ESTADISTICA DESCRIPTIVA

La estadística descriptiva es una ciencia que analiza series de datos (por ejemplo, edad de una población, peso de los animales de un rodeo, temperatura en los meses de verano, etc) y trata de extraer conclusiones sobre el comportamiento de estas variables.

Las variables pueden ser de dos tipos:

- Variables cualitativas o atributos: no se pueden medir numéricamente (por ejemplo: raza, categoría animal, 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 animales de un lote).
- Variables bidimensionales: recogen información sobre dos características de la población (por ejemplo: edad y altura de los animales de un lote).
- Variables pluridimensionales: recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los animales de un lote).

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 (puede ser 1, 2, 3....,etc, pero, por ejemplo, nunca podrá ser 3,45).
- 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 conceptos:

Individuo: cualquier elemento que porte información sobre el fenómeno que se estudia. Así, si estudiamos la altura de los animales de un lote, cada animal 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 porten 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 la ciudad (sería una labor muy compleja), sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

Distribución de frecuencia

La **distribución de frecuencia** es la representación estructurada, en forma de tabla, de toda la información que se ha recogido sobre la variable que se estudia.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acumulada	Simple	Acumulada
X1	n1	N1	f1 = n1 / n	f1
X2	n2	n1 + n2	f2 = n2 / n	f1 + f2
...	...	...	...	...
Xn-1	nn-1	n1 + n2 +..+ nn-1	fn-1 = nn-1 / n	f1 + f2 +..+fn-1
Xn	nn	Σ n	fn = nn / n	Σ f
Siendo **X** los distintos valores que puede tomar la variable.				
Siendo **n** el número de veces que se repite cada valor.				
Siendo **f** el porcentaje que la repetición de cada valor supone sobre el total				

Veamos un ejemplo:
Medimos la altura de los animales de un lote y obtenemos los siguientes resultados (cm):

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

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

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	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%

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 frecuencia muy extensa que aportaría muy poco valor a efectos de síntesis.

Distribuciones de frecuencia agrupada

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

Habitante	Estatura	Habitante	Estatura	Habitante	Estatura
x	x	x	x	x	x
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, preferimos agrupar los datos por intervalos, con lo que la información queda más resumida (se pierde, por tanto, algo de información), pero es más manejable e informativa:

Estatura	Frecuencias absolutas		Frecuencias relativas	
Cm	Simple	Acumulada	Simple	Acumulada
1,01 - 1,10	1	1	3,3%	3,3%
1,11 - 1,20	3	4	10,0%	13,3%
1,21 - 1,30	3	7	10,0%	23,3%
1,31 - 1,40	2	9	6,6%	30,0%
1,41 - 1,50	6	15	20,0%	50,0%
1,51 - 1,60	4	19	13,3%	63,3%
1,61 - 1,70	3	22	10,0%	73,3%
1,71 - 1,80	3	25	10,0%	83,3%
1,81 - 1,90	2	27	6,6%	90,0%
1,91 - 2,00	3	30	10,0%	100,0%

El número de tramos 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 puede que menos representativa e informativa sea la tabla.

MEDIDAS DE POSICIÓN CENTRAL

Las medidas de posición nos facilitan información sobre la serie de datos que estamos analizando. Estas medidas permiten conocer diversas características de esta serie de datos.
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 de como se distribuye el resto de los valores de la serie.

a) 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 de la muestra:

$$Xm = \frac{(X1 * n1) + (X2 * n2) + (X3 * n3) + ...+ (Xn-1 * nn-1) + (Xn * nn)}{n}$$

b) Media geométrica: se eleva cada valor al número de veces que se ha repetido. Se multiplican todo estos resultados y al producto final se le calcula la raíz "n" (siendo "n" el total de datos de la muestra).

$$X = (X_1^{n1} * X_2^{n2} * X_3^{n3} * .... * X_n^{nn})^{(1/n)}$$

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 presentan el problema de estar influido por los valores extremos, pero en cambio no utiliza en su cálculo toda la información de la serie de datos (no pondera cada valor por el número de veces que se ha repetido).

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 animales que vimos anteriormente.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	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%

Vamos a calcular los valores de las distintas posiciones centrales:

1.- Media aritmética:

$$Xm = \frac{(1,20*1) + (1,21*4) + (1,22 * 4) + (1,23 * 2) + ... + (1,29 * 3) + (1,30 * 3)}{30}$$

Por lo tanto:

Xm = 1,253

Por lo tanto, la estatura media de este grupo de animales es de 1,253 cm.

2.- Media geométrica:

X = ((1,20^ 1) * (1,21^4) * (1,22^ 4) * .....* (1,29^3)* (1,30^3)) ^ (1/30)

O sea :

Xm = 1,253

En este ejemplo la media aritmética y la media geométrica coinciden, pero no tiene siempre por qué ser así.

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.

4.- Moda:

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

MEDIDAS DE POSICIÓN NO CENTRAL

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

Cuartiles: son 3 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cuatro tramos iguales, en los que cada uno de ellos concentra el 25% de los resultados.

Deciles: son 9 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados.

Percentiles: son 99 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cien tramos iguales, en los que cada uno de ellos concentra el 1% de los resultados.

Ejemplo: Vamos a calcular los cuartiles de la serie de datos referidos a la estatura de un grupo de animales. Los deciles y centiles se calculan de igual manera, aunque haría falta distribuciones con mayor número de datos.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	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%

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.

3º cuartil: es el valor 1,28 cm, ya que entre este valor y el 2º cuartil se sitúa otro 25% de la frecuencia. Además, por encima suya queda el restante 25% de la frecuencia.

Atención: cuando un cuartil recae en un valor que se ha repetido más de una vez (como ocurre en el ejemplo en los tres cuartiles) la medida de posición no central sería realmente una de las repeticiones.

MEDIDAS DE DISPERSIÓN

Estudia la distribución de los valores 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 valor más bajo.

2.- Varianza: Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatoria de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. El sumatoria obtenida se divide por el tamaño de la muestra.

$$S^2_x = \frac{\Sigma (x_i - \bar{x}_m)^2 * n_i}{n}$$

La varianza siempre será mayor que cero. Mientras más se aproxima a cero, 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 estandar: Se calcula como raíz cuadrada de la varianza.

4.- Coeficiente de variación de Pearson: se calcula como cociente entre la desviación típica y la media.

Ejemplo: vamos a utilizar la serie de datos de la estatura de los animales de un lote y vamos a calcular sus medidas de dispersión.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	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%

1.- Rango: Diferencia entre el mayor valor de la muestra (1,30) y el menor valor (1,20). Luego el rango de esta muestra es 10 cm.

2.- Varianza: recordemos que la media de esta muestra es 1,253. Luego, aplicamos la fórmula:

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

Por lo tanto, la varianza es 0,0010

3.- Desviación típica: es la raíz cuadrada de la varianza.

$$\sigma = (S^2_x)^{(1/2)}$$

O sea que:

$$\sigma = (0,010)^{(1/2)} = 0,0320$$

4.- Coeficiente de variación de Pearson: se calcula como cociente entre la desviación típica y la media de la muestra.

Cv = 0,0320 / 1,253

Por lo tanto,

Cv = 0,0255

El interés del coeficiente de variación es que al ser un porcentaje permite comparar el nivel de dispersión de dos muestras. Esto no ocurre con la desviación típica, ya que viene expresada en las mismas unidas que los datos de la serie.

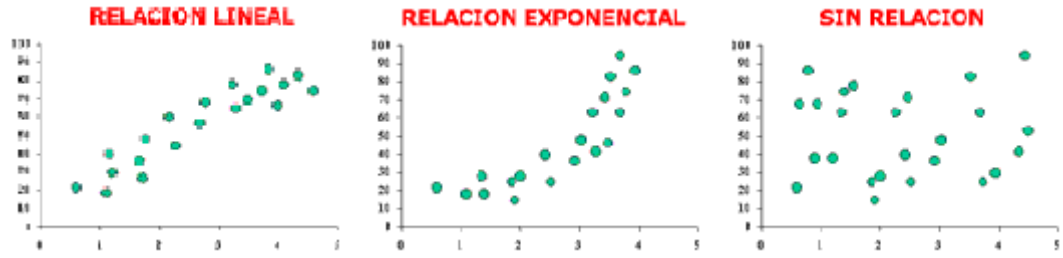
Por ejemplo, para comparar el nivel de dispersión de una serie de datos de la altura de los animales de un lote y otra serie con el peso de dichos animales, no se puede utilizar las desviaciones típicas (una viene vienes expresada en cm y la otra en kg). En cambio, sus coeficientes de variación son ambos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

COEFICIENTE DE CORRELACIÓN LINEAL

En una distribución bidimensional puede ocurrir que las dos variables guarden algún tipo de relación entre si.

Por ejemplo, si se analiza la estatura y el peso de los animales de un lote es muy posible que exista relación entre ambas variables: mientras más alto sea el animal, mayor será su peso.

El coeficiente de correlación lineal mide el grado de intensidad de esta posible relación entre las variables. Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representaramos en un gráfico los pares de valores de las dos variables la nube de puntos se aproximaría a una recta).



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

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

$$r = \frac{1/n * \Sigma (x_i - xm) * (y_i - ym)}{\left((1/n * \Sigma (x_i - xm)^2) * (1/n * \Sigma (y_i - ym)^2) \right)^{1/2}}$$

Es decir:
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 varianzas de "x" y de "y", y a este producto se le calcula la raíz cuadrada.

Los valores que puede tomar el **coeficiente de correlación "r"** son: $-1 < r < 1$
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: altura y peso: los animales más altos suelen pesar más.
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: peso y velocidad: los animales más gordos suelen correr menos.
Si "r" = 0, no existe correlación lineal entre las variables. Aunque podría existir otro tipo de correlación (parabólica, exponencial, etc.)

De todos modos, aunque el valor de "r" fuera próximo a 1 o -1, tampoco esto quiere decir obligatoriamente que existe una relación de causa-efecto entre las dos variables, ya que este resultado podría haberse debido al puro azar.

Ejemplo: vamos a calcular el coeficiente de correlación de la siguiente serie de datos de altura y peso de los animales de un establecimiento:

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

Aplicamos la fórmula:

$$r = \frac{(1/30) * (0,826)}{(((1/30)*(0,02568)) * ((1/30)*(51,366)))^{(1/2)}}$$

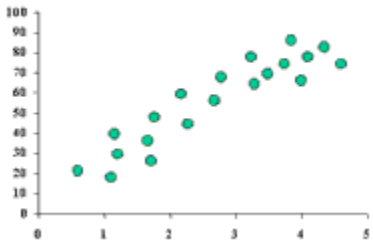
O sea,

R = 0,719

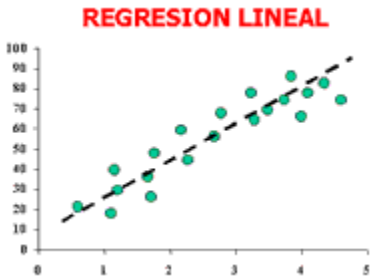
Por lo tanto, la correlación existente entre estas dos variables es elevada (0,7) y de signo positivo.

REGRESIÓN LINEAL

Representamos en un gráfico los pares de valores de una distribución bidimensional: la variable "x" en el eje horizontal o eje de abcisa, y la variable "y" en el eje vertical, o eje de ordenada. Vemos que la nube de puntos sigue una tendencia lineal:



El **coeficiente de correlación lineal** nos permite determinar si, efectivamente, existe relación entre las dos variables. Una vez que se concluye que sí existe relación, la **regresión** nos permite definir la recta que mejor se ajusta a esta nube de puntos.



Una recta viene definida por la siguiente fórmula:

$$y = a + bx$$

Donde "y" sería la variable dependiente, es decir, aquella que viene definida a partir de la otra variable "x" (variable independiente). Para definir la recta hay que determinar los valores de los parámetros "a" y "b":

El **parámetro "a"** es el valor que toma la variable dependiente "y", cuando la variable independiente "x" vale 0, y es el punto donde la recta cruza el eje vertical.
El **parámetro "b"** determina la pendiente de la recta, su grado de inclinación.

La **regresión lineal** nos permite calcular el valor de estos dos parámetros, definiendo la recta que mejor se ajusta a esta nube de puntos.

El **parámetro "b"** viene determinado por la siguiente fórmula:

$$b = \frac{1/n * \Sigma (x_i - xm) * (y_i - ym)}{1/n * \Sigma (x_i - xm)^2}$$

Es la covarianza de las dos variables, dividida por la varianza de la variable "x".
El **parámetro "a"** viene determinado por:

$$a = y_m - (b * x_m)$$

Es la media de la variable "y", menos la media de la variable "x" multiplicada por el parámetro "b" que hemos calculado.

Ejemplo: vamos a calcular la recta de regresión de la siguiente serie de datos de altura y peso de los animales de una clase. Vamos a considerar que la altura es la variable independiente "x" y que el peso es la variable dependiente "y" (podíamos hacerlo también al contrario):

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

El **parámetro "b"** viene determinado por:

$$b = \frac{(1/30) * 1,034}{(1/30) * 0,00856} = 40,265$$

Y el **parámetro "a"** por:

$$a = 33,1 - (40,265 * 1,262) = -17,714$$

Por lo tanto, la **recta** que mejor se ajusta a esta serie de datos es:

y = -17,714 + (40,265 * x)

Esta recta define un valor de la variable dependiente (peso), para cada valor de la variable independiente (estatura):

Estatura	Peso

1,20	30,6
1,21	31,0
1,22	31,4
1,23	31,8
1,24	32,2
1,25	32,6
1,26	33,0
1,27	33,4
1,28	33,8
1,29	34,2
1,30	34,6

PROBABILIDAD

La **probabilidad** mide la frecuencia con la que aparece un resultado determinado cuando se realiza un experimento.

Ejemplo: tiramos un dado al aire y queremos saber cual es la probabilidad de que salga un 2, o que salga un número par, o que salga un número menor que 4.

El experimento tiene que ser aleatorio, es decir, que pueden presentarse diversos resultados, dentro de un conjunto posible de soluciones, y esto aún realizando el experimento en las mismas condiciones. Por lo tanto, a priori no se conoce cual de los resultados se va a presentar:

Ejemplos: lanzamos una moneda al aire: el resultado puede ser cara o cruz, pero no sabemos de antemano cual de ellos va a salir.

En la Lotería de Navidad, el "Gordo" (primer premio) puede ser cualquier número entre el 1 y el 100.000, pero no sabemos a priori cual va a ser.

Hay experimentos que no son aleatorios y por lo tanto no se les puede aplicar las reglas de la probabilidad.

Ejemplo: en lugar de tirar la moneda al aire, directamente seleccionamos la cara. Aquí no podemos hablar de probabilidades, sino que ha sido un resultado determinado por uno mismo.
Antes de calcular las probabilidades de un experimento aleatorio hay que definir una serie de conceptos:

Suceso elemental: hace referencia a cada una de las posibles soluciones que se pueden presentar.

Ejemplo: al lanzar una moneda al aire, los sucesos elementales son la cara y la cruz. Al lanzar un dado, los sucesos elementales son el 1, el 2, .., hasta el 6.

Suceso compuesto: es un subconjunto de sucesos elementales.

Ejemplo: lanzamos un dado y queremos que salga un número par. El suceso "numero par" es un suceso compuesto, integrado por 3 sucesos elementales: el 2, el 4 y el 6
O, por ejemplo, jugamos a la ruleta y queremos que salga "menor o igual que 18". Este es un suceso compuesto formado por 18 sucesos elementales (todos los números que van del 1 al 18).

Al conjunto de todos los posibles sucesos elementales lo denominamos **espacio muestral**. Cada experimento aleatorio tiene definido su espacio muestral (es decir, un conjunto con todas las soluciones posibles).

Ejemplo: si tiramos una moneda al aíre una sola vez, el espacio muestral será cara o cruz.
Si el experimento consiste en lanzar una moneda al aire dos veces, entonces el espacio muestral estaría formado por (cara-cara), (cara-cruz), (cruz-cara) y (cruz-cruz).

Relación entre sucesos

Entre los sucesos compuestos se pueden establecer distintas relaciones:

a) **Un suceso puede estar contenido en otro:** las posibles soluciones del primer suceso también lo son del segundo, pero este segundo suceso tiene además otras soluciones suyas propias.

Ejemplo: lanzamos un dado y analizamos dos sucesos: a) que salga el número 6, y b) que salga un número par. Vemos que el suceso a) está contenido en el suceso b).
Siempre que se da el suceso a) se da el suceso b), pero no al contrario. Por ejemplo, si el resultado fuera el 2, se cumpliría el suceso b), pero no el el a).

b) **Dos sucesos pueden ser iguales:** esto ocurre cuando siempre que se cumple uno de ellos se cumple obligatoriamente el otro y viceversa.

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga número par, y b) que salga múltiplo de 2. Vemos que las soluciones coinciden en ambos casos.

c) **Unión de dos o más sucesos:** la unión será otro suceso formado por todos los elementos de los sucesos que se unen.

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga número par y b) que el resultado sea mayor que 3. El suceso unión estaría formado por los siguientes resultados: el 2, el 4, el 5 y el 6

d) **Intersección de sucesos:** es aquel suceso compuesto por los elementos comunes de dos o más sucesos que se intersectan.

Ejemplo: lanzamos un dado al aire, y analizamos dos sucesos: a) que salga número par, y b) que sea mayor que 4. La intersección de estos dos sucesos tiene un sólo elemento, el número 6 (es el único resultado común a ambos sucesos: es mayor que 4 y es número par).

e) **Sucesos incompatibles:** son aquellos que no se pueden dar al mismo tiempo ya que no tienen elementos comunes (su interesección es el conjunto vacio).

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga un número menor que 3, y b) que salga el número 6. Es evidente que ambos no se pueden dar al mismo tiempo.

f) **Sucesos complementarios:** son aquellos que si no se da uno, obligatoriamente se tiene que dar el otro.

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga un número par, y b) que salga un número impar. Vemos que si no se da el primero se tiene que dar el segundo (y viceversa).

Cálculo de probabilidades

Como hemos comentado anteriormente, la probabilidad mide la mayor o menor posibilidad de que se dé un determinado resultado (suceso) cuando se realiza un experimento aleatorio.
La probabilidad toma valores entre 0 y 1 (o expresados en tanto por ciento, entre 0% y 100%):
El valor cero corresponde al suceso imposible: lanzamos un dado al aire y la probabilidad de que salga el número 7 es cero.

El valor uno corresponde al suceso seguro: lanzamos un dado al aire y la probabilidad de que salga cualquier número del 1 al 6 es igual a uno (100%).

El resto de sucesos tendrá probabilidades entre cero y uno: que será tanto mayor cuanto más probable sea que dicho suceso tenga lugar.

¿Cómo se mide la probabilidad?

Uno de los métodos más utilizados es aplicando la **Regla de Laplace:** define la probabilidad de un suceso como el cociente entre casos favorables y casos posibles.

P(A) = Casos favorables / casos posibles

Veamos algunos ejemplos:

a) Probabilidad de que al lanzar un dado salga el número 2: el caso favorable es tan sólo uno (que salga el dos), mientras que los casos posibles son seis (puede salir cualquier número del uno al seis). Por lo tanto:
P(A) = 1 / 6 = 0,166 (o lo que es lo mismo, 16,6%)

b) Probabilidad de que al lanzar un dado salga un número par: en este caso los casos favorables son tres (que salga el dos, el cuatro o el seis), mientras que los casos posibles siguen siendo seis. Por lo tanto:
P(A) = 3 / 6 = 0,50 (o lo que es lo mismo, 50%)

c) Probabilidad de que al lanzar un dado salga un número menor que 5: en este caso tenemos cuatro casos favorables (que salga el uno, el dos, el tres o el cuatro), frente a los seis casos posibles. Por lo tanto:
P(A) = 4 / 6 = 0,666 (o lo que es lo mismo, 66,6%)

d) Probabilidad de que nos toque el "Gordo" de Navidad: tan sólo un caso favorable, el número que jugamos frente a 100.000 casos posibles. Por lo tanto:
P(A) = 1 / 100.000 = 0,00001 (o lo que es lo mismo, 0,001%)
Por cierto, tiene la misma probabilidad el número 45.264, que el número 00001, pero ¿cuál de los dos comprarías?

Para poder aplicar la **Regla de Laplace** el experimento aleatorio tiene que cumplir **dos requisitos:**

a) El número de resultados posibles (sucesos) tiene que ser finito. Si hubiera infinitos resultados, al aplicar la regla "casos favorables / casos posibles" el cociente siempre sería cero.
b) Todos los sucesos tienen que tener la misma probabilidad. Si al lanzar un dado, algunas caras tuvieran mayor probabilidad de salir que otras, no podríamos aplicar esta regla.

A la regla de Laplace también se le denomina **"probabilidad a priori"**, ya que para aplicarla hay que conocer antes de realizar el experimento cuales son los posibles resultados y saber que todos tienen las mismas probabilidades.

¿Y si el experimento aleatorio no cumple los dos requisitos indicados, qué hacemos?
En este caso podemos acudir a otro modelo de cálculo de probabilidades que se basa en la experiencia **(modelo frecuentista)**:

Cuando se realiza un experimento aleatorio un número muy elevado de veces, las probabilidades de los diversos posibles sucesos empiezan a converger hacia valores determinados, que son sus respectivas probabilidades.

Ejemplo: si lanzo una vez una moneda al aire y sale "cara", quiere decir que el suceso "cara" ha aparecido el 100% de las veces y el suceso "cruz" el 0%.

Si lanzo diez veces la moneda al aire, es posible que el suceso "cara" salga 7 veces y el suceso "cruz" las 3 restantes. En este caso, la probabilidad del suceso "cara" ya no sería del 100%, sino que se habría reducido al 70%.

Si repito este experimento un número elevado de veces, lo normal es que las probabilidades de los sucesos "cara" y "cruz" se vayan aproximando al 50% cada una. Este 50% será la probabilidad de estos sucesos según el modelo frecuentista.
En este modelo ya no será necesario que el número de soluciones sea finito, ni que todos los sucesos tengan la misma probabilidad.

Ejemplo: si la moneda que utilizamos en el ejemplo anterior fuera defectuosa (o estuviera trucada), es posible que al repetir dicho experimento un número elevado de veces, la "cara" saliera con una frecuencia, por ejemplo, del 65% y la "cruz" del 35%. Estos valores serían las probabilidades de estos dos sucesos según el modelo frecuentista.

A esta definición de la probabilidad se le denomina **probabilidad a posteriori**, ya que tan sólo repitiendo un experimento un número elevado de veces podremos saber cual es la probabilidad de cada suceso.

Probabilidad de sucesos

Al definir los sucesos hablamos de las diferentes relaciones que pueden guardar dos sucesos entre sí, así como de las posibles relaciones que se pueden establecer entre los mismos. Vamos a ver ahora cómo se refleja esto en el cálculo de probabilidades.

a) **Un suceso puede estar contenido en otro**: entonces, la probabilidad del primer suceso será menor que la del suceso que lo contiene.

Ejemplo: lanzamos un dado y analizamos dos sucesos: a) que salga el número 6, y b) que salga un número par. Dijimos que el suceso a) está contenido en el suceso b).
P(A) = 1/6 = 0,166
P(B) = 3 / 6 = 0,50

Por lo tanto, podemos ver que la probabilidad del suceso contenido, suceso a), es menor que la probabilidad del suceso que lo contiene, suceso b).

b) **Dos sucesos pueden ser iguales**: en este caso, las probabilidades de ambos sucesos son las mismas.

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga número par, y b) que salga múltiplo de 2. Las soluciones coinciden en ambos casos.
P(A) = 3 / 6 = 0,50
P(B) = 3 / 6 = 0,50

c) **Intersección de sucesos**: es aquel suceso compuesto por los elementos comunes de los dos o más sucesos que se intersectan. La probabilidad será igual a la probabilidad de los elementos comunes.

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga número par, y b) que sea mayor que 3. La intersección de estos dos sucesos tiene dos elementos: el 4 y el 6.
Su probabilidad será por tanto:
$P(A \wedge B) = 2 / 6 = 0,33$

d) **Unión de dos o más sucesos**: la probabilidad de la unión de dos sucesos es igual a la suma de las probabilidades individuales de los dos sucesos que se unen, menos la probabilidad del suceso intersección

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga número par, y b) que el resultado sea mayor que 3. El suceso unión estaría formado por los siguientes resultados: el 2, el 4, el 5 y el 6.
P(A) = 3 / 6 = 0,50

P(B) = 3 / 6 = 0,50
P (A ∧ B) = 2 / 6 = 0,33
Por lo tanto,
P (A u B) = (0,50 + 0,50) - 0,33 = 0,666

e) **Sucesos incompatibles**: la probabilidad de la unión de dos sucesos incompatibles será igual a la suma de las probabilidades de cada uno de los sucesos (ya que su intersección es el conjunto vacío y por lo tanto no hay que restarle nada).

Ejemplo: lanzamos un dado al aire y analizamos dos sucesos: a) que salga un número menor que 3, y b) que salga el número 6.
La probabilidad del suceso unión de estos dos sucesos será igual a:
P(A) = 2 / 6 = 0,333
P(B) = 1 / 6 = 0,166
Por lo tanto,
P(A u B) = 0,33 + 0,166 = 0,50

f) **Sucesos complementarios:** la probabilidad de un suceso complementario a un suceso (A) es igual a 1 - P(A)

Ejemplo: lanzamos un dado al aire. el suceso (A) es que salga un número par, luego su complementario, suceso (B), es que salga un número impar.
La probabilidad del suceso (A) es igual a :
P(A) = 3 / 6 = 0,50
Luego, la probabilidad del suceso (B) es igual a:
P(B) = 1 - P(A) = 1 - 0,50 = 0,50
Se puede comprobar aplicando la regla de "casos favorables / casos posibles":
P(B) = 3 / 6 = 0,50

g) **Unión de sucesos complementarios:** la probabilidad de la unión de dos sucesos complementarios es igual a 1.

Ejemplo: seguimos con el ejemplo anterior: a) que salga un número par, y b) que salga un número impar. La probabilidad del suceso unión de estos dos sucesos será igual a:
P(A) = 3 / 6 = 0,50
P(B) = 3 / 6 = 0,50
Por lo tanto,
P(A ∪ B) = 0,50 + 0,50 = 1

Para aplicar la **Regla de Laplace**, el cálculo de los sucesos favorables y de los sucesos posibles a veces no plantea ningún problema, ya que son un número reducido y se pueden calcular con facilidad:

Por ejemplo: Probabilidad de que al lanzar un dado salga el número 2. Tan sólo hay un caso favorable, mientras que los casos posibles son seis.
Probabilidad de acertar al primer intento el horóscopo de una persona. Hay un caso favorable y 12 casos posibles.

Sin embargo, a veces calcular el número de casos favorables y casos posibles es complejo y hay que aplicar reglas matemáticas:

Por ejemplo: 5 matrimonios se sientan aleatoriamente a cenar y queremos calcular la probabilidad de que al menos los miembros de un matrimonio se sienten junto. En este caso, determinar el número de casos favorables y de casos posibles es complejo.

Las reglas matemáticas que nos pueden ayudar son el cálculo de **combinaciones**, el cálculo de **variaciones** y el cálculo de **permutaciones**.

a) Combinaciones:
Determina el número de subgrupos de 1, 2, 3, etc. elementos que se pueden formar con los "n" elementos de una nuestra. Cada subgrupo se diferencia del resto en los elementos que lo componen, sin que influya el orden.

Por ejemplo, calcular las posibles combinaciones de 2 elementos que se pueden formar con los números 1, 2 y 3.
Se pueden establecer 3 parejas diferentes: (1,2), (1,3) y (2,3). En el cálculo de combinaciones las parejas (1,2) y (2,1) se consideran idénticas, por lo que sólo se cuentan una vez.

b) Variaciones:
Calcula el número de subgrupos de 1, 2, 3, etc. elementos que se pueden establecer con los "n" elementos de una muestra. Cada subgrupo se diferencia del resto en los elementos que lo componen o en el orden de dichos elementos (es lo que le diferencia de las combinaciones).

Por ejemplo, calcular las posibles variaciones de 2 elementos que se pueden establecer con los números 1, 2 y 3.
Ahora tendríamos 6 posibles parejas: (1,2), (1,3), (2,1), (2,3), (3,1) y (3,3). En este caso los subgrupos (1,2) y (2,1) se consideran distintos.

c) Permutaciones:
Calcula las posibles agrupaciones que se pueden establecer con todos los elementos de un grupo, por lo tanto, lo que diferencia a cada subgrupo del resto es el orden de los elementos.
Por ejemplo, calcular las posibles formas en que se pueden ordenar los números 1, 2 y 3.
Hay 6 posibles agrupaciones: (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2) y (3, 2, 1)

Combinaciones, Variaciones y Permutaciones

a) Combinaciones:
Para calcular el número de combinaciones se aplica la siguiente fórmula:

$$C_{m,n} = \frac{m!}{n! * (m - n)!}$$

El termino **" n ! "** se denomina "factorial de n" y es la multiplicación de todos los números que van desde "n" hasta 1.

Por ejemplo: 4 ! = 4 * 3 * 2 * 1 = 24
La expresión **"Cm,n"** representa las combinaciones de "m" elementos, formando subgrupos de "n" elementos.

Ejemplo: C10,4 son las combinaciones de 10 elementos agrupándolos en subgrupos de 4 elementos:

$$C_{10,4} = \frac{10!}{4! * (10 - 4)!} = \frac{10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1}{(4 * 3 * 2 * 1) * (6 * 5 * 4 * 3 * 2 * 1)} = 210$$

Es decir, podríamos formar 210 subgrupos diferentes de 4 elementos, a partir de los 10 elementos.

b) Variaciones:
Para calcular el número de variaciones se aplica la siguiente fórmula:

$$V_{m,n} = \frac{m!}{(m - n)!}$$

La expresión **"Vm,n"** representa las variaciones de "m" elementos, formando subgrupos de "n" elementos. En este caso, como vimos en la lección anterior, un subgrupo se diferenciará del resto, bien por los elementos que lo forman, o bien por el orden de dichos elementos.

Ejemplo: V10,4 son las variaciones de 10 elementos agrupándolos en subgrupos de 4 elementos:

$$V_{10,4} = \frac{10!}{(10-4)!} = \frac{10*9*8*7*6*5*4*3*2*1}{(6*5*4*3*2*1)} = 5.040$$

Es decir, podríamos formar 5.040 subgrupos diferentes de 4 elementos, a partir de los 10 elementos.

c) Permutaciones:

Para calcular el número de permutaciones se aplica la siguiente fórmula:

$$P_m = m!$$

La expresión **"Pm"** representa las permutaciones de "m" elementos, tomando todos los elementos. Los subgrupos se diferenciaran únicamente por el orden de los elementos.

Ejemplo: P10 son las permutaciones de 10 elementos:

$$P_{10} = 10! = 10*9*8*7*6*5*4*3*2*1 = 3.628.800$$

Es decir, tendríamos 3.628.800 formas diferentes de agrupar 10 elementos.
Vamos a analizar ahora que ocurriría con el cálculo de las combinaciones, de las variaciones o de las permutaciones en el **supuesto** de que al formar los subgrupos **los elementos pudieran repetirse**.

Por ejemplo: tenemos bolas de 6 colores diferentes y queremos formar subgrupos en los que pudiera darse el caso de que 2, 3, 4 o todas las bolas del subgrupo tuvieran el mismo color. En este caso no podríamos utilizar las fórmulas que vimos en la lección anterior.

a) Combinaciones con repetición:
Para calcular el número de combinaciones con repetición se aplica la siguiente fórmula:

$$C'_{m,n} = \frac{(m+n-1)!}{n!*(m-1)!}$$

Ejemplo: C'10,4 son las combinaciones de 10 elementos con repetición, agrupándolos en subgrupos de 4, en los que 2, 3 o los 4 elementos podrían estar repetidos:

$$C'_{10,4} = \frac{13!}{4!*9!} = \frac{13*12*11*10*9*8*7*6*5*4*3*2*1}{(4*3*2*1)*(9*8*7*6*5*4*3*2*1)} = 715$$

Es decir, podríamos formar 715 subgrupos diferentes de 4 elementos.

b) Variaciones con repetición:
Para calcular el número de variaciones con repetición se aplica la siguiente fórmula:

$$V'_{m,n} = m^n$$

Ejemplo: V'10,4 son las variaciones de 10 elementos con repetición, agrupándolos en subgrupos de 4 elementos:

$$V'_{10,4} = 10^{\wedge 4} = 10.000$$

Es decir, podríamos formar 10.000 subgrupos diferentes de 4 elementos.

c) Permutaciones con repetición:
Para calcular el número de permutaciones con repetición se aplica la siguiente fórmula:

$$P'^{x1,x2,\ldots xk}_{m} = \frac{m!}{x_1!*x_2!*\ldots*x_k!}$$

Son permutaciones de "m" elementos, en los que uno de ellos se repite " x1 " veces, otro " x2 " veces y así ... hasta uno que se repite " xk " veces.

Ejemplo: Calcular las permutaciones de 10 elementos, en los que uno de ellos se repite en 2 ocasiones y otro se repite en 3 ocasiones:

$$P'_{10}{}^{2,3} = \frac{10!}{2! * 3!} = 302.400$$

Es decir, tendríamos 302,400 formas diferentes de agrupar estos 10 elementos.

1.- Ejercicio

Calcular la probabilidad de, en una carrera de 12 caballos, acertar los 3 que quedan primeros (sin importar cual de ellos queda primero, cual segundo y cual tercero).

Solución:
Se aplica la **Regla de Laplace**. El **caso favorable** es tan sólo uno: los 3 caballos que entran en primer lugar. Los **casos posibles** se calculan como combinaciones de 12 elementos tomados de 3 en 3 (es decir, determinamos todos las posibles alternativas de 3 caballos que pueden entrar en las 3 primeras posiciones). Como el orden de estos 3 primeros caballos no importa, utilizamos combinaciones en lugar de variaciones.

Por lo tanto, los casos posibles son:

$$C_{12,3} = \frac{12!}{3! * (12 - 3)!} = 220$$

Por lo que la probabilidad de acertar los 3 caballos ganadores es:

$$P(A) = \frac{1}{220} = 0{,}00455$$

Algo mayor que en las quinielas.... Eso sí, se paga menos.

2.- Ejercicio
Y si hubiera que acertar, no sólo los 3 caballos que ganan, sino el orden de su entrada en meta.

Solución:
El **caso favorable** sigue siendo uno: los 3 caballos que entran en primer lugar, colocados en su orden correspondiente.
Los **casos posibles** se calculan ahora como variaciones (ya que el orden influye) de 12 elementos tomados de 3 en 3 (calculamos todas las posibles maneras en que los 12 caballos podrían ocupar las 3 primeras posiciones.

$$V_{12,3} = \frac{12!}{(12 - 3)!} = 1.320$$

Por lo que la probabilidad de acertar los 3 caballos ganadores es:

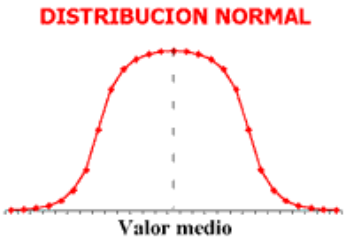
$$P(A) = \frac{1}{1.320} = 0{,}00076$$

Menor que en el ejemplo 3º. Ya no vale acertar que 3 caballos entran en primer lugar, sino que tenemos que acertar el orden de su entrada.

LA DISTRIBUCION NORMAL

Es el **modelo de distribución más utilizado** en la práctica, ya que multitud de fenómenos se comportan según una distribución normal.

Esta distribución de caracteriza porque los valores se distribuyen formando una **campana de Gauss**, en torno a un valor central que coincide con el valor medio de la distribución:



Un 50% de los valores están a la derecha de este valor central y otro 50% a la izquierda
Esta distribución viene definida por **dos parámetros**:

X: N (μ, σ 2)

μ : es el valor medio de la distribución y es precisamente donde se sitúa el centro de la curva (de la campana de Gauss).

σ 2 : es la varianza. Indica si los valores están más o menos alejados del valor central: si la varianza es baja los valores están próximos a la media; si es alta, entonces los valores están muy dispersos.

Cuando la media de la distribución es 0 y la varianza es 1se denomina **"normal tipificada**", y su ventaja reside en que hay tablas donde se recoge la probabilidad acumulada para cada punto de la curva de esta distribución.

Además, **toda distribución normal se puede transformar en una normal tipificada:**

Ejemplo: una variable aleatoria sigue el modelo de una distribución normal con media 10 y varianza 4. Transformarla en una normal tipificada.
X: N (10, 4)

Para transformarla en una normal tipificada **se crea una nueva variable** (Y) que será **igual a la anterior (X) menos su media y dividida por su desviación típica** (que es la raíz cuadrada de la varianza)

$$Y = \frac{X - \mu}{\sigma}$$

En el ejemplo, la nueva variable sería:

$$Y = \frac{X - 10}{2}$$

Esta nueva variable se distribuye como una normal tipificada, permitiéndonos, por tanto, conocer la probabilidad acumulada en cada valor.
Y: N (0, 1)

La **distribución normal tipificada** tiene la ventaja, como ya hemos indicado, de que las probabilidades para cada valor de la curva se encuentran recogidas en una tabla.

X	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5723
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7090	0,7224

0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7813	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8416	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861

¿Cómo se lee esta tabla?
La columna de la izquierda indica el valor cuya probabilidad acumulada queremos conocer. La primera fila nos indica el segundo decimal del valor que estamos consultando.

Ejemplo: queremos conocer la probabilidad acumulada en el valor 2,75.Entonces buscamos en la columna de la izquierda el valor 2,7 y en la primera fila el valor 0,05. La casilla en la que se interseccionan es su probabilidad acumulada (0,99702, es decir 99.7%).

Atención: la tabla nos da la probabilidad acumulada, es decir, la que va desde el inicio de la curva por la izquierda hasta dicho valor. No nos da la probabilidad concreta en ese punto. En una distribución continua en el que la variable puede tomar infinitos valores, la probabilidad en un punto concreto es prácticamente despreciable.

Ejemplo: Imaginemos que una variable continua puede tomar valores entre 0 y 5. La probabilidad de que tome exactamente el valor 2 es despreciable, ya que podría tomar infinitos valores: por ejemplo: 1,99, 1,994, 1,9967, 1,9998, 1999791, etc.

Veamos **otros ejemplos**:
Probabilidad acumulada en el valor 0,67: la respuesta es 0,7486
Probabilidad acumulada en el valor 1,35: la respuesta es 0,9115
Probabilidad acumulada en el valor 2,19: la respuesta es 0,98574
Veamos ahora, como podemos **utilizar esta tabla con una distribución normal**:

Ejemplo: el salario medio de los empleados de una empresa se distribuye según una distribución normal, con media 5 mil pesos y desviación típica mil pesos. Calcular el porcentaje de empleados con un sueldo inferior a 7 mil pesos.
Lo primero que haremos es transformar esa distribución en una normal tipificada, para ello se crea una nueva variable (Y) que será igual a la anterior (X) menos su media y dividida por la desviación típica

$$Y = \frac{X - \mu}{\sigma}$$

En el ejemplo, la nueva variable sería:

$$Y = \frac{X - 5}{1}$$

Esta nueva variable se distribuye como una normal tipificada. La variable Y que corresponde a una variable X de valor 7 es:

$$Y = \frac{7 - 5}{1} = 2$$

Ya podemos consultar en la tabla la probabilidad acumulada para el valor 2 (equivalente a la probabilidad de sueldos inferiores a 7 mil pesos.). Esta probabilidad es 0,97725
Por lo tanto, el porcentaje de empleados con salarios inferiores a 7 mil pesos. es del 97,725%.

Ejercicio 1º: La renta media de los habitantes de un país es de 4 millones de pesos/año, con una varianza de 1,5. Se supone que se distribuye según una distribución normal. Calcular:
a) Porcentaje de la población con una renta inferior a 3 millones de pesos.
b) Renta a partir de la cual se sitúa el 10% de la población con mayores ingresos.
c) Ingresos mínimo y máximo que engloba al 60% de la población con renta media.

a) Porcentaje de la población con una renta inferior a 3 millones de pesos.
Lo primero que tenemos que hacer es calcular la normal tipificada:

$$Y = \frac{X - 4}{1,22}$$

(*) Recordemos que el denominador es la desviación típica (raíz cuadrada de la varianza)
El valor de Y equivalente a 3 millones de pesos es -0,816.
P (X < 3) = P (Y < -0,816)
Ahora tenemos que ver cuál es la probabilidad acumulada hasta ese valor. Tenemos un problema: la tabla de probabilidades sólo abarca valores positivos, no obstante, este problema tiene fácil solución, ya que la distribución normal es simétrica respecto al valor medio.
Por lo tanto:
P (Y < -0,816) = P (Y > 0,816)
Por otra parte, la probabilidad que hay a partir de un valor es igual a 1 (100%) menos la probabilidad acumulada hasta dicho valor:
P (Y > 0,816) = 1 - P (Y < 0,816) = 1 - 0,7925 (aprox.) = 0,2075
Luego, el 20,75% de la población tiene una renta inferior a 3 millones de pesos.

b) Nivel de ingresos a partir del cual se sitúa el 10% de la población con renta más elevada.
Vemos en la tabla el valor de la variable tipificada cuya probabilidad acumulada es el 0,9 (90%), lo que quiere decir que por encima se sitúa el 10% superior.
Ese valor corresponde a Y = 1,282 (aprox.). Ahora calculamos la variable normal X equivalente a ese valor de la normal tipificada:

$$1,282 = \frac{X - 4}{1,22}$$

Despejando X, su valor es 5,57. Por lo tanto, aquellas personas con ingresos superiores a 5,57 millones de pesos constituyen el 10% de la población con renta más elevada.

c) Nivel de ingresos mínimo y máximo que engloba al 60% de la población con renta media
Vemos en la tabla el valor de la variable normalizada Y cuya probabilidad acumulada es el 0,8 (80%). Como sabemos que hasta la media la probabilidad acumulada es del 50%, quiere decir que entre la media y este valor de Y hay un 30% de probabilidad.

Por otra parte, al ser la distribución normal simétrica, entre -Y y la media hay otro 30% de probabilidad. En definitiva, el segmento (-Y, Y) engloba al 60% de población con renta media.

El valor de Y que acumula el 80% de la probabilidad es 0,842 (aprox.), por lo que el segmento viene definido por (-0,842, +0,842). Ahora calculamos los valores de la variable X correspondientes a estos valores de Y.

Los valores de X son 2,97 y 5,03. Por lo tanto, las personas con ingresos superiores a 2,97 millones de pesos e inferiores a 5,03 millones de pesos constituyen el 60% de la población con un nivel medio de renta.

Ejercicio 2º: La vida media de los habitantes de un país es de 68 años, con una varianza de 25. Se hace un estudio en una pequeña ciudad de 10.000 habitantes:
a) ¿Cuántas personas superarán previsiblemente los 75 años?
b) ¿Cuántos vivirán menos de 60 años?

a) Personas que vivirán (previsiblemente) más de 75 años
Calculamos el valor de la normal tipificada equivalente a 75 años

$$Y = \frac{75 - 68}{5} = 1,4$$

Por lo tanto
P (X > 75) = (Y > 1,4) = 1 - P (Y < 1,4) = 1 - 0,9192 = 0,0808
Luego, el 8,08% de la población (808 habitantes) vivirán más de 75 años.

b) Personas que vivirán (previsiblemente) menos de 60 años
Calculamos el valor de la normal tipificada equivalente a 60 años

$$Y = \frac{60 - 68}{5} = -1,6$$

Por lo tanto
P (X < 60) = (Y < -1,6) = P (Y > 1,6) = 1 - P (Y < 1,6) = 0,0548
Luego, el 5,48% de la población (548 habitantes) no llegarán probablemente a esta edad.

Ejercicio 1º: El consumo medio anual de cerveza de los habitantes de una país es de 59 litros, con una varianza de 36. Se supone que se distribuye según una distribución normal.
a) Si usted presume de buen bebedor, ¿cuántos litros de cerveza tendría que beber al año para pertenecer al 5% de la población que más bebe?.
b) Si usted bebe 45 litros de cerveza al año y su mujer le califica de borracho ¿qué podría argumentar en su defensa?

a) 5% de la población que más bebe.
Vemos en la tabla el valor de la variable tipificada cuya probabilidad acumulada es el 0,95 (95%), por lo que por arriba estaría el 5% restante.
Ese valor corresponde a Y = 1,645 (aprox.). Ahora calculamos la variable normal X equivalente a ese valor de la normal tipificada:

$$1,645 = \frac{X - 58}{6}$$

Despejando X, su valor es 67,87. Por lo tanto, tendría usted que beber más de 67,87 litros al año para pertenecer a ese "selecto" club de grandes bebedores de cerveza.

b) Usted bebe 45 litros de cerveza al año. ¿Tiene problemas con la bebida?
Vamos a ver en que nivel de la población se situaría usted en función de los litros de cerveza consumidos.
Calculamos el valor de la normal tipificada correspondiente a 45 litros:

$$Y = \frac{45 - 58}{6} = -2,2$$

Por lo tanto
P (X < 45) = (Y < -2,2) = P (Y > 2,2) = 1 - P (Y < 2,2) = 0,0139
Luego, tan sólo un 1,39% de la población bebe menos que usted. Estadísticamente es un argumento suficiente como para ser considerado un bebedor social.

Ejercicio 2º: A un examen de oposición se han presentado 2.000 aspirantes. La nota media ha sido un 5,5, con una varianza de 1,5.

a) Tan sólo hay 100 plazas. Usted ha obtenido un 7,7. ¿Sería oportuno ir organizando una fiesta para celebrar su éxito?
b) Va a haber una 2ª oportunidad para el 20% de notas más altas que no se hayan clasificados. ¿A partir de que nota se podrá participar en este "repechaje"?

a) Ha obtenido usted un 7,7
Vamos a ver con ese 7,7 en que nivel porcentual se ha situado usted, para ello vamos a comenzar por calcular el valor de la normal tipificada equivalente.

$$Y = \frac{7,7 - 5,5}{1,049} = 2,1$$

A este valor de Y le corresponde una probabilidad acumulada (ver tablas) de 0,98214 (98,214%), lo que quiere decir que por encima de usted tan sólo se encuentra un 1,786%.
Si se han presentado 2.000 aspirante, ese 1,786% equivale a unos 36 aspirantes. Por lo que si hay 100 plazas disponibles, tiene usted suficientes probabilidades como para ir organizando la "mejor de las fiestas".

b) "Repechaje" para el 20% de los candidatos
Vemos en la tabla el valor de la normal tipificada que acumula el 80% de la probabilidad, ya que por arriba sólo quedaría el 20% restante.
Este valor de Y corresponde a 0,842 (aprox.). Ahora calculamos el valor de la normal X equivalente:

$$0,842 = \frac{X - 5,5}{1,049}$$

Despejamos la X y su valor es 6,38. Por lo tanto, esta es la nota a partir de la cual se podrá acudir a la "repechaje".

TEOREMA CENTRAL DEL LÍMITE

El **Teorema Central del Límite** dice que si tenemos un grupo numeroso de variables independientes y todas ellas siguen el mismo modelo de distribución (cualquiera que éste sea), la suma de ellas se distribuye según una **distribución normal**.

Ejemplo: la variable "tirar una moneda al aire" sigue la distribución de Bernouilli. Si lanzamos la moneda al aire 50 veces, la suma de estas 50 variables (cada una independiente entre si) se distribuye según una distribución normal.

Este teorema se aplica tanto a suma de variables discretas como de variables continuas.
Los parámetros de la distribución normal son:

Media: **n** * μ (media de la variable individual multiplicada por el número de variables independientes)

Varianza: **n** * $\sigma 2$ (varianza de la variable individual multiplicada por el número de variables individuales)

Veamos un **ejemplo**:
Se lanza una moneda al aire 100 veces, si sale cara le damos el valor 1 y si sale cruz el valor 0. Cada lanzamiento es una variable independiente que se distribuye según el modelo de Bernouilli, con media 0,5 y varianza 0,25.
Calcular la probabilidad de que en estos 100 lanzamientos salgan más de 60 caras.
La variable suma de estas 100 variables independientes se distribuye, por tanto, según una distribución normal.

Media = 100 * 0,5 = 50
Varianza = 100 * 0,25 = 25
Para ver la probabilidad de que salgan más de 60 caras calculamos la variable normal tipificada equivalente:

$$Y = \frac{60 - 50}{5,0} = 2,00$$

(*) 5 es la raiz cuadrada de 25, o sea la desviación típica de esta distribución
Por lo tanto:
P (X > 60) = P (Y > 2,0) = 1- P (Y < 2,0) = 1 - 0,9772 = 0,0228
Es decir, la probabilidad de que al tirar 100 veces la moneda salgan más de 60 caras es tan sólo del 2,28%

Ejercicio 1.
La renta media de los habitantes de un país se distribuye uniformemente entre 4,0 millones de pesos y 10,0 millones pesos. Calcular la probabilidad de que al seleccionar al azar a 100 personas la suma de sus rentas supere los 725 millones de pesos.

Cada renta personal es una variable independiente que se distribuye según una función uniforme. Por ello, a la suma de las rentas de 100 personas se le puede aplicar el **Teorema Central del Límite.**

La **media** y **varianza** de cada variable individual es:
$\mu = (4 + 10) / 2 = 7$
$\sigma\,\mathbf{2} = (10 - 4)^2 / 12 = 3$
Por tanto, la suma de las 100 variables se distribuye según una normal cuya **media** y **varianza** son:

Media: **n *** $\mu = 100 * 7 = 700$
Varianza: **n *** $\sigma 2 = 100 * 3 = 300$

Para calcular la probabilidad de que la suma de las rentas sea superior a 725 millones, comenzamos por calcular el valor equivalente de la variable normal tipificada:

$$Y = \frac{725 - 700}{17,3} = 1,44$$

Luego:
P (X > 725) = P (Y > 1,44) = 1 - P (Y < 1,44) = 1 - 0,9251 = 0,0749
Es decir, la probabilidad de que la suma de las rentas de 100 personas seleccionadas al azar supere los 725 millones de pesos es tan sólo del 7,49%

Ejercicio 2.
En una asignatura del colegio la probabilidad de pasar al frente en cada clase es del 10%. A lo largo del año tienes 100 clases de esa asignatura. ¿Cuál es la probabilidad de tener que pasar al frente más de 15 veces?

Se vuelve a aplicar el **Teorema Central del Límite**.
Salir a la pizarra es una variable independiente que sigue el modelo de distribución de Bernouilli:

"Pasar al frente", le damos el valor 1 y tiene una probabilidad del 0,10
"No pasar al frente", le damos el valor 0 y tiene una probabilidad del 0,9
La **media** y la **varianza** de cada variable independientes es:
$\mu = 0,10$
$\sigma\,\mathbf{2} = 0,10 * 0,90 = 0,09$
Por tanto, la suma de las 100 variables se distribuye según una normal cuya **media** y **varianza** son:

Media: **n *** $\mu = 100 * 0,10 = 10$
Varianza: **n *** $\sigma 2 = 100 * 0,09 = 9$

Para calcular la probabilidad de pasar al frente más de 15 veces, calculamos el valor equivalente de la variable normal tipificada:

$$Y = \frac{15 - 10}{3,0} = 1,67$$

O sea:
P (X > 15) = P (Y > 1,67) = 1 - P (Y < 1,67) = 1 - 0,9525 = 0,0475
Es decir, la probabilidad de tener que pasar más de 15 veces a la pizarra a lo largo del curso es tan sólo del 4,75%

Ejercicio 3.
Un día visitamos el Casino y decidimos jugar en la ruleta. Nuestra apuesta va a ser siempre al negro y cada apuesta de 500 ptas. Llevamos 10.000 ptas. y queremos calcular que probabilidad tenemos de que tras jugar 80 veces consigamos doblar nuestro dinero.

Cada jugada es una variable independiente que sigue el modelo de distribución de Bernouilli.
"Salir negro", le damos el valor 1 y tiene una probabilidad del 0,485
"No salir negro", le damos el valor 0 y tiene una probabilidad del 0,515

(*) La probabilidad de "no salir negro" es mayor ya que puede salir rojo o el cero
La **media** y **varianza** de cada variable individual es:

$\mu = 0,485$
$\sigma^2 = 0,485 * 0,515 = 0,25$

A la suma de las 80 apuestas se le aplica el **Teorema Central del Límite**, por lo que se distribuye según una normal cuya **media** y **varianza** son:

Media: **n *** $\mu = 80 * 0,485 = 38,8$
Varianza: **n *** $\sigma^2 = 80 * 0,25 = 20$

Para doblar nuestro dinero el negro tiene que salir al menos 20 veces más que el rojo (20 * 500 = 10.000), por lo que tendrá que salir como mínimo 50 veces (implica que el rojo o el cero salgan como máximo 30 veces).

Comenzamos por calcular el valor equivalente de la variable normal tipificada:

$$Y = \frac{50 - 38,8}{4,5} = 2,50$$

Luego:
P (X > 50) = P (Y > 2,50) = 1 - P (Y < 2,50) = 1 - 0,9938 = 0,0062
Es decir, la probabilidad de doblar el dinero es tan sólo del 0,62%.

Ejercicio 4.
El precio de una acción en bolsa se mueve aleatoriamente entre 10 dólares. y 20 dólares., con la misma probabilidad en todo el tramo. Hemos dado la orden a nuestro broker de que nos compre paquetes de 1.000 acciones cada día durante las próximas 40 sesiones.

Una vez ejecutada la orden, tenemos un total de 40.000 acciones. A final de año vendemos todas las acciones al precio de 13 ptas./acción, recibiendo 520.000 dólares. Calcular la probabilidad de que ganemos dinero en esta operación.

El precio de cada paquete comprado es una variable aleatoria independiente que se distribuye uniformemente entre 10.000 dólares y 20.000 dólares. Su **media** y **varianza** son:

$\mu = (10.000 + 20.000) / 2 = 15.000$
$\sigma^2 = (20.000 - 10.000)^2 / 12 = 833,3$

El precio total de los 40 paquetes comprados se distribuye según una distribución normal cuya **media** y **varianza** son:

Media: **n *** μ = 40 * 15.000 = 600.000
Varianza: **n *** $\sigma 2$ = 40 * 833,3 = 33.333,3

Para estimar la probabilidad de que ganemos dinero, calculamos el valor equivalente de la variable normal tipificada:

$$Y = \frac{520.000 - 600,000}{33.333,3} = 2,40$$

Luego:
P (X > 520.000) = P (Y > 2,40) = 1 - P (Y < 2,40) = 1 - 0,9918 = 0,0082
Por tanto, la probabilidad de que ganemos dinero con la operación es tan sólo del 0,82% .

ESTADISTICA INFERENCIAL

Hasta ahora, hemos estudiado estadística descriptiva, una serie de procedimientos y técnicas, que permitían un conocimiento descriptivo de las características básicas de una *población*.

Pero en general, no podremos casi nunca tratar con la totalidad de la población. Ya sea porque la población a estudiar es muy grande, ya sea por motivos económicos, de falta de personal cualificado, o para una mayor rapidez en la recogida y presentación de los datos, lo que se suele hacer es obtener los datos, de tan sólo una muestra de la población.

En consecuencia, deberemos contentarnos con utilizar muestras, que sean capaces de revelarnos algo acerca de la población de las que han sido extraídas. La Estadística inferencial se ocupa de extender o extrapolar a toda una población, informaciones obtenidos de una muestra, así como de la toma de decisiones basada en dicha información.

TEORIA DE LAS MUESTRAS

El objetivo va a ser a partir de ahora, el tratamiento estadístico de muestras.
¿Pero bajo que condiciones, resulta apropiada una muestra?.Existen una serie de factores que inciden en la respuesta de esta pregunta, y que resultan fundamentales en estadística inferencial.

Una primera cuestión, es el tamaño que ha de tener. Parece evidente, que a mayor tamaño, más se acercaran los parámetros que calculemos, a los de la población (y es cierto siempre que se tenga en cuenta la representatividad de la muestra, que es un aspecto que desarrollaremos ahora). En la práctica real, el número de elementos de una muestra está determinado por una serie de factores: grado de fiabilidad deseado, dificultad en la elección de los elementos que la compongan, tiempo necesario para la elección, gastos originados,...

La segunda y más importante cuestión es ¿cómo deben ser elegidos los elementos que la compongan?. Para ser válidas ,las muestras han de ser representativas, esto es, si queremos inferir de los resultados de una muestra, en ella se ha de reproducir en igual porcentaje el carácter estudiado, que en la población total. Por tanto, será necesario, que en el momento de la elección de los elementos de la muestra, verifiquemos que todos los elementos de la población tiene igual probabilidad de ser elegidos para la muestra.

Cuando no se tienen en cuenta estos dos principios básicos, las inferencias realizadas son deficientes. Existe una variedad de "mentiras estadísticas", procedentes de afirmaciones basadas en pequeñas muestras , o en muestras no representativas. Así por ejemplo ,si se dice "7 de cada 10 dentistas consultados recomiendan el dentífrico X", no debemos inferir que el

70% de los dentistas los recomiendan, hasta saber de que forma fueron elegidos los dentistas consultados ,y cuántos fueron en total.

Las consideraciones referentes al tamaño de la muestra, se estudiarán más adelante. Las referentes a la forma de elegir la muestra, serán estudiadas ahora.

TIPOS DE MUESTREOS

Existen básicamente dos tipos de muestreo, los aleatorios y los no aleatorios.

En los primeros, el aspecto principal, es que todos los miembros de la muestra han sido elegidos al azar, de forma que cada miembro de la población tuvo igual oportunidad de salir en la muestra. Este tipo de muestreo, que es el más consistente, es al mismo tiempo el que resulta más costoso, y el que utilizaremos siempre en el desarrollo de los próximos epígrafes.
Los centros oficiales como el INE, utilizan siempre muestreos aleatorios.

Los segundos, carecen del grado de representatividad de los primeros, pero permiten un gran ahorro en los costes. Se eligen los elementos, en función de que sean representativos, según la opinión del investigador. Es el método que utilizan generalmente las empresas privadas, y presenta el inconveniente de que la precisión de los resultados no es muy grandes, y es difícil medir el error de muestreo.

MUESTREOS ALEATORIOS SIMPLE

Su utilización es muy sencilla, una vez que todos los elementos de la población han sido identificados y numerados (y éste es probablemente su mayor inconveniente). A partir de aquí, decidido el tamaño n de la muestra, los elementos que la compongan se han de elegir aleatoriamente entre los N de la población.

Si queremos elegir una muestra formada por 40 elementos de una población de 600, iremos tomando cifras aleatorias de tres en tres. Si la cifra considerada es menor de 600, ya tendremos elegido un elemento de la muestra. Siguiendo este proceso, y saltándonos las cifras superiores a 600, podremos elegir todos los elementos que compondrán la muestra.

SISTEMÁTICO

Es análogo al anterior, aunque resulta más cómoda la elección de los elementos. Si hemos de elegir 40 elementos de un grupo de 600, se comienza por calcular el cociente 600/40 que nos dice que existen 40 grupos de 15 elementos entre los 600. Se elige un elemento de salida entre los 15 primeros, y suponiendo que sea el k-simo, el resto de los elementos serán los k-simos de cada grupo. En concreto, si el elemento de partida es el número 6, los restantes serán los que tengan los números: 15+6 ,2x15+6,.....,39x15+6

Este procedimiento simplifica enormemente la elección de elementos, pero puede dar al traste con la representatividad de la muestra, cuando los elementos se hayan numerados por algún criterio concreto, y los k-simos tienen todos una determinada característica, que haga conformarse una muestra no representativa.

ESTRATIFICADO

A veces nos interesa, cuando las poblaciones son muy grandes, dividir éstas en subpoblaciones o estratos, sin elementos comunes, y que cubran toda la población. Una vez hecho esto podemos elegir, por muestreo aleatorio simple, de cada estrato, un número de elementos igual o proporcional al tamaño del estrato.

Este procedimiento tiene la gran ventaja de que se puede obtener una mayor precisión en poblaciones no homogéneas (aunque en este curso no estudiaremos los métodos necesarios)
Si decidiéramos hacer una encuesta sobre la incidencia del tabaco en nuestro centro, podríamos razonar de la siguiente forma:

Una facultad tiene 2000 alumnos, 720 en 3º, 700 en 4º, 340 en 1º y 240 en 2º.
Si deseamos tomar una muestra de 100 alumnos, para analizar la incidencia del tabaco en la adolescencia, bastaría tomar un número igual de alumnos de cada estrato, es decir 25.

Si embargo, si lo que se quiere es hacer una encuesta para conocer la opinión que tiene el alumnado sobre una medida que ha tomado el Consejo Escolar, es más representativo elegir de cada estrato, y en número proporcional a su tamaño, los elementos que compondrán la muestra. Si 3º representa al 36% del alumnado, el 36% de la muestra (es decir 36 alumnos) se elegirán de este estrato por muestreo aleatorio simple, 35 para 4º, y así hasta completar los 100 elementos de la muestra.

POR CONGLOMERADOS
A veces, para simplificar los procesos de toma de datos, se empieza por elegir ciertos conglomerados (que pueden ser bloques de viviendas, municipios, urnas electorales,...) y dentro de ellos se realiza el muestreo aleatorio.

TOMA DE DATOS: LA ENCUESTA
Una vez decidido el tamaño y la forma de elegir la muestra, aparece el problema de cómo realizar la toma de datos. La encuesta es el instrumento idóneo para este fín.
Se debe establecer en primer lugar el objetivo de la encuesta, desmenuzando el problema a investigar, eliminando lo que resulte superfluo, y centrándonos en los aspectos más relevantes.
A partir de aquí, se elabora un cuestionario, formado por un conjunto de preguntas que han de ser respondidas por los encuestados.

De la calidad de éste último depende en gran parte el resultado del trabajo. Existen una serie de factores que se han de tener en cuenta a la hora de redactar el cuestionario, entre los que destacan los siguientes:

• Las preguntas han de ser pocas (no más de 30) y cortas.
• Cerradas (es decir que aparezcan todas las posibles repuestas). Si preguntamos a un encuestado si le gustan las matemáticas, no podemos dejar que aparezcan respuestas de todo índole, sino que responda de acuerdo a una escala numérica o de valor. Por ejemplo podemos valorar su gusto de 1 a 5, o bien : Nada, Poco, Normal, Mucho, Muchísimo.
• Numéricas o al menos codificables (es decir que podamos traducir las respuestas a números, por ejemplo asignando números del 1 al 5 a las respuestas del apartado anterior).
• Deben ser redactadas de forma concreta y precisa (sin palabras abstractas o ambiguas), de manera que las repuestas puedan ser inequívocas.

A partir de aquí, debe ser realizado el "trabajo de campo", es decir las entrevistas previstas, por medio de los encuestadores. Este trabajo también ha de hacerse bajo unas ciertas condiciones, que garanticen que las respuestas sean fidedignas.

Una vez recopilados todos los datos, se procede a tabularlos, y describirlos, utilizando las técnicas que ya conoces de cursos anteriores.

Si repasamos nuestros conocimientos sobre el Teorema Central del Límite, veremos que:

Imagina que tienes una población con media μ y desviación típica σ. y que extraes aleatoriamente todas las posibles muestras, todas ellas de tamaño n. Si obtuvieras las medias de todas estas muestras, y las consideras una distribución de datos (la distribución muestral de medias), comprobarías que:

a) La media de los datos, es la media μ de la poblaci**ó**n , es decir la media de las medias de las muestras, es igual que la media de la población.

b) Estas medias se distribuyen alrededor de la media de la población, con una desviación típica (llamada *desviación típica de la media*) igual a la de la población dividida por la raíz de n, es decir, la d.t. de la media es

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

c) La distribución de las medias muestrales, es una distribución de tipo "normal", siempre que la población de procedencia lo sea, o incluso si no lo es, siempre que el tamaño de las muestras sea 30 o mayor.

En consecuencia, "si una población tiene media μ y d.t. σ , y tomamos muestras de tamaño n (de tamaño al menos 30, o cualquier tamaño, si la población es "normal"), las medias de estas muestras siguen aproximadamente la distribución

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \textbf{(1)}$$

Además, cuanto mayor es el valor de n, mejor es la aproximación "normal".

Hemos nombrado un concepto importante: la d.t. de la media $\frac{\sigma}{\sqrt{n}}$, que es el grado de variabilidad de las medias muestrales. Cuanto menor sea, más ajustadas a la media de la población serán las medias que obtengamos de una muestra. De su propia definición, es fácil darse cuenta de que cuanto mayor es el tamaño de la muestra, menor es este grado de variabilidad, y por tanto más similar a la media de la población será la media obtenida de la muestra.

EJEMPLO:
Una compañía aérea sabe que el equipaje de sus pasajeros tiene como media 25 kg. con una d.t. de 6 kg. Si uno de sus aviones transporta a 50 pasajeros, el peso medio de los equipajes de dicho grupo estará en la distribución muestral de medias

$$N\left(25, \frac{6}{\sqrt{50}}\right) = N(25;0'84)$$

La probabilidad de que el peso medio para estos pasajeros sea superior a 26 kg sería:

$$p(X > 26) = p\left(Z > \frac{26 - 25}{0'84}\right) = p(Z > 1'18) = 0'1190 \approx 11'9\%$$

Si el avión no debe cargar más de 1300 kg en sus bodegas, la media del conjunto de los 50 pasajeros no debe superar los

$$\frac{1300}{50} = 26\text{kg}$$

En consecuencia en un 11,9% de los casos los aviones de esta compañía superan el margen de seguridad.

Hemos estudiado ya el T.C.L., que nos permite conocer de que forman se distribuyen las medias de las muestras de una población.

Ahora invertiremos el caso: se selecciona una muestra de una población de la que se desconoce la media, y se calcula la media muestral. A partir de aquí haremos una inferencia sobre la media poblacional, con base en la media muestral.

Imaginemos que preguntamos a una muestra de 40 alumnos, por el recorrido en km. que tienen que hacer todos los días para llegar al instituto, y que la media de tal muestra es de 3 km. Las dos preguntas siguientes responden a las dos formas de inferencia que estudiaremos en este curso:

1º.- Si nos habían dicho que la media de distancia de todo el instituto era el año pasado de 3,8 km, ¿es significativamente diferente esta media?, o lo que es lo mismo, ¿podemos decir que la media del instituto ha cambiado este año, o por el contrario la diferencia de medias es normal y se debe al azar al elegir los elementos de la muestra?
Esta pregunta implica una **decisión**, que podremos tomar a través de los denominados **test de contraste de hipótesis.**

2º.- Tomando como base la muestra (es decir si suponemos que desconocemos la distancia media), ¿qué estimación puede hacerse sobre la media poblacional (es decir la de todo el Instituto) ?

Esta pregunta implica una **estimación**, que aprenderemos a hacer ahora.

ESTIMACIÓN

Llamaremos así al procedimiento utilizado cuando se quiere conocer las características de un parámetro poblacional, a partir del conocimiento de la muestra.

Imaginemos que hemos hecho la encuesta a la que se aludía en el apartado anterior, y queremos saber cual es la verdadera media del instituto. Podemos hacer una primera aproximación, utilizando la media muestral $\bar{x} = 3$ km. Sin embargo , este valor está sesgado debido a que solo representa a una muestra.

Podríamos decir que la media buscada es próxima a 3, pero ¿cuánto de próxima?. ¿Digamos que 200 metros más o menos?. Esto significaría que la media estaría entre 2,8 y 3,2. Esto último se denomina estimar por intervalo, y es el método que ahora vamos a ver.

INTERVALO DE CONFIANZA
Se llama así a un intervalo en el que sabemos que está un parámetro, con un nivel de confianza específico

Si dijéramos que la media se encuentra en el intervalo (2,8 , 3,2) con un nivel de confianza del 95%, lo que decimos es que si hiciéramos muestras de tamaño 40, y fuéramos contabilizando sus medias, a la larga, en el 95% de los casos, la media calculada estaría en dicho intervalo.

Además, al valor 0,2 (200 metros), que mide la mitad de la anchura del intervalo, se le denomina error máximo de la estimación. Lo anteriormente argumentado se expresa en términos estadísticos como:

"A un nivel de confianza del 95%, la media poblacional es 3 km, con un error máximo de estimación de $\pm 0{,}2$ km."

Por lo tanto:

NIVEL DE CONFIANZA
Probabilidad de que el parámetro a estimar se encuentre en el intervalo de confianza.

Los valores que se suelen utilizar para el nivel de confianza son el 95%, 99% y 99,9%

ERROR DE ESTIMACIÓN MÁXIMO
Es el radio de anchura del intervalo de confianza.

Este valor nos dice en qué margen de la media muestral se encuentra la media poblacional al nivel de confianza asignado.

Durante este curso aprenderemos a realizar estimaciones sobre la media y la proporción de una característica en una población. La estimación de otros parámetros poblacionales, tales como la desviación típica, quedará fuera de nuestro estudio.

Estimación de la media de una población

Para estimar la media poblacional por medio de intervalos de confianza, será necesario recordar que el Teorema Central del Límite nos daba información de como se hallaban distribuidas las medias muestrales: "normalmente" con una media igual a la de la población original μ (que es la que ahora tratamos de conocer) y desviación típica

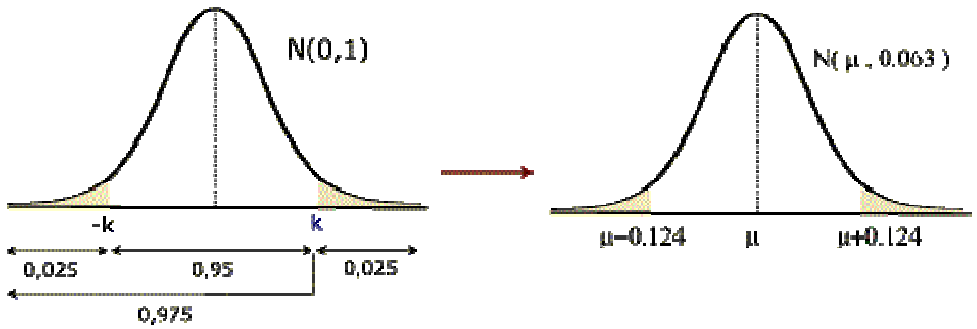
$$\frac{\sigma}{\sqrt{n}}$$

Supongamos que hemos analizado la muestra ya nombrada de media $\bar{x} = 3$ Km., y que sabemos que la desv. típica de la población es de σ=0,4 km., y que nos planteamos estimar la media de todo el instituto, con un nivel de confianza del 95% .El proceso para realizar la estimación es el siguiente:

Sabemos por el T.C.L. que las medias muestrales se distribuyen según

$$N\left(\mu, \frac{0,4}{\sqrt{40}}\right) = N(\mu, 0'063)$$

La siguiente figura nos ilustrará:



Hallamos el valor k de forma que p(-k<Z<k)=0,95 , o lo que es lo mismo p(Z<k)=0,975. Consultando nuestra tabla de la distribución normal, encontraremos que k=1.96 .

Este valor nos dice que la medias muestrales se encuentran en un 95% de los casos como máximo a 1.96 desviaciones típicas de la media buscada, es decir, nuestra media $\bar{x} = 3$, en un 95% de los casos, dista de la media poblacional menos de 1,96.0,063=0,124 km.

Si tomamos un intervalo con centro en dicha media muestral , y radio 0,124, en un 95% de los casos la media buscada estará dentro del intervalo.

Encontramos por tanto que a un nivel de confianza del 95%, la media poblacional es de 3 km. con un error máximo de

$$E = k\frac{\sigma}{\sqrt{n}} = 0,124\,km$$

o lo que es lo mismo, existe una probabilidad del 95%, de que la media buscada se encuentre en el intervalo de confianza (3-0,124 , 3+0,124) = (2,876 , 3,124).

Así pues en general para un proceso de estimación de la media, el intervalo de confianza será:
($\bar{x}$ - E , $\bar{x}$ + E)

siendo $\bar{x}$ la media de la muestra, y $E = k\frac{\sigma}{\sqrt{n}}$ el error de estimación.

TAMAÑO DE LA MUESTRA

Pero imaginemos ahora, que nos disponemos a elegir una muestra para poder determinar con un 95% de confianza la media, con un margen de error de 50 metros. Desde luego hará falta una muestra mayor para tener tan poco margen de error ¿Cuál deberá ser el tamaño de la muestra para conseguirlo? .

Despejando en

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

obtenemos que

$$n = \left(\frac{k\sigma}{E}\right)^2$$

Como k=1,96 , E=0,05 y σ=0,4 calculando obtendremos que n=245,8 es decir, redondeando, hará falta una muestra correspondiente a 246 estudiantes para que el margen de error sea de tan sólo 50 metros.

De la expresión del tamaño de la muestra, se deduce muy fácilmente, que deberá ser mayor cuanto mayor sea:

a) El nivel de confianza asignado
b) El grado de variabilidad de los datos originales

Por el contrario, cuanto mayor sea el tamaño de la muestra, menor será el error de la estimación.

Estimación de la proporción de una población

Como recordarás, la distribución binomial B(n,p), nos permite conocer como se distribuye el número de éxitos, correspondiente a un experimento realizado n veces, y en el que la probabilidad de éxito en cada experimento es p. Dicha distribución tiene media y desviación típica:

$$\mu = np$$
$$\sigma = \sqrt{npq}$$

Supongamos que sea X la variable que mide el número de éxitos. Ya sabes que los posibles valores de X son 0,1,2,...,n. Si utilizaramos la nueva variable,

$$Y = \frac{X}{n}$$

ésta tomaría los valores correspondientes a las proporciones (en tanto por uno) de éxito.
Si por ejemplo n=200, se tendría:

X=0 , (0 éxitos) equivale a Y=0 (es decir un 0% de éxitos)
X=1 , (1 éxito) equivale a Y=0,005 (es decir 0,5% de éxitos)
X=2 , Y=0,01 (es decir 2 éxitos equivalen a un 1% de éxitos)
....
X=n , Y=1 (n éxitos = 100% de éxitos)

Dividiendo por n, obtendremos la media y desviación típica de la variable Y que representa la proporción de éxitos:

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

$$\mu = \frac{np}{n} = p$$

Si además np>5, nq>5, utilizando la aproximación normal a la binomial, podremos afirmar que las proporciones de éxito para un experimento binomial de n pruebas con probabilidad de éxito p en cada prueba, se distribuyen según:

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

Distribución muestral de proporciones

Imaginemos que sabemos que la proporción del alumnado de un centro que es favorable a realizar una huelga es del 60%. Cuando elegimos a un alumno, y nos preguntamos si es favorable a la huelga, es como si realizaramos una prueba binomial con probabilidad de éxito p=0,6.

Cuando elijamos muestras aleatorias de digamos 70 alumnos, el número de ellos favorable a la huelga, deberá seguir una distribución B(70, 0´6), o bien, la proporción de ellos que es favorable se debe distribuir según

$$N\left(0'6, \sqrt{\frac{0{,}6 \times 0{,}4}{70}}\right) = N(0'6, 0'058)$$

(Debe notarse que en este caso, n=70, p=0,6, q=0,4 y por tanto np>5, nq>5), o lo que es lo mismo, las proporciones que vayamos encontrando para muestras de tamaño 70, se irán distribuyendo de forma "normal" alrededor del 60%, con una desviación típica del 5,8%.

Por tanto, si en una población, una determinada característica de tipo binomial (es decir la población se divide entre los que la tienen y los que no), se presenta en una proporción p, al tomar muestras de tamaño n, las proporciones p' obtenidas, se distribuirán según

$$N\left(p, \sqrt{\frac{pq}{n}}\right)$$

(a partir de este momento supondremos siempre que np>5,nq>5). A esta distribución se la denomina *distribución muestral de proporciones.*

EJEMPLO:
En una empresa está establecido que si una máquina opera correctamente, como máximo un 5% de su producción es defectuoso. Si se elige aleatoriamente una muestra de 40 artículos producidos por una máquina y 15 de ellos son defectuosos, ¿existe razón para pensar que la máquina está averiada?.

Las proporciones muestrales para muestras de tamaño 40 en una máquina normal se distribuyen según

$$N\left(0'05, \sqrt{\frac{0{,}05 \times 0{,}95}{40}}\right) = N(0'05, 0'034)$$

, es decir se distribuyen de forma "normal" alrededor del 5% con una d.t. del 3'4%.
En consecuencia, la probabilidad de valores como el registrado

$$\frac{15}{40} = 0'375 \approx 37'5\%$$

resulta ser:

$$p(Y > 0'375) = p\left(Z > \frac{0'375 - 0'05}{0'034}\right) = 0$$

y podemos asegurar "estadísticamente" que la máquina está averiada.

Ahora que sabemos como se distribuyen las proporciones muestrales, por un proceso similar al utilizado para estimar la media poblacional, podremos realizar estimaciones sobre la proporción poblacional de un carácter, conociendo la proporción en una muestra.

Estimación de una proporción

Imaginemos que hemos tomado una muestra aleatoria de 500 personas, y que les preguntamos si creen que el Presidente del Gobierno debe dimitir, obteniendo el SÍ un 70%. Supongamos que nos planteamos un intervalo de confianza del 90% para poder estimar el porcentaje p de toda la población que diría SÍ.

Según todo lo dicho, las proporciones del SÍ en las muestras, se distribuirán según:

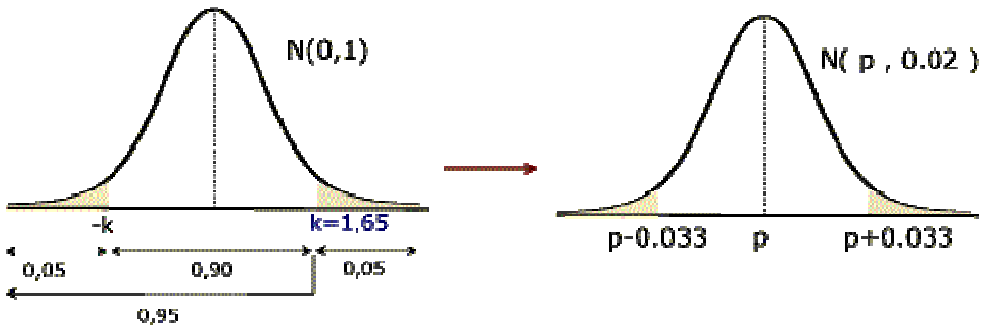
$$N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Como quiera que no conocemos la verdadera proporción p, no podemos conocer la desviación típica de la distribución muestral

$$\sqrt{\frac{pq}{n}}$$

por lo que utilizaremos como sustituto para p, la proporción muestral p'=0,7, que causará poco cambio en los resultados finales.

En consecuencia, las proporciones muestrales, siguen la distribución N(p,0,02) (Nota: puesto que utilizamos tantos por uno, deberemos utilizar en los cálculos una precisión de al menos centésimas, mejorando el resultado si precisamos más)



Llevando a cabo los mismos pasos que en el caso de la estimación de medias, vemos que un 90% de las proporciones muestrales que se obtengan estarán a como máximo 1,65 desviaciones típicas de p, es decir a

$$\pm k\sqrt{\frac{p'q'}{n}} = \pm 0'033$$

y en consecuencia, si suponemos que p' es una de tales proporciones (y será acertado suponerlo en un 90% de los casos), la verdadera proporción quedará siempre en el intervalo (p'-0'033 , p'+0'033)=(0'667,0'733).

Esto lo podemos expresar como: "Con un nivel de confianza del 90%, la proporción ciudadanos que creen que el Presidente del Gobierno debe dimitir es de un 70%, con un error máximo de $\pm$ 3,3 % "

La estimación de proporciones es de gran importancia en la vida cotidiana, dado que influyen por ejemplo en la programación de la tv, los productos que consumimos, las leyes que se legislan,.....

En los periódicos, revistas, televisión y los informativos de radio, es muy corriente que se den informes de encuestas.

Sin embargo frecuentemente, se dan porcentajes, sin ninguna indicación del grado de confianza, el margen de error o el tamaño de la muestra. Sin conocer estos datos, no podemos tener una idea clara de la calidad de los resultados obtenidos, por lo que deberías siempre de tratar de conocer la ficha técnica de estos estudios.

Tamaño de la muestra

Como ya sabemos, el error máximo depende del tamaño de la muestra: a muestras mayores corresponden errores menores.

Normalmente, cuando queremos hacer una estimación, con un determinado margen de confianza, nos plantearemos que el error máximo tenga un determinado valor.

Imaginemos por ejemplo que queremos conocer el porcentaje de alumnos de nuestro centro , que es favorable a hacer la Fuga de San Diego el día 12 de Noviembre (este carácter se considerará como éxito) en contraposición con los que la quieren hacer en otra fecha. Nos marcamos un nivel de confianza del 90%, y queremos que el error máximo no sobrepase el 10%.

Puesto que el error máximo es

$$E = k\sqrt{\frac{pq}{n}}$$

,
el tamaño de la muestra habrá de ser

$$n = \frac{k^2 pq}{E^2}$$

Existe un problema: no conocemos p, ni tan siquiera el valor p' de la muestra puesto que aún no ha sido realizada la encuesta (a no ser que por anteriores sondeos, pueda tenerse un valor fiable para p).

Si se tiene información previa sobre el valor de p, puede utilizarse, pero si no, se utilizará inicialmente p=0,5, pues se puede demostrar que para este valor se obtiene el máximo valor del tamaño de la muestra (mirar grafico siguiente) y en consecuencia, quedará asegurado que el error es como máximo del 10%

En este caso concreto, tomando E=0,1 , p=0,5 , k=1,65, obtendremos que $n=68,08 \approx 69$ es el tamaño de la muestra que debemos tomar.

Aunque el error máximo fijado es del 10%, en la práctica resultará en general más pequeño, a medida que la verdadera proporción p se aleje del valor 0,5.

En particular, si en lugar de tomar inicialmente p=0,5 , hubiéramos supuesto que p=0,95 , el error máximo que cometeríamos utilizando 68 personas en la muestra sería: E= 0,043, es decir un 4,3%. Una vez estimado p, podremos reajustar el margen de error cometido.

En la práctica normalmente no dispondremos de información previa sobre el valor de p, y deberemos partir de p=0,5.

EJEMPLO:

Imagina que queremos estimar con un error máximo del 3%, el porcentaje de audiencia de un programa de TV, y queremos un 95% de confianza para nuestros resultados. No disponemos de información previa sobre el posible valor de p. ¿Cuántos telespectadoeres deberán ser encuestados?

Para un nivel de confianza del 95% deberemos tomar k=1,96.

Puesto que desconocemos p , tomaremos p=0,5, con lo que n=1068 (redondeado).

Tenemos pues un 95% de confianza en que el porcentaje que encontremos se halle a menos de tres puntos porcentuales de la proporción exacta. Teniendo en cuenta que este número de telespectadores es muy pequeño respecto del total de telespectadores, nos daremos cuenta de la potencia del método de estimación.

TEST DE CONTRASTE DE HIPOTESIS

INTRODUCCIÓN

Veremos ahora la forma de tomar una decisión en base a datos estadísticos, controlando el margen de error que podemos cometer.

Supongamos que una empresa privada, decide otorgar una premio a aquellos centros, en los que la nota media de una prueba realizada por los alumnos supere los 7 puntos.

Como no puede (por razones económicas, de tiempo, disponibilidad, etc) realizar la prueba en todos los alumnos en cada facultad, decide elegir una muestra aleatoria de 45 alumnos de cada facultad, y que sean ellos los que realicen la prueba.

Imagina que en nuestra facultad, se han obtenido los siguientes resultados: $\bar{x} = 7'9 \quad s = 2'95$ (recuerda que s podía considerarse un buen sustituto de la desviación típica de la población, y que por tanto a partir de ahora asumiremos que σ=2'95)

Ahora, la empresa se plantea la siguiente duda, ¿puede afirmar con seguridad que la media de la facultad es superior a 7, o por el contrario el resultado obtenido se debe al azar en la elección de la muestra?.

Nuestra facultad, dado su convencimiento de merecer el premio, propone el siguiente proceso: Para probar que **" la media μ es superior a 7 "** (1), supondremos en principio lo contrario, es decir que **" la media es menor o igual que 7 "** (2), y veremos en términos probabilísticos la posibilidad de que esto último ocurra. Llegan al acuerdo de que si la probabilidad de que " la media sea menor o igual a 7 " es menor del 5%, se aceptará la hipótesis de la facultad y se concederá el premio.

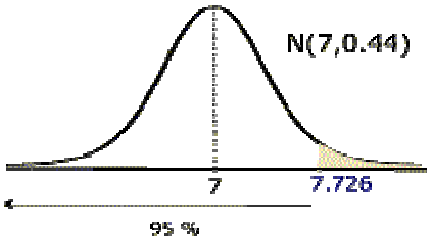
La facultad argumenta lo siguiente:

Si la hipótesis (2) fuera cierta, es decir, la media menor o igual a 7,en el caso extremo la media sería 7, y la distribución muestral de medias sería N(7, 0'44).

Si esto es así, en como mínimo (*) el 95% de los casos, la media muestral habría de ser menor que el valor t=7,726 para el que se verifica que

$p(\bar{X} > t) = 0'05$

Este valor t se obtiene buscando en primer lugar la puntuación típica k para la que p(Z<k)=0,95 , que resulta ser k=1,65. Los valores que se encuentran a más de 1,96 desviaciones de la media, es decir, superiores a t=7+1,65x0,44=7,726 son los que forman la región crítica, es decir las notas medias que tienen una probabilidad de producirse menor del 5%.



Podría ocurrir que la hipótesis (2) fuera cierta y la media muestral 7,9 perteneciera a esa distribución y fuera un valor correspondiente a la región crítica (y la probabilidad de que ello ocurra es del 5%), o bien que lo que ocurra realmente, es que (2) sea falsa, y la media obtenida pertenezca a una distribución muestral con media μ superior (por ejemplo 7,5), con lo cual tal valor no sería tan raro.

En estadística, "se apuesta" a lo que tiene mayor probabilidad de ocurrir, por lo que se considera que la segunda elección es la correcta. (aunque nunca podremos saber si lo que realmente sucede es esto).

Puesto que suponiendo que la media poblacional es como máximo 7 en al menos 95 de cada 100 muestras la media muestral debería de ser menor que 7,726, y dado que la media muestral obtenida fue 7,9 (que se encuentra en la región crítica), el centro concluye que:

"Con un nivel de significación del 5%, (probabilidad de equivocarnos al rechazar que la media pueda ser menor o igual a 7), existe evidencia suficiente de que la media del centro es superior a 7 ".

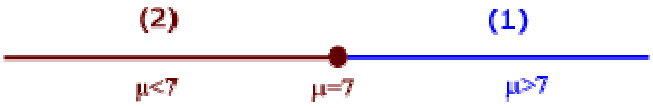
Si el nivel de significación fuera menor, la región crítica disminuiría, y tendremos más confianza en una decisión de rechazo de la hipótesis nula (**)

Si hubiéramos obtenido de la muestra que $\bar{x} = 7'7$, al nivel de significación especificado no podríamos rechazar que realmente la media del centro fuera inferior a 7, es decir., "no existiría evidencia suficiente de que la media fuera superior a 7". Es evidente que al no rechazar que la media poblacional sea menor o igual a 7, también estaríamos arriesgándonos a cometer un error.

En cualquier caso, lo que hacemos es tomar una decisión, una vez vistas las evidencias (datos obtenidos de la muestra), y asumido un margen de error para nuestra decisión.

ELEMENTOS DE LOS TESTS DE HIPÓTESIS

El proceso que hemos descrito en el apartado anterior se denomina **"test de contraste de hipótesis"**, y ahora detallaremos de forma más precisa, los elementos que intervienen en él.
En primer lugar se han de hacer dos hipótesis (1) y (2) que barran el conjunto de posibilidades para la media (o en general el parámetro poblacional sobre el que se quiere tomar una decisión). En el caso estudiado fue:



A la hipótesis (2) que en principio se consideró cierta, se la denomina **hipótesis nula (H_0)** ,por ser el punto de partida, y siempre ha de incluir una igualdad . Esta es la hipótesis que se trata de contrastar, de forma que al final del proceso, la rechazaremos o no.

A la hipótesis (1) que es complementaria de la (2), se la denomina **hipótesis alternativa (H_A)**
El rechazo de la hipótesis nula lleva emparejado la aceptación de la hipótesis alternativa.

Cuando se lleva a cabo un test de contraste de hipótesis, se ha de comenzar por establecer las hipótesis nula y alternativa, recordando que la hipótesis nula ha de contener obligatoriamente una igualdad.

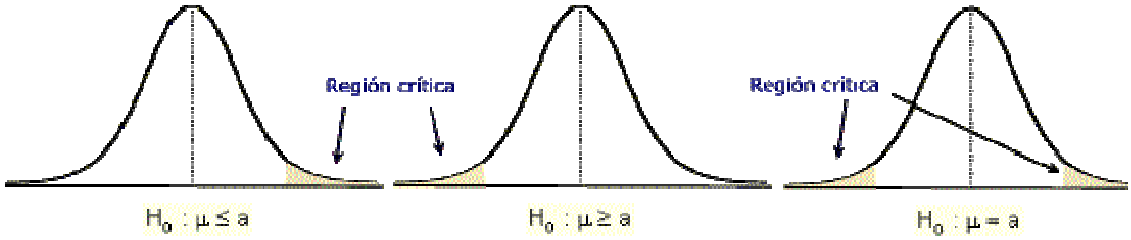
Por lo general, se establece como hipótesis alternativa, la que trata de probar algo que significa un cambio sobre lo que se encuentra preestablecido (por resultados anteriores al test o por inercia) y que está representado por la hipótesis nula. La hipótesis nula es siempre conservadora, frente a la alternativa que propugna el cambio.

Establecidas las hipótesis nula y alternativa, Se toma la muestra, y se calculan los datos necesarios para el contraste, en nuestro caso, la media, y la desviación típica muestral
En segundo lugar se establece el nivel de significación que es la probabilidad de que rechacemos la hipótesis nula, siendo en realidad cierta. Utilizaremos la letra α para

denominarlo. Este nivel de significación es la cantidad de error que nos podemos permitir, y su elección depende en cada caso de la persona que realiza el test. Los más usuales son 10%, 5%, 1% , 0,1%. Se le denomina error de tipo I

Puede también ocurrir que no rechacemos la hipótesis nula, y sea en realidad falsa. Este tipo de error denominado de tipo II y denotado con la letra β, es un error que va directamente ligado al valor α.

Para este nivel de significación habrá de estudiarse la región crítica asociada. En el caso anterior, dado que la hipótesis nula establece que la media es igual o inferior a 7, la región crítica queda a la derecha. Cuando la hipótesis nula establezca que la media es igual o superior a un valor, la región crítica quedará a la izquierda. Por último, si la hipótesis nula establece que la media tiene un valor determinado, la región crítica se habrá de establecer a ambos lados, de forma que el área total que ocupen las dos subregiones sea igual al nivel de significación:



Se estudia para el nivel de significación dado, si se puede rechazar o no la hipótesis nula. Esto se hace viendo si la media obtenida se encuentra dentro de la región crítica asociada al nivel de significación, o si por el contrario, está fuera.

Si "se rechaza la hipótesis nula", la conclusión debe ser redactada:
"Existe evidencia suficiente al nivel de significación α para indicar que ..(significado de la hipótesis alternativa)"

Si por el contrario la decisión es "no se puede rechazar la hipótesis nula", la conclusión debería ser redactada:

"No existe suficiente evidencia al nivel de significación α que indique que ...(significado de la hipótesis alternativa)"

Veremos ahora varios ejemplos que nos ilustrarán sobre el proceso y los diferentes casos que pueden presentarse.

EJEMPLO 1:

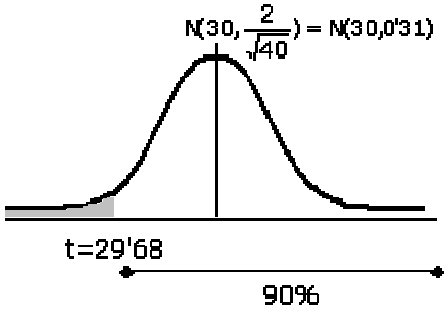
El instituto cree poder probar que la edad media de los alumnos del turno de Noche es inferior a los 30 años. Se ha tomado una muestra de 40 alumnos, y ha resultado que la media es 29,5 y la desviación típica muestral es s=2.

Se deberá en primer lugar establecer las hipótesis nula y alternativa, que deberían ser:

$H_0 : \mu \geq 30$
$H_A : \mu < 30$

En segundo lugar elegimos nivel de significación. Dado que no es demasiado grave equivocarse, se elige un nivel del 10%.

Razonando de forma similar al ejemplo anterior, la región crítica correspondiente a un 10% de significación, sería la que correspondiese a la figura:



Donde

$$t = 30 - 1'28 \times 0'31 = 29'68$$

y k=1'28 es la puntuación típica asociada a un 10% de significación.

Puesto que la media muestral 29,5 está dentro de la región crítica, tendremos que rechazar la hipótesis nula, y por tanto:

"A un nivel de significación del 10%, existe evidencia suficiente de que la media de edad en el turno de noche es inferior a 30 años"

EJEMPLO 2:

Un estudiante, ha leído en la prensa, que el coste medio de un menú en las cafeterías de Las Palmas es de 500 pesos. Como no está conforme, hace un test de hipótesis, para tratar de probar que no es así.

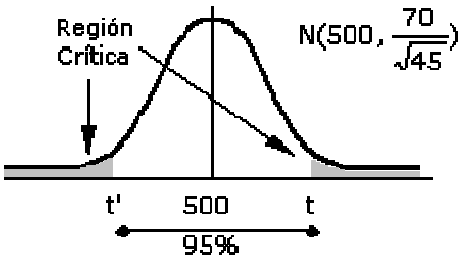
Establece como hipótesis:

H_0: $\mu = 500$

H_A: $\mu \neq 500$

Fija un nivel de significación del 5%, y obtiene una muestra aleatoria de 45 cafeterías, obteniendo como media 518 pesos, y s=70 pesos.

La región crítica asociada a este nivel de significación para las hipótesis planteadas sería:



Ahora k=1'96 y por tanto $$t = 500 + 1{,}96 \times \frac{70}{\sqrt{45}} = 520'45$$

, y $$t' = 500 - 1{,}96 \times \frac{70}{\sqrt{45}} = 479{,}55$$

En consecuencia, no puede rechazarse a este nivel de significación la hipótesis nula y por tanto:

"A un nivel de significación del 5% no existe evidencia suficiente de que la media de precios sea diferente de 500 pesos."

De hecho, esto no significa que sea cierta la hipótesis nula, sino sólo que no se puede rechazar a este nivel de significación. Si hubiéramos tomado un nivel de significación del 10%, la región crítica correspondiente habría estado delimitada por los valores 482,78 y 517,22, con lo que habríamos rechazado la hipótesis nula para ese nivel de significación.

De la misma forma que hemos estado realizando tests sobre medias, pueden ser realizados tests sobre otros parámetros de una población. En particular resulta muy interesante hacerlo sobre una proporción en una determinada población. Veremos ahora un ejemplo de como hacerlo:

EJEMPLO 3:

Antonio dice a Luis que al menos un 15% de los alumnos del Instituto, tiene una moto. Como discrepan, Luis realiza una encuesta aleatoria a 200 compañeros del Instituto, y encuentra que 18 de ellos tiene moto. A un nivel de significación del 10%, ¿cual de los dos tiene estadísticamente la razón?

Establecemos la hipótesis nula y alternativa.

$$\begin{cases} H_0: & p \geq 0.15 \\ H_A: & p < 0.15 \end{cases}$$

Encontramos que la proporción buscada en la muestra es p'= 18/200=0,09.
Supongamos que H_0 es cierta, y que por tanto en el peor de los casos sería p=0,15. Sabemos que si así fuera, las proporciones muestrales, se habrían de distribuir según:

$$N(0'15, \sqrt{\frac{0'15 \times 0'85}{200}}) = N(0'15, 0'0252)$$

Puesto que a un nivel de significación del 10%, la región crítica es la correspondiente a valores menores que k=0,15-1,28x0,0252=0,118 , ésta la forman los porcentajes inferiores al 11,8%. El porcentaje obtenido en la muestra queda dentro de esta región y por tanto rechazamos la hipótesis nula, redactando la conclusión como:

"A un nivel de significación del 10%, existe suficiente evidencia de que la proporción de alumnos con bicicleta es inferior al 15%".

Aunque el resultado dé la razón a Luis, podemos habernos equivocado (con una probabilidad del 10%), . Si hubiera sido otro el resultado, y le hubiéramos dado la razón a Antonio, también podríamos habernos equivocado (recuerda el error de tipo II).

EJERCICIOS

1.- La directiva del Club Las Palmas, alega que en una escala de 1 a 10, sobre satisfacción de los socios, la puntuación que obtiene el club es mayor o igual a 5. Tú has hecho una encuesta a 50 socios elegidos al azar y obtienes una puntuación media de 4,5 (s=0,5). ¿Podrías afirmar ante un periodista, que a un nivel de significación del 1%, existe evidencia suficiente de que la puntuación media es menor?

2.- Una compañía de seguros calcula las primas de seguro de incendios en función de la distancia a la estación de bomberos. Para un barrio, estiman que como media, esta distancia es superior a 5 km. Según los miembros de la asociación de vecinos en cambio, la media no llega a 5 km, por lo que hacen una encuesta aleatoria de 64 viviendas del barrio, obteniendo

como media 4,5 km. Suponiendo que s=2,5 km, ¿proporciona la muestra suficiente evidencia para respaldar la opinión de los vecinos, a un nivel de significación del 5%?

3.- Has sido nombrado director de personal de una gran compañía, y se requiere de tí que establezcas el número medio por empleado de días de baja laboral. Has realizado un estudio basado en 40 empleados elegidos aleatoriamente, y obtienes una media de 16 días por año, con una desviación típica muestral de 2,4 días. ¿Podrías decir a tus superiores que la media es de 18 días con un nivel de significación del 5%?

4.- La intendencia departamental estima que el nº medio de hijos para las familias que residen en un determinado barrio es menor o igual a 1,54. Sin embargo una asociación de vecinos, no está de acuerdo, y solicita tus servicios para conseguir demostrar que ello no es así. Describe el proceso que seguirías para poder conseguirlo.

5.- Se asegura, que el peso medio de las ovejas de un lote es de 54,4 kg. Uno de los productores no cree que esto sea correcto, por lo que reúne una muestra aleatoria de 100 ovejas, obteniendo una media muestral de peso de 53,75 kg (s=5,4kg). A un nivel de significación del 5%, ¿ se puede rechazar que el peso medio sea de 54,4 kg. ?

6.- Un fabricante de lámparas utilizadas por un gran Centro Comercial, asegura que la vida útil de sus lámparas es por lo menos de 1.600 horas. El Jefe de mantenimiento del Centro Comercial, que no estaba de acuerdo, hizo un seguimiento sobre la duración de 100 lámparas seleccionadas aleatoriamente. ¿Respalda una media muestral de 1.562,3 horas su parecer de que la duración efectiva es menor que 1.600 horas a un nivel de significación del 2%? (Supóngase que la desviación típica poblacional es de 150 horas) (P.A.U. 1996)

7.- Una empresa comercializa una bebida refrescante, en un envase en cuya etiqueta se puede leer: "Contenido 250 cc". La Dirección de Defensa del Consumidor toma aleatoriamente 36 envases, y estudia el contenido medio, obteniendo una media de 234 cc y una desviación típica muestral de 18 cc. ¿Puede afirmarse con un 1% de significación que se está estafando al público? (Consideraremos estafa que el contenido sea menor que el expresado en la etiqueta) (P.A.U. 1996).

INDICE