



**ntic**  
master  
**School**

UNIVERSIDAD  
COMPLUTENSE  
DE MADRID



# Master Big Data and Business Analytics

Conrado M. Manuel García

Departamento de Estadística y Ciencia de los Datos de  
la Universidad Complutense de Madrid.

# Estadística Descriptiva

En Estadística buscamos encontrar relaciones entre variables para hacer predicciones. Esto también ocurre en el caso de otras ciencias. La singularidad de la Estadística radica en que los procesos a los que se dirige son aleatorios, no deterministas.

Para conseguir sus objetivos, la Estadística se divide en tres cuerpos doctrinales: Estadística Descriptiva, Cálculo de Probabilidades e Inferencia.

La Estadística Descriptiva sirve para resumir la información contenida en un conjunto de datos. No persigue objetivos prospectivos.

La Inferencia trata de proyectar la información de las muestras a toda la población. También sirve para la toma de decisiones a partir de los datos disponibles.

En el medio de ambas surge el Cálculo de Probabilidades, la más matematizada de las tres disciplinas, cuyo objetivo básico es desarrollar distribuciones de probabilidad que sirvan para controlar la variabilidad y para medir la bondad de ajuste de los modelos, así como el margen de error en la toma de decisiones.



En Estadística Descriptiva se obtiene información utilizando variables estadísticas, que pueden ser de dos tipos:

- **Variables cualitativas o atributos:** no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo).
- **Variables cuantitativas:** tienen valor numérico (edad, precio de un producto, ingresos anuales).

Las **variables** también se pueden clasificar en:

- **Variables unidimensionales:** sólo recogen información sobre una característica (por ejemplo: edad de los alumnos de una clase).
- **Variables bidimensionales:** recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase).
- **Variables pluridimensionales:** recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

Por su parte, las **variables cuantitativas** se pueden clasificar en discretas y continuas:

- **Discretas:** sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos.
- **Continuas:** pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 80,3 km/h, 94,57 km/h...etc.

Cuando se estudia el comportamiento de una variable hay que distinguir los siguientes elementos:

- **Individuo:** cualquier elemento portador de información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los niños de una clase, cada alumno es un individuo; si estudiamos el precio de la vivienda, cada vivienda es un individuo.
- **Población:** conjunto de todos los individuos (personas, objetos, animales, etc.) que interesan desde el punto de vista de la información sobre el fenómeno que se estudia. Por ejemplo, si estudiamos el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.
- **Muestra:** subconjunto que seleccionamos de la población. Así, si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de ella (sería una labor muy compleja), sino que se suele seleccionar un subgrupo (muestra) que deberá ser representativo si se desea que la información sea veraz.

La **distribución de frecuencias** es la representación estructurada, en forma de tabla, de la información recogida sobre la variable que se estudia.

Variable (Valor)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
$X_1$	$k_1$	$k_1$	$f_1 = k_1 / n$	$f_1$
$X_2$	$k_2$	$k_1 + k_2$	$f_2 = k_2 / n$	$f_1 + f_2$
...	...	...	...	...
$X_{n-1}$	$k_{n-1}$	$k_1 + k_2 + \dots + k_{n-1}$	$f_{n-1} = k_{n-1} / n$	$f_1 + f_2 + \dots + f_{n-1}$
$X_n$	$k_n$	$\sum k_i$	$f_n = k_n / n$	$\sum f_i$

### Ejemplo:

Medimos la altura de los niños de una clase y obtenemos los siguientes resultados (cm):

Alumno	Estatura	Alumno	Estatura	Alumno	Estatura
Alumno 1	1,25	Alumno 11	1,23	Alumno 21	1,21
Alumno 2	1,28	Alumno 12	1,26	Alumno 22	1,29
Alumno 3	1,27	Alumno 13	1,30	Alumno 23	1,26
Alumno 4	1,21	Alumno 14	1,21	Alumno 24	1,22
Alumno 5	1,22	Alumno 15	1,28	Alumno 25	1,28
Alumno 6	1,29	Alumno 16	1,30	Alumno 26	1,27
Alumno 7	1,30	Alumno 17	1,22	Alumno 27	1,26
Alumno 8	1,24	Alumno 18	1,25	Alumno 28	1,23
Alumno 9	1,27	Alumno 19	1,20	Alumno 29	1,22
Alumno 10	1,29	Alumno 20	1,28	Alumno 30	1,21



Veamos **un ejemplo**:

Medimos la altura de los niños de una clase y obtenemos los siguientes resultados (cm).

Si presentamos esta información estructurada obtendríamos la siguiente **tabla de frecuencias**:

Variable (Valor)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%

Si los valores que toma la variable son muy diversos y cada uno de ellos se repite muy pocas veces, entonces conviene agruparlos por intervalos, ya que de otra manera obtendríamos una tabla de frecuencias muy pequeñas con poco valor de síntesis

Supongamos que medimos la estatura de los habitantes de un edificio y obtenemos los siguientes resultados (cm):

Habitante	Estatura	Habitante	Estatura	Habitante	Estatura
Habitante 1	1,15	Habitante 11	1,53	Habitante 21	1,21
Habitante 2	1,48	Habitante 12	1,16	Habitante 22	1,59
Habitante 3	1,57	Habitante 13	1,60	Habitante 23	1,86
Habitante 4	1,71	Habitante 14	1,81	Habitante 24	1,52
Habitante 5	1,92	Habitante 15	1,98	Habitante 25	1,48
Habitante 6	1,39	Habitante 16	1,20	Habitante 26	1,37
Habitante 7	1,40	Habitante 17	1,42	Habitante 27	1,16
Habitante 8	1,64	Habitante 18	1,45	Habitante 28	1,73
Habitante 9	1,77	Habitante 19	1,20	Habitante 29	1,62
Habitante 10	1,49	Habitante 20	1,98	Habitante 30	1,01

Si presentáramos esta información en una tabla de frecuencia obtendríamos una tabla de 30 líneas (una para cada valor), cada uno de ellos con una frecuencia absoluta de 1 y con una frecuencia relativa del 3,3%. Esta tabla nos aportaría escasa información.

En lugar de ello, es preferible agrupar los datos por intervalos, con lo que se obtiene un mejor resumen, aunque se pierde algo de información. El número de intervalos en los que se agrupa la información es una decisión que debe tomar el analista: la regla es que mientras más tramos se utilicen menos información se pierde, pero también menos resumen se conseguirá.

Estatura Cm	Frecuencias absolutas		Frecuencias relativas
	Simple	Acumulada	Simple
1,01 - 1,10	1	1	3,3%
1,11 - 1,20	3	4	10,0%
1,21 - 1,30	3	7	10,0%
1,31 - 1,40	2	9	6,6%
1,41 - 1,50	6	15	20,0%
1,51 - 1,60	4	19	13,3%
1,61 - 1,70	3	22	10,0%
1,71 - 1,80	3	25	10,0%
1,81 - 1,90	2	27	6,6%
1,91 - 2,00	3	30	10,0%

## Medidas de Posición

Las medidas de posición nos facilitan también información resumida sobre la serie de datos que estamos analizando. Estas medidas permiten conocer diversas características de ella. Las **medidas de posición** son de dos tipos:

- a) **Medidas de posición central:** informan sobre los valores medios de la serie de datos.
- b) **Medidas de posición no centrales:** informan sobre cómo se distribuye el resto de los valores de la serie.

## Medidas de posición central

Las principales medidas de posición central son las siguientes:

1. **Media:** es el valor medio ponderado de la serie de datos. Se pueden calcular diversos tipos de media, siendo las más utilizadas:
  - a) **Media aritmética:** se calcula multiplicando cada valor por el número de veces que se repite. La suma de todos estos productos se divide por el total de datos.
  - b) **Media geométrica:** se eleva cada valor al número de veces que se ha repetido. Se multiplican todos estos resultados y al producto final se le calcula la raíz "n-ésima" (siendo "n" el total de datos de la muestra).

La media geométrica se suele utilizar en series de datos como tipos de interés anuales, inflación, etc., donde el valor de cada año tiene un efecto multiplicativo sobre el de los años anteriores. En todo caso, la media aritmética es la medida de posición central más utilizada.

Lo más positivo de la media es que en su cálculo se utilizan todos los valores de la serie, por lo que no se pierde ninguna información.

Sin embargo, presenta el problema de que su valor (tanto en el caso de la media aritmética como geométrica) se puede ver muy influido por valores extremos, que se aparten en exceso del resto de la serie. Estos valores anómalos podrían condicionar en gran medida el valor de la media, perdiendo ésta representatividad.

**2. Mediana:** es el valor de la serie de datos que se sitúa justamente en el centro de la muestra (un 50% de valores son inferiores y otro 50% son superiores).

No presenta el problema de estar influida por los valores extremos, pero en cambio no utiliza en su cálculo toda la información de la serie de datos, sino sólo el orden.

**3. Moda:** es el valor que más se repite en la muestra.

**Ejemplo:** vamos a utilizar la tabla de distribución de frecuencias con los datos de la estatura de los alumnos para obtener las medidas de posición central.

Variable (Valor)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%



**Media Aritmética:**  $\frac{1.20x1+1.21x4+\cdots+1.30x3}{30} = 1.253$

**Media Geométrica:**  $\sqrt[30]{1.20x1.21^4x \dots x1.30^3}$

**Mediana:**

La mediana de esta muestra es 1,26 cm, ya que por debajo está el 50% de los valores y por arriba el otro 50%. Esto se puede ver al analizar la columna de frecuencias relativas acumuladas.

En este ejemplo, como el valor 1,26 se repite en 3 ocasiones, la media se situaría exactamente entre el primer y el segundo valor de este grupo, ya que entre estos dos valores se encuentra la división entre el 50% inferior y el 50% superior.

**Moda:**

Hay 3 valores que se repiten en 4 ocasiones: el 1,21, el 1,22 y el 1,28, por lo tanto esta serie tiene 3 modas.

## Medidas de posición no centrales

Las medidas de posición no centrales permiten conocer otros valores característicos de la distribución que no son los centrales. Entre otros indicadores, se suelen utilizar los que dividen a la muestra en tramos iguales:

**Cuartiles:** son 3 valores que dividen la serie de datos, ordenada de forma creciente, en cuatro tramos iguales. En cada uno de ellos se concentra el 25% de los resultados. El segundo cuartil coincide con la mediana, dada la definición.

**Deciles:** son 9 valores que dividen la serie de datos, ordenada de forma creciente, en diez tramos iguales, en el sentido de que en cada uno se concentra el 10% de los resultados.

**Percentiles:** Análogamente son 99 valores que dividen la muestra en 100 partes iguales. En cada una de ellas se encuentra el 1% de los datos.

**Ejemplo:** Vamos a calcular los cuartiles de la serie de datos referidos a la estatura del grupo de alumnos

Variable (Valor)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
x	X	x	x	x
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%

**1º cuartil:** es el valor 1,22 cm, ya que por debajo suya se sitúa el 25% de la frecuencia (tal como se puede ver en la columna de la frecuencia relativa acumulada).

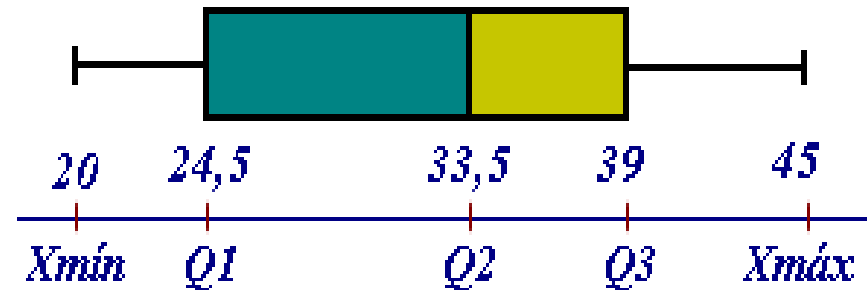
**2º cuartil:** es el valor 1,26 cm, ya que entre este valor y el 1º cuartil se sitúa otro 25% de la frecuencia. El segundo cuartil, como se ha dicho, es la mediana también.

**3º cuartil:** es el valor 1,28 cm, ya que entre este valor y el 2º cuartil se sitúa otro 25% de la frecuencia.

## Diagrama de Caja-Bigotes

Los diagramas de **Caja-Bigotes** (boxplots o box and whiskers) son una presentación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría.

Para su realización se representan **los tres cuartiles** y los valores **mínimo** y **máximo** de los datos, sobre un rectángulo, alineado horizontal o verticalmente. Para el conjunto de datos (edades) 36 25 37 24 39 20 36 45 31 31 39 24 29 23 41 40 33 24 34 40, se tendría



La parte izquierda de la caja es mayor que la de la derecha; ello quiere decir que las edades comprendidas entre el 25% y el 50% de la población está más dispersa que entre el 50% y el 75%.

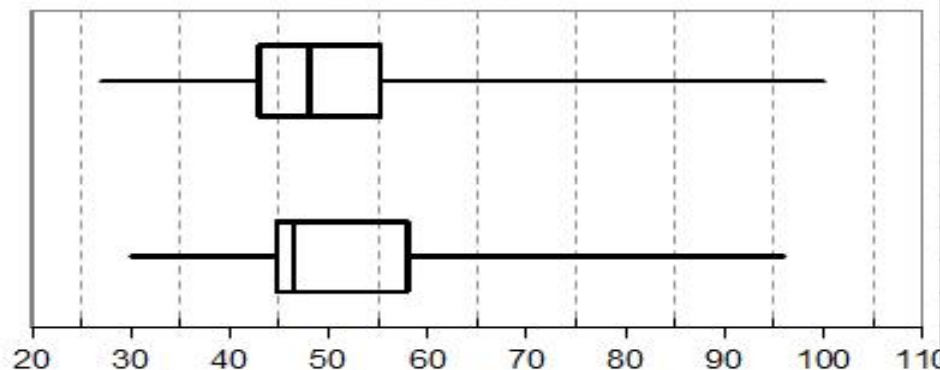
El bigote de la izquierda ( $X_{\min}$ ,  $Q_1$ ) es más corto que el de la derecha; por ello el 25% de los más jóvenes están más concentrados que el 25% de los mayores.

El **rango intercuartílico** es  $Q_3 - Q_1 = 14,5$ ; es decir, el 50% de la población está comprendido en 14,5 años.

## COMPARACIÓN DE DOS TEMPORADAS DE LA LIGA BBVA

Minimo	Q1	Mediana	Q3	Máximo
30,00	44,75	46,50	58,00	96,00
27,00	43,00	48,00	55,25	100,00

Gráfico de caja y bigote para dos variables



	Temporada 2010-11	Temporada 2011-12
1	96	100
2	92	91
3	71	61
4	62	58
5	58	56
6	58	55
7	58	54
8	49	52
9	47	50
10	47	49
11	46	47
12	46	47
13	45	47
14	45	46
15	45	43
16	44	43
17	44	42
18	43	41
19	35	37
20	30	27



## Medidas de Dispersión

Estudian la variabilidad de los datos de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.

Existen diversas **Medidas de Dispersión**. Entre las más utilizadas podemos destacar las siguientes:

**1.- Rango:** mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el menor de ellos.

**2.- Varianza:** Mide la distancia existente entre los valores de la serie y la media. Se calcula como la suma de las diferencias cuadráticas entre cada valor y la media, multiplicadas por el número de veces que se ha repetido dicho valor. Dicha suma se divide por el tamaño de la muestra.

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

La varianza siempre será mayor que cero. Cuanto más próxima a cero esté, más concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

**3.- Desviación típica:** Se calcula como raíz cuadrada de la varianza. Así se consigue tener las mismas unidades que la media.

**4.- Coeficiente de variación de Pearson:** se calcula como el cociente entre la desviación típica y la media. El interés del coeficiente de variación es que al ser adimensional permite comparar el nivel de dispersión de dos muestras. Esto no ocurre con la desviación típica, ya que viene expresada en las mismas unidades que los datos de la serie.

Por ejemplo, para comparar el nivel de dispersión de una serie de datos de la altura de los alumnos de una clase y otra serie con el peso de dichos alumnos, no se puede utilizar las desviaciones típicas (una viene expresada en cm y la otra en kg). En cambio, sus coeficientes de variación son ambos adimensionales (incluso se pueden convertir en porcentajes), por lo que sí son comparables.

**Ejemplo:** vamos a obtener las medidas de dispersión para el conjunto de datos que venimos utilizando con la estatura de los alumnos de una clase

Variable (Valor)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%

**Rango:** Diferencia entre el mayor valor (1.30) y el menor (1.20). Por tanto, 0.10.

**Varianza:** recordemos que la media de esta muestra es 1.253. Luego, aplicando la fórmula:

$$S_x^2 = \frac{((1,20-1,253)^2 * 1) + ((1,21-1,253)^2 * 4) + ((1,22-1,253)^2 * 4) + \dots + ((1,30-1,253)^2 * 3)}{30}$$

Por lo tanto, la varianza es 0,0010.

**Desviación típica:** la raíz de 0.0010, es decir, 0.0316.

**Coeficiente de variación de Pearson:**  $0.0316/1.253=0.025$

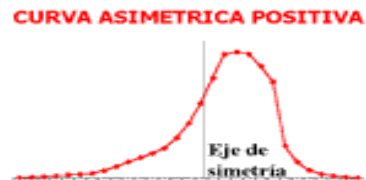
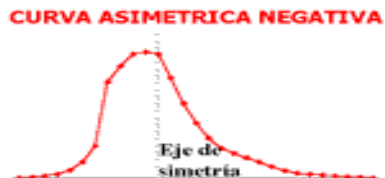
## Medidas de Forma.

Las **medidas de forma**, como su nombre indica, permiten conocer aspectos de la forma que tiene la curva que representa la serie de datos de la muestra. En concreto, se estudian habitualmente las siguientes características de la curva:

- a) **Asimetría:** mide si la curva tiene una forma simétrica, es decir, si respecto al centro de la misma (centro de simetría) los tramos de curva que quedan a derecha e izquierda son similares.
- b) **Curtosis:** mide si los valores de la distribución están más o menos concentrados alrededor de del valor medio de la muestra.

# Asimetría

Como se ha dicho, el concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a izquierda y derecha de un valor central (media aritmética).



Para medir el nivel de asimetría se utiliza el llamado **Coeficiente de Asimetría de Fisher**, definido como:

$$CA_F = \frac{\sum_{i=1}^N (x_i - \bar{x})^3}{N \cdot S_x^3}$$

siendo  $\bar{x}$  la media y  $S_x$  la desviación típica

En función del signo, el coeficiente se interpreta de la siguiente manera,

**g1 = 0** (distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media)

**g1 > 0** (distribución asimétrica positiva; existe mayor concentración de valores a la derecha de la media que a su izquierda)

**g1 < 0** (distribución asimétrica negativa; existe mayor concentración de valores a la izquierda de la media que a su derecha)

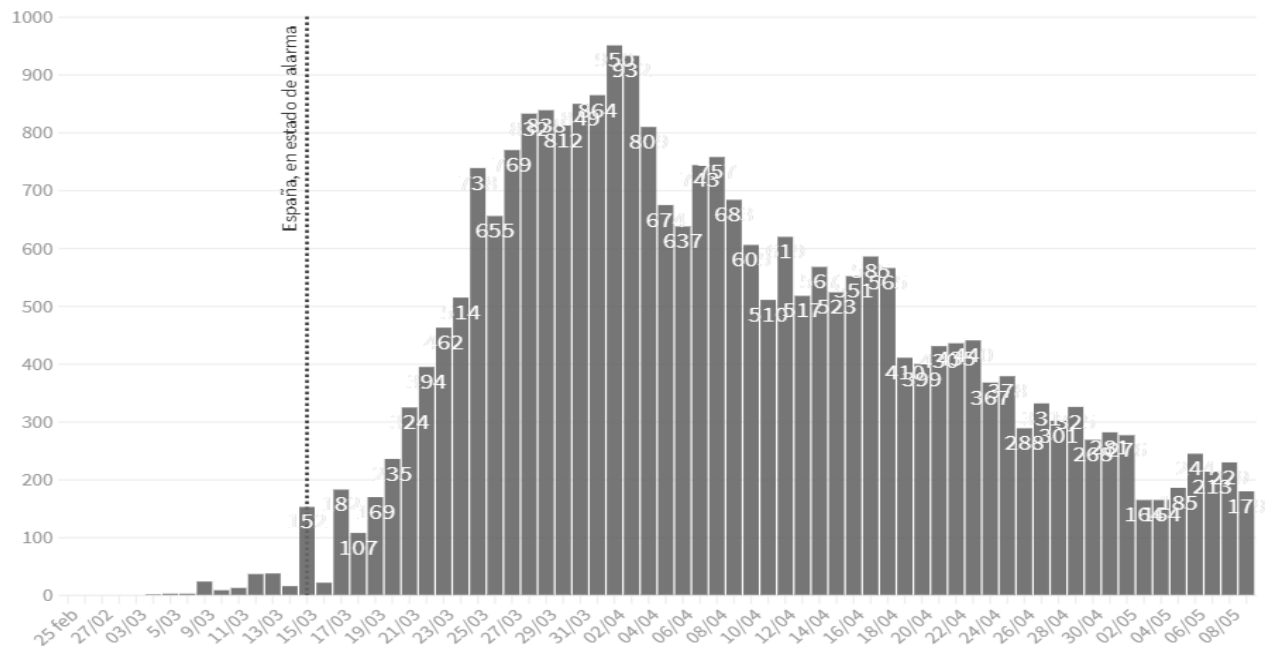
**Ejemplo:** Calculemos el Coeficiente de Asimetría de Fisher de la serie de datos referidos a la estatura de un grupo de alumnos.

El coeficiente resulta -0.119, de donde se deduce que la muestra presenta una distribución asimétrica negativa (más valores a la izquierda de la media que a su derecha).

Variable (Valor)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%



## Fallecidos diarios con coronavirus en España



Fuente: Elaboración propia, [Ministerio de Sanidad](#) • Sanidad realizó un ajuste a partir del 17 de abril para validar los datos recopilados por las autoridades de Salud Pública de Cataluña. Esta variación afectará a la serie global hasta que el Ministerio ofrezca las nuevas cifras corregidas.

## Curtosis

**El coeficiente de Curtosis** analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución.

Se definen 3 tipos de distribuciones según su grado de curtosis:

**Distribución mesocúrtica:** presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).

**Distribución leptocúrtica:** presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

**Distribución platicúrtica:** presenta un reducido grado de concentración alrededor de los valores centrales de la variable.



El coeficiente de Curtosis,  $g_2$  se define de la siguiente forma:

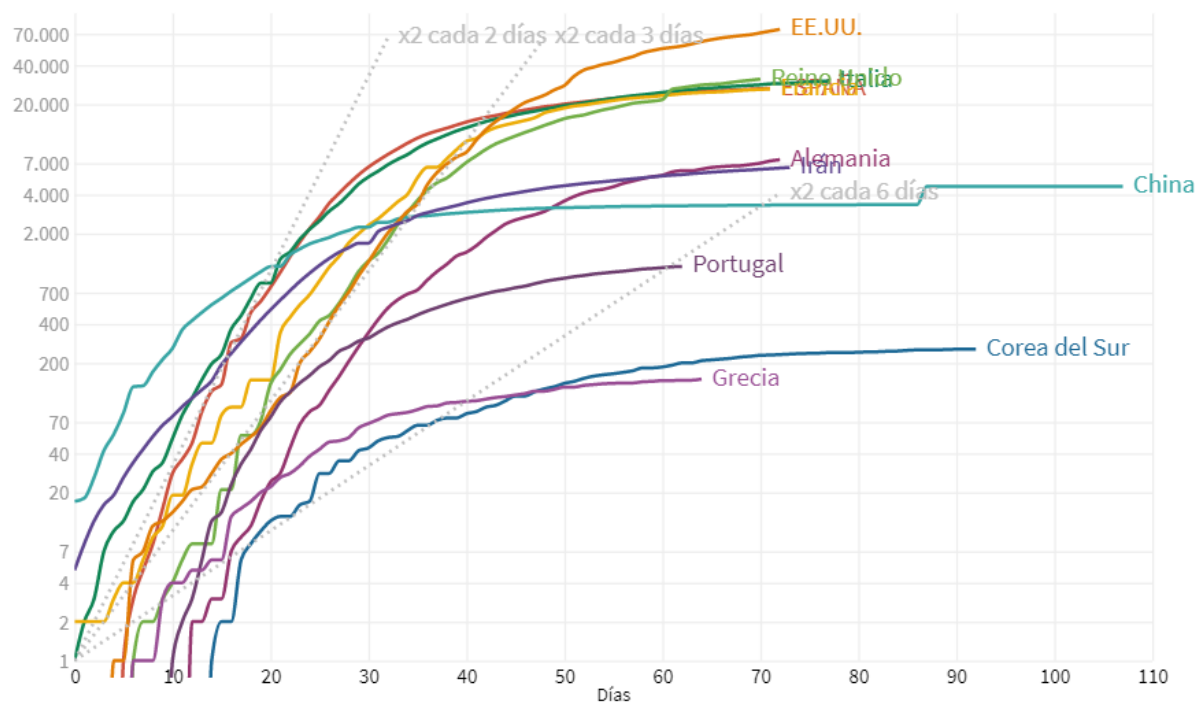
$$\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n s^2} - 3$$

Si coincide con 0 la distribución es mesocúrtica, si es menor que cero será platicúrtica y si es mayor que cero, leptocúrtica.

**Ejemplo:** Vamos a calcular el Coeficiente de Curtosis de la serie de datos referidos a la estatura de un grupo de alumnos. Aplicando la fórmula se obtiene -1.428 lo que revela una distribución platicúrtica.

Variable (Valor)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
x	x	x	x	x
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%

## Muertes por coronavirus en España y otros países



FUENTE: Elaboración propia, [Ministerio de Sanidad](#), [Ministero della Salute](#), [JHU CSSE](#) • La fecha de inicio del brote corresponde al primer día con 20 casos registrados en cada país.

## Distribuciones bidimensionales

Las distribuciones bidimensionales son aquellas en las que se estudian al mismo tiempo dos variables de cada elemento de la población, por ejemplo, peso y altura de un grupo de estudiantes; superficie y precio de las viviendas de una ciudad; potencia y velocidad de una gama de coches deportivos.

Para representar los datos obtenidos se utiliza una **tabla de correspondencias o tabla de contingencia**:

X / Y	y <sub>1</sub>	y <sub>2</sub>	.....	y <sub>m-1</sub>	y <sub>m</sub>
x <sub>1</sub>	n <sub>1,1</sub>	n <sub>1,2</sub>		n <sub>1,m-1</sub>	n <sub>1,m</sub>
x <sub>2</sub>	n <sub>2,1</sub>	n <sub>2,2</sub>		n <sub>2,m-1</sub>	n <sub>2,m</sub>
.....					
x <sub>n-1</sub>	n <sub>n-1,1</sub>	n <sub>n-1,2</sub>		n <sub>n-1,m-1</sub>	n <sub>n-1,m</sub>
x <sub>n</sub>	n <sub>n,1</sub>	n <sub>n,2</sub>		n <sub>n,m-1</sub>	n <sub>n,m</sub>

**Ejemplo:** Medimos el peso y la estatura de los alumnos de una clase y obtenemos los siguientes resultados:

Alumno	Estatura	Peso	Alumno	Estatura	Peso	Alumno	Estatura	Peso
Alumno 1	1,25	32	Alumno 11	1,25	31	Alumno 21	1,25	33
Alumno 2	1,28	33	Alumno 12	1,28	35	Alumno 22	1,28	32
Alumno 3	1,27	31	Alumno 13	1,27	34	Alumno 23	1,27	34
Alumno 4	1,21	34	Alumno 14	1,21	33	Alumno 24	1,21	34
Alumno 5	1,22	32	Alumno 15	1,22	33	Alumno 25	1,22	35
Alumno 6	1,29	31	Alumno 16	1,29	31	Alumno 26	1,29	31
Alumno 7	1,30	34	Alumno 17	1,30	35	Alumno 27	1,30	34
Alumno 8	1,24	32	Alumno 18	1,24	32	Alumno 28	1,24	33
Alumno 9	1,27	32	Alumno 19	1,27	31	Alumno 29	1,27	35
Alumno 10	1,29	35	Alumno 20	1,29	33	Alumno 30	1,29	34

Esta información se puede representar de un modo más organizado en la siguiente tabla de correspondencias:

Estatura / Peso	31 kg	32 kg	33 kg	34 kg	35 kg
1,21 cm	0	0	1	2	0
1,22 cm	0	1	1	0	1
1,23 cm	0	0	0	0	0
1,24 cm	0	2	1	0	0
1,25 cm	1	1	1	0	0
1,26 cm	0	0	0	0	0
1,27 cm	2	1	0	2	1
1,28 cm	0	1	1	0	1
1,29 cm	3	0	1	1	1
1,30 cm	0	0	0	2	1



## Distribuciones marginales

Al analizar una distribución bidimensional, uno puede centrar su estudio en el comportamiento de una de las variables, con independencia de como se comporta la otra. Estaríamos así en el análisis de una **distribución marginal**.

De cada distribución bidimensional se pueden deducir **dos distribuciones marginales**: una correspondiente a la variable  $X$ , y otra correspondiente a la variable  $Y$ .

## Marginal de X

X	$n_{i.}$
$x_1$	$n_{1.}$
$x_2$	$n_{2.}$
.....	...
$x_{n-1}$	$n_{n-1.}$
$x_n$	$n_{n.}$

## Marginal de Y

Y	n.j
y <sub>1</sub>	n.1
y <sub>2</sub>	n.2
.....	...
y <sub>m-1</sub>	n.m-1
y <sub>m</sub>	n.m
x <sub>n-1</sub>	n <sub>n-1.</sub>
x <sub>n</sub>	n <sub>n.</sub>

Estatura / Peso	31 kg	32 kg	33 kg	34 kg	35 kg
1,21 cm	0	0	1	2	0
1,22 cm	0	1	1	0	1
1,23 cm	0	0	0	0	0
1,24 cm	0	2	1	0	0
1,25 cm	1	1	1	0	0
1,26 cm	0	0	0	0	0
1,27 cm	2	1	0	2	1
1,28 cm	0	1	1	0	1
1,29 cm	3	0	1	1	1
1,30 cm	0	0	0	2	1

## Distribución marginal de la variable X (estatura)

Obtenemos la siguiente tabla de frecuencia:

Variable (Estatura)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
1,21	3	3	10,0%	10,0%
1,22	3	6	10,0%	20,0%
1,23	0	6	0,0%	20,0%
1,24	3	9	10,0%	30,0%
1,25	3	12	10,0%	40,0%
1,26	0	12	0,0%	40,0%
1,27	6	18	20,0%	60,0%
1,28	3	21	10,0%	70,0%
1,29	6	27	20,0%	90,0%
1,30	3	30	10,0%	100,0%

## Distribución marginal de la variable Y (peso)

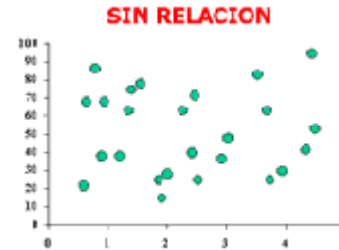
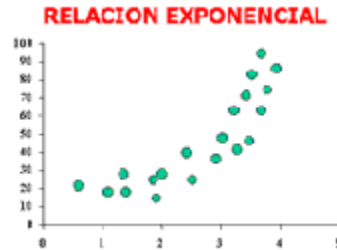
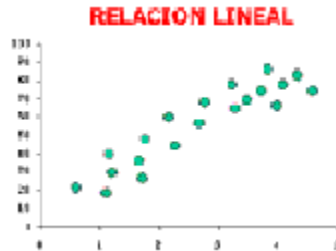
Obtenemos la siguiente tabla de frecuencia:

X

Variable (Peso)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple	Acumulada
31	6	6	20,0%	20,0%
32	6	12	20,0%	40,0%
33	6	18	20,0%	60,0%
34	7	25	23,3%	83,3%
35	5	30	16,6%	100,0%

## Coeficiente de Correlación lineal

**El coeficiente de correlación lineal** mide el grado de intensidad en la dependencia lineal entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables la nube de puntos tiene forma longitudinal).



No obstante, puede que exista una relación que no sea lineal, sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación las variables, por lo que convendría utilizar otro tipo de coeficiente más apropiado.

Para analizar, por tanto, si se puede utilizar el coeficiente de correlación lineal, lo mejor es representar los pares de valores en un gráfico y ver qué forma tienen.

El **coeficiente de correlación lineal**, **r**, se calcula aplicando la siguiente fórmula:

$$\frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$



**Numerador:** se denomina **covarianza** y se calcula de la siguiente manera: en cada par de valores (x,y) se multiplica la "x" menos su media, por la "y" menos su media. Se suma el resultado obtenido de todos los pares de valores y este resultado se divide por el tamaño de la muestra.

**Denominador** se calcula el producto de las desviaciones típicas de X y de Y.

El **coeficiente de correlación** se mueve entre -1 y 1.

**Si  $r > 0$** , la correlación lineal es positiva (si sube el valor de una variable sube el de la otra). La correlación es tanto más fuerte cuanto más se aproxime a 1.

Por ejemplo: ingresos y consumo.

**Si  $r < 0$** , la correlación lineal es negativa (si sube el valor de una variable disminuye el de la otra). La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.

Por ejemplo: tipos de interés y consumo.

**Si  $r = 0$** , no existe correlación lineal entre las variables. Aunque podría existir otro tipo de correlación (parabólica, exponencial, etc.)

**Ejemplo:** El coeficiente de correlación de nuestra serie de datos de altura y peso de los alumnos de una clase vale 0.79.

Alumno	Estatura	Peso	Alumno	Estatura	Peso	Alumno	Estatura	Peso
Alumno 1	1,25	32	Alumno 11	1,25	33	Alumno 21	1,25	33
Alumno 2	1,28	33	Alumno 12	1,28	35	Alumno 22	1,28	34
Alumno 3	1,27	34	Alumno 13	1,27	34	Alumno 23	1,27	34
Alumno 4	1,21	30	Alumno 14	1,21	30	Alumno 24	1,21	31
Alumno 5	1,22	32	Alumno 15	1,22	33	Alumno 25	1,22	32
Alumno 6	1,29	35	Alumno 16	1,29	34	Alumno 26	1,29	34
Alumno 7	1,30	34	Alumno 17	1,30	35	Alumno 27	1,30	34
Alumno 8	1,24	32	Alumno 18	1,24	32	Alumno 28	1,24	31
Alumno 9	1,27	32	Alumno 19	1,27	33	Alumno 29	1,27	35
Alumno 10	1,29	35	Alumno 20	1,29	33	Alumno 30	1,29	34



## Cálculo de Probabilidades

Podemos definir la probabilidad como una medida de la incertidumbre asociada a los resultados aleatorios. En el Cálculo de Probabilidades se construyen modelos probabilísticos para medir esta incertidumbre, partiendo de la premisa de que lo real es siempre más complejo y multiforme que cualquier modelo que se pueda construir. Estos modelos teóricos a los que hacemos referencia se reducen en muchos casos (o incluyen en su formulación) a funciones de probabilidad. La teoría de la probabilidad tiene su origen en el estudio de los juegos de azar, que impulsaron los primeros estudios sobre cálculo de probabilidades en el siglo XVI, aunque no es hasta el siglo XVIII cuando se aborda la probabilidad desde una perspectiva matemática con la demostración del teorema de Bernoulli según el cual, al aumentar el número de ensayos, la frecuencia de un suceso tiende a aproximarse a un número fijo denominado su probabilidad. Es la justificación matemática de la Ley Empírica del azar. La definición de probabilidad basada en este hecho es conocida como frecuentista.

En el siglo XX el matemático ruso Andrei Nikolaevich Kolmogorov (1903-1987) formula la teoría axiomática de la probabilidad cerrando el debate existente entre los diferentes enfoques y definiciones preexistentes. Para Kolmogorov la probabilidad es una función que asigna a cada posible resultado de un experimento aleatorio un valor no negativo, de forma que se cumpla la propiedad aditiva y tal que el conjunto de todos los resultados posibles tenga probabilidad unitaria. La definición axiomática establece las reglas que deben cumplir las probabilidades, aunque no asigna valores concretos.



## Variable aleatoria

Uno de los conceptos más importantes de la teoría de probabilidades es el de variable aleatoria que, intuitivamente, puede definirse como cualquier característica medible que toma diferentes valores con probabilidades determinadas. Toda variable aleatoria posee una distribución de probabilidad que describe su comportamiento.

La variable es discreta si toma una cantidad a lo sumo numerable de valores. Y es continua cuando puede tomar cualquier valor en un intervalo.

Una forma usual de describir la distribución de probabilidad de una variable aleatoria es mediante la denominada función de probabilidad en el caso de variables discretas y de la función de densidad en el caso de variables continuas. Además la función de distribución recoge las probabilidades acumuladas


Los científicos han desarrollado modelos de distribuciones de probabilidad para representar el comportamiento teórico de diferentes fenómenos aleatorios que aparecen en el mundo real.



## La distribución Normal

La distribución normal es, sin duda, la distribución más importante del Cálculo de probabilidades y de la Estadística. Fue descubierta, como aproximación de la distribución binomial, por Abraham De Moivre (1667-1754) y publicada en 1733 en su libro *The Doctrine of Chances*; estos resultados fueron ampliados por Pierre-Simon Laplace (1749-1827), quién también realizó aportaciones importantes. En 1809, Carl Friedrich Gauss (1777-1855) publicó un libro sobre el movimiento de los cuerpos celestes donde asumía errores normales. Por este motivo esta distribución también es conocida como distribución Gaussiana.

La importancia de la distribución normal queda totalmente enfatizada por ser la distribución límite de numerosas variables aleatorias, discretas y continuas, como se demuestra a través del teorema central del límite. Las consecuencias de estos teoremas implican la casi universal presencia de la distribución normal en todos los campos de las ciencias empíricas: biología, medicina, psicología, física, economía, etc. Además, de ella derivan, entre otras, tres distribuciones (ji-cuadrado, t de Student y F de Snedecor), de importancia clave en el campo de la contrastación de hipótesis estadísticas.

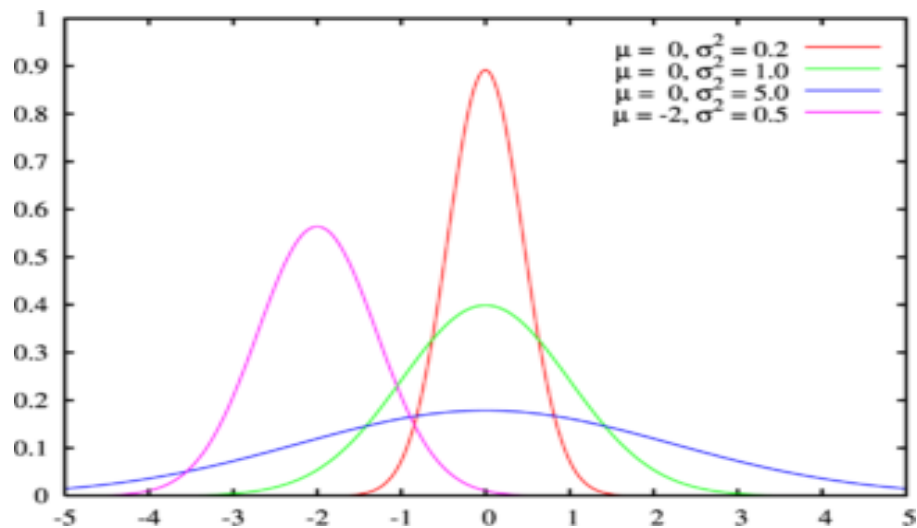


La distribución normal queda totalmente definida mediante dos parámetros: la media,  $\mu$ , y la desviación estándar o desviación típica,  $\sigma$ . Su función de densidad es simétrica respecto a la media y la desviación estándar nos indica el mayor o menor peso de las colas de la curva que se conoce como campana de Gauss. Esta distribución se denota por  $N(\mu, \sigma)$ . Cuando la distribución normal tiene media 0 y desviación típica 1 recibe el nombre de distribución normal tipificada o estandarizada. Restando a una variable Normal su media y dividiéndola por la varianza se obtiene una normal tipificada.

La función de densidad es:

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right], \quad x \in R$$

En la normal, la media es mediana y moda. El gráfico compara diferentes curvas normales.






## Distribución Gamma

Se denota por  $\text{Gamma}(a,p)$  siendo  $a$  el parámetro de escala y  $p$  el de forma. La distribución gamma aparece asociada al estudio de tiempos de vida.

Se puede caracterizar del modo siguiente: si se está interesado en la ocurrencia de un evento generado por un proceso de Poisson de media  $\lambda$ , la variable que mide el tiempo transcurrido hasta obtener  $n$  ocurrencias del evento sigue una distribución gamma con parámetros  $a = n\lambda$  (escala) y  $p = n$  (forma).

Cuando  $p$  es un número entero positivo se tiene un caso particular de la distribución gamma que se denomina distribución de Erlang. Otros casos particulares de la distribución gamma, son la distribución exponencial ( $\text{Gamma}(\lambda, 1)$ ) y la distribución ji-cuadrado ( $\text{Gamma}(1/2, n/2)$ ).

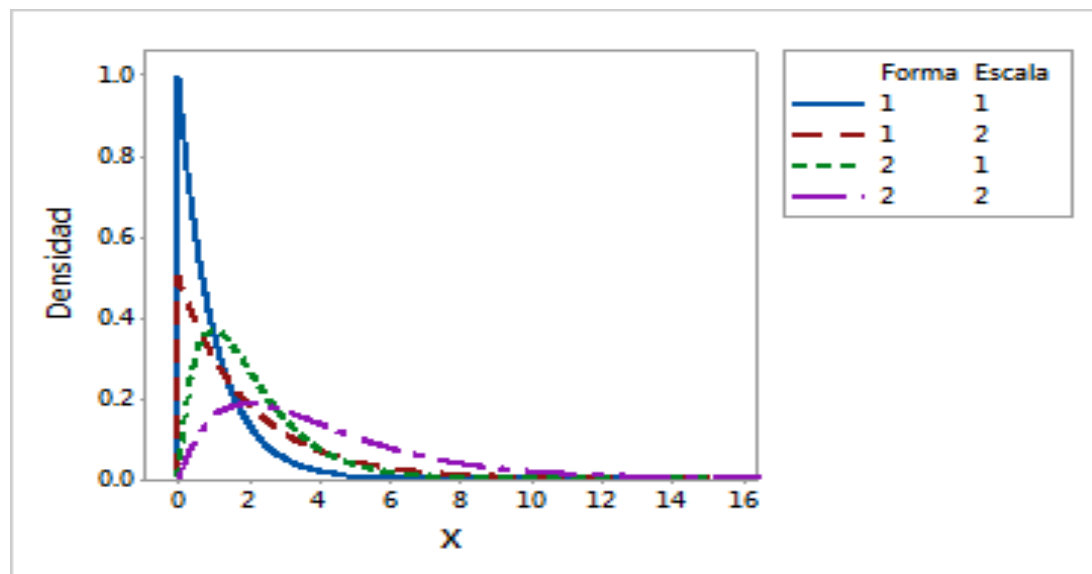




La función de densidad viene dada por

$$f(x) = \frac{a^p}{\Gamma(p)} x^{p-1} \exp[-ax], \quad x > 0$$
$$EX = p/a$$
$$VarX = p/a^2$$

La distribución Gamma es reproductiva respecto del segundo parámetro, respecto de la forma.





## Distribución t de Student y distribución F

Siempre que se divide una normal tipificada entre la raíz de una chi-cuadrado dividida, a su vez, por los grados de libertad, la distribución resultante es una t de Student, con tantos grados de libertad como la chi-cuadrado asociada.

El cociente entre dos distribuciones chi-cuadrado corregidas por sus grados de libertad es una distribución F con dos parámetros, llamados grados de libertad, que son los de la chi-cuadrado del numerador y denominador, respectivamente.

## Inferencia Estadística

En inferencia tratamos de obtener información sobre las poblaciones (variables aleatorias) a partir de los resultados muestrales. En la llamada Inferencia paramétrica la información se dirige a los parámetros poblacionales. En la inferencia no paramétrica puede tener un sentido más amplio, como la independencia de las observaciones, el tipo de distribución subyacente o la igualdad de distribuciones aunque no se conozcan éstas. El tipo de muestreo utilizado en inferencia es el llamado muestreo aleatorio simple en el que cualquier individuo de la población tiene la misma probabilidad que los demás de participar en la muestra.



## Técnicas de la inferencia

En inferencia paramétrica disponemos de tres técnicas para obtener información sobre los parámetros desconocidos:

**Estimación puntual:** consiste en obtener un pronóstico individualizado del parámetro o parámetros desconocidos. Lógicamente la teoría se dirige a desarrollar procedimientos que garanticen buenas propiedades de la estimación.

**Intervalos de confianza:** se busca un rango de valores en el que exista una “probabilidad” alta prefijada por el investigador de encontrar el verdadero valor del parámetro.


**Contrastes o tests de hipótesis:** son reglas de decisión para elegir entre diferentes valores del parámetro desconocido.



## Intervalos de confianza

La estimación puntual de un parámetro desconocido en una distribución, independientemente del criterio de optimalidad elegido, nos deja con la incertidumbre de saber realmente, ante una realización muestral concreta, la discrepancia entre nuestro pronóstico y el verdadero valor del parámetro.

El concepto de intervalo de confianza viene a reducir esta incertidumbre pues nos proporcionará un rango de valores entre los que se encuentra el parámetro desconocido con una confianza alta, controlada por el investigador.



**Definición:** Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_n$  de una población  $X$  con función de probabilidad o función de densidad  $f(x, \theta)$ , llamaremos intervalo de confianza para estimar  $\theta$ , de nivel de confianza  $1 - \alpha$ , a un par de estadísticos  $(U, V)$ , estimadores por defecto y por exceso de  $\theta$ , respectivamente, tales que 
$$p(U < \theta < V) = 1 - \alpha.$$

Algunas precisiones importantes sobre esta definición son las siguientes:

- i) el nivel de confianza lo fija el investigador, siendo valores usuales 0.95, 0.90 y 0.99. Por defecto se consideran intervalos de nivel 0.95.
- ii) la interpretación es frecuentista.
- iii) mayor nivel de confianza lleva asociada mayor longitud del intervalo y, por tanto, menor precisión en la estimación.



## **Métodos de obtención de intervalos de confianza**

Método de la función pivote

Método de Neyman





## Método de la función pivote

Supongamos que existe una función de la muestra y del parámetro desconocido pero cuya distribución no depende del parámetro desconocido. Es la llamada función pivote. Eligiendo los límites de probabilidad en dicha distribución y despejando, a continuación el parámetro, se obtienen los límites de confianza buscados.

Es especialmente apropiado para los intervalos de confianza de parámetros en poblaciones normales.




## Intervalo de confianza para la media de una distribución Normal con varianza conocida.

Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_n$  de una población  $X$  con distribución  $N(\mu, \sigma)$ , con  $\mu$  desconocido y  $\sigma$  conocido, para calcular el intervalo de confianza para  $\mu$  de nivel  $1 - \alpha$  la función pivote es

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

que sigue la distribución  $N(0,1)$ . Entonces existe  $z_{\alpha/2}$  tal que

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$



y despejando se obtiene

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

con lo que el intervalo resultante es

$$IC_{\alpha}(\mu) = \left[ \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Como se puede apreciar, la longitud del intervalo crece con la variabilidad  $\sigma$  y con el nivel de confianza, mientras que se reduce si se incrementa el tamaño muestral.

**Ejemplo:** De una población normal  $N(\mu, \sigma = 2.5)$  se extrae una m.a.s. de tamaño 30, siendo  $\sum_{i=1}^n x_i = 77$ . Obtener un intervalo de confianza para la media de la población al nivel de confianza 0.95.

La media de la realización muestral es  $\frac{\sum_{i=1}^n x_i}{30} = \frac{77}{30} = 2.57$ .

Teniendo en cuenta que la población de la que procede la muestra es normal con varianza conocida y que el intervalo pedido es de nivel

$1 - \alpha = 0.95$ , es decir,  $\alpha = 0.05$ ,

$$P\left(\bar{X} - z_{0.05/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{0.05/2} \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\begin{aligned}
 IC_{0.95}(\mu) &= \left[ \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = \\
 &= \left[ 2.57 - 1.96 \frac{2.5}{\sqrt{30}}, 2.57 + 1.96 \frac{2.5}{\sqrt{30}} \right] = [1.68; 3.46]
 \end{aligned}$$

**Ejemplo:** Consideremos el precio (en euros) de alquiler de las casas en Madrid. Supongamos que sabemos que sigue una distribución normal con media  $\mu$  desconocida y varianza conocida,  $\sigma^2=1600$  euros<sup>2</sup>. Para estimar el alquiler medio en Madrid se toma una muestra de 100 viviendas de alquiler y se registra el precio que pagan sus inquilinos.

Llamemos  $X$  a esta variable aleatoria que mide el precio del alquiler. Sabemos que  $X \rightarrow N(\mu, \sigma = 40)$ . Supongamos que la realización muestral de los precios de alquiler de 100 viviendas,  $x_1, \dots, x_{100}$ , arroja un valor medio de  $\bar{x} = 710$  euros. Partiendo de estos datos, queremos encontrar un intervalo de confianza al 95% para  $\mu$ .

Puesto que  $X \rightarrow N(\mu, \sigma = 40)$  se tiene que

$$\bar{X} \rightarrow N\left(\mu, \sqrt{\frac{1600}{100}}\right) = N(\mu, 4)$$

y por tanto,

$$\frac{\bar{X} - \mu}{4} \rightarrow N(0,1)$$

En consecuencia,

$$P\left(-1.96 < \frac{\bar{X} - \mu}{4} < 1.96\right) = 0.95$$

y despejando  $\mu$  en esta expresión obtenemos

$$P\left(\bar{X} - 1.96 \times 4 < \mu < \bar{X} + 1.96 \times 4\right) = 0.95$$

Con lo que sustituyendo el valor de la media muestral se obtienen los límites de confianza  $(702.16, 717.84)$ .

## Intervalo de confianza para la media de una distribución Normal con varianza desconocida.


Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_n$  de una población  $X$  con distribución  $N(\mu, \sigma)$ , con  $\mu$  y  $\sigma$  desconocidos, para calcular el intervalo de confianza para  $\mu$  de nivel  $1 - \alpha$  la función pivote es

$$\frac{\bar{X} - \mu}{S / \sqrt{n-1}}$$

que sigue la distribución  $t_{n-1}$ . Entonces existe  $t_{n-1, \alpha/2}$  tal que

$$P \left( -t_{n-1, \alpha/2} \leq \frac{\bar{X} - \mu}{S / \sqrt{n-1}} \leq t_{n-1, \alpha/2} \right) = 1 - \alpha$$






y despejando se obtiene

$$P\left(\bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n-1}}\right) = 1 - \alpha$$

con lo que el intervalo resultante es

$$IC_{\alpha}(\mu) = \left[ \bar{X} - t_{n-1, \alpha/2} \frac{S}{\sqrt{n-1}}, \bar{X} + t_{n-1, \alpha/2} \frac{S}{\sqrt{n-1}} \right]$$

Como se puede apreciar, la longitud del intervalo crece con la variabilidad y con el nivel de confianza, mientras que se reduce si se incrementa el tamaño muestral.



**Ejemplo:** La Oficina de Control de Calidad de una gran superficie quiere controlar el peso medio de las latas de sardinas de un cierto fabricante. Toma una muestra de 10 latas cuyo peso medio resulta ser 998 gr y cuya desviación típica es de 5 gr. Asumiendo que el peso de las latas está normalmente distribuido obtengamos los límites de confianza al nivel del 95% para el peso medio de las latas.

Las observaciones proceden de una población normal con media y varianza Desconocidas. Dado que el nivel de confianza es  $1 - \alpha = 0.95$ , el valor de la t de Student para  $\alpha/2 = 0.025$  y 9 grados de libertad es  $t_{9,0.025} = 2.262$ .



Por tanto

$$\begin{aligned} IC_{\alpha}(\mu) &= \left[ \bar{x} - t_{\alpha/2, n-1} \frac{s_x}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{s_x}{\sqrt{n}} \right] = \\ &= \left[ 998 - 2.262 \frac{5}{\sqrt{9}}, 998 + 2.262 \frac{5}{\sqrt{9}} \right] = [994.23; 1001.76] \end{aligned}$$

## Intervalo de confianza para la varianza de una distribución Normal con media desconocida.

Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_n$  de una población  $X$  con distribución  $N(\mu, \sigma)$ , con  $\mu$  y  $\sigma$  desconocidos, para calcular el intervalo de confianza para  $\sigma$  de nivel  $1 - \alpha$  la función pivote es

$$\frac{n S^2}{\sigma^2}$$

que sigue la distribución  $\chi^2_{n-1}$ . Entonces, existen  $\chi^2_{n-1, \alpha/2}$  y  $\chi^2_{n-1, 1-\alpha/2}$  tales que

$$P\left(\chi^2_{n-1, 1-\alpha/2} \leq \frac{n S^2}{\sigma^2} \leq \chi^2_{n-1, \alpha/2}\right) = 1 - \alpha$$




y despejando se obtiene

$$P\left(\frac{nS^2}{\chi_{n-1, \alpha/2}^2} \leq \sigma^2 \leq \frac{nS^2}{\chi_{n-1, 1-\alpha/2}^2}\right) = 1 - \alpha$$

con lo que el intervalo resultante es

$$IC_{\alpha}(\sigma^2) = \left[ \frac{nS^2}{\chi_{n-1, \alpha/2}^2}, \frac{nS^2}{\chi_{n-1, 1-\alpha/2}^2} \right]$$

Como se puede apreciar, la longitud del intervalo crece con el nivel de confianza.



Ejemplo: Consideremos de nuevo la situación del control de calidad del peso de las latas de sardinas. Y obtengamos un intervalo de confianza para la varianza de la distribución, de nivel 0.95 a partir de las mismas observaciones, es decir, 10 latas cuyo peso medio resulta ser 998 gr y cuya desviación típica es de 5 gr. En este caso  $\chi^2_{9,0,025} = 19.023$  y  $\chi^2_{9,0,975} = 2.700$  con lo que el intervalo resultante será

$$IC_{\alpha}(\sigma^2) = \left[ \frac{ns^2}{\chi^2_{9,0,025}}, \frac{ns^2}{\chi^2_{9,0,975}} \right] = (13.14, 92.593)$$

## Intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas conocidas.

Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_{n_1}$  de una población  $X$  con distribución  $N(\mu_1, \sigma_1)$ , y otra muestra independiente de la anterior  $Y_1, Y_2, \dots, Y_{n_2}$  de una población  $Y \sim N(\mu_2, \sigma_2)$  con  $\mu_1, \mu_2$  desconocidos y  $\sigma_1, \sigma_2$  conocidos, para calcular el intervalo de confianza para  $\mu_1 - \mu_2$  de nivel  $1 - \alpha$  la función pivote es

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}}$$

que sigue la distribución  $N(0,1)$ . Entonces existe  $z_{\alpha/2}$  tal que

$$P \left( -z_{\alpha/2} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}} \leq z_{\alpha/2} \right) = 1 - \alpha$$




Y despejando se obtiene

$$P\left(\bar{X}-\bar{Y}-z_{\alpha/2}\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)}\leq\mu_1-\mu_2\leq\bar{X}-\bar{Y}+z_{\alpha/2}\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)}\right)=1-\alpha$$

Con lo que

$$IC_{\alpha}(\mu_1-\mu_2)=\left(\bar{X}-\bar{Y}-z_{\alpha/2}\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)},\bar{X}-\bar{Y}+z_{\alpha/2}\sqrt{(\sigma_1^2/n_1)+(\sigma_2^2/n_2)}\right)$$





**Ejemplo:** Supongamos que las duraciones de los neumáticos para coche de dos marcas Premium están normalmente distribuidas con varianzas respectivas de  $4 \cdot 10^6$  y  $9 \cdot 10^6 \text{ kms}^2$ . Elegidos al azar 9 vehículos calzados con neumáticos de la primera marca se obtuvo una duración media de 35000 Kms, mientras que 14 vehículos (independientes de los anteriores) y también elegidos al azar, calzados con neumáticos de la segunda marca registraron una duración media de 37000 kms. Se desea obtener un intervalo de confianza para la diferencia de las duraciones (teóricas) de los neumáticos de ambas marcas, al 95%. Aplicando directamente la fórmula obtenida, tenemos,

$$\begin{aligned} IC_{\alpha}(\mu_1 - \mu_2) &= \left( \bar{x} - \bar{y} - z_{0,025} \sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)}, \bar{x} - \bar{y} + z_{0,025} \sqrt{(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)} \right) = \\ &= \left( 35000 - 37000 - 1.96 \sqrt{(4 \cdot 10^6 / 9) + (9 \cdot 10^6 / 14)}, 35000 - 37000 + 1.96 \sqrt{(4 \cdot 10^6 / 9) + (9 \cdot 10^6 / 14)} \right) = \\ &= (-4043.765, 43.765). \end{aligned}$$

## Intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas desconocidas pero iguales.

Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_{n_1}$  de una población  $X$  con distribución  $N(\mu_1, \sigma^2)$ , y otra muestra independiente de la anterior  $Y_1, Y_2, \dots, Y_{n_2}$  de una población  $Y \sim N(\mu_2, \sigma^2)$  con  $\mu_1, \mu_2$  y  $\sigma^2$  desconocidos, para calcular el intervalo de confianza para  $\mu_1 - \mu_2$  de nivel  $1 - \alpha$  la función pivote es

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(n_1 S_X^2 + n_2 S_Y^2)[(1/n_1) + (1/n_2)]}} \sqrt{n_1 + n_2 - 2}$$

que sigue la distribución  $t_{n_1+n_2-2}$ . Entonces existe  $t_{n_1+n_2-2, \alpha/2}$  tal que

$$P \left( -t_{n_1+n_2-2, \alpha/2} \leq \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(n_1 S_X^2 + n_2 S_Y^2)[(1/n_1) + (1/n_2)]}} \sqrt{n_1 + n_2 - 2} \leq t_{n_1+n_2-2, \alpha/2} \right) = 1 - \alpha$$




Y despejando


$$P\left(\bar{X}-\bar{Y}-t_{n_1+n_2-2,\alpha/2}\frac{\sqrt{(n_1S_X^2+n_2S_Y^2)[(1/n_1)+(1/n_2)]}}{\sqrt{n_1+n_2-2}}\leq\mu_1-\mu_2\leq\bar{X}-\bar{Y}+t_{n_1+n_2-2,\alpha/2}\frac{\sqrt{(n_1S_X^2+n_2S_Y^2)[(1/n_1)+(1/n_2)]}}{\sqrt{n_1+n_2-2}}\right)=1-\alpha$$

Por tanto,

$$IC_\alpha(\mu_1-\mu_2)=\left(\bar{X}-\bar{Y}-t_{n_1+n_2-2,\alpha/2}\frac{\sqrt{(n_1S_X^2+n_2S_Y^2)[(1/n_1)+(1/n_2)]}}{\sqrt{n_1+n_2-2}},\bar{X}-\bar{Y}+t_{n_1+n_2-2,\alpha/2}\frac{\sqrt{(n_1S_X^2+n_2S_Y^2)[(1/n_1)+(1/n_2)]}}{\sqrt{n_1+n_2-2}}\right)$$



**Ejemplo:** Retomando el ejemplo anterior, supongamos ahora que las duraciones de los neumáticos para coche de dos marcas Premium están normalmente distribuidas pero que desconocemos las varianzas aunque podemos suponer que son iguales. Elegidos al azar 9 vehículos calzados con neumáticos de la primera marca se obtuvo una duración media de 35000 Kms, mientras que 14 vehículos (independientes de los anteriores) y también elegidos al azar, calzados con neumáticos de la segunda marca registraron una duración media de 37000 kms. Si las varianzas respectivas de ambas muestras son  $4 \cdot 10^6$  y  $9 \cdot 10^6 \text{ kms}^2$  se desea obtener un intervalo de confianza para la diferencia de las duraciones (teóricas) de los neumáticos de ambas marcas, al 95%. Aplicando directamente la fórmula obtenida, tenemos,



$$IC_{\alpha}(\mu_1 - \mu_2) = \left( \bar{x} - \bar{y} - t_{n_1+n_2-2, \alpha/2} \frac{\sqrt{(n_1 s_X^2 + n_2 s_Y^2)[(1/n_1) + (1/n_2)]}}{\sqrt{n_1 + n_2 - 2}}, \bar{x} - \bar{y} + t_{n_1+n_2-2, \alpha/2} \frac{\sqrt{(n_1 s_X^2 + n_2 s_Y^2)[(1/n_1) + (1/n_2)]}}{\sqrt{n_1 + n_2 - 2}} \right) =$$

$$= \left( 35000 - 37000 - t_{21, 0.025} \frac{\sqrt{(9 \times 4 \times 10^6 + 14 \times 9 \times 10^6)[(1/9) + (1/14)]}}{\sqrt{21}}, 35000 - 37000 + t_{21, 0.025} \frac{\sqrt{(9 \times 4 \times 10^6 + 14 \times 9 \times 10^6)[(1/9) + (1/14)]}}{\sqrt{21}} \right)$$

$$IC_{\alpha}(\mu_1 - \mu_2) = (-4468.25, 468.25)$$

En este caso  $t_{21, 0.025} = 2.08$  y el intervalo resultante es


## Intervalo de confianza aproximado (con muestra grande) para una proporción.

Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_n$  de una población  $X$  con distribución  $B(1, p)$ , y  $p$  desconocido, para calcular el intervalo de confianza para  $p$  de nivel  $1 - \alpha$  la función pivote es

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

que sigue, en muestras grandes, aproximadamente la distribución  $N(0,1)$ . En este caso  $\bar{X}$  representa la frecuencia relativa de éxitos. Entonces existe  $z_{\alpha/2}$  tal que

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$




En este caso  $\bar{X}$  representa la frecuencia relativa de éxitos.  
Entonces existe  $z_{\alpha/2}$  tal que

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

De donde se obtiene, después de estimar  $p$  en los límites de confianza, el intervalo aproximado

$$IC_{\alpha}(p) = \left( \bar{X} - z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right)$$



**Ejemplo:** En un control de calidad de una empresa que fabrica tornillos se quiere determinar un intervalo de confianza para la proporción de tornillos defectuosos que se está produciendo en un momento dado, con nivel de confianza del 95%. Examinados al azar 100 tornillos se detectó que 3 de ellos no eran útiles para su uso. Entonces, aplicando, la fórmula anterior se tiene

$$\begin{aligned} IC_{\alpha}(p) &= \left( \bar{x} - z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right) = \\ &= \left( 0.03 - 1.96 \sqrt{\frac{0.03 \times 0.97}{100}}, 0.03 + 1.96 \sqrt{\frac{0.03 \times 0.97}{100}} \right) = \\ &= (-0.0034, 0.0064) \end{aligned}$$



## Intervalo de confianza aproximado (con muestra grandes) para la diferencia de dos proporciones.

Dada una muestra aleatoria simple  $X_1, X_2, \dots, X_{n_1}$  de una población  $X$  con distribución  $B(1, p_1)$ , y otra muestra aleatoria simple, independiente de la anterior,  $Y_1, Y_2, \dots, Y_{n_2}$  de una población  $Y$  con distribución  $B(1, p_2)$ ,  $p_1$  y  $p_2$  desconocidos, para calcular el intervalo de confianza para  $p_1 - p_2$  de nivel  $1 - \alpha$  la función pivote es

$$\frac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$


que sigue, en muestras grandes, aproximadamente la distribución  $N(0,1)$ .  
Entonces existe  $z_{\alpha/2}$  tal que

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \bar{Y} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha$$



Y, por tanto, el intervalo de confianza resultante es

$$IC_{\alpha}(p_1 - p_2) = \left( \bar{X} - \bar{Y} - z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n_1} + \frac{\bar{Y}(1-\bar{Y})}{n_2}}, \bar{X} - \bar{Y} + z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n_1} + \frac{\bar{Y}(1-\bar{Y})}{n_2}} \right)$$



**Ejemplo:** Se desea analizar la diferencia entre la proporción de europeístas en dos países de la UE. Elegidos al azar 1000 ciudadanos en el primero de ellos se declararon europeístas 923, mientras que en una muestra aleatoria de 1200 ciudadanos del segundo país se encontraron 1023 europeístas. De acuerdo con estos datos, un intervalo de confianza al 95% para la diferencia entre las proporciones teóricas de europeístas en ambos países viene dado por

$$\begin{aligned} IC_{\alpha}(p_1 - p_2) &= \left( \bar{x} - \bar{y} - z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n_1} + \frac{\bar{y}(1-\bar{y})}{n_2}}, \bar{x} - \bar{y} + z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n_1} + \frac{\bar{y}(1-\bar{y})}{n_2}} \right) = \\ &= \left( 0.923 - 0.853 - 1.96 \sqrt{\frac{0.923 \times 0.077}{1000} + \frac{0.853 \times 0.147}{1200}}, 0.923 - 0.853 + 1.96 \sqrt{\frac{0.923 \times 0.077}{1000} + \frac{0.853 \times 0.147}{1200}} \right) = \\ &= (0.014, 0.126) \end{aligned}$$



## CONTRASTE DE HIPOTESIS

Un contraste o test de hipótesis es una regla de decisión para elegir entre dos hipótesis, una llamada nula, y notada  $H_0$ , y otra llamada alternativa,  $H_1$ .

La regla se basa en elegir un conjunto de realizaciones muestrales para las que rechazaremos la hipótesis nula, por no ser verosímiles bajo ésta, es decir, no ser verosímiles si la hipótesis nula es cierta.

Las realizaciones muestrales para las que se rechaza la hipótesis nula constituyen la llamada región crítica. Obtener la regla de decisión equivale a determinar la región crítica.

Al tomar la decisión se pueden cometer dos tipos de error, el llamado error de tipo I, que se produce si rechazamos  $H_0$  y, sin embargo, es cierta; y el error de tipo II que se comete al mantener  $H_0$  cuando en realidad es falsa.

Es imposible controlar simultáneamente ambos tipos de error. Cuando disminuye la probabilidad de cometer uno de ellos se incrementa la probabilidad del otro.

Habitualmente se fija una probabilidad máxima de error tipo I y luego se elige la región, entre todas las que cumplen esa condición, para la que el error de tipo II es mínimo.

# Tests clásicos para una muestra

## Contraste unilateral para la media de una población normal con varianza conocida

Supongamos que  $X_1, X_2, \dots, X_n$  es una muestra aleatoria simple de una población normal con media  $\mu$ , desconocida, y varianza  $\sigma^2$  conocida.

Queremos realizar el siguiente contraste unilateral:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Bajo la hipótesis nula (es decir, si la hipótesis nula es cierta), se sabe que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$$

Por tanto, fijado el nivel de significación  $\alpha$  (máxima probabilidad de cometer el error tipo I), rechazaremos la hipótesis nula, si el valor del estadístico anterior para la realización muestral obtenida es significativamente grande, es decir,

$$\text{Si } z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha \quad \text{se rechaza } H_0.$$

$$\text{Equivalentemente, si } p_{valor} = P\left(Z > z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \quad \text{es tal que}$$

$$p_{valor} < \alpha \quad \text{entonces, se rechaza } H_0 .$$

En el caso de los precios de los alquileres donde la media de 100 observaciones era 710 y la varianza (observada) era 1600, si se deseara contrastar con tales datos que el precio medio de todos los alquileres (no de los de la muestra) es de 600 euros frente a que es superior a 600, tendríamos que

$$z=(710-600)/4=27.5$$

que es un valor muy significativo. Para un 5% de nivel de significación el valor crítico en la normal es 1.645. Como 27.5 es mucho mayor que 1.645 rechazamos la hipótesis de que los alquileres en promedio cuestan 600 euros.

Aquí claramente el p-valor,  $p(Z>27.5)=0$ , lo que nos lleva a la misma decisión.

## Contraste bilateral para la media de una población normal con varianza conocida

Supongamos que  $X_1, X_2, \dots, X_n$  es una muestra aleatoria simple de una población normal de media  $\mu$ , desconocida, y varianza  $\sigma^2$  conocida.

Queremos realizar el siguiente contraste bilateral:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Bajo la hipótesis nula (es decir, si la hipótesis nula es cierta), se sabe que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0,1)$$



Entonces, una vez fijado el nivel de significación, se rechaza la hipótesis nula (en favor de la alternativa) si el valor del estadístico es anormalmente grande o pequeño, es decir:

$$\text{si } z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{\alpha/2} \quad \text{ó si } z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < -z_{\alpha/2} \quad \text{se rechaza } H_0$$

$$\text{Equivalentemente, si } p_{valor} = P\left(|Z| > \left| z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| \right) \quad \text{es tal que}$$

$$p_{valor} < \alpha \quad \text{entonces, se rechaza } H_0 .$$

## Contraste unilateral para la media de una población normal con varianza desconocida

Supongamos que  $X_1, X_2, \dots, X_n$  es una muestra aleatoria simple de una población normal de media  $\mu$ , desconocida, y varianza  $\sigma^2$  desconocida. Queremos realizar el siguiente contraste unilateral

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Bajo la hipótesis nula (es decir, si la hipótesis nula es cierta), se sabe que

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n-1}}} \rightarrow t_{n-1}$$

Por tanto, fijado el nivel de significación  $\alpha$ , rechazaremos la hipótesis nula, si el valor de tal estadístico para la realización muestral obtenida es significativamente grande, es decir:

$$\text{si} \quad t = \frac{\frac{\bar{x} - \mu_0}{s}}{\sqrt{n-1}} > t_{n-1, \alpha} \quad \text{se rechaza } H_0.$$

$$\text{Equivalentemente, si} \quad p_{valor} = P \left( t_{n-1} > t = \frac{\frac{\bar{x} - \mu_0}{s}}{\sqrt{n-1}} \right) \text{ es tal que}$$

$$p_{valor} < \alpha \quad \text{entonces, se rechaza } H_0.$$

## Contraste bilateral para la media de una población normal con varianza desconocida

Supongamos que  $X_1, X_2, \dots, X_n$  es una muestra aleatoria simple de una población normal con media  $\mu$  (conocida) y varianza  $\sigma^2$  desconocida.

Queremos realizar el siguiente contraste bilateral:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Bajo la hipótesis nula (es decir, si la hipótesis nula es cierta), se sabe que

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n-1}}} \rightarrow t_{n-1}$$

Entonces, fijado el nivel de significación, se rechaza la hipótesis nula si el valor del estadístico es anormalmente grande o pequeño, es decir:

$$\text{Si } t = \frac{\frac{\bar{x} - \mu_0}{S}}{\sqrt{n-1}} > t_{n-1, \alpha/2} \quad \text{o si } t = \frac{\frac{\bar{x} - \mu_0}{S}}{\sqrt{n-1}} < -t_{n-1, \alpha/2}$$

se rechaza  $H_0$ .

$$\text{Equivalentemente, si } p_{valor} = P\left(|t_{n-1}| > \left| \frac{\frac{\bar{x} - \mu_0}{S}}{\sqrt{n-1}} \right| \right) \text{ es tal que}$$

$p_{valor} < \alpha$  entonces, se rechaza  $H_0$

## Contraste unilateral para la varianza de una población normal

Supongamos que  $X_1, X_2, \dots, X_n$  es una muestra aleatoria simple de una población normal de media  $\mu$  y varianza  $\sigma^2$  desconocidas. Queremos realizar el siguiente contraste unilateral

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 < \sigma_0^2$$

Bajo la hipótesis nula, sabemos que

$$\chi^2 = \frac{nS^2}{\sigma_0^2} \rightarrow \chi_{n-1}^2$$

Por tanto, fijado el nivel de significación, rechazaremos la hipótesis nula si para la realización muestral el valor del estadístico es significativamente pequeño. Por tanto,

$$\text{si } \chi^2 = \frac{ns^2}{\sigma_0^2} < \chi_{n-1, 1-\alpha}^2 \quad \text{se rechaza la hipótesis nula}$$

$$\text{Equivalentemente, si } p_{valor} = P\left(\chi_{n-1}^2 < \frac{ns^2}{\sigma_0^2}\right) \quad \text{es tal que}$$

$$p_{valor} < \alpha \quad \text{entonces, se rechaza } H_0$$

## Contraste bilateral para la varianza de una población normal

Supongamos que  $X_1, X_2, \dots, X_n$  es una muestra aleatoria simple m.a.s. de una población normal de media  $\mu$  y varianza  $\sigma^2$  desconocidas. Queremos realizar el siguiente contraste bilateral

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_1 : \sigma^2 \neq \sigma_0^2$$

Bajo la hipótesis nula sabemos que

$$\chi^2 = \frac{nS^2}{\sigma_0^2} \rightarrow \chi_{n-1}^2$$



Entonces, una vez fijado el nivel de significación, se rechaza la hipótesis nula si el valor del estadístico es grande o pequeño, es decir:

$$\text{Si } \chi^2 = \frac{ns^2}{\sigma_0^2} > \chi_{n-1, \alpha/2}^2 \quad \text{ó si} \quad \chi^2 = \frac{ns^2}{\sigma_0^2} < \chi_{n-1, 1-\alpha/2}^2 \quad \text{se rechaza } H_0$$

# Tests clásicos para dos muestras

## Comparación de medias para muestras

### independientes con varianzas iguales

Se desea comparar las medias de dos variables aleatorias normalmente distribuidas y con varianzas iguales. Para ello disponemos de dos muestras aleatorias independientes  $X_1, \dots, X_m$  e  $Y_1, \dots, Y_n$ . Las hipótesis son

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Bajo la hipótesis nula, se sabe que

$$\frac{\bar{X} - \bar{Y}}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} \rightarrow t_{m+n-2} \quad \text{donde} \quad S_c = \frac{mS_x^2 + nS_y^2}{m+n-2}$$

Entonces, una vez fijado el nivel de significación, se rechaza la hipótesis nula si el valor del estadístico es grande o pequeño, es decir,

$$\text{Si } \frac{|\bar{X} - \bar{Y}|}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} > t_{m+n-2, \alpha/2} \text{ se rechaza } H_0$$

$$\text{Equivalentemente, si } p_{valor} = P \left( |t_{m+n-2}| > \frac{|\bar{X} - \bar{Y}|}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} \right) \text{ es tal que}$$

$$p_{valor} < \alpha \text{ entonces, se rechaza } H_0$$

En el caso en el cual el tamaño de las muestras son grandes, tenemos que:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} \rightarrow N(0,1)$$

Entonces, si  $\frac{|\bar{X} - \bar{Y}|}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} > z_{\alpha/2}$  se rechaza  $H_0$

Equivalentemente, si  $p_{valor} = P \left( |Z| > \frac{|\bar{X} - \bar{Y}|}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} \right)$  es tal que  $p_{valor} < \alpha$

entonces, se rechaza  $H_0$

## Comparación de medias para muestras dependientes apareadas

Queremos comparar las medias de dos variables aleatorias normalmente distribuídas. Disponemos de dos muestras aleatorias  $X_1, \dots, X_m$  e  $Y_1, \dots, Y_m$  que además están relacionadas por ser los individuos comunes,

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Definimos  $D = X - Y \rightarrow N\left(\mu_1 - \mu_2, \sqrt{\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}}\right)$  y las diferencias  $(X_1 - Y_1, \dots, X_m - Y_m)$  constituyen una m.a.s. de la población  $X - Y$

La media muestral  $\bar{D} = \bar{X} - \bar{Y} \rightarrow N\left(\mu_1 - \mu_2, \sqrt{(\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy})/n}\right)$

Por tanto:  $\sqrt{n} \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy})}} \rightarrow N(0,1)$

Si alguno de los elementos de la matriz de covarianzas no fuese conocido el estadístico anterior no es calculable y utilizaríamos el de Student:

$$\sqrt{n-1} \frac{\bar{D} - (\mu_1 - \mu_2)}{S_D} \rightarrow t_{n-1}$$

Entonces, una vez fijado el nivel de significación, se rechaza la hipótesis nula si el valor del estadístico es grande o pequeño, es decir:

$$\text{si } \left| \sqrt{n-1} \frac{\bar{D}}{s_D} \right| > t_{n-1, \alpha/2} \quad \text{se rechaza } H_0$$

## Comparación de varianzas para muestras independientes

Se desea comparar las varianzas de dos poblaciones normales y disponemos de dos muestras aleatorias independientes  $X_1, \dots, X_n$  e  $Y_1, \dots, Y_m$ :

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Bajo la hipótesis nula (es decir, si la hipótesis nula es cierta), se sabe que

$$\frac{\frac{\hat{S}_x^2}{\sigma_1^2}}{\frac{\hat{S}_y^2}{\sigma_2^2}} = \frac{\hat{S}_x^2}{\hat{S}_y^2} \rightarrow F_{n-1, m-1}$$



Hemos de buscar en las tablas dos valores  $F_{c1}$  y  $F_{c2}$  tales que:

$$P(F_{n-1,m-1} < F_{c1}) = \alpha / 2 \quad \text{y} \quad P(F_{n-1,m-1} > F_{c2}) = \alpha / 2$$

por tanto la región de aceptación será

$$F_{c1} \leq \frac{\hat{S}_x^2}{\hat{S}_y^2} \leq F_{c2}$$

# INFERENCIA NO PARAMETRICA

- Introducción
- Contraste de bondad de ajuste  $\chi^2$  de Pearson
- Contraste de bondad de ajuste de Kolmogorov – Smirnov
- Contraste de independencia
- Contraste de homogeneidad

# Introducción

Se introducen a continuación algunos de los contrastes no paramétricos más usados en Estadística. Cuando se habla de estadística paramétrica, se supone un conocimiento de la distribución de la población. En estos casos, lo único que falta por determinar son los parámetros asociados a esta distribución. Por otra parte, cuando se habla de estadística no paramétrica, que debería considerarse como algo previo a la estadística paramétrica, no siempre se tiene conocimiento acerca de la distribución de la variable aleatoria en estudio.

En el caso de la estadística no paramétrica, los objetivos son completamente distintas. Por ejemplo, uno habitual y que supone un paso previo a la estadística paramétrica es el de determinar si la población en la que se está interesado se distribuye según alguna distribución conocida. Los contrastes para dar respuesta a esta pregunta se concocen como Contrastes de Bondad de Ajuste. (Pearson y Kolmogorov-Smirnov).

## **Contraste de bondad de ajuste $\chi^2$ de Pearson**

Se aplica tanto a variables aleatorias continuas como discretas y permite decidir si una variable aleatoria se distribuye según una cierta distribución.

Se trata, por tanto, de contrastar si una muestra procede de una distribución F o no. Las hipótesis son

$$\begin{cases} H_0 : X \equiv F \\ H_1 : X \neq F \end{cases}$$

## Caso discreto

Consideremos una variable aleatoria  $X$  discreta que toma los valores  $a_1 \dots a_k$  con probabilidades  $p_1 \dots p_k$ . Sea  $X_1, \dots, X_n$  una muestra aleatoria simple de  $X$  de tamaño  $n$ .

Sean  $O_1 \dots O_k$  las frecuencias observadas respectivamente de cada valor  $a_1 \dots a_k$ , es decir,  $O_i$  es el número de veces que aparece  $a_i$  en la muestra.

Sean  $e_i = np_i$  las frecuencias esperadas bajo la hipótesis nula.

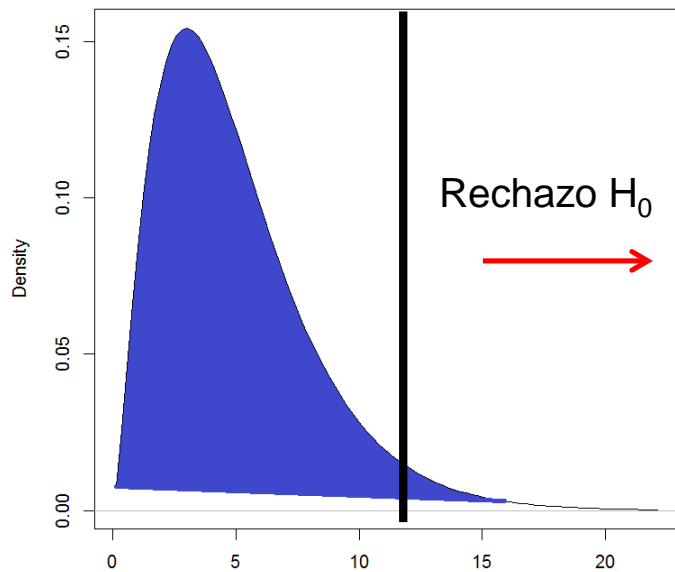
El estadístico del contraste es:

$$\chi^2 = \sum_{i=1}^K \frac{(o_i - e_i)^2}{e_i} \rightarrow \chi_{k-1}^2$$

Bajo  $H_0$ , el estadístico  $\chi^2$  sigue una distribución chi-cuadrado (o ji-cuadrado) con  $k-1$  grados de libertad, si la hipótesis nula está completamente especificada.

En otro caso,  $\chi^2$  se distribuye según una variable ji-cuadrado de  $k-r-1$  grados de libertad, si es necesario estimar  $r$  parámetros para la especificación de la distribución de  $X$  en  $H_0$ . Por ejemplo, si es una normal y no conocemos media ni varianza habrá que estimar dos parámetros, y  $r=2$ .

Decisión: Rechazar  $H_0$  si:  $\chi^2 > \chi^2_{k-1, \alpha}$





**Ejemplo:** Fernando Quevedo, Medwave 2011 Dic, 11(12):e5266doi:10,5867/medwave.2011.12.5266

Consideremos el siguiente ejemplo en el que tratamos de determinar si una muestra de grupos sanguíneos correspondiente a un conjunto de 150 donantes se ajusta a la distribución en la población de los diferentes grupos sanguíneos. Se admite que la distribución en la población (y ésta es la hipótesis nula) viene dada por las siguientes probabilidades (en porcentaje):

Grupo	Frecuencia esperada
AB	2,0%
A	30,5%
B	9,3%
O	58,2%

En la muestra se han observado las siguientes frecuencias:

Grupo	Frecuencia observada
AB	4
A	48
B	15
O	83

Las frecuencias esperadas bajo la hipótesis nula aparecen en la siguiente tabla:

Grupo	Frec. oi	Frec. ei
AB	4	3,00
A	48	45,75
B	15	13,95
0	83	87,30
Total	150	150,00

Y el estadístico de contraste resulta ser:

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(4-3)^2}{3} + \frac{(48-45,75)^2}{45,75} + \frac{(15-13,95)^2}{13,95} + \frac{(83-87,3)^2}{87,3} = 0,73$$

Ahora debemos comparar el valor obtenido (0,73) con el correspondiente valor crítico obtenido de la tabla de la distribución chi-cuadrado para 3 grados de libertad (la variable tipos sanguíneos tiene 4 modalidades), y el nivel de significación especificado.

**TABLA 3-Distribución Chi Cuadrado  $\chi^2$**

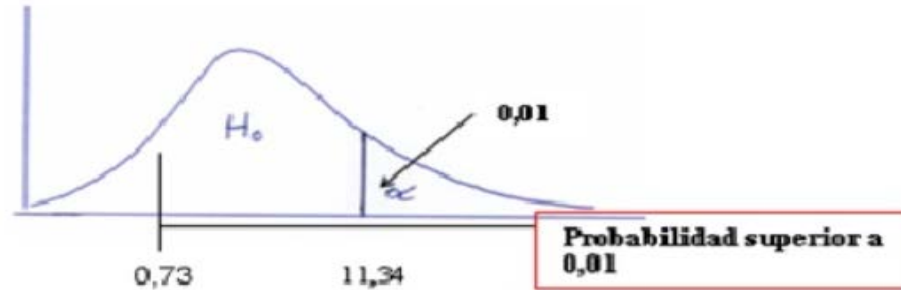
P = Probabilidad de encontrar un valor mayor o igual que el chi cuadrado tabulado, v = Grados de Libertad

v/p	0,001	0,0025	0,005	0,01	0,025	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
1	10,8274	9,1404	7,8794	6,6349	5,0239	3,8415	2,7055	2,0722	1,6424	1,3233	1,0742	0,8735	0,7083	0,5707	0,4549
2	13,8150	11,9827	10,5965	9,2104	7,3778	5,9915	4,6052	3,7942	3,2189	2,7726	2,4079	2,0996	1,8326	1,5970	1,3863
3	16,2660	14,3202	12,8381	11,3449	9,3484	7,8147	6,2514	5,3170	4,6416	4,1083	3,6649	3,2831	2,9462	2,6430	2,3660
4	18,4662	16,4238	14,8602	13,2767	11,1433	9,4877	7,7794	6,7449	5,9886	5,3853	4,8784	4,4377	4,0446	3,6871	3,3567
5	20,5147	18,3854	16,7496	15,0863	12,8325	11,0705	9,2363	8,1152	7,2893	6,6257	6,0644	5,5731	5,1319	4,7278	4,3515
6	22,4575	20,2491	18,5475	16,8119	14,4494	12,5916	10,6446	9,4461	8,5581	7,8408	7,2311	6,6948	6,2108	5,7652	5,3481
7	24,3213	22,0402	20,2777	18,4753	16,0128	14,0671	12,0170	10,7479	9,8032	9,0371	8,3834	7,8061	7,2832	6,8000	6,3458
8	26,1239	23,7742	21,9549	20,0902	17,5345	15,5073	13,3616	12,0271	11,0301	10,2189	9,5245	8,9094	8,3505	7,8325	7,3441
9	27,8767	25,4625	23,5893	21,6660	19,0228	16,9190	14,6837	13,2880	12,2421	11,3887	10,6564	10,0060	9,4136	8,8632	8,3428
10	29,5870	27,1110	25,1881	23,2093	20,4832	18,3070	16,0927	14,5330	13,4470	12,5480	11,7807	11,0071	10,4732	9,9022	9,3418

Podemos observar que en la intersección de la fila de los tres grados de libertad y la columna del nivel de significación 0.01 (si utilizamos éste, por ejemplo), el valor que aparece es 11.3449.

Como nuestro valor del estadístico es mucho menor que 11.3449, podemos concluir que los datos no muestran evidencia en contra de la hipótesis nula, o dicho de otra manera, podemos admitir que la muestra de donantes se ajusta a la distribución poblacional de los grupos sanguíneos.

Gráficamente:



## Caso continuo

Para este caso se divide el rango de la variable aleatoria  $X$  en  $k$  clases ( $k \geq 4$ ).

Sean  $O_i$  = frecuencia absoluta de los datos observados en la clase  $i$ ,  $i \in \{1, \dots, k\}$  (al menos 3 en cada clase).

Sean  $e_i = np_i$  las frecuencias esperadas, donde  $p_i$  es la probabilidad de que la variable aleatoria  $F$  tome valores en la clase  $i$  (es decir,  $p\{\text{clase } i\}$ ).

Una vez calculadas las frecuencias observadas y las frecuencias esperadas para cada clase, se repite el procedimiento descrito en el caso discreto.

**Observaciones:**

- a. El test presenta mayor ajuste si  $n \geq 30$ ,  $k \geq 4$ ,  $O_i \geq 3$  ;  $e_i \geq 5$  para cualquier  $i$ .
- b. Conviene calcular por separado los sumandos de  $\chi^2$ , de modo que podamos observar datos atípicos y excluirllos.

Ejemplo: <https://groomch14.wixsite.com/literature-blog-1/single-post/2016/04/06/Pruebas-de-la-bondad-de-ajuste> (recuperado el 16 abril de 2020)

Supongamos que se dispone de los datos de duración, en años, de una muestra de 40 baterías de móvil. La duración de las baterías se ha agrupado en 7 clases, obteniéndose las siguientes frecuencias:

I	Clase (duración)	Frecuencia observada ( $o_i$ )
1	1.45-1.95	2
2	1.95-2.45	1
3	2.45-2.95	4
4	2.95-3.45	15
5	3.45-3.95	10
6	3.95-4.45	5
7	4.45-4.95	3

Se desea contrastar si tal duración se ajusta a una distribución normal con media 3.5 años y desviación típica 0.7 años. Esta es, entonces, la hipótesis nula.

Calculando la probabilidad correspondiente a cada intervalo, se tiene,

$$p_1 = P(X \leq 1.95) = P(Z \leq (1.95 - 3.5)/0.7) = 0.0136$$

$$p_2 = P(1.95 \leq X \leq 2.45) = P((1.95 - 3.5)/0.7 \leq Z \leq (2.45 - 3.5)/0.7) = 0.0532$$

$$p_3 = P(2.45 \leq X \leq 2.95) = P((2.45 - 3.5)/0.7 \leq Z \leq (2.95 - 3.5)/0.7) = 0.135$$

... (etc)

y las frecuencias esperadas serán,

$$e_1 = p_1 n = 0.0136 (40) \approx 0.5$$

$$e_2 = p_2 n = 0.0532 (40) \approx 2.1$$

$$e_3 = p_3 n = 0.135 (40) \approx 5.4$$

... (etc)



En la siguiente tabla tenemos todas las frecuencias observadas y esperadas,

Duración de año	Frecuencia observada ( $o_i$ )	Frecuencia Esperada
1.45-1.95	2	0.5
1.95-2.45	1	2.1
2.45-2.95	4	5.4
2.95-3.45	15	10.3
3.45-3.95	10	10.7
3.95-4.45	5	7
4.45-4.95	3	3.5

Como las frecuencias esperadas en varias clases son menores que 5, unimos clases adyacentes, obteniendo 4 clases

Duración (años)	Frecuencia observada ( $o_i$ )	Frecuencia Esperada
1.45-1.95	7	8
2.95-3.45	15	10.3
3.45-3.95	10	10.7
3.95-4.45	8	10.5

- Entonces el estadístico chi-cuadrado vale:

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \left[ \frac{(7 - 8)^2}{8} + \frac{(15 - 10.3)^2}{10.3} + \frac{(10 - 10.7)^2}{10.7} + \frac{(8 - 10.5)^2}{10.5} \right] : 2.92$$

- En este caso para un nivel de significación de 0.05 se tendría

$$\alpha = 0.05, v = k - 1 = 3, \Rightarrow \chi^2_{0.05} = 7.815$$

Rechazar  $H_0$  si  $\chi^2 > 7.815$

---

- Y como el valor observado del estadístico, 2.92 no supera el valor crítico obtenido a partir de la distribución chi-cuadrado, concluiríamos que los datos no muestran evidencia en contra de la hipótesis nula de normalidad.

# Kolmogorov-Smirnov

## Contraste de bondad de ajuste Kolmogorov-Smirnov

Este contraste es usado principalmente en variables aleatorias continuas. Se trata de determinar si la distribución de un conjunto de datos proviene de una distribución  $F_0$  o no:

$$\begin{cases} H_0 : F \equiv F_0 \\ H_1 : F \neq F_0 \end{cases}$$

Consideramos  $X$  v.a. y una muestra aleatoria simple  $X_1, \dots, X_n$ :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1 & \text{si } x_i \leq x \\ 0 & \text{en otro caso} \end{cases}$$

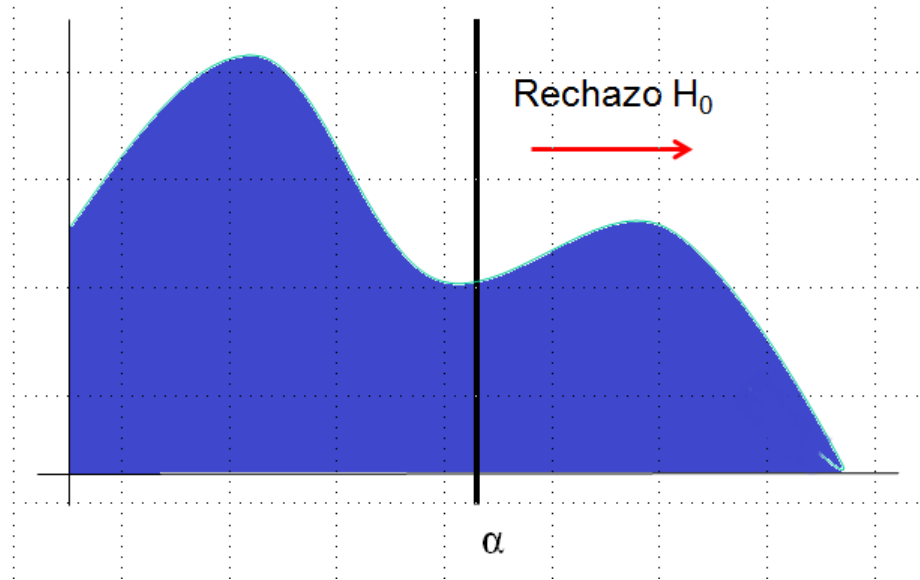
El estadístico de contraste se construye a partir de:

$$D_n^+ = \max(F_n(x_i) - F_0(x_i))$$

$$D_n^- = \max(F_0(x_{i-1}) - F_n(x_i))$$

$$D = \sup_{1 \leq i \leq n} |F_n(x) - F_0(x)|$$

Decisión: Rechazar  $H_0$  si:  $D > D_\alpha$



Siendo  $D_\alpha$  el valor crítico obtenido a partir de la distribución de Kolmogorov-Smirnov.

### Test de Kolmogorov-Smirnov sobre Bondad de Ajuste

<i>n</i>	<i>Nivel de significación <math>\alpha</math></i>						
	0.20	0.10	0.05	0.02	0.01	0.005	0.001
1	0.90000	0.95000	0.97500	0.99000	0.99500	0.99750	0.99900
2	0.68337	0.77639	0.84189	0.90000	0.92929	0.95000	0.96838
3	0.56481	0.63604	0.70760	0.78456	0.82900	0.86428	0.90000
4	0.49265	0.56522	0.62394	0.68887	0.73424	0.77639	0.82217
5	0.44698	0.50945	0.56328	0.62718	0.66853	0.70543	0.75000
6	0.41037	0.46799	0.51926	0.57741	0.61661	0.65287	0.69571
7	0.38148	0.43607	0.48342	0.53844	0.57581	0.60975	0.65071
8	0.35831	0.40962	0.45427	0.50654	0.54179	0.57429	0.61368
9	0.33910	0.38746	0.43001	0.47960	0.51332	0.54443	0.58210
10	0.32260	0.36866	0.40925	0.45562	0.48893	0.51872	0.55500
11	0.30829	0.35242	0.39122	0.43670	0.46770	0.49539	0.53135
12	0.29577	0.33815	0.37543	0.41918	0.44905	0.47672	0.51047
13	0.28470	0.32549	0.36143	0.40362	0.43247	0.45921	0.49189
14	0.27481	0.31417	0.34890	0.38970	0.41762	0.44352	0.47520
15	0.26589	0.30397	0.33750	0.37713	0.40420	0.42934	0.45611
16	0.25778	0.29472	0.32733	0.36571	0.39201	0.41644	0.44637
17	0.25039	0.28627	0.31796	0.35528	0.38086	0.40464	0.43380
18	0.24360	0.27851	0.30936	0.34569	0.37062	0.39380	0.42224
19	0.23735	0.27136	0.30143	0.33685	0.36117	0.38379	0.41156
20	0.23156	0.26473	0.29408	0.32866	0.35241	0.37451	0.40165
21	0.22517	0.25858	0.28724	0.32104	0.34426	0.36588	0.39243
22	0.22115	0.25283	0.28087	0.31394	0.33666	0.35782	0.38382
23	0.21646	0.24746	0.27491	0.30728	0.32954	0.35027	0.37575
24	0.21205	0.24242	0.26931	0.30104	0.32286	0.34318	0.36787

**Ejemplo.** Se desea analizar si las 10 observaciones siguientes proceden de una distribución exponencial de media 2: 0.406, 2.343, 0.538, 5.088, 5.587, 2.563, 0.023, 3.334, 3.491, 1.267.

- Ordenamos las observaciones de menor a mayor y obtenemos la función de distribución de la hipótesis nula en dichos valores y también la función de distribución empírica.

$i$	$Y(i)$	$F(y_i)$	$i/n$	$(i-1)/n$	$i/n - F(y_i)$	$F(y_i) - (i-1)/n$
1	0.023	0.0114	0.1	0.0	0.0886	0.0114
2	0.406	0.1838	0.2	0.1	0.0162	0.0838
3	0.538	0.2359	0.3	0.2	0.0641	0.0359
4	1.267	0.4693	0.4	0.3	-0.0693	0.1693
5	2.343	0.6901	0.5	0.4	-0.19801	0.2901
6	2.563	0.7224	0.6	0.5	-0.1224	0.2224
7	3.334	0.8112	0.7	0.6	-0.1112	0.2112
8	3.491	0.8254	0.8	0.7	-0.0254	0.1254
9	5.088	0.9214	0.9	0.8	-0.0214	0.1214
10	5.587	0.9388	0.10	0.9	0.0612	0.0388



- Existen diferentes correcciones del estadístico en función de la información que se tenga sobre la hipótesis nula

	Forma modificada de $D$	Area del extremo superior				
		0.15	0.1	0.05	0.025	0.01
$F(y)$ especificada	$(D)(\sqrt{n} + 0.12 + 0.11 / \sqrt{n})$	1.138	1.22	1.358	1.48	1.626
$F(y)$ normal						
$\mu$ y $\sigma_2$ desconocidas	$(D)(\sqrt{n} - 0.01 + 0.85 / \sqrt{n})$	0.775	0.819	0.895	0.955	1.035
$F(y)$ exponencial						
$\theta$ desconocida	$(D - 0.2/n)\sqrt{n + 0.26 + 0.5/\sqrt{n}}$	0.926	0.99	1.094	1.19	1.3

- En este ejemplo

$$(D)\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) = (0.2901)(3.317) = 0.9623$$

- Y, por tanto, como es menor que el valor crítico proporcionado por la distribución K-S no existe evidencia en contra de que los datos sigan una distribución exponencial.
- Como siempre los paquetes estadísticos nos proporcionan el p-valor, es decir la probabilidad de que bajo la distribución K-S tome un valor más extremo que el observado en la muestra. Si dicha probabilidad es pequeña, menor que el nivel de significación, existe evidencia en contra de la hipótesis nula.

## **Asociación en tablas de contingencia. Contraste de independencia.**

El término asociación debe ser entendido como posibilidad de predicción de la modalidad que presenta una variable cualitativa si se conoce la modalidad que presentan las demás, en un cierto individuo. No debe utilizarse correlación como término equivalente pues la correlación está asociada indefectiblemente a las variables cuantitativas.

Para entender el concepto de asociación entre variables cualitativas, consideremos el siguiente ejemplo sencillo en el que se cruzan las variables  $X$ ="Ser o no alcohólico" e  $Y$ ="Ser o no cirrótico", medidas sobre 100 individuos.

<b>X</b>	<b>Y</b>	
	<b>Cirrótico</b>	<b>No cirrótico</b>
<b>Alcohólico</b>	<b>35</b>	<b>10</b>
<b>No Alcohólico</b>	<b>5</b>	<b>50</b>

- Si existiera independencia (falta de asociación) entre ambas variables,

$$p(\text{alcohólico y cirrótico})=p(\text{alcohólico}).p(\text{cirrótico})$$

(dado que bajo independencia, la probabilidad de la intersección es el producto de probabilidades), y entonces estimaríamos esas probabilidades a partir de las frecuencias relativas correspondientes, es decir:

- La frecuencia relativa de alcohólicos y cirróticos,  $35/100=0.35$ , por un lado, y el producto de la frecuencia relativa de alcohólicos por la frecuencia relativa de cirróticos, por otro, es decir,  $(45/100) \times (40/100)=0.18$ .
- Si la independencia es cierta, 35 (frecuencia observada en la celda) debe ser similar (salvo perturbaciones aleatorias) a  $45 \times 40 / 100 = 18$  (que es la llamada frecuencia esperada en la celda, bajo independencia).
- Y similarmente, para las otras 3 celdas.

- Ese exceso de la frecuencia observada en la celda noroeste, 35, respecto a la esperada bajo independencia, 18, se percibe como debido a la asociación entre ambas modalidades. Cuanto mayor sea la diferencia entre ambas cantidades, mayor es la asociación existente.
- La Estadística es una ciencia de fronteras, y el problema es determinar en qué momento esa diferencia es lo suficientemente significativa para poder rechazar la independencia y afirmar que los datos muestran evidencia en contra de ella.
- A tal efecto, la prueba chi-cuadrado de Fisher establece cuándo la independencia entre ambas variables puede ser rechazada, proporcionando además las celdas (cruce de modalidades) con asociación.

En la siguiente tabla aparece, para cada celda, la frecuencia observada (recuento). Además se tienen las frecuencias esperadas en las diferentes celdas y los residuos estandarizados. Para cada celda el residuo estandarizado es la diferencia entre frecuencia observada y esperada, dividida dicha diferencia por la raíz de la frecuencia esperada.

**Tabla cruzada Alcohólico\*Cirrótico**

			Cirrótico		Total
			,00	1,00	
Alcohólico	,00	Recuento	35	10	45
		Recuento esperado	18,0	27,0	45,0
		Residuo estandarizado	4,0	-3,3	
	1,00	Recuento	5	50	55
		Recuento esperado	22,0	33,0	55,0
		Residuo estandarizado	-3,6	3,0	
Total		Recuento	40	60	100
		Recuento esperado	40,0	60,0	100,0

- Si realmente no existiera asociación, los residuos tipificados deberían ser valores con distribución normal tipificada y, por tanto, con nivel de significación del 5%, moverse entre -1.96 y 1.96.
- Sin embargo, todos los residuos tipificados están fuera de ese rango, incluso fuera del rango para niveles de significación más restrictivos.
- Para la celda noroeste (los que son alcohólicos y cirróticos), el residuo tipificado es 4, revelando una intensa asociación. Es decir, existe una incidencia significativamente mayor de la cirrosis en el grupo de los alcohólicos. Similarmente, el valor de -3.3 correspondiente a la celda ( $X=0, Y=1$ ) muestra una asociación fuertemente negativa entre alcohólico y no cirrótico. Por tanto, en el grupo de los alcohólicos, los no cirróticos se presenta en mucha menor medida que lo que daría la independencia.

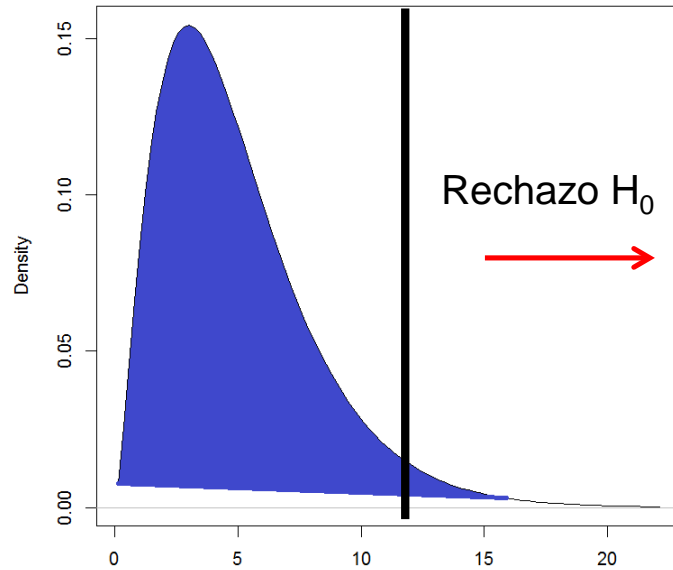


- El ejemplo propuesto es sencillo, pero cuando hay más celdas, puede ocurrir que no se detecte asociación en todas y, por otro lado sería conveniente saber si se puede mantener la independencia..
- El estadístico chi-cuadrado de Fisher se obtiene como la suma de los cuadrados de los residuos tipificados. Y se contrasta con una chi-cuadrado con un número de grados de libertad que coincide con  $(\text{número de filas de la tabla}-1) \times (\text{número de columnas de la tabla}-1)$ .
- De hecho, ahí nace el concepto de grado de libertad. Son (como en un sudoku) las celdas de la tabla que se pueden rellenar libremente, si los totales marginales están fijos.
- Para tablas 2x2 es conveniente utilizar la corrección por continuidad de Yates, pues con un solo grado de libertad, puede detectarse falta de ajuste del estadístico chi-cuadrado la chi con un grado de libertad.

Formalmente el estadístico de contraste de la prueba chi cuadrado es:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left( n_{ij} - \frac{n_{i.} n_{.j}}{n} \right)^2}{\frac{n_{i.} n_{.j}}{n}} \rightarrow \chi_{(r-1)(s-1)}^2$$

Decisión: Rechazar  $H_0$  si:  $\chi^2 > \chi^2_{(r-1)(s-1), \alpha}$



## Prueba chi-cuadrado para el ejemplo de los alcohólicos y los cirróticos

Como se aprecia, la prueba de Fisher, tanto sin la corrección por continuidad, como con ella, nos conduce a rechazar la hipótesis de independencia, dado que el p-valor muestra una fuerte evidencia en contra de ella.

Pruebas de chi-cuadrado					
	Valor	df	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	48,653 <sup>a</sup>	1	,000		
Corrección de continuidad <sup>b</sup>	45,833	1	,000		

UNIVERSIDAD  
COMPLUTENSE  
DE MADRID

