- **Poblaciones Finitas**

  En investigación social casi siempre nos referimos a poblaciones finitas. Por tanto las formulas anteriores aplicadas a poblaciones finitas deben incorporar el factor de corrección para poblaciones finitas:

Factor de Corrección para Poblaciones Finitas

$$\boxed{\left(\frac{N-n}{N-1}\right)}$$

- **Poblaciones Infinitas**

  Una población infinita es un conjunto de elementos que no pueden definirse mediante enumeración. Cuando la relación entre N y n (N/n) es grande, se desprecia el factor de corrección para poblaciones finitas:

  Cuando $\boldsymbol{N} > 20\boldsymbol{n}$ (En la práctica cuando $\boldsymbol{N} > 100.000$)

Errores Típicos y Fórmulas del Tamaños Muestral para los estadísticos de la Media y de la Proporción (Poblaciones Finitas)

	Media	**Proporción***
Parámetro	μ	P
Error Típico **Error Estándar**	$\sigma_\mu = \frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$	$\sigma_p = \sqrt{\frac{PQ}{n}\left(\frac{N-n}{N-1}\right)}$
La distribución se considera como la Normal cuando:	$n > 30$	$n > 30$
Tamaño Muestral n	$n = \frac{Z^2\sigma^2 N}{e^2(N-1)+Z^2\sigma^2}$	$n = \frac{Z^2 NPQ}{e^2(N-1)+Z^2 PQ}$

Muestras Grandes y Pequeñas

Muestras Grandes → $\boldsymbol{N > 30}$ → La distribución muestral se a ajusta a la **NORMAL**

Muestras Pequeñas → $\boldsymbol{N < 30}$ → La distribución muestral se a ajusta a la ***t* de STUDENT**

Regla de aproximación a la Desviación típica cuando se desconoce la de la Población

$$\sigma \sim \frac{Recorrido}{4} = \frac{V_{máx} - V_{mín}}{4}$$

V: valores de la muestra o población.

Diseño de Muestras

Marco

El registro físico de los elementos de la población se denomina marco. En algunos casos es un listado para seleccionar elementos de una población. En la mayoría de las ocasiones no pude obtenerse un marco completo y actualizado de la población, o la elaboración es costosa o presenta problemas de manipulado.

Procedimiento de Muestreo

Dadas las dificultades de realizar muestras aleatorias simples, en las que todos los elementos de una población tienen la misma probabilidad de ser seleccionados, se recurre a distintos sistemas de muestreo que tienen en cuenta lo que ya conocemos sobre la población a investigar. Al incorporar información existente podremos realizar diseños muestrales más eficientes, esto es, obtendremos la misma información con menor coste que si usáramos métodos aleatorios puros. Además, al incorporar información existente, podremos mejorar la representatividad de la muestra.

Método de Selección de la Muestra

a. Muestreo Aleatorio

- **Estratificado**

  Afijación
  - **Uniforme**
  - **Proporcional**
  - **Neyman** (proporcional a la Varianza)

- **Por Conglomerados**

b. Muestreo por Cuotas

Muestreo Estratificado

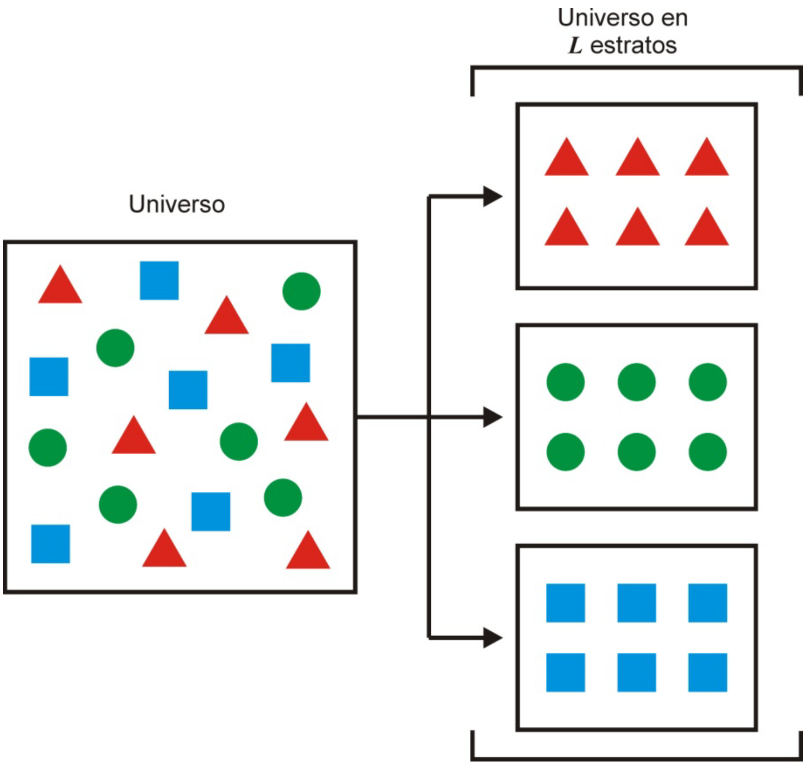
Estratos

Son subconjuntos o grupos del universo, de forma que todos los elementos pertenecen a uno de los grupos y sólo a uno.

- **Diseño Uniforme:** deseamos conocer resultados para cada uno de los estratos.
- **Diseño Proporcional:** aprovechamos la ventaja de la estratificación pare reducir el número de extracciones necesarias para la estimación del conjunto.
- **Afijación:** es el reparto de los elementos de la muestra entre los distintos estratos.

Para la aplicación del diseño estratificado se necesita definir los diferentes estratos y disponer de informaciones que nos permitan estimar las varianzas en cada uno de ellos.

Descomposición esquemática de una población en estratos



Elementos del muestreo estratificado

	L	Número de estratos	
Población	N_i	Población en el estrato i	$N = \sum_{i=1}^{L} N_i$
	W_i	Proporción de la población en el estrato i o peso del estrato	$W_i = \frac{N_i}{N}$
Muestra	n_i	Tamaño de la muestra en el estrato i	$n = \sum_{i=1}^{L} n_i$
	w_i	Proporción de la muestra en el estrato i	$w_i = \frac{n_i}{n}$

Varianza de la Población Estratificada $(pq)_{st}$

$$(pq)_{st} = \frac{1}{N}\sum_{i=1}^{L} p_i q_i N_i = \sum_{i=1}^{L} p_i q_i W_i$$

Procedimientos de Afijación

Número de entrevistas en el estrato i una vez establecido el valor de n

Afijación Uniforme	$n_i = \frac{n}{L}$	Se hace el mismo número de entrevistas en cada estrato.
Afijación Proporcional al tamaño	$n_i = n \cdot W_i$	El número de entrevistas se reparte considerando el peso que tienen los estratos en la población.
Afijación de Neyman (proporcional a la varianza)	$n_i = n \frac{N_i \sqrt{p_i q_i}}{\sum_{i=1}^{L} N_i \sqrt{p_i q_i}}$	Las entrevistas se distribuyen de forma que se concentran relativamente en los estratos de mayor varianza.

Ponderación en diseños no proporcionales al tamaño

Cuando se utilizan criterios de afijación no porporcionales al tamaño de la población (por ejemplo el diseño estratificado uniforme), los resultados obtenidos para el conjunto total necesitan ser corregidos. Esto se realiza con la **Ponderación**: el procedimiento de ajuste del peso que tienen los estratos en la muestra al peso que les corresponde en la población.

Ponderar las unidades muestrales equivale a darles el peso que, por el estrato al que pertenecen, les corresponde en la población. Para ello se aplica en cada estrato un **coeficiente de ponderación** que atribuye el peso correspondiente a cada estrato en el conjunto de la población.

$$W_i = \frac{N_i}{N} = \frac{N_i}{\sum N_i}$$

$$w_i = \frac{n_i}{n} = \frac{n_i}{\sum n_i}$$

Coeficiente de Ponderación

$$u_i = \frac{W_i}{w_i}$$

Los cuestionarios ponderados se representan con el signo «’»: $n_i' = n_i u_i$

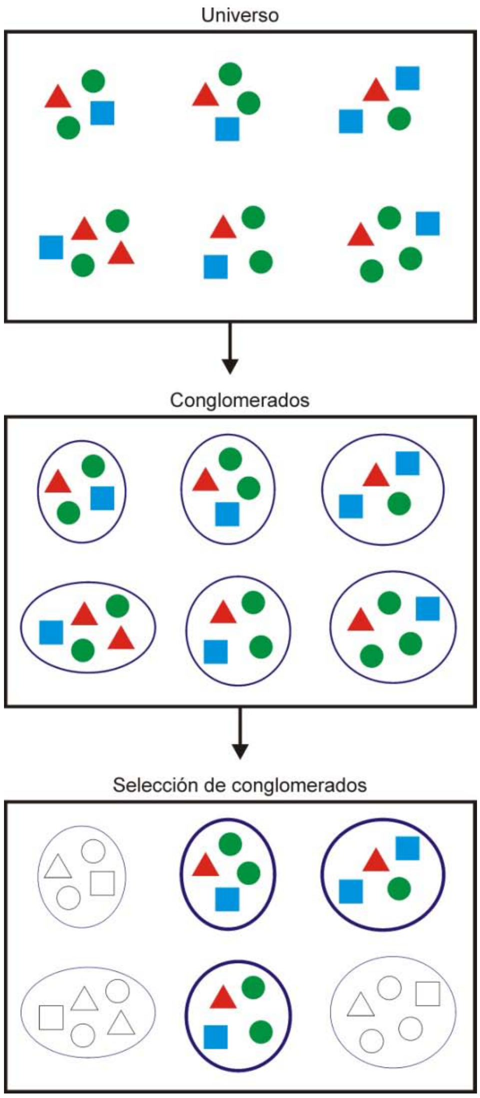
Muestreo por Conglomerados

El diseño estratificado mejora la aplicación del muestreo respecto al muestreo aleatorio, pero no resuelve de forma efectiva la cuestión de la obtención del marco. Si no tenemos un listado del universo, es difícil tener un listado para sus estratos. El **diseño por conglomerados** resuelve de forma práctica estas cuestiones y permite reducir los costes y tiempo de ejecución.

Los **conglomerados** son también subconjuntos de la población. La diferencia es que mientras los estratos se construyen buscando homogeneidad entre sus elementos, en los conglomerados se definen los grupos buscando que los elementos en su interior sean lo más diversos posibles. Una vez definidos los conglomerados, se hace una muestra de ellos y se investiga únicamente los que han sido seleccionados. De hecho, se trata de una técnica de selección, no de unidades sino de grupos de unidades y en general se compone de varias etapas.

Conglomerado → Unidad Colectiva

Exposición esquemática del muestreo por conglomerados



En el diseño de conglomerados, los **tamaños muestrales n son mayores** que en el muestreo aleatorio simple y el muestreo estratificado, debido a que los conglomerados tienen varianzas más elevadas que el conjunto poblacional.

En la práctica se utilizan **métodos de selección combinados**. La población se divide en estratos y dentro de los distintos estratos se realiza un diseño de conglomerados.

Muestreo por Cuotas

Aunque los diseños derivados del muestreo aleatorio (estratos y conglomerados) permiten la obtención de datos representativos de la población, así como la inferencia de los resultados obtenidos mediante muestras al conjunto del universo bajo criterios conocidos de error, resultan todavía procedimientos costosos. En algunos casos, cuando la investigación se refiere a opiniones, pueden emplearse otros procedimientos para obtener muestras representativas.

La principal técnica alternativa es el **muestreo por cuotas**. Se trata de, en vez de confiar al azar la selección de las unidades, desarrollar un procedimiento de selección que reproduzca con la máxima fidelidad posible la propia estructura de la población. No obstante, para el muestreo por cuotas es necesario tener un conocimiento preciso sobre la población, pues sería la única forma válida de reproducir una miniatura de la población a estudiar.

Gracias, por ejemplo, a los censos podemos conocer algunas variables como la edad, sexo y nivel de estudios. De forma conjunta, estas tres variables están muy relacionadas con las opiniones políticas y con las actitudes. Siempre que tengamos un conocimiento amplio de la distribución de variables que están relacionadas con las características a investigar, podremos utilizar el muestreo por cuotas.

La diferencia respecto al muestreo estratificado es que la localización de los entrevistados no sigue un procedimiento probabilístico, sino que se seleccionan los primeros elementos localizados que cumplan con las características a estudiar. Sin embargo este procedimiento tiene mayor riesgo de introducir sesgos, y deben aplicarse algunas normas para darle un mínimo carácter aleatorio:

1. **Distribuir al máximo las entrevistas entre los encuestadores**: a un mismo entrevistador se le asigna diferentes perfiles de entrevistados para que no pueda obtenerlos en un mismo lugar.
2. **Utilizar el sistema de «barrido cuotas»**: una vez asignadas las entrevistas, a cada entrevistador se le proporciona unas rutas que debe realizar buscando los perfiles a investigar.

Muestreo Mixto

En la práctica, en muchas encuestas sociológicas y de opinión se utiliza un **procedimiento mixto** en la selección de la muestra. La **primera fase** del diseño muestral se realiza mediante **técnicas probabilísticas** (selección de municipios dentro de estratos por tamaño poblacional); y la **fase final** de selección del entrevistado se realiza mediante **muestreo por cuotas.**

La estimación de las Varianzas

Media → Estimador insesgado

Desviación Típica → Estimador sesgado

Cuasivarianza → Estimador insesgado

Cuasivarianza Muestral

$$\boxed{S_{n-1}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Sobre grados de libertad $(n-1)$

Para tamaños $(\boldsymbol{n} > 90)$ → **Varianza = Cuasivarianza**

Cuasidesviación Muestral

$$\boxed{S_{n-1} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}}$$

Sobre grados de libertad $(n-1)$

Varianza Muestral a partir del estadístico de la Cuasivarianza

$$\boxed{\sigma^2 = \frac{N-1}{N} S_{n-1}^2}$$

Error Típico cuando estimamos la Varianza a través de la propia muestra

Media	Proporción
$\sigma_{\bar{x}} = \sqrt{\frac{S_{n-1}^2}{n}\left(\frac{N-n}{N}\right)}$	$\sigma_p = \sqrt{\frac{pq}{n-1}\left(\frac{N-n}{N}\right)}$

Inferencia Estadística

Estimación de Medias y Proporciones

$$Z = \frac{e}{\frac{\sigma}{\sqrt{n}}} \quad ; \quad e = Z\frac{\sigma}{\sqrt{n}} \quad ; \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Error Estadístico

$$\boxed{e = Z\sigma_{\bar{x}}}$$

Errores e Intervalos para la Media y la Proporción estimando la Varianza a partir de la muestra

Estadístico	Error Típico (Poblaciones Infinitas)	Error Típico (Poblaciones Finitas)	Intervalo
MEDIA	$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$	$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$	$\bar{x} \pm Z\sigma_{\bar{x}}$
PROPORCIÓN	$\sigma_p = \sqrt{\frac{pq}{n-1}}$	$\sigma_p = \sqrt{\frac{pq}{n-1}}\sqrt{\frac{N-n}{N}}$	$p \pm Z\sigma_p$

Notas

- $\boldsymbol{s}$ → **Cuasidesviación**
- Fórmulas para **Poblaciones Finitas** cuando ($\boldsymbol{N} < 20\boldsymbol{n}$)
- Para **Muestras Pequeñas** ($\boldsymbol{n} < 30$) se usa $\boldsymbol{t}$ **de Student** en vez de Z.
- En el caso de las **Proporciones** se usa la **Varianza** que nos proporciona la muestra $\boldsymbol{pq}$ y no la poblacional $\boldsymbol{PQ}$.

Intervalos en el caso de Muestras Pequeñas

Estimación de Medias (muestras pequeñas)

- En la práctica si ($\boldsymbol{n} < 20$) se usa la distribución $\boldsymbol{t}$ **de Student**
- La Varianza (Desviación Típica) hay que obtenerla a partir de la Cuasivarianza (Cuasidesviación)

$$\boxed{S_{n-1}^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}} \qquad \boxed{S_{n-1} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}}$$

Estimación de Proporciones (muestras pequeñas)

- **Dificultades** para $(\boldsymbol{n} < 30)$ o las proporciones son muy próximas a "0" o "1" $(\boldsymbol{p\ o\ q \sim 0\ o\ 1})$
- No se usa la fórmula del Error Típico si $\boldsymbol{np} < 5$ **o** $\boldsymbol{nq} < 5$. Como alternativa se aplicaría el Intervalo de Wilson

Intervalo de Wilson

$$\frac{p + \frac{Z^2}{2n} \pm Z\sqrt{\frac{pq}{n} + \frac{Z^2}{4n^2}}}{1 + \frac{Z^2}{n}}$$

Diferencias de Estadísticos

En el campo de la investigación social es frecuente el uso de **estadísticos compuestos** que resultan de la **combinación de estadís**ticos. Puede interesarnos conocer la **diferencia** de medias de los salarios de hombres y mujeres, o la razón entre dos cantidades que han sido obtenidas mediante una muestra: por ejemplo, si en una encuesta se ha preguntado por el número de horas trabajadas y el salario, podemos calcular la razón entre la media de horas trabajadas y el salario medio, de donde se obtendrá un estadístico nuevos que será el cociente entre dos estadísticos que ya disponíamos en la muestra.

Cuando producimos nuevos estadísticos a partir de las combinaciones de otros estadísticos tenemos dos situaciones diferentes:

1. Que las observaciones que comparamos procedan de **muestras independientes**, por ejemplo si queremos estimar la diferencia de salario medio entre hombres y mujeres. Al hablar de muestras independientes no nos referimos a dos encuestas diferentes, sino que las unidades muestrales no tienen relación entre sí. Para comparar los salarios de hombres y mujeres, los hombres y las mujeres pertenecen a la misma muestra, sin embargo son subconjuntos que no guardan relación entre sí, es decir, las observaciones sobre los salarios de los hombres no afectan a las observaciones de las mujeres. Por consiguiente, ambos conjuntos de datos se consideran muestras independientes.
2. Que sean **muestras relacionadas entre sí**. Esta situación tiene distintos orígines. En unos casos, sobre la misma unidad muestral tomamos medidas repetidas en el tiempo, como sucede en los estudios de panel (se realiza la misma pregunta a la misma persona en fechas diferentes y se comparan). En otros casos, sobre la misma unidad muestral medimos características o variables distintas pero que están relacionadas entre sí. Por ejemplo, se pregunta a una persona, por ejemplo, sobre su valoración del presidente del gobierno y también del líder de la oposición. Para calcular un estadístico de diferencia entre ambas valoraciones tenemos que tener en cuenta que existe relación entre las repuestas. Seguramente quien valore alto al presidente del gobierno, lo haga bajo sobre el líder de la oposición.

Situaciones de dependencia e independencia entre muestras

1. Muestras Independientes: aquellas en que las observaciones se realizan sobre unidades muestrales distintas (una sola variable para comparar distintos grupos de individuos).
2. Muestras Relacionadas: aquellas en que las observaciones (variables) se refieren a la misma unidad muestral (se comparan varias variables sobre un solo conjunto de individuos).

Determinación del tipo de situación de construcción de estadísticos compuestos

	Una Variable	**Dos Variables**
Misma Unidad Muestral		Dependiente
Distintas Unidades Muestrales	Independiente	

Diferencia de Medias en casos de muestras independientes

Error Típico

$$\sigma_{(\mu_1-\mu_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Error Estadístico

$$e = Z\sigma_{(\mu_1-\mu_2)}$$

Diferencia de Proporciones en casos de muestras independientes

Error Típico

$$\sigma_{(p_1-p_2)} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

Error Estadístico

$$e = Z\sigma_{(p_1-p_2)}$$

Diferencia de Estadísticos en casos de muestras relacionadas

Error Típico

$$\sigma_{\overline{D}} = \frac{S_D}{\sqrt{n-1}}$$

S_D: Desviación Típica de la variable generada como diferencia de medias.

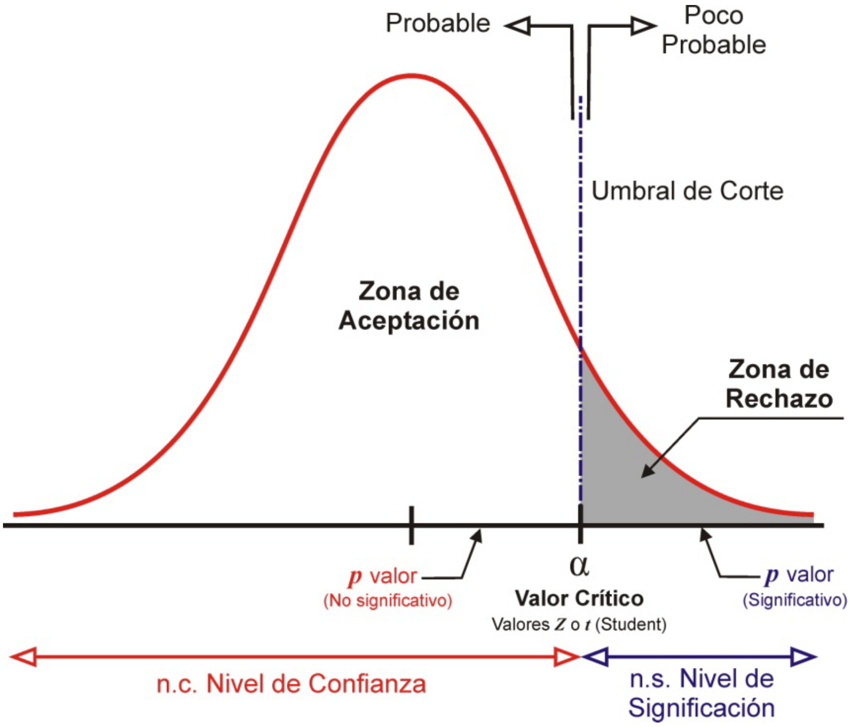
Error Típico a partir de la Cuasidesviación

$$\sigma_{\overline{D}} = \frac{S_D(n-1)}{\sqrt{n}}$$

Test de significación

Un resultado estadístico es **significativo** cuando **no se debe al azar**. En estadística, *significante* no se interpreta como en el lenguaje normal, no significa que el estadístico sea relevante o importante, ni siquiera que se considere verdadero. Un estadístico es significativo porque el investigador lo considera fiable.

Los **test de significación** son pruebas que se realizan para contrastar la información empírica con la teórica. Para ello se fija un **umbral de corte** a partir del cual los valores obtenidos en la muestra empírica se consideran poco probables. Si el resultado obtenido está al lado del umbral de corte donde los resultados se consideran probables, decimos que está dentro de la **zona de aceptación** de la hipótesis nula, pero si, por el contrario, el resultado cae del lado del umbral de corte donde se considera poco probable, decimos que entra dentro de la **zona de rechazo** de la hipótesis nula.



Hipótesis Nula (H_0)

En estadísticas la **hipótesis nula** (H_0) es la hipótesis de partida (también denominada hipótesis inicial) que considera las diferencias no significativas. La **hipótesis nula** (H_0) puede ser rechazada o no rechazada, pero no podemos probarla, salvo que estudiemos todos los elementos de la población y la mayoría de las veces eso no es posible.

Un resultado puede ser estadísticamente significativo aunque la diferencia sea muy pequeña y pueda parecer que no tiene importancia; por esta razón en los tets de significación se debe indicar el **efecto de la talla estadística**, esto es, el tamaño de la muestra. En muestras de tamaño grande pequeñas diferencias pueden ser consideradas a través de los test como significativas.

Pruebas de hipótesis

La evidencia que se necesita para aceptar que un acontecimiento se ha producido por azar es el **nivel de significación** (**n.s.** / *significante level* en inglés) o el **valor crítico** $\propto$. Si el $\boldsymbol{p}$ ***valor*** es pequeño, entonces debe rechazarse la **hipótesis nula** ($\boldsymbol{H_0}$), es falsa o inusual.

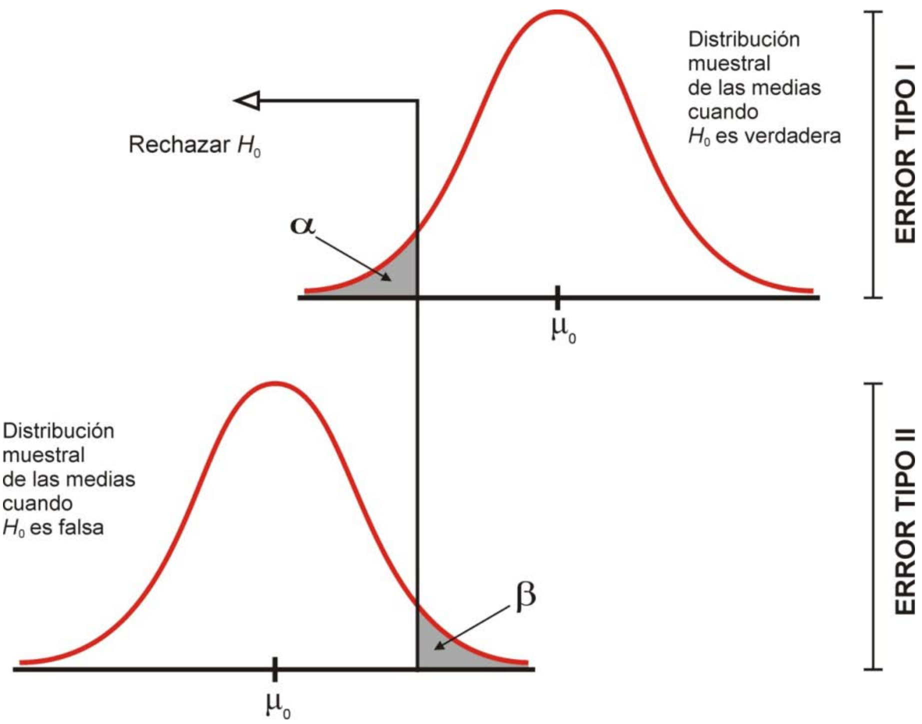
Según los trabajos de Neyman y Pearson, el planteamiento requiere tanto de la **hipótesis nula** ($\boldsymbol{H_0}$) como de una **hipótesis alternativa** ($\boldsymbol{H_1}$), que deben ser definidas e investigadas mediante repetición de procedimientos de muestreo; si no, se corre el riesgo de rechazar un resultado que cae fuera de la zona de aceptación y, sin embargo, ser verdadero. Es el **valor crítico** $\propto$ el que nos da la probabilidad de equivocarnos al rechazar ese valor.

Cuando se realizan pruebas de hipótesis existen dos tipos de errores que se pueden cometer:

1. **Error de Tipo I**: rechazar la hipótesis nula cuando es verdadera.
2. **Error de Tipo II**: aceptar la hipótesis nula cuando es falsa.

Tipos de error en los test de hipótesis

	H_0 verdadera	H_0 falsa
Aceptar H_0	Correcto	**Error Tipo II**
Rechazar H_0	**Error Tipo I**	Correcto



Nivel de significación del test o valor crítico $\propto$

Es la probabilidad que se está dispuesto a aceptar de rechazar erróneamente la hipótesis nula. Se trata de una probabilidad establecida. Habitualmente, como niveles de significación se usan el 5% y el 1%.

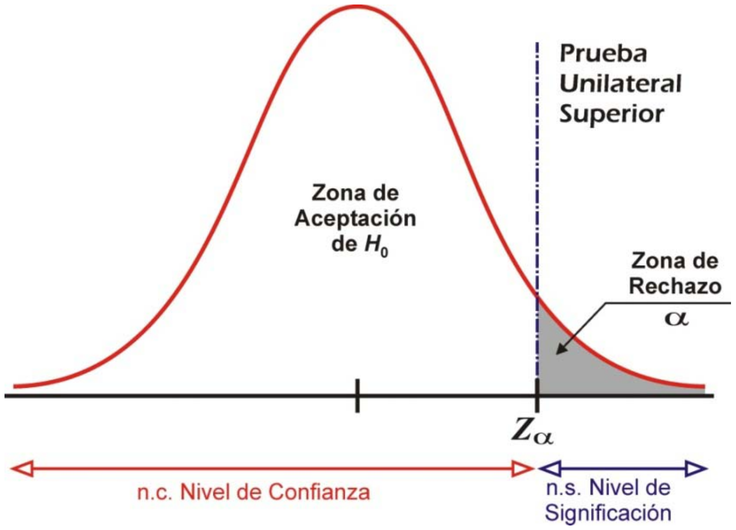
p valor

Es la probabilidad de obtener un resultado al menos tan extremo como el que realmente se ha obtenido en el experimento o en una muestra (valor del estadístico calculado), suponiendo que la hipótesis nula es cierta. Es fundamental tener en cuenta que el $p\ valor$ está basado en la aceptación de la hipótesis de partida (o hipótesis nula). A su vez, es la probabilidad de equivocarnos al aceptar que las diferencias (entre valor empírico y valor teórico) son significativas. Cuanto más pequeña sea esta probabilidad más seguro se está de no equivocarse.

Dos tipos de test

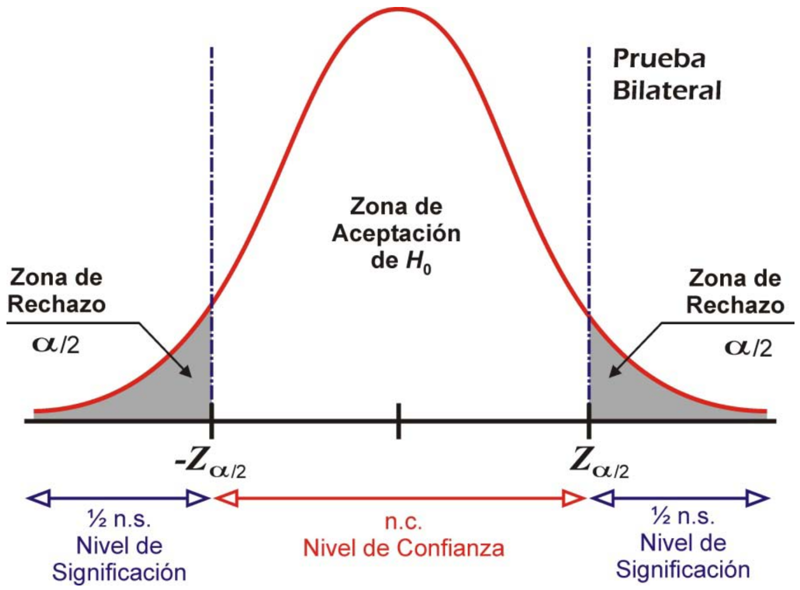
Test de una cola (Unilateral)

Cuando la hipótesis de partida se enuncia la dirección: que un valor “es mayor que…” o “menor que…” se realiza la prueba de una sola cola.



Test de dos colas (Bilateral)

Cuando la hipótesis de partida no hace ninguna especificación.



Procedimiento del test de significación o pruebas de hipótesis

1. Establecer el **valor crítico** $\propto$ (el error que se está dispuesto a aceptar).
2. **Calcular el estadístico.**
3. **Comparar** el estadístico (en valores de $\boldsymbol{Z}$ o $\boldsymbol{t}$ de Student) con el umbral fijado (también en valores de $\boldsymbol{Z}$ o $\boldsymbol{t}$ de Student) para el valor crítico $\propto$.
4. **Resultado:**
   - **a.** Si el estadístico es más alto que el valor crítico, queda por tanto en la zona de rechazo de la hipótesis nula y la diferencia es significativa. La probabilidad de equivocarse al rechazar H_0 es pequeña y el $p\ valor$ es menor que la relación encontrada por azar. **[** $\boldsymbol{p\ valor}\ < valor\ crítico\ \alpha$ **]**
   - **b.** Si el estadístico es más bajo que el valor crítico, al contrario, queda en la zona de aceptación de la hipótesis nula y la diferencia no es significativ. La probabilidad de rechazar H_0 es muy alta y el $p\ valor$ es mayor que el valor crítico. **[** $\boldsymbol{p\ valor\ > valor\ crítico\ \alpha}$ **]**

Valores de Z para los niveles de significación (n.s.) 5% y 1%

Tipo de Test	n.s. = 5%	n.s. = 1%
Test de una cola Unilateral	$1{,}65\ Z$	$2{,}33\ Z$
Test de dos colas Bilateral	$1{,}96\ Z$	$2{,}58\ Z$

Contrastes de hipótesis (muestra vs población de referencia)

Contraste para una media

Hipótesis de partida o nula $H_0: \bar{x} = \mu$

Hipótesis alternativa $H_1: \bar{x} \neq \mu$

Cálculo del estadístico en valor de $\boldsymbol{t}$ (Student)

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{N-1}}}$$

Cálculo del estadístico en valor de $\boldsymbol{Z}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

$$Z_e = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Contraste para una proporción

Hipótesis de partida o nula $H_0: p = P$

Hipótesis alternativa $H_1: p < P$

Cálculo en valor de Z

$$\boxed{\sigma_p = \sqrt{\frac{p \cdot q}{n}}}$$

$$\boxed{Z_e = \frac{p - P}{\sigma_p}}$$

Comparaciones (muestra vs muestra)

Comparación de medias

$$\boxed{Z_e = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_1}}}$$

$$\sigma_{\bar{x}_1 - \bar{x}_1} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} \qquad \sigma_{\bar{x}_1} = \frac{s_{X_1}}{\sqrt{N_1}} \qquad \sigma_{\bar{x}_2} = \frac{s_{X_2}}{\sqrt{N_2}}$$

Comparación de proporciones

Proporción Conjunta

$$\boxed{p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}}$$

$$\boxed{q = 1 - p}$$

$$\boxed{\sigma_p = \sqrt{p \cdot q \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\boxed{Z_e = \frac{p_1 - p_2}{\sigma_p}}$$

Análisis de Varianza (ANOVA)

Procedimiento de cálculo ANOVA

<u>1. Suma de cuadrados total</u>

$$\boxed{SC_t = \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}^2 - \frac{\left(\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}\right)^2}{N}}$$

$\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}^2$ = Sumatorio de los sumatorios de las observaciones.

$\left(\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}\right)^2$ = Sumatorio de los sumatorios de los cuadrados de las observaciones.

x_{ij} = Valor de la observación i en cada grupo j.

n_j = Cantidad de observaciones en los grupos j.

N = Total de observaciones ij.

2. Suma de los cuadrados entre grupos

$$SC_{ent} = \sum_{j=1}^{k} \frac{\left(\sum_{j=1}^{k} x_{ij}\right)^2}{n_j} - \frac{\left(\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}\right)^2}{N}$$

$\sum_{j=1}^{k} \frac{\left(\sum_{j=1}^{k} x_{ij}\right)^2}{n_j}$ = Sumatorio de los sumatorios de las observaciones al cuadrado dividido por el número de observaciones del grupo.

$\frac{\left(\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}\right)^2}{N}$ = Sumatorio de los cuadrados de las observaciones dividido por el total de observaciones.

3. Suma de cuadrados dentro de los grupos

$$SC_d = SC_t - SC_{ent}$$

4. Grados de Libertad

Grados de libertad total

$$gl_t = N - 1$$

Grados de libertad entre grupos

$$gl_{ent} = k - 1$$

Grados de libertad dentro de los grupos

$$gl_d = (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1)$$

También:

$$gl_d = gl_t - gl_{ent}$$

5. Estimación de la varianza

Varianza entre grupos

$$V_{ent} = \frac{SC_{ent}}{gl_{ent}}$$

Varianza dentro de los grupos

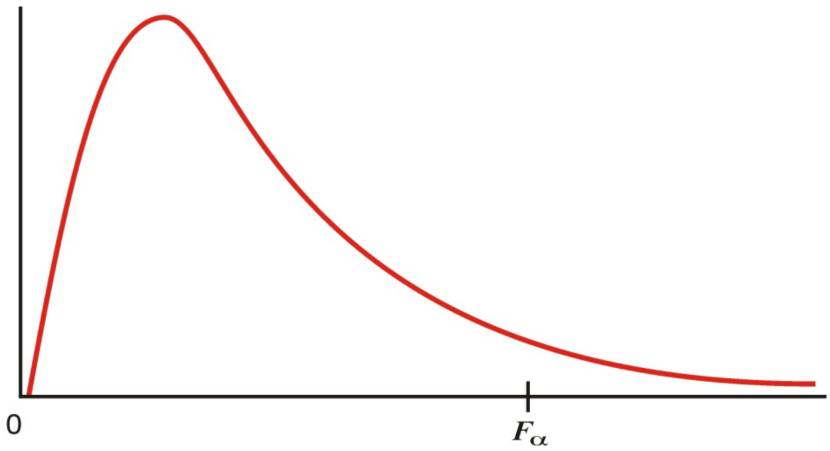
$$V_d = \frac{SC_d}{gl_d}$$

6. Contraste de la F (Estimador F de Snedecor)

$$F = \frac{V_{ent}}{V_d}$$

Distribución F de Snedecor

La distribución F de Snedecor es una distribución de probabilidad asociada a la normal no simétrica:



Diseño aleatorizado con un factor

1. Suma de cuadrados total

2. Suma de cuadrados debida a los tratamientos (factores)

3. Suma de los cuadrados debida a los bloques

4. Cálculo de la suma de cuadrados debida al error

Varianza Factor

$$V_{fact} = \frac{SC_{fact}}{gl_{fact}}$$

Error

$$error = \frac{SC_e}{gl_{error}}$$

Contraste de F

$$F = \frac{V_{fact}}{error}$$

Regresión y correlación lineal

La **regresión** tiene por objeto definir una función matemática que se ajuste lo mejor posible a los datos observados. Cuando se realiza sobre la relación de dos variables mediante el ajuste de una línea recta, hablamos de **regresión lineal simple**, y la manera más habitual e idónea de representarla es sobre los **diagramas de dispersión**.

La Covarianza

En una distribución bivariada contamos con una medida de dispersión que tiene en cuenta las dos variables a la vez. Para ello es necesario hacer uso de la Covarianza, que se obtiene del producto de las diferencias de x e y a sus medias.

Covarianza

$$Cov(x,y) = S_{xy} = \frac{\sum(x_i - \bar{x}) \cdot \sum(y_i - \bar{y})}{n}$$

Interpretación de la Covarianza

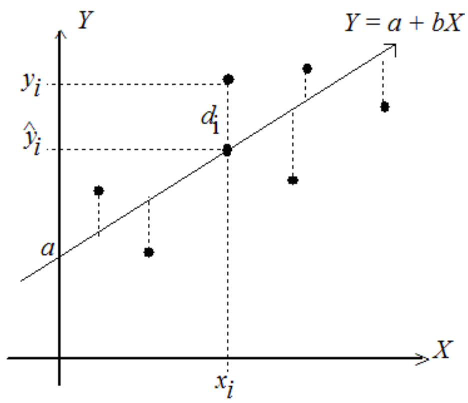
El valor de la covarianza nos informa de la existencia (o no) de dependencia lineal entre las variables. Si no hay relación lineal entre las dos variables, la covarianza será igual a

- Si $\boldsymbol{S_{xy}} = \mathbf{0}$: **No hay relación** lineal entre ambas variabes.
- Si $\boldsymbol{S_{xy}} \neq \mathbf{0}$: **Hay relación**. Mayor cuanto mayor sea la covarianza S_{xy}
- Si $\boldsymbol{S_{xy}} > \mathbf{0}$ la relación de dependencia lineal es **positiva**. Para grandes valores de X se obtienen grandes valores de Y
- Si $\boldsymbol{S_{xy}} < \mathbf{0}$ la relación de dependencia lineal es **negativa**. Para grandes valores de X se obtienen pequeños valores de Y.

Errores o residuos

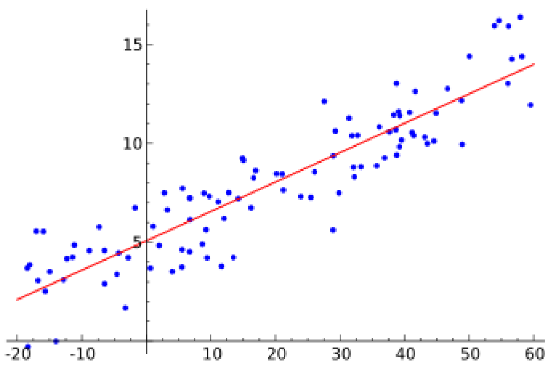
Los errores o residuos son la diferencia entre los valores reales y los de la recta ajustada a los datos. Por cada valor de x tenemos dos valores de y, el de la ecuación y el real observado. La suma de los residuos sirve para evaluar el ajuste final.

$$\boxed{errores\ o\ residuos = \varepsilon_i = (y_i - \hat{y}_i)}$$



Ecuación de la recta de regresión

$$\boxed{\hat{Y}_i = a + bX_i + \varepsilon_i}$$



Fórmulas de ajuste de la recta de regresión

Los parámetros de la recta de regresión se ajustan por el método de los **mínimos cuadrados**.

Covarianza

$$S_{xy} = \frac{\sum(x_i - \bar{x}) \cdot \sum(y_i - \bar{y})}{n} = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y}$$

Varianza de X

$$S_x^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Pendiente de la recta de regresión

$$b = \frac{S_{xy}}{S_x^2}$$

-> La covarianza medida en unidades de la varianza de X.

Punto de corte de la recta con el eje de la variable dependiente

$$a = \hat{y} - b\bar{x}$$

Una vez se han calculado los parámetros de la función de la recta de regresión lineal, la correlación sirve para cuantificar la bondad del ajuste de la recta a la nube de puntos.

Coeficiente de correlación de Pearson r

$$r = \frac{S_{xy}}{S_x S_y}$$

$$r = \frac{\sum(x_i - \bar{x}) \cdot \sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

$$r = \frac{\sum Z_x Z_y}{n}$$

$$Z_x = \frac{x - \bar{x}}{S_x} \quad ; \quad Z_y = \frac{y - \bar{y}}{S_y}$$

Interpretación del valor de r

El valor de $\boldsymbol{r}$ varía entre **-1** y **+1**

- Si $\boldsymbol{r} = \mathbf{0}$: No hay **correlación**. Las rectas de regresión son paralelas a los ejes.
- Si $\boldsymbol{r} > 0$: Hay **correlación positiva**. Al aumentar una variable, al aumentar una variable la otra también aumenta. La recta de regresión tiene **pendiente positiva**.
- Si $\boldsymbol{r} < 0$: Hay **correlación negativa**. Al aumentar una variable, al aumentar una variable la otra disminuye. La recta de regresión tiene **pendiente negativa**.
- Si $\boldsymbol{r} = \mathbf{1}$ o $\boldsymbol{r} = -\mathbf{1}$: Todos los puntos están contenidos en la recta de regresión.

Aunque el coeficiente de correlación se alto, no significa necesariamente que el ajuste sea óptimo. Puede existir una relación fuerte pero no lineal. Por ello, es imprescindible siempre acompañar el cálculo del **coeficiente de correlación $\boldsymbol{r}$** con el **gráfico** del **diagrama de dispersión**, con el fin de comprobar que se cumple la relación lineal entre las variables.

El **coeficiente de correlación $\boldsymbol{r}$** es **simétrico**. Tiene el mismo valor tanto si se trata de cuantificar con los mismos datos el ajuste de la recta de Y en X como de la recta de X en Y.

Coeficiente de determinación $\boldsymbol{r^2}$

Es el coeficiente de correlación de Pearson al cuadrado $\boldsymbol{r^2}$. Expresa la reducción proporcional del error que se comete al estimar los valores de la variable dependiente Y a partir de la recta de regresión. Se puede interpretar también como la probabilidad de la varianza total en una variable que es explicada por la otra variable en el modelo lineal. Al igual que r, $\boldsymbol{r^2}$ es **simétrico** y su valor es el mismo tanto si tomamos la variable independiente X por la dependiente Y y viceversa.

Coeficiente de determinación $\boldsymbol{r^2}$

$$r^2 = \frac{Variación\ explicada}{Variación\ total} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

$\sum(\hat{y} - \bar{y})^2$: Diferencia entre los valores de y obtenidos de la ecuación lineal y la media.

$\sum(y - \bar{y})^2$: Diferencia entre los valores observados y la media.

Para cada observación

$$\hat{y}_i = a + bx_i + \varepsilon_i$$

$\boldsymbol{\varepsilon_i}$: Residuo o diferencia entre el valor observado y_i y el estimado $\hat{y}$. La parte del valor observado de y no explicada por el modelo

Residuo

$$\varepsilon_i = (y_i - \hat{y}_i)$$

Suma de cuadrados total

$$\sum(y_i - \bar{y})^2 = \sum[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

Sólo una parte de la variación de Y es explicada por el modelo, y el valor de $\boldsymbol{r^2}$ indica qué parte de la variación total supone la variación explicada por la recta de regresión.

El **coeficiente de determinación $\boldsymbol{r^2}$** también puede obtenerse mediante el producto de las pendientes de las **dos rectas de regresión**:

$$r^2 = b \cdot b'$$

Tablas de contingencia

Las tablas de contingencia sirven para comparar grupos y observar cómo se distribuye una variable en el seno de otra variable. Para ver cómo se agrupa la población simultáneamente en dos variables, se realiza un ***cruce de variables*** y obtenemos una ***tabla de contingencia***.[2]

Tabla de contingencia Nivel de Estudios * Actitud hacia los impuestos

		Actitud hacia los impuestos			
		Bajar los impuestos y gastar menos en prestaciones	Gastar más en prestaciones aumentando los impuestos	Ns/Nc.	Total
Nivel de Estudios	Sin Estudios o Primarios	571	441	275	1287
	Secundarios	251	283	112	646
	Superiores	151	293	83	527
Total		973	1017	470	2460

Fuente: Barómetro CIS 2011. Estudio 2911. Tabla de elaboración propia.

La *distribución total de las filas* se denomina ***distribución marginal de filas***, y la *distribución total de columnas* se denomina ***distribución marginal de columnas***. En la práctica se suele referirse a ellas como ***marginales***, que son las distribuciones totales de cada una de las variables cruzadas. La ***distribución conjunta de las variables*** se observa en las casillas centrales.

La anterior tabla de contingencia se ha realizado con frecuencias absolutas, con lo que la información contenida en la tabla no se presenta fácil para el análisis. Para una mejor comparación de la distribución entre las dos variables se hace necesario obtener la tabla con las frecuencias relativas (en porcentajes).

Análisis Bivariable

La relación entre dos variables se establece con el esquema explicativo: *la distribución de una variable es explicada por la distribución de otra variable.* En otras palabras, una variable es

[2] Es importante destacar la diferencia entre ***variables de registro*** y ***variables de análisis***. Las ***variables de registro*** son las variables originales que se emplean para registra la información, que suelen presentarse con mucho detalle (asimiladas a términos abstractos o administrativos de registro) y sirven como punto de partida para la investigación. Las ***variables de análisis*** son las variables construidas con pertinencia empírica y/o teórica para el análisis sociológico.

determinada por otra. Por tanto, tenemos **variable independiente** y **variable dependiente**. Los cambios de la variable independiente explican los de la dependiente.

$$\boxed{X \longrightarrow Y}$$

Desde el **enfoque matemático**, cualquiera de las dos variables puestas en relación puede ejercer indistintamente el papel de independiente como de dependiente. Pero desde el **enfoque empírico**, estos papeles se definen contrastando que una de las variables es antecedente (temporalmente) de la otra o presenta más estabilidad temporal o se considera más básica; y esta será la variable independiente. Podría darse el caso que dos variables en relación teórica no tengan ninguna relación de dependencia, y por consiguiente ambas variables serían independientes.

Regla de Zeisel

Para el análisis de la relación de dos variables en una tabla de contingencia se calculan los porcentajes en la dirección de la variable independiente y se comparan en la dirección de la variable dependiente.

Como regla general colocamos los datos de **la variable independiente en las filas**. Según De Miguel (1997): «La mejor disposición es la de **porcentajes horizontales** porque el ojo humano compara mejor las relaciones de arriba abajo»[3]. Por tanto, conviene por sistema emplear esta regla en todos los casos, calcular los porcentajes en la dirección horizontal y comparar las relaciones entre ambas variables en la dirección vertical.

Tabla de contingencia Nivel de Estudios * Actitud hacia los impuestos

		Actitud hacia los impuestos			
		Bajar los impuestos y gastar menos en prestaciones	Gastar más en prestaciones aumentando los impuestos	Ns/Nc.	Total
Nivel de Estudios	Sin Estudios o Primarios	44,4%	34,3%	21,4%	100,0%
	Secundarios	38,9%	43,8%	17,3%	100,0%
	Superiores	28,7%	55,6%	15,7%	100,0%
Total		39,6%	41,3%	19,1%	100,0%

Fuente: Barómetro CIS 2011. Estudio 2911. Tabla de elaboración propia.

[3] De Migel, A. 1997. Manual del perfecto sociólogo, Madrid, Espasa, p. 67.

Análisis Trivariable (Tercera Variable. Variable de Control)

Con las tablas de contingencia de dos variables se parte del supuesto de que el resto de variables se mantienen constantes. Sin embargo, en la realidad esto no ocurre nunca. La experiencia empírica demuestra que múltiples variables afectan al comportamiento de otra; y muchas no se pueden controlar cuando no hay registro. Es posible que algunas no aparezcan cuando se está estudiando una encuesta estadística; otras se tienen que construir a partir de las variables de registro.

La introducción de una **tercera variable** puede alterar las supuestas relaciones halladas en la tabla de contingencia de dos variables, de manera que la relación bivariable anterior se desvanezca. A esta tercera variable la llamamos **variable de control**, y se introduce segmentando los datos del cruce de las dos variables anteriores en función de esta tercera variable.

Tabla de contingencia Nivel de Estudios * Actitud hacia los impuestos * Sexo

Sexo			Actitud hacia los impuestos: Bajar los impuestos y gastar menos en prestaciones	Gastar más en prestaciones aumentando los impuestos	Ns/Nc.	Total
Hombre	Nivel de Estudios	Sin Estudios o Primarios	41,3%	41,6%	17,2%	100,0%
		Secundarios	36,8%	44,2%	19,0%	100,0%
		Superiores	27,5%	57,2%	15,3%	100,0%
	Total		37,4%	45,2%	17,4%	100,0%
Mujer	Nivel de Estudios	Sin Estudios o Primarios	47,4%	27,1%	25,5%	100,0%
		Secundarios	41,3%	43,3%	15,4%	100,0%
		Superiores	29,5%	54,4%	16,1%	100,0%
	Total		41,6%	37,6%	20,8%	100,0%

Fuente: Barómetro CIS 2011. Estudio 2911. Tabla de elaboración propia.

Paradoja de Simpson

Una asociación entre dos variables desaparece o cambia de sentido cuando los datos son desagregados por grupos, esto es, cuando se controla el efecto de una tercera variable.

La variable de control $\boldsymbol{Z}$ que se introduce para observar el comportamiento en una relación bivariable original afecta a ésta si:

1. Se relaciona con la variable independiente o causal (X).

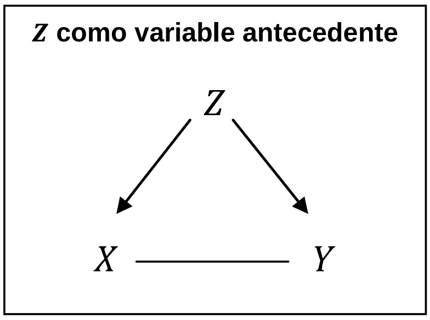
2. Se relaciona con la variable dependiente o efecto (Y)

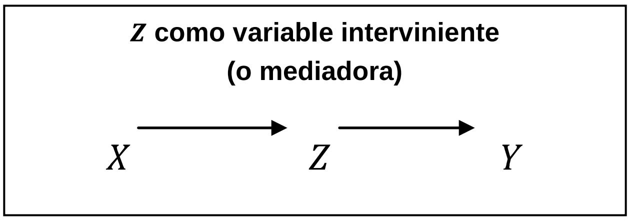
3. Cuando segmentamos la población según los gruos de la variable de control (Z), se observa que las relaciones entre las variables originales $(X\ e\ Y)$ son de menor intensidad que la que manifestaban antes de introducir la tercera variable. Pero también cabe la posibilidad de que la tercera variable afecte a las dos variables $(X\ e\ Y)$ sin alterar la relación original establecida por éstas.

En los modelos donde una tercera variable altera una determinada relación entre dos variables, existen dos casos en la forma en que actúa la tercera variable:

1. (Z) Variable Antecedente: Z actua sobre X y sobre Y.

2. (Z) Variable Interviniente (o mediadora): Z se interpone entre X e Y, esto es, X actua sobre Z y ésta sobre Y.





Todas las interacciones posibles se deben fundamentar en modelos de interacción de pertinencia sociológica.

La interacción entre tres variables se puede resumir en una tabla, pero la forma más óptima de analizar las relaciones entre las tres variables es segmentando la población entre los valores de una de las variables, la que se considere más básica o primordial. De esta manera se puede observar cómo se comporta las otras dos variables en cada una de las categorías (o valores) de la primera.

En muchas ocasiones resulta interesante dualizar (o dicotomizar) las variables que se cruzan. La visibilidad de las confrontaciones duales facilita el análisis.

Ji-cuadrado (χ^2)

La prueba de **Ji-cuadrado** (χ^2) es un **test** que afecta a la distribución de frecuencias de los diferentes grupos que componen una población y que son generados por un cruce de variables.

Con la **prueba de Ji-cuadrado** (χ^2) se pretende conocer en términos probabilísticos si el conjunto de las frecuencias relativas (proporciones) de todos los grupos generados en una población (por el cruce de variables) se distribuye forma a aleatoria (al azar), sin diferencias «significativas» (significación estadística) entre ellas (en conjunto). Igual que en todas las pruebas estadísticas, se contrasta la ***distribución real*** de los datos (a partir de un estadístico) con una ***distribución teórica***, en este caso, la **Ji-cuadrado** (χ^2), en donde definimos un ***punto crítico*** a partir del cual las diferencias se consideran «significativas» entre las proporciones del conjunto de los grupos poblacionales (condensadas en el estadístico Ji-cuadrado). Por tanto, por un lado tenemos el ***estadísitco* (Ji-cuadrado)** asociado a una tabla de contingencia, y por otro, una ***distribución teórica*** de este estadístico.

La prueba de Ji-cuadrado sólo tiene sentido cuando se trabaja con datos muestrales extraídos de forma aleatoria, esto es, en distribuciones (empíricas) de probabilidad.

Con la prueba de Ji-cuadrado la especificidad (empírica/teórica) concreta de las categorías desaparece. Para Ji-cuadrado es irrelevante si la prueba la hacemos sobre variables de edad, nivel de estudios, definiciones ideológicas, etc. Ji-cuadrado vale lo mismo para categorías sociológicas, médicas, criminológicas, biológicas, etc. **Ji-cuadrado no distingue la especificidad de las categorías consideradas.** Se trata de un índice que resume toda la distribución teniendo en cuenta sólo el número de categorías y las frecuencias que se dan en éstas.

Frecuencia teórica (o esperada según la lógica de Ji-cuadrado)

$$Frecuencia\ esperada = \frac{Total\ fila\ x\ Total\ columna}{Total\ tabla}$$

$$f_{e_{ij}} = \frac{N_{fila\ ij} \times N_{columna\ ij}}{N_{Total}}$$

Se podría decir que las frecuencias esperadas en cada una de las casillas de una tabla de contingencia son aquellas que cabe esperar en el caso de que no exista relación entre las variables, esto es, independencia estadística o probabilística entre las variables. **Dos variables son independientes estadísticamente si la probabilidad de que nos aparezca una categoría de una variable no depende de la distribución de probabilidad de la otra variable.**

Cálculo de residuos

Los residuos son las diferencias entre las frecuencias observadas y las esperadas en cada celda. Los residuos son «errores» del muestreo aleatorio.

Residuo

$$Residuo = f_o - f_e$$

Para poder comparar las «distancias» entre las frecuencias observadas y las esperadas dentro de cada celda, se ha de **estandarizar** estas distancias.

Residuos Estandarizados

$$Residuo\ Estandarizado = \frac{f_o - f_e}{\sqrt{f_e}}$$

χ^2 de cada celda

$$\chi^2 = \frac{(f_o - f_e)^2}{f_e}$$

$$\chi_{ij}^2 = \frac{\left(f_{o\,ij} - f_{e\,ij}\right)^2}{f_{e\,ij}}$$

Coeficiente Ji-cuadrado χ^2 de toda la tabla (Ejemplo Tabla 2x2)

$$\chi^2 = \chi_{11}^2 + \chi_{12}^2 + \chi_{21}^2 + \chi_{22}^2 = \sum_{i=1}^{i=2}\sum_{j=1}^{j=2}\chi_{ij}^2$$

Tabla $(r \times s)$ para la fórmula general del índice χ^2

n_{11}	n_{12}	n_{13}	…	n_{1j}	…	n_{1s}	$\boldsymbol{N_{1\cdot}}$
n_{21}	n_{22}	n_{23}	…	n_{2j}	…	n_{2s}	$\boldsymbol{N_{2\cdot}}$
n_{31}	n_{32}	n_{33}	…	n_{3j}	…	n_{3s}	$\boldsymbol{N_{3\cdot}}$
…	…	…	…	…	…	…	…
n_{i1}	n_{i2}	n_{i3}	…	n_{ij}	…	n_{is}	$\boldsymbol{N_{i\cdot}}$
…	…	…	…	…	…	…	…
n_{r1}	n_{r2}	n_{r3}	…	n_{rj}	…	n_{rs}	$\boldsymbol{N_{r\cdot}}$
$\boldsymbol{N_{\cdot 1}}$	$\boldsymbol{N_{\cdot 2}}$	$\boldsymbol{N_{\cdot 3}}$	…	$\boldsymbol{N_{\cdot j}}$	…	$\boldsymbol{N_{\cdot s}}$	$\boldsymbol{N_{\cdot\cdot}}$

Fórmula suma de frecuencias columna genérica $(\boldsymbol{j})$

$$N_{\cdot j} = n_{1j} + n_{2j} + n_{3j} + \cdots + n_{ij} + \cdots + n_{rj} = \sum_{i=1}^{j=r} n_{ij}$$

Fórmula suma de frecuencias fila genérica $(\boldsymbol{i})$

$$N_{\cdot j} = n_{i1} + n_{i2} + n_{i3} + \cdots + n_{ij} + \cdots + n_{is} = \sum_{j=1}^{j=s} n_{ij}$$

Sumatorio de todas las celdas

$$f_{e\,ij} = \sum_{i=1}^{j=r}\sum_{j=1}^{j=s} n_{ij}$$

Frecuencia esperada de una celda cualquiera $(\boldsymbol{i}, \boldsymbol{j})$

$$f_{e_{ij}} = \frac{N_{i.} \times N_{.j}}{N_{..}}$$

Coeficiente Ji-cuadrado $\boldsymbol{\chi}^2$ de la Tabla

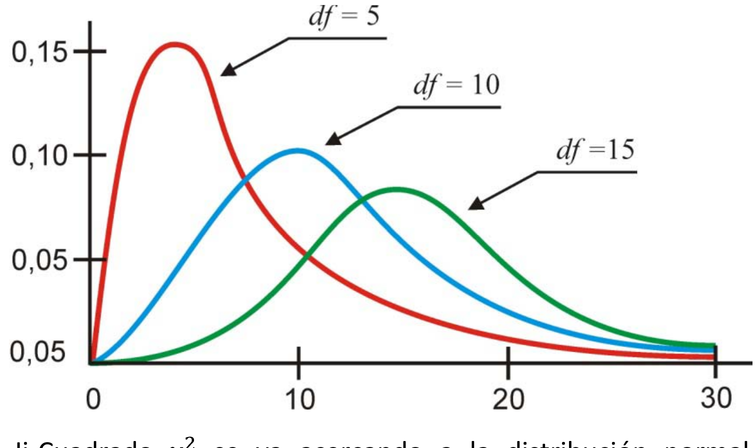
$$\chi^2 = \sum_{i=1}^{j=r} \sum_{j=1}^{j=s} {\chi_{ij}}^2 = \sum_{i=1}^{j=r} \sum_{j=1}^{j=s} \frac{\left(f_{o\,ij} - f_{e\,ij}\right)^2}{f_{e\,ij}}$$

Interpretación de Ji-cuadrado $\boldsymbol{\chi}^2$en tablas bivariables

- Si $\boldsymbol{\chi}^2$ toma valores entre $\boldsymbol{0}$ $\boldsymbol{e}$ $\infty \rightarrow$ **Índice o Coeficiente de relación.**
- Si $\boldsymbol{\chi}^2 = \boldsymbol{0}$: **No hay relación** entre las variables.
- χ^2 de una tabla concreta no vale para comparar otras tablas distintas. Para ello existen otros coeficientes de contingencia: *V de Cramer, Coeficiente de contingencia* o (φ).

La **prueba o test de Ji-cuadrado $\boldsymbol{\chi}^2$** es un ***contraste de proporciones múltiples.*** La *distribución teórica* asociada a esta prueba es la **distribución $\boldsymbol{\chi}^2$**. Se trata de una distribución continua de probabilidad, una distribución muestral de la varianza derivada de la distribución normal. Al igual que la distribución t de Student, depende de un parámetro: ***grados de libertad***. Por tanto, existe una distribución distinta para cada número de grados de libertad. La **media** de la distribución Ji-cuadrado χ^2 es igual a sus grados de libertad, su **varianza** es dos veces sus grados de libertad, y sus **valores** oscilan entre $\boldsymbol{0}$ $\boldsymbol{e}$ ∞.

Distintas distribuciones Ji-cuadrado χ^2 para distintos *grados de libertad (df)*



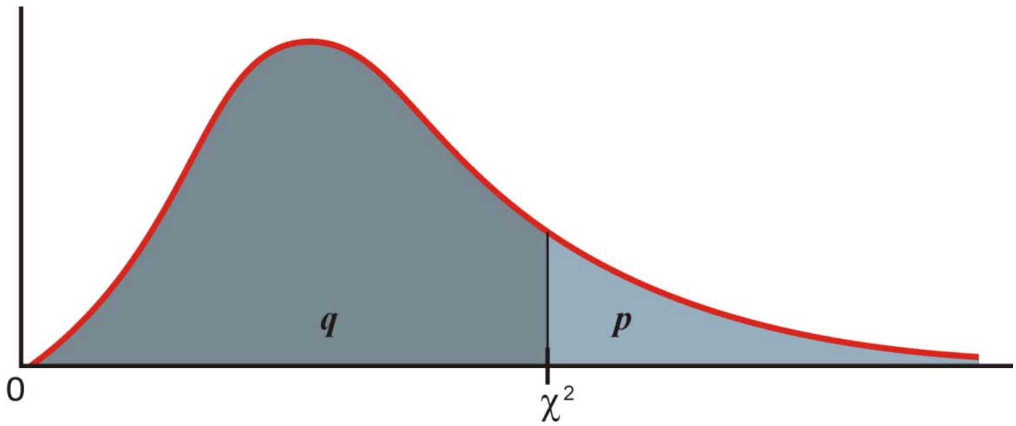
La distribución Ji-Cuadrado χ^2 se va acercando a la distribución normal a medida que aumentan sus grados de libertad.

Número de *grados de libertad* de una tabla bidimensional $(\boldsymbol{r} \times \boldsymbol{s})$

$$\mathrm{gl}_{rs} = (r-1) \times (s-1)$$

Prueba de la distribución Ji-Cuadrado χ^2 con los correspondientes grados de libertad

Como todas las distribuciones de probabilidad, la distribución Ji-Cuadrado χ^2 describe un área bajo la curva que es igual a 1, esto es, donde encontramos el 100% de los casos. Cualquier valor de χ^2 deja a su izquierda un porcentaje de casos que asignaremos a la proporción $\boldsymbol{q}$, y a su derecha el resto de casos que asignaremos a la proporción $\boldsymbol{p}$. Entre 0 y un valor cualquiera de χ^2 encontraremos el $\mathbf{100} \cdot \boldsymbol{q}\%$ de los casos. Y entre el valor χ^2 e ∞ encontraremos el resto, el $\mathbf{100} \cdot \boldsymbol{p}\%$. Y sabemos que $\boldsymbol{p} + \boldsymbol{q} = \mathbf{1}$.



El valor crítico y nivel de significación de χ^2

La $\boldsymbol{p}$ de la Tabla Ji-cuadrado, distribución de χ^2 (ver Anexo 5) nos indica el **valor crítico de $\boldsymbol{\chi^2}$** a partir del cual todo valor mayor entre en la zona de rechazo de la hipótesis nula (H_0), que considera las variables cruzadas en la tabla son independientes estadísticamente, es decir, no hay relación entre ellas. Por tanto, la $\boldsymbol{p}$ es el ***nivel de significación*** que usamos para la prueba de la distribución Ji-cuadrado χ^2.

Anexo 1: Tabla Z. Distribución Normal Estándar

Tabla *Z*. Distribución normal estándar

Áreas entre 0 y Z_x

Para conocer el valor entre $Z = 0$ y $Z = 1,96$, seleccionamos en la primera columna, la fila $Z = 1,9$ y la columna 0,06. El valor es: 0,4750.

Para valores de Z negativos téngase en cuenta que el área entre $-Z$ y 0 es igual que entre 0 y $+Z$.

Z_x	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998

Fuente: Camarero Rioja, L. et al. 2010. *Estadística para la investigación social*. Ibergarceta

Anexo 2: Tabla *t*. Distribución *t* de Student

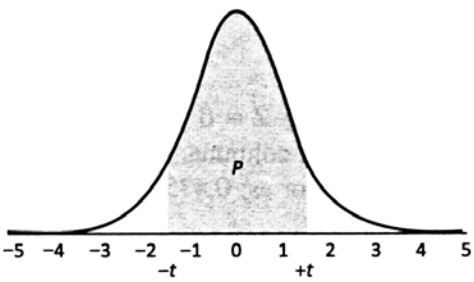


Tabla *t*.
Distribución *t* de Student

Valores de *t*, que en el intervalo [−*t*; +*t*], comprenden la probabilidad señalada

[Para cada tamaño de la muestra «*n*», indicado por el número de grados de libertad («*n* − 1»), tenemos una distribución *t*. Cada línea es, por tanto, una distribución distinta.]

	Probabilidad										
Grados de libertad	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**0,95**	**0,98**	**0,99**	**0,995**	**0,998**	**0,999**
1	1,000	1,376	1,963	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,767
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,690
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,659
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
50	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
80	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
100	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,390
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	2,860	3,160	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Fuente: Camarero Rioja, L. et al. 2010. *Estadística para la investigación social*. Ibergarceta

Anexo 3: Tabla *F*. Distribución *F* de Fisher (N.s. = 0,01)

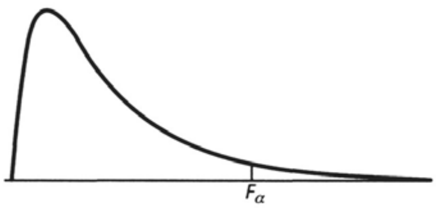


Tabla.
Distribución *F* de Fisher

N.s. = 0,01

gl_d \ gl_{ent}	1	2	3	4	5	6	8	12	24	∞
1	4.052	4.999	5.403	5.625	5.764	5.859	5.981	6.106	6.234	6.366
2	98,49	99,01	99,17	99,25	99,30	99,33	99,36	99,42	99,46	99,50
3	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,60	26,12
4	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	13,93	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,27	9,89	9,47	9,02
6	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,31	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,47	6,07	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,67	5,28	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,11	4,73	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,71	4,33	3,91
11	9,65	7,20	6,22	5,67	5,32	5,07	4,74	4,40	4,02	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,78	3,36
13	9,07	6,70	5,74	5,20	4,86	4,62	4,30	3,96	3,59	3,16
14	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,43	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,18	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65
18	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,00	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36
22	7,94	5,72	4,82	4,31	3,99	3,76	3,45	3,12	2,75	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2,70	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,66	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,99	2,62	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,58	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,55	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,23	2,90	2,52	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,20	2,87	2,49	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,47	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,66	2,29	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2,12	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,66	2,34	1,95	1,38
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,79	1,00

Fuente: Almazán, A. et al. 2011. *Análisis estadístico para la investigación social*. Ibergarceta

Anexo 4: Tabla *F*. Distribución *F* de Fisher (N.s. = 0,05)

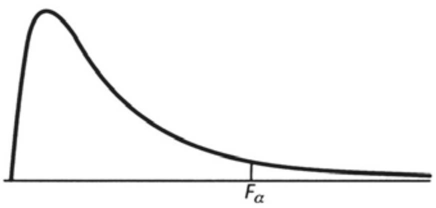


Tabla.
Distribución *F* de Fisher
N.s. = 0,05

gl_d \ gl_{ent}	1	2	3	4	5	6	8	12	24	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,40
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,14
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,71
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Fuente: Almazán, A. et al. 2011. *Análisis estadístico para la investigación social*. Ibergarceta

Anexo 5: Tabla Ji-cuadrado. Distribución de χ^2

Distribución de χ^2

gl	Nivel de significación													
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,000157	0,000628	0,00393	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,0201	0,0404	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,665	4,624	6,251	7,815	9,837	11,345	16,268
4	0,297	0,429	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,465
5	0,554	0,752	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,517
6	0,872	1,134	1,635	2,204	2,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	1,646	2,032	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,90	26,125
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	2,558	3,059	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	4,107	4,765	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697

Fuente: Almazán, A. et al. 2011. *Análisis estadístico para la investigación social*. Ibergarceta