Teoría de muestras

ESTADÍSTICA DESCRIPTIVA
VARIABLES ALEATORIAS
TEORÍA DE MUESTRAS I

Matemáticas 2º de Bachillerato Ciencias Sociales

TEMA 1.- INTRODUCCIÓN A LA ESTADÍSTICA DESCRIPTIVA

1.- CONCEPTOS BÁSICOS

La ***estadística descriptiva*** es una ciencia que analiza series de datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, 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: nacionalidad, color de la piel, sexo).
- *Variables cuantitativas*: tienen valor numérico (edad, precio de un producto, ingresos anuales).

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 niños de una clase, cada alumno es un individuo; si estudiamos el precio de la vivienda, cada vivienda es un individuo.
- ***Población*****:** conjunto de todos los individuos (personas, objetos, animales, etc.) que 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.

2.- TABLAS DE FRECUENCIAS

La ***distribución o tabla de frecuencias*** 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(X_i)	**Simple(n_i)**	**Acumulada(N_i)**	**Simple(f_i)**	**Acumulada(F_i)**
X_1	n1	n1	f1 = n1 / n	f1
X_2	n2	n1 + n2	f2 = n2 / n	f1 + f2
...	...	...	...	...
X_{n-1}	n-1	n1 + n2 +...+ n-1	FN-1 = n-1 / n	f1 + f2 +...+fn-1
X_n	n	S n	fn = n / n	S 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 el número de hermanos (variable discreta) de 25 alumnos de una clase, obteniendo:

Variable	**Frecuencias absolutas**		**Frecuencias relativas**	
X_i	**(n_i)**	**(N_i)**	**(f_i)**	**(F_i)**
0	10	10	40%	40%
1	8	18	32%	72%
2	4	22	16%	88%
3	2	24	8%	94%
4	1	25	4%	100%

Si la variable es continua o toma muchos y diversos valores, conviene agruparlos por intervalos.

Por ***ejemplo***:

Se ha recabado información sobre la paga semanal que reciben 40 jóvenes de un instituto, obteniendo:

	Variable	Frecuencias absolutas		Frecuencias relativas	
Intervalos	(marca de clase) X_i	(n_i)	(N_i)	(f_i)	(F_i)
[0, 5)	2.5	1	1	0.025	0.025
[5, 10)	7.5	1	2	0.025	0.050
[10, 15)	12.5	3	5	0.075	0.125
[15, 20)	17.5	3	8	0.075	0.200
[20, 25)	22.5	3	11	0.075	0.2775
[25, 30)	27.5	6	17	0.15	0.425
[30, 35)	32.5	7	24	0.175	0.600
[35, 40)	37.5	10	34	0.250	0.850

3.- MEDIDAS DE CENTRALIZACIÓN

Son medidas (valores) que informan sobre los valores medios de la serie de datos. La más habitual es la media aritmética:

La ***media aritmética*** es el valor obtenido al sumar todos los datos y dividir el resultado entre el número total de datos:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{x_1 n_1 + x_2 n_{21} + x_3 n_3 + \ldots + x_n n_{n1}}{N}$$

Los x_i son los valores de la variable en las discretas o las marcas de clase en las continuas agrupadas por intervalos.

Ejemplo:

En un test realizado a un grupo de 42 personas se han obtenido las puntuaciones que muestra la tabla. Calcula la puntuación media.

Puntuación	x_i	n_i	$x_i \cdot n_i$
[10, 20)	**15**	**1**	**15**
[20, 30)	**25**	**8**	**200**
[30,40)	**35**	**10**	**350**
[40, 50)	**45**	**9**	**405**
[50, 60	**55**	**8**	**440**
[60,70)	**65**	**4**	**260**
[70, 80)	**75**	**2**	**150**
		42	**1 820**

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{15 \cdot 1 + 25 \cdot 8 + ... + 75 \cdot 2}{42} = \frac{1820}{42} = 43.33$$

4.- MEDIDAS DE DISPERSIÓN

Estudian la distribución de los valores de la variable, analizando si éstos se encuentran más o menos concentrados, o más o menos dispersos.

La más habitual es la ***varianza***.
Mide la distancia existente entre los valores de la serie y la media. Se calcula como sumatorio de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. El sumatorio obtenido se divide por el tamaño de la muestra:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot n_i}{N}$$

Para simplificar el cálculo de la varianza, se usa la fórmula:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{N} - \overline{x}^2$$

Calculamos la varianza de la tabla anterior:

Puntuación	x_i	n_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
[10, 20)	**15**	**1**	**15**	**225**
[20, 30)	**25**	**8**	**200**	**5000**
[30,40)	**35**	**10**	**350**	**12 250**
[40, 50)	**45**	**9**	**405**	**18 225**
[50, 60	**55**	**8**	**440**	**24 200**
[60,70)	**65**	**4**	**260**	**16 900**
[70, 80)	**75**	**2**	**150**	**11 250**
		42	**1 820**	**88 050**

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 \cdot n_i}{N} - \bar{x}^2 = \frac{88050}{42} - (43.33)^2 = 218.94$$

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.

El problema de la varianza es que no viene expresada en las mismas unidades de la variable, sino en cuadrados. Así, la varianza del ejemplo anterior serían 218.94 puntos cuadrados, lo cual no tiene mucho sentido

Para solucionarlo, se define la ***desviación típica*** como la raíz cuadrada positiva de la desviación típica, es decir:

$$S = +\sqrt{S^2}$$

En el caso anterior: $S = +\sqrt{S^2} = \sqrt{218.94} = 14.79$ puntos

Cuanta más pequeña sea la desviación típica mayor será la concentración de datos alrededor de la media

Además, se puede ver que en una distribución simétrica, la mayoría de los datos (aproximadamente el 68%) están en el intervalo

$$(\bar{x} - s, \bar{x} + s)$$

Es decir, si vemos el ejemplo anterior:
$\bar{x} = 43.33$; $s = 14.79$

Lo que significa que la mayoría de las personas a las que se les ha hecho el test ha obtenido una puntuación comprendida en el intervalo (43.33 – 14.79 , 43.33 + 14.79) = (28.54 , 58.12)

Es decir, lo “normal” es obtener en ese test una puntuación dentro de ese intervalo, siendo “raro” una persona que obtenga por ejemplo un 65, o una que obtenga un 20.

EJERCICIOS

1.- En una población de 25 familias se ha observado el número de vehículos que tienen obteniéndose los siguientes datos:
0, 1, 2, 3, 1, 0, 1, 1, 1, 4, 3, 2, 2, 1, 1, 2, 2, 1, 1, 1, 2, 1, 3, 2, 1

a) Obtén la tabla de frecuencias correspondiente
b) Calcula e interpreta su media y su desviación típica

2.- La siguiente tabla representa la duración (en horas) de un determinado número de pilas eléctricas:

Duración	[0,10)	[10,20)	[20,30)	[30,40)	[40,50]
Nº pilas	8	9	4	***a***	4

a) Calcular el valor de ***a*** sabiendo que la duración media de las pilas es de 21 horas
b) Calcular e interpretar la desviación típica

3.- Dada la siguiente distribución referida a los beneficios anuales de un conjunto de empresas.

Beneficio (Miles €)	Nº empresas
230-280	5
280-330	7
330-580	14
580-630	9
630-780	3

a) Calcula el beneficio medio de dicho conjunto de empresas
b) Una empresa que gana 570.000 €, ¿entra dentro de lo normal? ¿Y una que gana 400.000 €?

4.- a) Completar los datos que faltan en la siguiente tabla estadística:

x_i	n_i	N_i	f_i
1	4		0,08
2	4		
3		16	0,16
4	7		0,14
5	5	28	
6		38	
7	7	45	
8			

b) Calcula la media y la distribución típica de esta distribución

5.- Los pacientes que acuden a una consulta médica se distribuyen, según la edad, en una tabla:

X(edad)	[0, 10)	[10, 20)	[20,30)	[30, 40)	[40, 50)	[50,60)
N (frecuencia)	7	10	30	18	12	3

¿Cuál es el intervalo de edad en el que se encuentran la mayoría de los pacientes de la consulta?

TEMA 2.- VARIABLES ALEATORIAS

1.- DEFINICIÓN

Consideremos un experimento aleatorio y sea E el espacio muestral asociado.
Llamamos ***variable aleatoria*** o variable estocástica, X, a toda aplicación que asocia a cada elemento del espacio muestral, E, un número real:

Ejemplo 1. Si lanzamos tres monedas al aire y X es el número de caras que salen, los valores que toma X son 0, 1, 2 y 3.

Ejemplo 2. Si de una camada de 6 cachorros se cuenta el nº de hembras que se "obtienen" la variable aleatoria toma los valores x =0, x=1,....x =6

Ejemplo 3. Al extraer una bombilla de una población y observar si es o no defectuosa, X tomaría los valores 1 y 0 según sea o no defectuosa.

En los ejemplos anteriores se habla de ***variable aleatoria discreta*** (toma valores discretos)

Ejemplo 4. Si se toma como X la estatura de los alumnos de un determinado curso puede tomar todos los valores (dentro de unos límites, es decir, dentro de un intervalo)

Ejemplo 5. Si se toma como variable la duración de las bombillas de una determinada marca puede tomar todos los valores de un intervalo.

En estos casos últimos se hablará de ***variable aleatoria continua***.

Estudiaremos dos modelos de variable aleatoria: uno discreto $\rightarrow$ la binomial
otro continuo $\rightarrow$ la normal

2.- VARIABLE ALEATORIA DISCRETA

Es aquella que sólo puede tomar determinados valores aislados.
Por ejemplo: X = "suma de las caras superiores en el lanzamiento de dos dados"

Se llama ***función de probabilidad*** de una v.a. discreta a la aplicación que asocia a cada valor x de la v.a. X su probabilidad p: $f(x_i) = P(X = x_i)$
Los valores que toma una v.a. discreta X y sus correspondientes probabilidades suelen disponerse en una tabla con dos filas o dos columnas llamada tabla de distribución de probabilidad:

X	$x_1 \quad x_2 \quad x_3 \qquad x_n$
$P(X = x_i)$	$p_1 \quad p_2 \quad p_3 \qquad p_n$

En toda función de probabilidad se verifica que $p_1 + p_2 + p_3 + \qquad + p_n = 1$

***Ejemplo 1*:**
La v.a. "número de caras en el lanzamiento de tres monedas" tiene la siguiente función de probabilidad:

X = Nº de caras	0	1	2	3
$f(x_i) = P(X = x_i)$	$1/8$	$3/8$	$3/8$	$1/8$

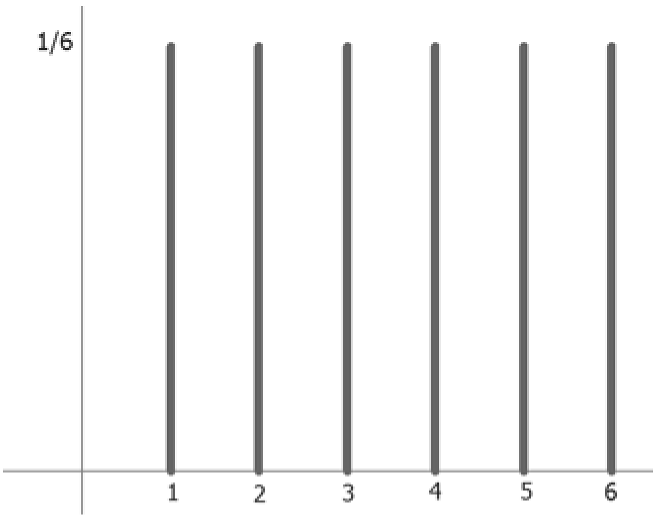
***Ejemplo 2*:**

La v.a. "puntuación obtenida al lanzar un dado" tiene la siguiente función de probabilidad:

X = Puntuación	1	2	3	4	5	6
$f(x_i) = P(X = x_i)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

Ejercicio:
Calcular la función de probabilidad de la variable X = "suma de las caras superiores en el lanzamiento de dos dados"

La representación de una distribución discreta de probabilidad es un diagrama de barras. En el caso del lanzamiento de un dado sería:



Se llama ***función de distribución*** de la v.a. discreta X a la función que asocia a cada valor de la v.a. la probabilidad acumulada hasta ese valor, es decir, $F(x_i) = p(X \leq x_i)$

***Ejemplo 1*:**
La v.a. "número de caras en el lanzamiento de tres monedas" tiene la siguiente función de distribución:

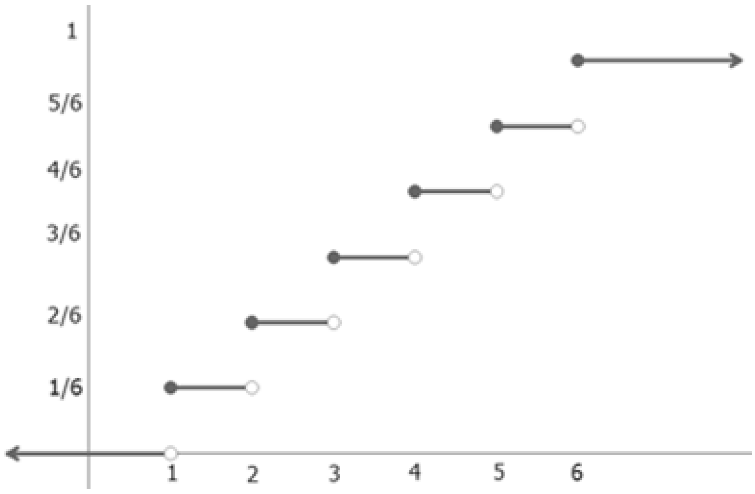
X = Nº de caras	0	1	2	3
$F(x_i) = p(X \leq x_i)$	$1/8$	$4/8$	$7/8$	1

***Ejemplo 2*:**

La v.a. "puntuación obtenida al lanzar un dado" tiene la siguiente función de probabilidad:

X = Puntuación	1	2	3	4	5	6
$F(x_i) = p(X \le x_i)$	$1/6$	$2/6$	$3/6$	$4/6$	$5/6$	1

La representación de una función de distribución es una gráfica escalonada:



Se llama ***media o esperanza matemática*** de una v.a. discreta X, que toma los valores $x_1, x_2, x_3 .... x_n$ con probabilidades $p_1, p_2, p_3 .... p_n$ al valor de la siguiente expresión:

$$E(X) = \mu = \sum x_i . p_i$$

La ***varianza*** viene dada por la siguiente fórmula:

$\sigma^2 = \sum (x_i - \mu)^2 . p_i$, o bien $\sigma^2 = \sum x_i^2 . p_i - \mu^2$

La ***desviación típica*** , σ , es la raíz cuadrada de la varianza.

Ejemplo:
Calcular la esperanza matemática, la varianza, y la desviación típica, de la distribución de probabilidad de las puntuaciones obtenidas al lanzar un dado.

Solución:

x_i	p_i	$x_i \cdot p_i$	$x_i^2 \cdot p_i$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$
3	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{9}{6}$
4	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{16}{6}$
5	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{25}{6}$
6	$\frac{1}{6}$	**1**	**6**
		$\frac{21}{6}$	$\mathbf{\frac{91}{6}}$

$$\mu = \frac{21}{6} = 3.5$$

$$\sigma^2 = \frac{91}{6} - (3.5)^2 = 2.9167$$

$$\sigma = \sqrt{2.9167} = 1.7078$$

Ejercicios:

1.- La distribución de probabilidad de una v.a. X viene dada por la siguiente tabla:

x_i	1	2	3	4	5
p_i	0,1	0,3		0,2	0,3

¿Cuánto vale P(X=3)?
Calcula la media y la varianza.

2.- Un jugador lanza un dado corriente. Si sale número primo, gana tantos cientos de euros como marca el dado, pero si no sale número primo, pierde tantos cientos de euros como marca el dado. Determinar la función de probabilidad y la esperanza matemática del juego.

3.- DISTRIBUCIÓN BINOMIAL

Un experimento sigue el modelo de la ***distribución binomial o de Bernouilli*** si:

1. En cada prueba del experimento sólo son posibles **dos resultados**: el suceso A (**éxito**) y su contrario $\overline{A}$.
2. La **probabilidad del suceso A es constante**, es decir, que no varía de una prueba a otra. Se representa por **p**.
3. El **resultado** obtenido en cada prueba es **independiente** de los resultados obtenidos anteriormente.

La ***variable aleatoria binomial***, **X**, expresa el ***número de éxitos obtenidos*** en cada prueba del experimento.

La variable binomial es una variable aleatoria discreta, sólo puede tomar los valores 0, 1, 2, 3, 4, ..., n suponiendo que se han realizado n pruebas.

Por *ejemplo*, la variable X = "número de caras obtenidas al lanzar 6 monedas" es una variable binomial, donde *n = 6* y *A = "salir cara"*

Cuando una variable sigue una distribución binomial se suele representar por:

$X \to B(n, p)$

Donde *n* es el número de pruebas de que consta el experimento y *p* es la probabilidad de éxito en cada prueba

La probabilidad de $\overline{A}$ es, lógicamente, *1-p*, y se representa por *q*.

La ***función de probabilidad*** **de la distribución binomial** es:

$$p(X=k)=\binom{n}{k}p^k \cdot q^{n-k}$$

donde:

- ***n*** es el número de pruebas.
- ***k*** es el número de éxitos.
- ***p*** es la probabilidad de éxito.
- ***q*** es la probabilidad de fracaso.
- El ***número combinatorio***

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

Nota: **n!** se lee "n factorial" y es: $n!=n\cdot(n-1)\cdot(n-2)\cdot(n-3)\cdot....\cdot 1$

Así:

$$\binom{7}{4}=\frac{7!}{4!\cdot 3!}=\frac{7\cdot \cancel{6}\cdot 5\cdot \cancel{4!}}{\cancel{4!}\cdot \cancel{3\cdot 2}}=35$$

Ejemplo 1: Calcular la probabilidad de que una familia que tiene 4 hijos, 3 de ellos sean varones.

Solución:
Si llamamos éxito a *A* = "ser varón", *p* = 0'5, *n* = 4 y *q* = 0'5, la variable X = "número de hijos varones" se distribuye según una binomial *B(4,0'5)*.

Entonces $P(X=3)=\binom{4}{3}.0.5^3.0,5^1 = 4\cdot(0.5)^4\ \frac{1}{4}$

Ejemplo 2: La última novela de un autor ha tenido un gran éxito, hasta el punto de que el 80% de los lectores ya la han leído. Un grupo de 4 amigos son aficionados a la lectura:

a) ¿Cuál es la probabilidad de que del grupo hayan leído la novela dos personas?

b) ¿Y de que lo hayan leído como mucho dos personas?

Solución:
Si llamamos éxito a *A* = "haber leído la novela", *p* = 0'8, *n* = 4 y *q* = 0'2, la variable X = "número de personas que la han leído" se distribuye según una binomial *B(4,0'8)*.

a) $P(X=2)=\binom{4}{2}(0'8)^2(0'2)^2 = \frac{4\cdot 3}{2}\cdot 0'64\cdot 0'04 = 0'1536$

b) $p(X\le 2)=p(X=0)+p(X=1)+p(X=2)=$

$$=\binom{4}{0}0.8^4+\binom{4}{1}0.8^3\cdot 0.2+\binom{4}{2}0.8^2\cdot 0.2^2=0.9728$$

En una distribución binomial *B(n,p)* se puede demostrar que:

Media: $\mu = n.p$

Varianza: $\sigma^2 = n.p.q$

Desviación típica: $\sigma = \sqrt{n.p.q}$

Así, en el ejemplo anterior, *B(4,0'8)*:

$\mu = n.p = 4\cdot 0'8 = 3'2$; $\sigma^2 = n.p.q = 4\cdot 0'8\cdot 0'2 = 0'64$; $\sigma = \sqrt{n.p.q} = \sqrt{0'64} = 0'8$

Ejercicios:

1.- Se tiene una moneda trucada de modo que la probabilidad de sacar cara es cuatro veces la de sacar cruz. Se lanza 6 veces la moneda. Calcula las siguientes probabilidades:
  a) Obtener dos veces cruz.
  b) Obtener a lo sumo dos veces cruz.

2.- La probabilidad de que un alumno de 1º de Bachillerato repita curso es de 0,3. Elegimos 20 alumnos al azar. ¿Cuál es la probabilidad de que haya exactamente 4 alumnos repetidores?

3.- Un jugador de tenis tiene una probabilidad de ganar una partida de 0,25. Si juega 4 partidas, calcula la probabilidad de que gane más de la mitad

4.- Una máquina fabrica tornillos y se ha comprobado que el 2% de los mismos son defectuosos. Si se vende en paquetes de de 29, ¿cuál es la probabilidad de que al comprar un paquete haya en el mismo dos defectuosos?

5.- ¿Cuál es la probabilidad de obtener cuatro veces el número 3 al lanzar un dado ocho veces?

6.- Un agente de seguros vende pólizas a cinco personas de la misma edad y que disfrutan de buena salud. Según las tablas actuales, la probabilidad de que una persona en estas condiciones viva 30 años o más es 2/3. Hállese la probabilidad de que, transcurridos 30 años, vivan:

  a) Las cinco personas
  b) Al menos tres personas
  c) Exactamente dos personas

4.- VARIABLE ALEATORIA CONTINUA

Es aquella que puede tomar infinitos valores dentro de un intervalo de la recta real. Por ejemplo, la duración de las bombillas de una determinada marca y modelo.

En el caso de variables aleatorias continuas no tiene sentido plantearse probabilidades de resultados aislados, por ejemplo, probabilidad de que una bombilla dure 100 horas, 22 minutos y 16 segundos. La probabilidad sería 0 (1 caso favorable entre infinitos casos posibles)

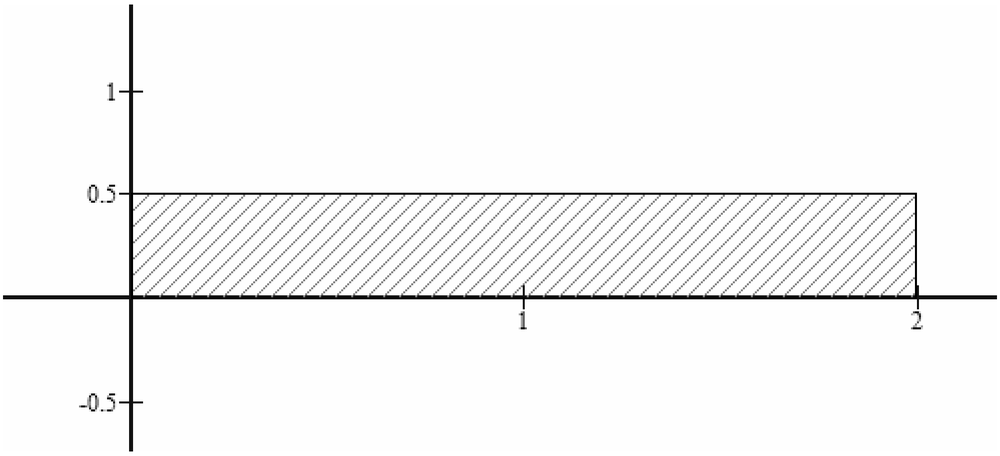
El interés de estas probabilidades está en conocer la probabilidad correspondiente a un intervalo. Dicha probabilidad se conoce mediante una curva llamada función de densidad y suponiendo que bajo dicha curva hay un área de una unidad.

Conociendo esta curva, basta calcular el área correspondiente para conocer la probabilidad de un intervalo cualquiera.

La ***función de densidad*** de una v.a. continua es una función que cumple las siguientes condiciones:

- Sólo puede tomar valores comprendidos entre 0 y 1: $0 \leq f(x) \leq 1$
- El área encerrada bajo la curva es igual a la unidad.

Por ejemplo, la función $f(x) = \frac{1}{2}$, $x \in [0,2]$, cuyo dibujo sería:



corresponde a una función de densidad puesto que es positiva y menor que 1 y es fácil ver que el área marcada es 1 (base x altura = 2 x 0'5 = 1)

Si X es una v.a. continua que tiene a ésta como función de densidad, para calcular probabilidades en intervalos basta con calcular el área correspondiente a dicho intervalo, es decir:

$P(a < X < b)$ = Área entre la gráfica de *f* y el eje OX en el intervalo *[a,b]*

En nuestro ejemplo:

$P(0 < X < 1) = 0'5$ (pues es el área del rectángulo correspondiente a ese intervalo)

$P(1 < X < 1'5) = 0'25$ (por el mismo motivo)

Como en el caso de la v.a. discreta, la ***función de distribución*** de una v.a. continua proporciona la probabilidad acumulada hasta un determinado valor de la variable, es decir, $F(x) = p(X \leq x)$ (Corresponde al área acumulada hasta ese valor de la variable)

Así, con la función del ejemplo:

$F(1'5) = p(X \leq 1'5) =$ área desde el 0 hasta el 1'5 = 0'75

Para calcular la ***media*** y la ***varianza*** de una v.a. continua existe cierta correspondencia con la variable aleatoria discreta:

Variable aleatoria discreta	Variable aleatoria continua
$\mu = \sum x_i . p_i$	$\mu = \int_a^b x. f(x).dx$
$\sigma^2 = \sum x_i^2 p_i - \mu^2$	$\sigma^2 = \int_a^b x^2 f(x)dx - \mu^2$

Lo que es $\sum$ pasa a ser $\int$ y lo que es p_i pasa a ser $f(x)$

Como podemos ver, el concepto de $\int$ se escapa al temario de este curso, con lo que no calcularemos medias y varianzas de v.a. continuas.

5.- DISTRIBUCIÓN NORMAL

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

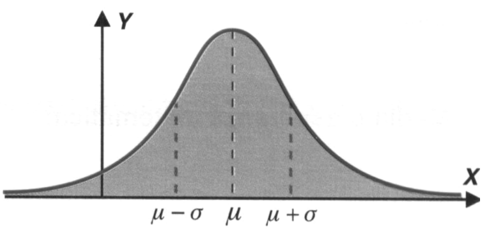
Ejemplos:
- La variable peso en una población de personas de la misma edad y sexo.
- La variable altura de la población citada.
- etc.

Una variable aleatoria continua, X, sigue una ***distribución normal*** de ***media μ*** y ***desviación típica σ***, y se designa por ***N(μ, σ)***, si se cumplen las siguientes condiciones:

- La variable puede tomar cualquier valor: (-∞, +∞)
- La *función de densidad*, es la expresión en términos de ecuación matemática de la ***curva de Gauss***:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

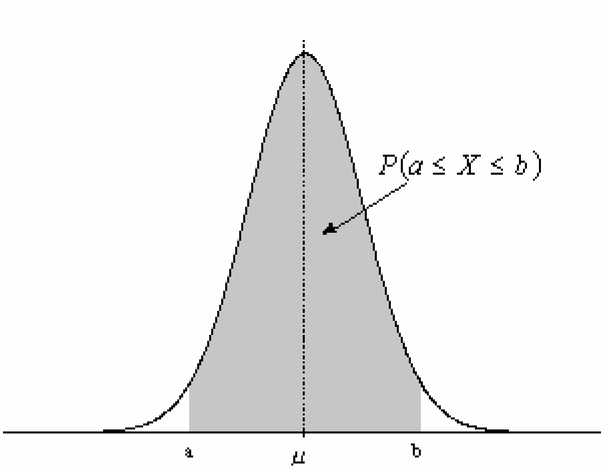
Su gráfica es:



El área del recinto determinado por la función y el eje de abscisas es igual a la unidad.

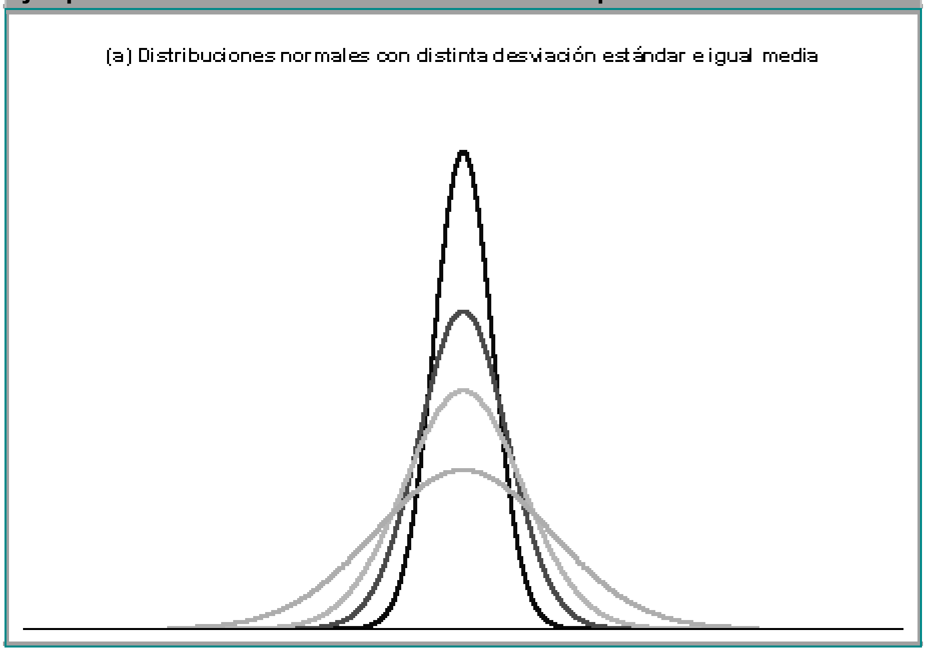
Al ser ***simétrica*** respecto al eje que pasa por **x = μ**, deja un área igual a 0.5 a la izquierda y otra igual a 0.5 a la derecha.

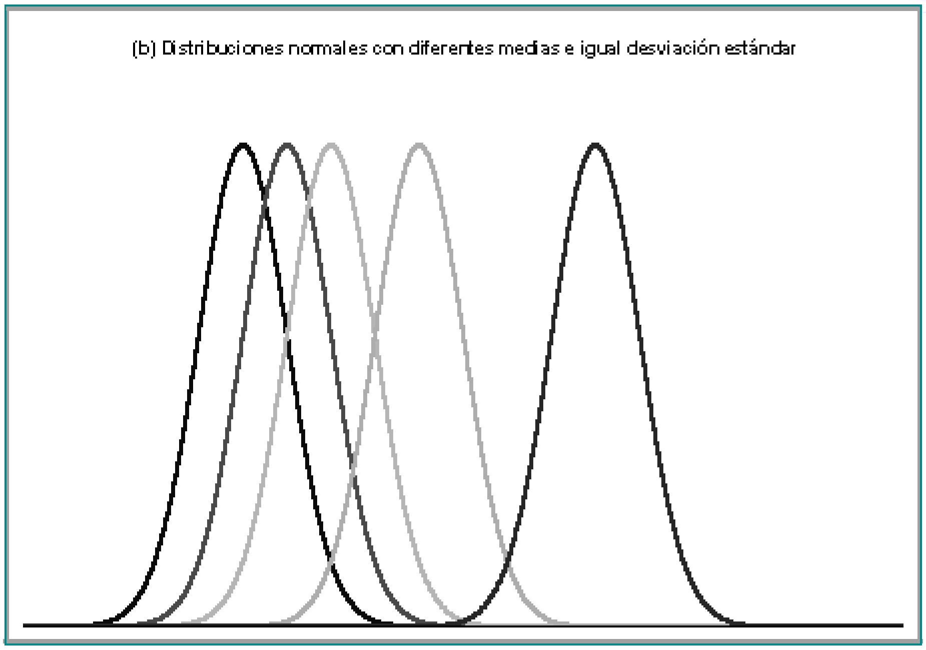
La probabilidad equivale al área encerrada bajo la curva:



La forma de la campana de Gauss depende de los parámetros μ y σ. La media indica la posición de la campana, de modo que para diferentes valores de μ la gráfica es desplazada a lo largo del eje horizontal. Por otra parte, la desviación estándar determina el grado de apuntamiento de la curva. Cuanto mayor sea el valor de σ, más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución.

Ejemplos de distribuciones normales con diferentes parámetros.





Como se deduce de este último apartado, no existe una única distribución normal, sino una familia de distribuciones con una forma común, diferenciadas por los valores de su media y su varianza. De entre todas ellas, la más utilizada es la ***distribución normal estándar, reducida o tipificada***, que corresponde a una distribución de media 0 y varianza 1, es decir, $N(0,1)$.

Su función de densidad es:

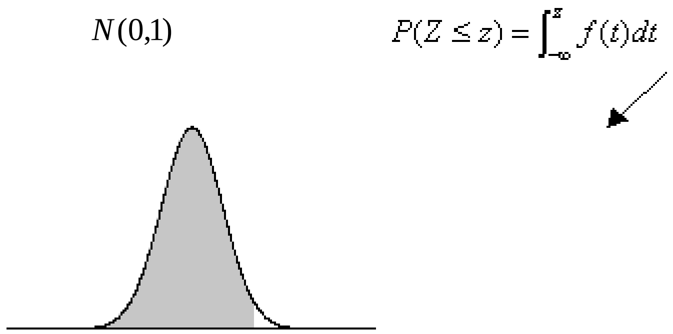
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Cuando una variable sigue una distribución $N(0,1)$ se le llama *Z*.

Existen unas tablas que permiten calcular probabilidades en distribuciones normales reducidas.

La ***tabla*** nos da las probabilidades de P(z ≤ k), siendo **z** la variable tipificada.

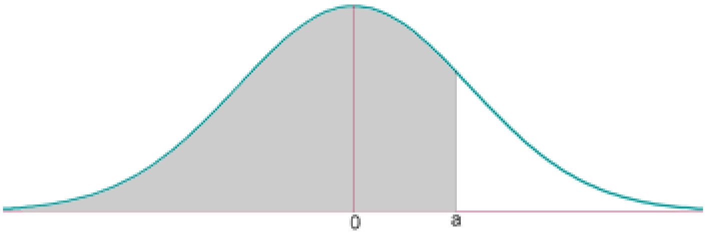
Unidades y décimas en la columna de la izquierda.
Centésimas en la fila de arriba.



Segunda cifra decimal del valor de z										
z	**0.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.4878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

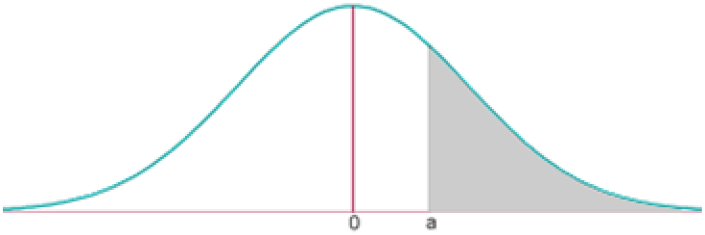
Veamos cómo se calculan las distintas probabilidades en una $N(0,1)$

P(Z ≤ a)



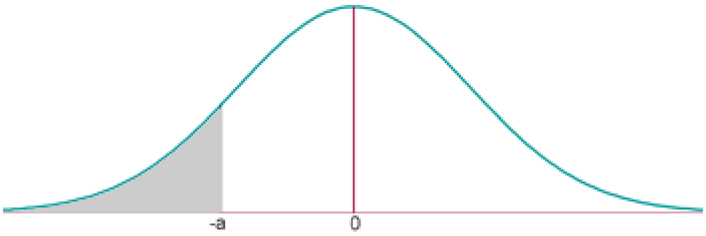
Ej: $P(Z \leq 1.47) = 0.9292$

P(Z > a) = 1 - P(Z ≤ a)



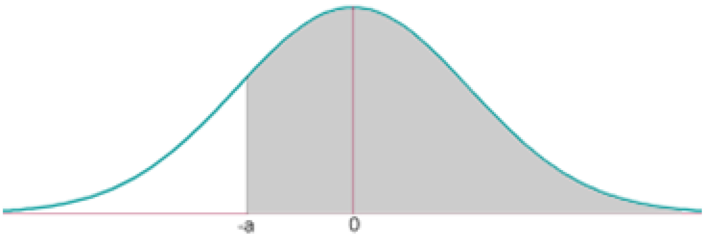
Ej: $P(Z > 1.47) = 1 - P(Z \leq 1.47) = 1 - 0.9292 = 0.0708$

P(Z ≤ −a) = P(Z > a) = 1 − P(Z ≤ a)



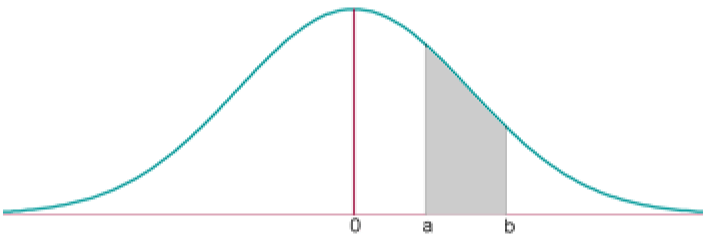
Ej: $P(Z \leq -1.47) = 1 - P(Z \leq 1.47) = 1 - 0.9292 = 0.0708$

P(Z > −a) = P(Z ≤ a)



Ej: $p(Z > 1.47) = p(Z \leq 1.47) = 0.9292$

P(a ≤ Z ≤ b) = P(Z ≤ b) − P(Z ≤ a)



Ej: P(0.45 ≤ Z ≤ 1.47) = P(Z ≤ 1.47) − P(Z ≤ 0.45) =
= 0.9292 − 0.6736 = 0.2556

Hay ocasiones en que queremos calcular el valor de la variable al que le corresponde determinada probabilidad.

Nos encontramos con el caso inverso a los anteriores, conocemos el valor de la probabilidad y se trata de hallar el valor de la abscisa. Ahora tenemos que buscar en la tabla el valor que más se aproxime a *k*.

Ej: P(Z ≤ *k*) = 0’87 □ buscando *dentro* de la tabla □ *k* = 1’13 (aprox.)

Ejercicios:

1.- En una $N(0,1)$, calcular:

a) $p(Z \leq 1{,}23)$

b) $p(Z \geq 1{,}24)$

c) $p(Z \leq -0{,}72)$

d) $p(0{,}5 \leq Z \leq 1{,}76)$

e) $p(-1{,}25 \leq Z \leq -0'73)$

2.- Calcular *k* para que en una $N(0,1)$ se cumpla:

a) P(Z ≤ *k*) = 0’67

b) P(Z > *k*) = 0’05

c) P(Z ≤ *k*) = 0’23

Tipificación de una Distribución Normal

Obviamente no existen tablas para todas las distribuciones normales *N(μ, σ)*, por lo que habrá que transformarla en una $N(0,1)$ para poder usar la tabla.
Al proceso de transformar una variable normal cualquiera *N(μ, σ)* en una $N(0,1)$ se le llama ***tipificación de la variable***.

El cambio de variable que hay que hacer es el siguiente:

Si $X \rightarrow N(\mu,\sigma) \quad \Rightarrow \quad Z = \dfrac{X-\mu}{\sigma} \rightarrow N(0,1)$

Es decir, tipificar una variable normal cualquiera consiste en restarle su media y dividirla por su desviación típica, con lo que se convierte en una $N(0,1)$.

Ejemplo 1:

Sea X una variable *N(8,2)*. Calcular:

a) $P(X \leq 6'5)$ b) $P(7'2 < X \leq 8'3)$

Solución:

a) $P(X \leq 6'5)$ = tipificando = $P\left(\frac{X-8}{2} \leq \frac{6'5-8}{2}\right) = P(Z \leq -0'75) =$

(Z ya es una $N(0,1)$) = $P(Z > 0'75) = 1 - P(Z \leq 0'75) =$ buscando en la tabla = 1 – 0’7734 = 0’2266

b) $P(7'2 < X \leq 8'3) = P\left(\frac{7'2-8}{2} < \frac{X-8}{2} \leq \frac{8'3-8}{2}\right) = P(-0'4 < Z \leq 0'15) =$

$= P(Z \leq 0'15) - P(Z < -0'4) = (*)$

Calculamos aparte:

$$P(Z < -0'4) = P(Z > 0'4) = 1 - P(Z \leq 0'4) = 1 - 0'6554 = 0'3446$$

Luego:

(*) = 0’5596 - 0’3446 = 0’215

Ejemplo 2:

La duración media de un lavavajillas es de 15 años y su desviación típica 0,5. Sabiendo que su vida útil se distribuye normalmente, halla la probabilidad de que al adquirir un lavavajillas dure más de 16 años.

Solución:

Es una distribución normal de media 15 y desviación típica 0,5, es decir, *N(15; 0,5).*

$p(X \geq 16) = p(\frac{X-15}{0'5} \geq \frac{16-15}{0,5}) = p(Z \geq 2) = 1 - p(Z < 2) =$

$= 1 - 0'9772 = 0'0228$

Nota: eso significa que un 2’28% de los lavavajillas de ese modelo duran más de 16 años

Ejercicios:

1.- La nota media de las pruebas de acceso correspondientes a los estudiantes que querían ingresar en una facultad era 5,8 y la desviación típica 1,75. Fueron admitidos los de nota superior a 6.
¿Cuál fue el porcentaje de admitidos si la distribución es normal?

2.- En las etiquetas de las cajas de unos determinados tornillos está indicado que tienen un diámetro comprendido entre 1’09 y 1’11 mm. Si dicho diámetro es una variable normal con media 1’10 y desviación típica 0’005 mm, ¿qué porcentaje de tornillos no cumple con la especificación de las etiquetas?

Aproximación de la distribución binomial mediante la normal

Cuando *n* es grande y *p* está próximo a 0,5 el comportamiento de una distribución binomial *B(n, p)* es aproximadamente igual a una distribución normal, $N(np, \sqrt{npq})$

Esto permite sustituir el estudio de una $B(n, p)$ por el de una $N(np, \sqrt{npq})$.

Suele considerarse que la aproximación es buena cuando *np>5* y *nq>5*

Ejemplo

En una ciudad una de cada tres familias posee teléfono. Si se eligen al azar 90 familias, calcular la probabilidad de que entre ellas haya por lo menos 30 tipos se han teléfono.

Solución:

$$n = 90 \qquad p = \frac{1}{3} \qquad q = \frac{2}{3}$$

$$n \cdot p > 5 \qquad n \cdot q > 5$$

$$B\left(90, \frac{1}{3}\right) \to N\left(90 \cdot \frac{1}{3}, \sqrt{90 \cdot \frac{1}{3} \cdot \frac{2}{3}}\right) = N(30,\ 4.47)$$

$$p(X > 30) = p\left(Z > \frac{30-30}{4.47}\right) = p(Z > 0) = 1 - p(Z \leq 0) = 0.5$$

Ejercicios:

1.- Se lanza una moneda correcta al aire 400 veces. Calcula la probabilidad de obtener un número de caras comprendido entre 180 y 210, ambos inclusive.

2.- El 35% de una población está afectada por la gripe. Si se eligen 30 personas al azar, ¿Cuál es la probabilidad de que más de 10 tengan gripe?

EJERCICIOS:

1.- Considera una variable aleatoria discreta X cuya distribución de probabilidad es la siguiente:

x_i	1	2	3
$P(X = x_i)$	k	0,45	k

a) Calcula el valor de k
b) Halla la función de probabilidad
c) Calcula su media y su desviación típica

Sol: k = 0’275 ; $\mu = 2$; $\sigma = 0'74$

2.- La probabilidad de que un estudiante obtenga el título de arquitecto es 0,3. Calcula la probabilidad de que un grupo de 7 estudiantes matriculados en primer curso:
a) Ninguno de los 7 finalice la carrera.
b) Finalicen los 7.
c) Al menos 2 acaben la carrera.
d) Sólo finalice uno la carrera.

Sol: 0,082; 0,00021; 0,671; 0,2471

3.- Una determinada raza de perros tiene 4 cachorros en cada camada. Si la probabilidad de que un cachorro sea macho es de 0,55, se pide:
a) La probabilidad de que en una camada dos exactamente sean hembras
b) Probabilidad de que en una camada al menos dos sean hembras.

Sol: 0,3675; 0,609

4.- Si de seis a siete de la tarde se admite que un número de teléfono de cada cinco está comunicando, ¿cuál es la probabilidad de que, cuando se marquen 10 números de teléfono elegidos al azar, sólo comuniquen dos?

Sol:0’302

5.- La probabilidad de que un hombre acierte en el blanco es 1/4. Si dispara 10 veces ¿cuál es la probabilidad de que acierte exactamente en tres ocasiones? ¿Cuál es la probabilidad de que acierte por lo menos en una ocasión?

Sol: 0’25 ; 0’9437

6.- En una urna hay 30 bolas, 10 rojas y el resto blancas. Se elige una bola al azar *y* se anota si es roja; el proceso se repite, devolviendo la bola, 10 veces. Calcular la media y la desviación típica.

Sol: 3’33; 1’49

7.- Si Z es la distribución normal tipificada, hállese:
a) P(Z>̣0,63) b) P(-0’65<Z<0,65) c) P(1<Ẓ1,32)

Sol: a) 0,2643; b) 0,5156; c) 0,0653

8.- Si Z es la normal tipificada, hállese k en los siguientes casos:
a) P(Ẓ<k) = 0,7823; b) P(Ẓ>k) = 0,0838; c) P(Ẓ<K) = 0,9948; d) P(Z<K) = 0,2206.

Sol: a) k=0,78; b) k=1,38; c) k=2,56; d) k=-0,77

9.- En una finca agrícola dedicada a la producción de manzanas se ha comprobado que el peso de las manzanas sigue una distribución normal con media 100 g y desviación 10. A la hora de comercializarlas se toman para la clase A las comprendidas entre 80 y 120 gr. Hallar la probabilidad de que escogida una manzana al azar:

a) corresponda a la clase A
b) pese menos de 70 g
c) pese más de 120 g

Sol: a) 0'9544; b) 0'0013; c) 0'0228

10.- Una patrulla de tráfico realiza un control de alcoholemia en una carretera y llegan a la conclusión de que el nivel de alcohol en sangre de los conductores sigue una distribución N(0'25,0'1). Si el nivel de alcoholemia permitido es de 0'5 para los conductores expertos y 0'3 para los conductores novatos.

a) ¿Cuál es la probabilidad de que un conductor al azar tenga más de 0'5?
b) ¿Cuál es la probabilidad de que tenga más de 0'3?
c) ¿Cuál es la probabilidad de que un conductor diese positivo en la de 0'3 pero no en la de 0'5?

Sol: a) 0'0062; b) 0'3085; c) 0'3023

11.- Una empresa lleva a cabo una prueba para seleccionar nuevos empleados. Por la experiencia de pruebas anteriores, se sabe que las puntuaciones siguen una distribución normal de media 80 y desviación típica 25.
¿Qué porcentaje de candidatos obtendrá entre 75 y 100 puntos?

(Sol. 36,74%)

12.- Un profesor de matemáticas ha observado que las notas obtenidas por sus alumnos en los exámenes de Estadística siguen una distribución *N(6; 2,5)*.
Se han presentado al último examen 32 alumnos, ¿cuántos sacaron al menos un 7?.

(Sol. 11)

13.- La nota media de un examen tipo test fue de 40'3 y su desviación típica 3'3. Si las calificaciones siguen una distribución normal y se considera aprobado a los que superen 37 puntos.

a) ¿Cuál es el porcentaje de aprobados?

b) Si a ese examen se presentaron 400 personas ¿cuántas aprobaron?

c) Si quisiésemos que sólo aprobasen 50 personas ¿cuál tendría que ser la nota de corte?

Sol: a) 84'13%; b) 336; c) 44'1

14.- Las alturas de 200 estudiantes se distribuyen normalmente, con una media de 175 cm y una desviación típica de 10 cm. ¿Cuántos de estos estudiantes tienen altura:

a) mayor de 180 cm. b) menor de 165 cm.

b) c) entre 160 cm y 180 cm. d) igual a 182 cm.

Sol: a) 61,7; b) 31,74; c) 124,94; d) 0

15.- La cantidad de azúcar depositada en cada bolsa por una máquina envasadora automática sigue una distribución normal con media μ=1050 grs y desviación típica σ=50 grs.

a) Calcula el porcentaje de bolsas con un peso mayor a 1 Kg

b) Calcula el tanto por ciento de paquetes con un contenido que tiene un peso comprendido entre 900 y 1000 grs.

Sol: a) 84'13%; b) 15'74%

16.- Se ha elegido una muestra de 500 tornillos fabricados por una máquina. La media de los diámetros de dichos tornillos es de 2,9 mm y la desviación típica, de 1 mm. Un cliente considera que un tornillo es inservible si su diámetro es inferior a 2,85 mm o superior a 3,1 mm.

a) Sabiendo que los diámetros se distribuyen normalmente, hállese qué porcentaje de tornillos son defectuosos.

b) Si para otro cliente son válidos desde 2,75 hasta 3,15, ¿qué porcentaje son inservibles?

Sol: a) 90%; b) 84,17%

17.- El coeficiente de inteligencia de una población es una v.a. cuya distribución sigue una ley normal del tipo N(100,10). Calcúlese, según esos datos, qué porcentaje de personas cabe esperar que tengan coeficiente de inteligencia:

a) superior a 120 b) entre 90 y 120 c) inferior a 80;

d) Si se escogen 5000 personas al azar, ¿cuántas tendrán un coeficiente de inteligencia mayor de 125?

Sol: a) 2,28%; b) 81,85%; c) 2,28%; d) 31

18.- Los gastos diarios de una familia siguen una distribución normal, con media de 50 euros y desviación típica de 5 euros.

a) Calcular el porcentaje de días en los que los gastos son inferiores a 55 euros.

b) Calcular el porcentaje de días en que los gastos superan los 60 euros.

c) Calcular el porcentaje de días en los gastos son superiores a 50 euros e inferiores a 60 euros.

Sol: a) 84,13%; b) 2,28%; c) 47,72%

19.- Una empresa de transporte de viajeros afirma que el tiempo de retraso de sus viajes sigue una distribución normal, con un retraso medio de 5 minutos y desviación típica de 2 minutos. Calcular:

a) Probabilidad de que un viaje no tenga retraso

b) Probabilidad de que el próximo llegue con más de 5 minutos de retraso.

c) Probabilidad de que el próximo llegue con más de 10 minutos de retraso.

Sol: a) 0,0062; b) 0,5; c) 0,0062.

20.- Una de las pruebas de acceso a la Universidad para mayores de 25 años consiste en un test con 100 preguntas, cada una de las cuales tiene 4 posibles respuestas y sólo una correcta. Para superar esta prueba deben obtenerse, al menos, 30 respuestas correctas.

Si una persona contesta al azar, ¿Qué probabilidad tendrá de superar la prueba?

(Sol. Utilizando la aproximación a través de la normal: p= 0,1492)

TEMA 3.- TEORÍA DE MUESTRAS

1.- POBLACIÓN Y MUESTRA

Entendemos por ***población*** un conjunto de elementos que poseen una característica o propiedad común, y que constituyen la totalidad de los individuos de interés para nuestro estudio.

En particular, nos interesa obtener información acerca de algún valor que caracteriza a la población, como una media, una varianza, una mediana, etc. A estos valores que se refieren a la totalidad de la población se les denomina ***parámetros poblacionales*** y en su notación es común utilizar el alfabeto griego: μ (media poblacional), σ^2 (varianza poblacional), σ (desviación típica poblacional), etc.

Como las poblaciones en las que se pretende estudiar una determinada variable aleatoria son grandes, es muy caro o imposible estudiar a todos sus individuos; lo que se hace, es estudiar una ***muestra*** (una parte) de la población:

Una ***muestra*** es cualquier subconjunto de la población sobre el que se realizan estudios para obtener conclusiones acerca de las características de la población. Cualquier valor obtenido a partir de los datos de la muestra se denomina ***estadístico muestral***. Ejemplos de estadísticos muestrales son: $\bar{x}$ (media muestral), S^2 (varianza muestral), S (desviación típica muestral), etc.

La ***teoría del muestreo*** tiene por objetivo, el estudio de las relaciones existentes entre la distribución de un carácter en dicha población y las distribuciones de dicho carácter en todas sus muestras.

Las ventajas de estudiar una población a partir de sus muestras son principalmente:

- **Coste reducido:**

  Si los datos que buscamos los podemos obtener a partir de una pequeña parte del total de la población, los gastos de recogida y tratamiento de los datos serán menores. Por ejemplo, cuando se realizan encuestas previas a un referéndum, es más barato preguntar a 4.000 personas su intención de voto, que a 30.000.000;

- **Mayor rapidez:**

  Estamos acostumbrados a ver cómo con los resultados del escrutinio de las primeras mesas electorales, se obtiene una aproximación bastante buena del resultado final de unas elecciones, muchas horas antes de que el recuento final de votos haya finalizado;

- **Más posibilidades:**

  Para hacer cierto tipo de estudios, por ejemplo el de duración de cierto tipo de bombillas, no es posible en la práctica destruirlas todas para conocer su vida media, ya que no quedaría nada que vender. Es mejor destruir sólo una pequeña parte de ellas y sacar conclusiones sobre las demás.

De este modo se ve que al hacer estadística inferencial debemos enfrentarnos con dos problemas:

- Elección de la muestra *(muestreo)*, que es a lo que nos dedicaremos en este capítulo.
- Extrapolación de las conclusiones obtenidas sobre la muestra, al resto de la población *(inferencia).*

Es decir, el principal objetivo de la mayoría de los estudios, análisis o investigaciones, es hacer generalizaciones *"acertadas"* con base en muestras de poblaciones de las que se derivan tales muestras. Obsérvese la palabra *"acertadas"* porque no es fácil responder cuándo y en que condiciones las muestras permiten tales generalizaciones.

Por ejemplo si queremos calcular la cantidad de dinero que se gasta una persona en vacaciones, ¿tomaríamos como muestra lo que gastan los viajeros que lo hacen en primera clase? Es obvio que no, pero saber a que tipo de personas debemos incluir en nuestra muestra no es algo intuitivo ni evidente.

Sobre las muestras hay dos aspectos que resultan fundamentales: el tamaño (nº de elementos de la muestra) y la forma en que se realiza la selección de los individuos que la forman.

La elección de la muestra influirá de manera determinante en los resultados, por lo que hay que evitar muestras *"sesgadas"* (parciales y subjetivas) que produzcan errores incontrolables (aparte de los errores propios de sustituir el estudio de una población por el de una muestra).

El tipo de muestreo más importante es el ***muestreo aleatorio***, en el que todos los elementos de la población tienen la misma probabilidad de ser extraídos; Aunque dependiendo del problema y con el objetivo de reducir los costes o aumentar la precisión, otros tipos de muestreo pueden ser considerados como veremos a continuación.

2.- TIPOS DE MUESTREO

1.- Muestreo Aleatorio Simple

Consiste en seleccionar *n* elementos sin reemplazamiento de entre los *N* que componen la población, de tal modo que todas las muestras tengan la misma probabilidad de ser elegidas.

En la práctica, se enumeran los individuos y se sortean cuáles de ellos se elegirán. Si los individuos son, por ejemplo, tornillos que se encuentran en una caja, se eligen al azar por simple extracción.

2.- Muestreo Aleatorio Sistemático

Se elige un individuo al azar y a partir de él, a intervalos constantes, se eligen los demás hasta completar la muestra.

Por ejemplo si tenemos una población formada por 100 elementos y queremos extraer una muestra de 25 elementos, en primer lugar debemos establecer el intervalo de selección que será igual a 100/25=4. A continuación elegimos el elemento de arranque, tomando aleatoriamente un número entre el 1 y el 4, y a partir de él obtenemos los restantes elementos de la muestra:

2 , 6 , 10 , 14 , … , 98

3.- *Muestreo Aleatorio Estratificado*

Si conocemos que la población puede dividirse en partes o estratos, en relación con variables que pueden ser de interés en nuestro estudio, de modo que en cada uno de los estratos los elementos posean una gran homogeneidad respecto al carácter que se estudia (sexo, grupos de edad, nivel de estudios,…), se puede aumentar la precisión si muestreamos los estratos por separado. La forma de repartir los elementos de la muestra, determinando cuantos deben corresponder a cada estrato se denomina ***afijación***, y puede ser de varios tipos:

- ***Afijación uniforme***: Si se toma el mismo número de elementos en cada estrato.
- ***Afijación Proporcional***: si el número de elementos que se toma en cada estrato es proporcional al tamaño del estrato. Se utiliza cuando las varianzas de los estratos no difieren mucho entre si.

Los elementos de cada estrato se toman mediante muestreo aleatorio simple.

Ejemplo:

En una ciudad se quiere hacer un estudio para conocer qué tipo de actividades se realizan en el tiempo dedicado al ocio. Para ello van a ser encuestadas 300 personas elegidas al azar mediante muestreo aleatorio estratificado con afijación proporcional. Teniendo en cuenta que de un total de 15.000 habitantes, 7.500 son adultos, 3.000 ancianos y 4.500 niños, definir los estratos y el tamaño muestral correspondiente a cada estrato.

Solución:

Basta con hacer una simple regla de 3 para determinar el tamaño muestral de cada estrato:

Niños: 15.000 -------- 4.500

300 -------- x

$$x = \frac{4.500 * 300}{15.000} = 90 \text{ niños}$$

Adultos: $$x = \frac{7.500 * 300}{15.000} = 150 \text{ adultos}$$

Ancianos: $$x = \frac{3.000 * 300}{15.000} = 60 \text{ ancianos}$$

4.- Muestreo Por Conglomerados

En este tipo de muestreo, llamado ***muestreo por conglomerados***, se divide la población total en un número determinado de subdivisiones relativamente pequeñas y se seleccionan al azar algunas de estas subdivisiones o conglomerados, para incluirlos en la muestra total. Si estos conglomerados coinciden con áreas geográficas, este muestreo se llama también ***muestreo por áreas***.

Por ejemplo, supongamos que una gran empresa quiere estudiar los patrones variables de los gastos familiares de una ciudad como Granada. Al intentar elaborar los programas de gastos de una muestra de 1200 familias, nos encontramos con la dificultad de realizar un muestreo aleatorio simple, (es complicado tener una lista actualizada de todos los habitantes de una ciudad). Una manera de tomar una muestra en esta situación es dividir el área total (Granada en este caso) en áreas más pequeñas que no se solapen (Por ejemplo Distritos postales, manzanas etc...) En este caso seleccionaríamos algunas áreas al azar y todas las familias (o muestras de éstas) que residen en estos distritos postales o manzanas, constituirían la muestra definitiva.

Aunque las estimaciones basadas en el muestreo por conglomerados, por lo general no son tan fiables como las obtenidas por muestreos aleatorios simples del mismo tamaño, son más baratas. Volviendo al ejemplo anterior, es mucho más económico visitar a familias que viven en el mismo vecindario, que ir visitando a familias que viven en un área muy extensa.

En la práctica se pueden combinar el uso de varios de los métodos de muestreo que hemos analizados para un mismo estudio.

Ejercicios:

1.- Consideremos la población formada por 5 bolas contenidas en una urna y numeradas del 1 al 5. Obtener todas las muestras de tamaño 2 extraídas mediante muestreo aleatorio simple.

2.- En un instituto de enseñanza secundaria en que se ofertan los siguientes tipos de enseñanza:

- Ciclos de grado superior: 110 alumnos.
- Bachillerato: 162 alumnos.
- Ciclos de grado medio : 210 alumnos
- 2º ciclo de enseñanza secundaria obligatoria: 338 alumnos.

Se pretende valorar las faltas de ortografía que cometen los alumnos del centro mediante una prueba-dictado de un texto de 20 líneas; la prueba se pasará a una muestra de 50 alumnos, para minimizar el costo en tiempo y medios.

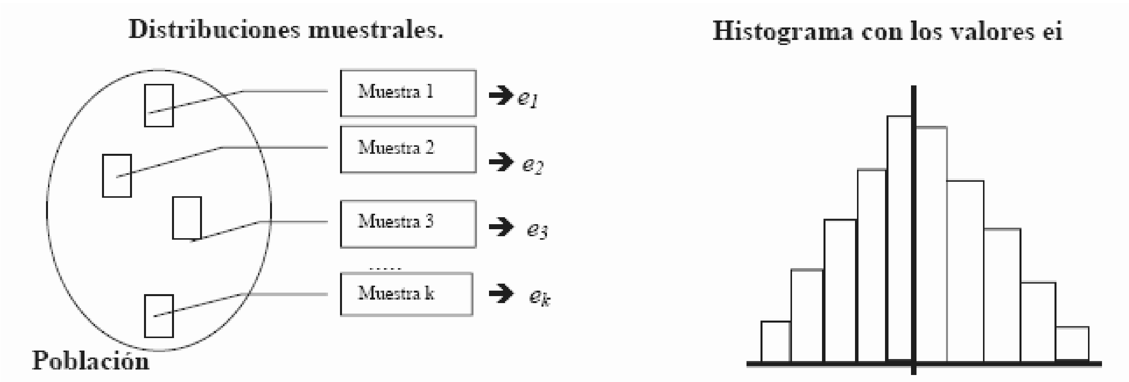
¿Cómo se obtiene una muestra adecuada a esta población?

3.- DISTRIBUCIÓN DE LAS MEDIAS MUESTRALES

La idea de ***distribución muestral*** es la siguiente:

Partimos de una población de tamaño *N*. Obtenemos *k* muestras (todas las posibles de tamaño *n*) y de cada una de ellas se calcula una medida (media, mediana, varianza, desviación típica,..) obteniendo *k* valores: $e_1, e_2, e_3, ...e_k$.

Estos valores pueden representarse mediante un histograma para observar su distribución:



Se puede observar que este histograma va adquiriendo forma de campana de Gauss, y a medida que *k* aumenta, se va pareciendo cada vez más a la forma de una distribución Normal.

Lo vemos con un ***ejemplo***:

Elaboraremos la distribución muestral de la media de una muestra aleatoria de tamaño n = 2 tomada sin reemplazo de la población finita de tamaño N = 5, cuyos elementos son: 3,5,7,9,11.

La media de esta población es: $\mu = \dfrac{3+5+7+9+11}{5} = 7$

y su desviación típica es:

$$\sigma^2 = \frac{3^2+5^2+7^2+9^2+11^2}{5} - 7^2 = 8 \quad \Rightarrow \sigma = \sqrt{8}$$

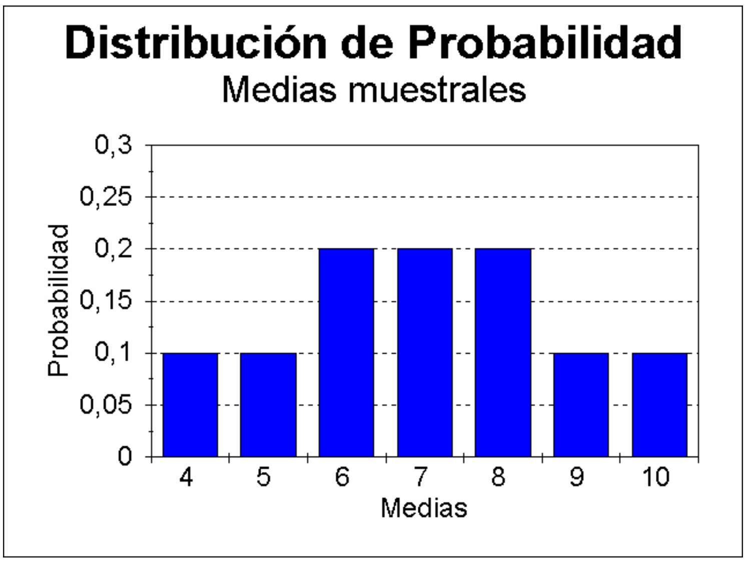
Ahora si tomamos una muestra aleatoria de tamaño n = 2 de esta población hay 20 posibilidades:

nº muestra	Muestras		$\overline{X}$ (media de la muestra)
1	3	5	4
2	3	7	5
3	3	9	6
4	3	11	7
5	5	3	4
6	5	7	6
7	5	9	7
8	5	11	8
9	7	3	5
10	7	5	6
11	7	9	8
12	7	11	9
13	9	3	4
14	9	5	7
15	9	7	8
16	9	11	10
17	11	3	7
18	11	5	8
19	11	7	9
20	11	9	10

Si hacemos la función de probabilidad de la variable $\overline{X}$ de las medias muestrales:

Media	Probabilidad
4	2/20
5	2/20
6	4/20
7	4/20
8	4/20
9	2/20
10	2/20

Cuyo histograma sería:



Si calculamos la media y la desviación típica de la distribución de las medias obtenemos que: $\mu_{\overline{X}} = 7$ y $\sigma_{\overline{X}} = \sqrt{3}$, luego la media μ_x coincide con la media de la población y la desviación típica ha disminuido.

Este ejemplo puede generalizarse para cualquier distribución según el siguiente teorema:

<u>Teorema Central del Límite</u>

Si una población tiene media μ y desviación típica σ, y tomamos muestras de tamaño n (n>30, ó cualquier tamaño si la población es "normal"), las medias de estas muestras siguen aproximadamente la distribución:

$$\overline{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Destacar que si la población de la que se obtienen las muestras es *normal*, las medias muestrales también se distribuyen según una distribución normal, independientemente del tamaño de la muestra.

Al saber cómo funcionan las medias muestrales, podemos obtener probabilidades relacionadas con ellas.

Ejemplo 1:

Se sabe que las bolsas de azúcar producidas por una máquina tiene una media de 500 g. y una desviación típica de 35 g. Dichas bolsas se empaquetan en cajas de 100 unidades.

a) ¿Cómo se distribuyen las medias de los pesos de las bolsas de cada caja?

b) Calcular la probabilidad de que la media de los pesos de las bolsas de un paquete sea inferior a 495 g.

Solución:

a) Tenemos $\mu = 500$, $\sigma = 35$, $n = 100$

Por el Teorema Central del Límite (aunque la variable peso no sea normal, tenemos n > 30), la variable $\overline{X}$ = "media de los pesos de la muestra" se distribuye según una normal:

$$\overline{X} \rightarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(500, \frac{35}{\sqrt{100}}\right) = N(500{,}3'5)$$

b) $$P(\overline{X} < 495) = tipificando = P\left(\frac{\overline{X} - 500}{3'5} < \frac{495 - 500}{3'5}\right) = P(Z < -1'43) =$$
$$= P(Z > 1'43) = 1 - P(Z < 1'43) = 0'0764$$

Lo que vendría a significar que en el 7'6 % de las cajas el peso medio de las bolsas es inferior a 495 g.

Ejemplo 2:

Las estaturas, en centímetros, de un grupo de soldados se distribuyen normalmente con media 173 y desviación típica 6.

a) Elegido un soldado al azar, ¿cuál es la probabilidad de que mida menos de 175 cm.?

b) Si se toma una muestra de 12 soldados, ¿cuál es la probabilidad de que su estatura media supere el 1’76?

Solución:

a) Cuidado, en este apartado no hay muestra, ni por tanto medias muestrales, sino un simple ejercicio de la normal.
Si X = “altura”, sabemos que sigue una N(173,6)
Luego

$$P(X < 175) = tipificando = P\left(\frac{X-173}{6} < \frac{175-173}{6}\right) =$$
$$= P(Z < 0'33) = 0'6293$$

b) Aquí si tenemos una muestra con *n = 12*
Aunque sea n < 30, como la población de partida es normal, podemos también aplicar el Teorema Central del Límite:

$$\overline{X} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(173, \frac{6}{\sqrt{12}}\right) = N(173{,}1'73)$$

Luego:

$$P(\overline{X} > 176) = tipificando)P\left(\frac{\overline{X}-173}{1'73} > \frac{176-173}{1'73}\right) = P(Z > 1'73) =$$
$$= 1 - P(Z < 1'73) = 1 - 0'9582 = 0'0418$$

Ejercicios:

1.- Se supone que la distribución de la temperatura del cuerpo humano en la población tiene de media 37º y de desviación típica 0’85º. Se elige una muestra de 105 personas. Hallar las probabilidades de que:

a) La media sea menor que 36’9º

b) La media esté comprendida entre 36’5º y 37’2º

2.- Las notas de cierto examen se distribuyen según una normal de media 5,8 y desviación típica 2,4. Hallar la probabilidad de que la media de una muestra tomada al azar de 16 estudiantes esté comprendida entre 5 y 7

3.- Consideremos la población formada por los cuatro elementos 0, 3, 4, 6. Hallar:

a) Todas las muestras posibles de tamaño 2 extraídas mediante muestreo aleatorio simple

b) La media y la desviación típica poblacionales

c) La media y la desviación típica de las medias muestrales

4.- DISTRIBUCIÓN DE LAS PROPORCIONES MUESTRALES

En lugar de calcular medias de las muestras, ahora vamos a trabajar con proporciones.

En una población, la proporción de individuos que poseen una determinada característica es *p*. (Llamaremos $q = 1 - p$)
Extraemos todas las posibles muestras de tamaño *n* que podemos extraer de esa población. La proporción de individuos de cada una de esas muestras con esa característica ser $\hat{p}$.
Llamaremos $\hat{P}$ a la variable aleatoria que toma los distintos valores de esas proporciones muestrales.

Si *n* es lo suficientemente grande ($n > 30$), se puede demostrar que la variable $\hat{P}$ sigue una distribución normal de parámetros:

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

(Esta fórmula proviene de la aproximación de una binomial por una normal)

Ejemplo 1:

Una nueva droga ha curado al 85% de los enfermos a los que se les ha aplicado. Si se toman muestras de 30 personas, ¿Cuál es la distribución de las proporciones muestrales? ¿Y si las muestras son de 100 personas? ¿Y si son de 1000?

Solución:
Tenemos $p = 0'85$, por lo que $q = 0'15$

Si $n = 30$:

$$\hat{P} \to N\left(p, \sqrt{\frac{pq}{n}}\right) \Rightarrow \hat{P} \to N\left(0'85, \sqrt{\frac{0'85 \cdot 0'15}{30}}\right) \Rightarrow \hat{P} \to N(0'85, 0'065)$$

Si $n = 100$:

$$\hat{P} \to N\left(p, \sqrt{\frac{pq}{n}}\right) \Rightarrow \hat{P} \to N\left(0'85, \sqrt{\frac{0'85 \cdot 0'15}{100}}\right) \Rightarrow \hat{P} \to N(0'85, 0'036)$$

Si $n = 1000$:

$$\hat{P} \to N\left(p, \sqrt{\frac{pq}{n}}\right) \Rightarrow \hat{P} \to N\left(0'85, \sqrt{\frac{0'85 \cdot 0'15}{1000}}\right) \Rightarrow \hat{P} \to N(0'85, 0'011)$$

Como ya vimos con las medias muestrales, al aumentar el tamaño de las muestras disminuye la varianza de las distribuciones muestrales.

Ejemplo 2:

Se sabe que el 15 % de los jóvenes entre 18 y 25 años son miopes.

a) ¿Cómo se distribuye la proporción de jóvenes miopes en muestras de 40 individuos?

b) ¿Cuál es la probabilidad de que en dicha muestra la proporción de miopes esté entre el 8 y el 22 %?

Solución:

a) Tenemos *p = 0’15*, por lo que *q = 0’85*. Además *n = 40*. Por tanto:

$$\hat{P} \to N\left(p, \sqrt{\frac{pq}{n}}\right) \Rightarrow \hat{P} \to N\left(0'15, \sqrt{\frac{0'15 \cdot 0'85}{40}}\right) \Rightarrow \hat{P} \to N(0'15, 0'0565)$$

b) Usando el apartado anterior:

$$P(0'08 < \hat{P} < 0'22) = tipificando = P\left(\frac{0'08 - 0'15}{0'0565} < \hat{P} < \frac{0'22 - 0'15}{0'0565}\right) =$$

$$= P(-1'24 < \hat{P} < 1'24) = P(\hat{P} < 1'24) - P(\hat{P} < -1'24) = (*)$$

Calculamos

$$P(\hat{P} < -1'24) = P(\hat{P} > 1'24) = 1 - P(\hat{P} < 1'24) = 1 - 0'8925 = 0'1075$$

Y por tanto:

(*) = 0’8925 – 0’1075 = 0’785

Esto significaría que en el 78’5 % de las muestras que extrajésemos de tamaño 40, la proporción de individuos miopes estaría entre el 8 y el 22%.

Ejercicios:

1.- Una máquina fabrica piezas de precisión y en su producción habitual tiene un 3% de piezas defectuosas. Se empaquetan en cajas de 200, ¿cuál es la probabilidad de encontrar entre 5 y 7 piezas defectuosas en una caja?

2.- Si tiramos una moneda no trucada 100 veces, ¿cuál es la probabilidad de que obtengamos más de 55 caras?

3.- Se ha realizado una encuesta a 20.000 universitarios sobre su actitud ante el botellón. De ellos, 13.200 son partidarios y el resto no. Se toman 100 muestras de 30 universitarios cada una. Hallar:

a) La distribución de la proporción muestral de partidarios del botellón

b) La probabilidad de que en una de estas muestras se manifiesten favorables al botellón más de 21 alumnos

c) ¿En cuántas muestras cabe esperar que más de 15 y menos de 19 estudiantes se muestren partidarios del botellón?

EJERCICIOS

1.- En el instituto de cierta localidad se imparten 4 niveles diferentes: 1º, 2º, 3º y 4º ESO, estando matriculados un total de 800 alumnos. En 1º hay 160 alumnos, en 2º hay 240 y en 3º hay 208. Explica cómo obtener una muestra de 50 alumnos mediante:

a) Muestreo aleatorio sistemático

b) Muestreo aleatorio estratificado con afijación proporcional

2.- En un barrio hay 4000 habitantes, distribuidos en cuatro urbanizaciones: el 12% viven en A, el 20% en B, el 36% en C y el 32% en D. Además, los porcentajes de mujeres en cada urbanización son: 50, 60, 66 y 75 % respectivamente. Se desea obtener una muestra de 50 habitantes que sea representativa en cuanto al sexo y a las urbanizaciones mediante muestreo aleatorio estratificado proporcional. ¿Cuántas personas (hombres y mujeres) de cada urbanización habrá que seleccionar?

3.- Dada la población {2, 4, 6, 8}:

a) Escribir todas las muestras posibles de tamaño 2 mediante muestreo aleatorio simple

b) Calcular la media y la varianza de las medias muestrales

c) Hacer lo mismo con todas las muestras posibles de tamaño 3

4.- La media de edad de los lectores de una determinada revista es de 17’2 años y la desviación típica de 2’3 años. Si elegimos muestras de 100 lectores:

a) ¿Cuál es la distribución de las medias muestrales?

b) ¿Cuál es la probabilidad de que la media de edad de la muestra esté comprendida entre 16’7 y 17’5 años?

5.- En una determinada población, los pesos se distribuyen según una normal de media 65 kg. y varianza 49. Si extraemos muestras de tamaño 64, ¿cuál es la probabilidad de que la media de los pesos de una de esas muestras sea mayor que 66’5 kg?

6.- La edad de los miembros de una determinada asociación se distribuye normalmente con media 52 años y desviación típica 3 años.

a) ¿Cuál es la probabilidad de que un miembro elegido al azar sea mayor de 60 años?

b) Si se toman muestras de tamaño 36, ¿Cuál es la probabilidad de que la edad media de dichas muestras sea superior a los 60 años?

7.- La edad de los alumnos de 2º Bachillerato de cierto instituto sigue una distribución *N(17’6 , 0’5)*. Los agrupamos al azar de 10 en 10 para una competición. ¿Cuál es la probabilidad de que en uno de esos grupos la edad media esté comprendida en el intervalo (17,18)?

8.- La duración (en horas) de un determinado tipo de pilas sigue una distribución normal con $\mu = 50$ y $\sigma = 5$. Empaquetamos las pilas en cajas de 16. ¿Cuál es la probabilidad de que la duración media de las pilas de una de las cajas sea inferior a 48 horas?

9.- En un instituto A las pruebas de acceso a la Universidad han obtenido una media de 5’8 con una desviación típica de 1’25, mientras que en otro instituto B la media ha sido de 5’6 con una desviación típica de 1’5. Si se toman al azar muestras de 100 personas de cada instituto, ¿en cuál de ellos es más probable que la nota media sea superior al 5’7?

10.- El peso de las truchas de una piscifactoría sigue una ley *N(200,50)*. Se toman muestras de 60 truchas y se calcula su peso medio. Hallar las probabilidades de que la media muestral:

a) Sea mayor que 210g
b) Sea menor que 185g
c) Esté entre 210 y 225g

11.- Una máquina produce el 5% de tornillos defectuosos. Si se empaquetan en cajas de 400 tornillos:

a) ¿Cuál es la distribución de la proporción de tornillos defectuosos en las cajas?
b) ¿Cuál es la probabilidad de que dicha proporción supere el 3%?
c) ¿Y de que esté entre el 3 y el 6%?

12.- En una población el 35% son fumadores. Se toman muestras de 60, 100 y 200 personas. ¿Cuál será la distribución de la proporción de fumadores en cada una de esas muestras?

13.- En el 77% de los hogares españoles hay teléfono móvil. Si se extrae una muestra de 500 viviendas, ¿Cuál es la probabilidad de que en más del 80% haya teléfono móvil? ¿Y si la muestra se toma de 1000 viviendas?

14.- Tras unas elecciones, se sabe que el candidato elegido obtuvo el 42% de los votos. Hallar la probabilidad de que de 1000 individuos elegidos al azar de entre los votantes hubiese obtenido dicho candidato más de 450 votos

15.- El 17% de familias de una determinada población tienen 3 o más hijos. Encuestadas 150 familias, ¿Cuál es la probabilidad de que más de 25 de ellas tengan 3 o más hijos?

16.- La duración de las baterías de determinados coches eléctricos se distribuye normalmente con media 225 y desviación típica 24 minutos. Si se eligen al azar 10 baterías, ¿Cuál es la probabilidad de que su duración media esté entre 220 y 235 minutos?

17.- Se ha determinado que 60% de los estudiantes de una universidad grande fuman cigarrillos. Se toma una muestra aleatoria de 800 estudiantes. Calcule la probabilidad de que la proporción de la muestra de la gente que fuma cigarrillos sea menor que 0.55.

18.- Se sabe que la verdadera proporción de los componentes defectuosos fabricados por una firma es de 4%, y encuentre la probabilidad de que una muestra aleatoria de tamaño 60 tenga:

a) Menos del 3% de los componentes defectuosos.

b) Más del 1% pero menos del 5% de partes defectuosas.

19.- Las estaturas de 1000 estudiantes están distribuidas aproximadamente en forma normal con una media de 174.5 centímetros y una desviación estándar de 6.9 centímetros. Si se extraen 200 muestras aleatorias de tamaño 25 sin reemplazo de esta población, determine:

a) El número de medias muestrales que caen entre 172’5 y 175’8 cm.
b) El número de medias muestrales que caen por debajo de 172 cm.

20.- Un medicamento para malestar estomacal tiene la advertencia de que algunos usuarios pueden presentar una reacción adversa a él, más aún, se piensa que alrededor del 3% de los usuarios tienen tal reacción. Si una muestra aleatoria de 150 personas con malestar estomacal usa el medicamento, encuentre la probabilidad de que la proporción de la muestra de los usuarios que realmente presentan una reacción adversa exceda el 4%.