Estadística

Ejercicios resueltos

Javier Blanco, Analista de Datos

www.linkedin.com/in/jblancop

ÍNDICE DE CONTENIDOS

1. SÍNTESIS DE LA INFORMACIÓN

Ejercicio I.1.

El número de crías por camada en una granja porcina para un periodo dado arroja los siguientes resultados: 4, 7, 2, 8, 6, 7, 2, 2, 9, 5, 5, 4, 5, 2, 6, 4, 7, 8, 4, 8. Se pide caracterizar esta distribución de valores de la forma más detallada posible.

a) Distribución de frecuencias

En primer lugar, resulta de utilidad ordenar la distribución en sentido creciente:

$$2, 2, 2, 2, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 7, 8, 8, 8, 9$$

Así, por ejemplo, se tiene que el primer valor, $x_1 = 2$, se repite cuatro veces, por lo que su frecuencia absoluta es $n_1 = 4$; por otro lado, el último, x_r, con $r = 7$, es 9 y sólo aparece una vez, por lo que $n_7 = 1$. De manera análoga para el resto de la serie se tiene:

i	x_i	n_i
1	2	4
2	4	4
3	5	3
4	6	2
5	7	3
6	8	3
7	9	1

x_i: Valores
n_i: Frecuencias absolutas

El sumatorio de frecuencias absolutas nos proporciona la frecuencia absoluta acumulada:

$$N_i = \sum_{j=1}^{i} n_j$$

Así, para el primer valor:

$$N_1 = \sum_{j=1}^{1} n_j = n_1 = 4$$

Para el segundo:

$$N_2 = \sum_{j=1}^{2} n_j = n_1 + n_2 = 4 + 4 = 8$$

Y así sucesivamente hasta el último, cuya frecuencia absoluta acumulada equivale al número total de observaciones, n:

$$N_r = N_7 = \sum_{j=1}^{7} n_j = n_1 + \cdots + n_7 = 4 + \cdots + 1 = 20 = n$$

El cociente entre la frecuencia absoluta de cada valor y el número total de observaciones nos da la frecuencia relativa:

$$f_i = \frac{n_i}{N_r} \equiv \frac{n_i}{n}$$

Para el primer valor:

$$f_1 = \frac{n_1}{N_7} \equiv \frac{n_1}{n} = \frac{4}{20} = 0{,}20$$

Y así hasta el último:

$$f_r = f_7 = \frac{n_7}{N_7} \equiv \frac{n_7}{n} = \frac{1}{20} = 0{,}05$$

Finalmente, de manera análoga a lo que se ha hecho con la frecuencia absoluta, se puede calcular la frecuencia relativa acumulada:

$$F_i = \sum_{j=1}^{i} f_j$$

Para el primer valor:

$$F_1 = \sum_{j=1}^{1} f_j = f_1 = 0{,}20$$

Para el segundo:

$$F_2 = \sum_{j=1}^{2} f_j = f_1 + f_2 = 0{,}20 + 0{,}20 = 0{,}40$$

Y así sucesivamente hasta el último, cuya frecuencia relativa acumulada equivale a la unidad:

$$F_r = F_7 = \sum_{j=1}^{7} f_j = f_1 + \cdots + f_7 = 0{,}20 + \cdots + 0{,}05 = 1{,}00$$

Se puede completar ahora la tabla esbozada inicialmente:

i	x_i	n_i	N_i	f_i	F_i
1	2	4	4	0,20	0,20
2	4	4	8	0,20	0,40
3	5	3	11	0,15	0,55
4	6	2	13	0,10	0,65
5	7	3	16	0,15	0,80
6	8	3	19	0,15	0,95
7	9	1	20	0,05	1,00

x_i: Valores
n_i: Frecuencias absolutas
N_i: Frecuencias absolutas acumuladas

f_i: Frecuencias relativas
F_i: Frecuencias relativas acumuladas

Y si se representa la distribución mediante un histograma:

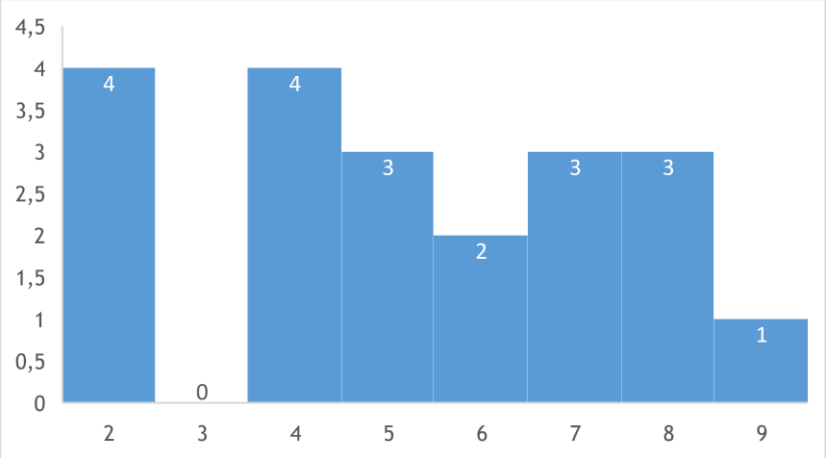

Figura I.1: Histograma de frecuencias absolutas

b) Medidas centrales

Dado que en el apartado anterior se han calculado las frecuencias relativas, la media aritmética se puede estimar de la siguiente manera:

$$\bar{x} = \sum_{i=1}^{r=7} x_i \cdot f_i = x_1 \cdot f_1 + \cdots + x_7 \cdot f_7 = 2 \cdot 0{,}20 + \cdots + 9 \cdot 0{,}05 = 5{,}25$$

x_i: Valores
f_i: Frecuencias relativas

Lo que querría decir que, en promedio, cada hembra pare 5 crías.

Para calcular la mediana y la moda se puede recurrir de nuevo a la serie ordenada de valores:

$$2, 2, 2, 2, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 7, 8, 8, 8, 9$$

La mediana parte en dos la distribución, dejando el $50\ \%$ de los datos tanto por arriba como por abajo; dado que hay un número par de observaciones, la mediana vendrá dada —por convenio— por la media aritmética de los dos valores centrales, que en este caso resultan ser sendos cincos; por tanto, 5.

Se trata de una distribución equilibrada por cuanto que media y mediana son similares.

La moda es el resultado más común; es decir, el de mayor frecuencia absoluta. Tanto el 2 como el 4 tienen $n_i = 4$, por lo que se podría considerar moda cualquiera de los dos.

c) Medidas de posición

A modo de ejemplo, puesto que las posibilidades son muy amplias, se van a calcular el tercer cuartil y el percentil veinte.

En general, el cuantil r de orden k es el valor x_i cuya frecuencia absoluta acumulada es la inmediatamente superior a:

$$N_i = \frac{r}{k} \cdot n$$

N_i: Frecuencia absoluta acumulada
r: Posición del cuantil
k: Orden del cuantil

n: Número total de observaciones

Para Q_3:

$$N_i = \frac{3}{4} \cdot 20 = 15$$

Si se consulta la tabla del primer apartado, la frecuencia acumulada inmediatamente superior a 15 es $N_5 = 16$; esto quiere decir que $x_5 = 7$ es el valor que deja tres cuartas partes (el $75\,\%$) de la distribución por debajo de él.

De manera análoga para P_{20}:

$$N_i = \frac{20}{100} \cdot 20 = 4$$

Dado que $N_1 = 4$ se puede considerar que el valor que deja veinte centésimas partes (el $20\,\%$) de la distribución por debajo de él es la media aritmética de x_1 y x_2; es decir, 3.

De la misma manera que se podrían calcular de forma gráfica estos cuantiles a partir de la lista ordenada de valores, se puede calcular analíticamente la mediana, que equivale al quinto decil —y también a Q_2 y P_{50}—.

Para D_5:

$$N_i = \frac{5}{10} \cdot 20 = 10$$

La frecuencia acumulada inmediatamente superior a 10 es $N_3 = 11$; esto quiere decir que $x_3 = 5$ es el valor que deja cinco décimas partes (el $50\,\%$) de la distribución por debajo de él, como ya se había comprobado previamente.

d) Medidas de dispersión

De nuevo, gracias a que se han calculado previamente las frecuencias relativas, se puede estimar la varianza de la siguiente forma:

$$S_X^2 = \sum_{i=1}^{r=7} (x_i - \bar{x})^2 \cdot f_i = (x_1 - \bar{x})^2 \cdot f_1 + \cdots + (x_7 - \bar{x})^2 \cdot f_7$$

$$S_X^2 = (2 - 5{,}25)^2 \cdot 0{,}20 + \cdots + (9 - 5{,}25)^2 \cdot 0{,}05 = 4{,}79$$

x_i: Valores
$\bar{x}$: Media aritmética
f_i: Frecuencias relativas

Dado que la varianza está en unidades al cuadrado, la desviación típica ofrece una interpretación más sencilla de la dispersión de los datos:

$$S_X = +\sqrt{S_X^2} = +\sqrt{5{,}92} = 2{,}19$$

En promedio, existe una distancia de $2{,}19$ unidades entre cada observación y la media, lo cual, de manera intuitiva y a tenor de los valores de la distribución, no parece desdeñable.

El rango, que es la distancia entre los valores extremos, sería:

$$R = x_{r=7} - x_1 = 9 - 2 = 7$$

Y el recorrido intercuartílico, que es la diferencia entre el tercer y el primer cuartil —calculado éste de manera análoga a aquel—:

$$R_I = Q_3 - Q_1 = 7 - 4 = 3$$

Es decir, el $50\,\%$ de los valores centrales se encuentran entre el 4 y el 7.

El coeficiente de variación, al ser adimensional, permite hacerse una mejor idea de la dispersión de los datos:

$$CV = \frac{S_X}{|\bar{x}|} = \frac{2{,}19}{5{,}25} = 0{,}42$$

S_X: Desviación típica
$\bar{x}$: Media aritmética

El resultado se encuentra en el rango correspondiente a valores medios, entre $0{,}1$ y $0{,}5$, si bien cerca del límite superior, a partir del cual se puede decir que la dispersión es alta y que la media no representa adecuadamente a la distribución.

e) Agrupación por intervalos

Aunque para este número de observaciones no es necesario, en algunos casos puede ser apropiado agrupar los valores por intervalos para facilitar los cálculos.

El número de intervalos debe ser:

$$k = \sqrt{n} = \sqrt{20} = 4{,}47 \cong 5$$

n: Número total de observaciones

Y su amplitud:

$$A = \frac{R}{\sqrt{n}} = \frac{7}{5} = 1{,}4 \cong 1{,}5$$

R: Rango

Nótese que en este caso también se redondea hacia arriba, al objeto de evitar problemas de solapamiento entre los valores de la distribución.

Los intervalos se toman cerrados por su izquierda y abiertos por su derecha, $[L_i, L_{i+1})$, y sus extremos se pueden calcular según la siguiente progresión aritmética:

$$L_i = L_1 + (i-1) \cdot A \quad i = 2, \dots, k+1$$

L_1 sería el valor mínimo de la distribución menos la mitad de la precisión, que en este caso se ha marcado en las décimas:

$$L_1 = 2 - \frac{1}{2} \cdot 0{,}1 = 1{,}95$$

Por tanto, el extremo superior del primer intervalo sería:

$$L_2 = 1{,}95 + (2-1) \cdot 1{,}5 = 3{,}45$$

El valor que caracteriza a cada intervalo, su marca de clase, es el punto intermedio:

$$x_i = \frac{L_i + L_{i+1}}{2}$$

En consecuencia, para el primer intervalo:

$$x_1 = \frac{L_1 + L_2}{2} = \frac{1{,}95 + 3{,}45}{2} = 2{,}7$$

Así, la tabla de frecuencias quedaría como sigue:

i	x_i	$[L_i, L_{i+1})$	n_i
1	2,7	[1,95; 3,45)	4
2	4,2	[3,45; 4,95)	4
3	5,7	[4,95; 6,45)	5
4	7,2	[6,45; 7,95)	3

5	8,7	[7,95; 9,45)	4

x_i: Marcas de clase
L_i: Límites inferiores de los intervalos
L_{i+1}: Límites superiores de los intervalos
n_i: Frecuencias absolutas

Y si se representa la distribución mediante un histograma:

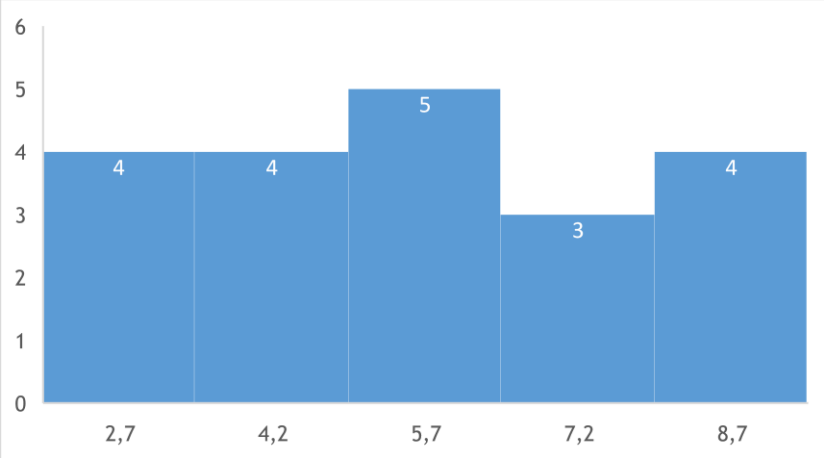

Figura I.2: Histograma de frecuencias absolutas agrupando por intervalos

Ejercicio I.2.

Dados dos trabajadores de distintas compañías del mismo sector, se quiere saber cuál de ellos ocupa mejor posición relativa dentro de su empresa atendiendo a los siguientes datos:

Empresa	X	$\bar{x}$	S_X
A	620 €	580 €	25 €
B	672 €	640 €	33 €

X: Sueldo del trabajador
$\bar{x}$: Sueldo medio de la empresa
S_X: Desviación típica

Se ha de proceder a la tipificación de las variables para facilitar su comparación:

$$Z_A = \frac{X_A - \overline{x_A}}{S_{X_A}} = \frac{(620 - 580)\ €}{25\ €} = 1{,}60$$

$$Z_B = \frac{X_B - \overline{x_B}}{S_{X_B}} = \frac{(672 - 640)\ €}{33\ €} = 0{,}97$$

Esto quiere decir que el trabajador de la empresa A ocupa mejor posición relativa que el de la B.

II. ANÁLISIS BIVARIADO

Ejercicio II.1.

Se ha clasificado el peso de los huevos, Y, de una especie de pez en función del peso de la madre, X, como refleja la tabla adjunta, y se pide estimar una serie de características de la distribución.

X/Y	$[25,27)$	$[27,29)$	$[29,31)$	$[31,33)$
$[500,550)$	15	11	18	0
$[550,600)$	12	14	0	12
$[600,650)$	0	3	7	18

a) Distribución del peso del huevo

Se trata de la distribución marginal de Y; para la primera columna:

$$n_{\cdot 1} = \sum_{i=1}^{r=3} n_{i1} = n_{11} + n_{21} + n_{31} \equiv n_{[25,27)} = n\big(y \in [25,27)\big) =$$
$$= n\big(x \in [500,550), y \in [25,27)\big) + n\big(x \in [550,600), y \in [25,27)\big) +$$
$$+ n\big(x \in [600,650), y \in [25,27)\big) = 15 + 12 + 0 = 27$$

n_{ij}: Frecuencias absolutas del par (x, y)

Y de manera análoga para el resto de columnas —y de filas; es decir, la distribución marginal del peso de la madre—, se tiene:

X/Y	$[25,27)$	$[27,29)$	$[29,31)$	$[31,33)$	$f(X)$
$[500,550)$	15	11	18	0	44
$[550,600)$	12	14	0	12	38
$[600,650)$	0	3	7	18	28
$f(Y)$	27	28	25	30	110

b) Distribución del peso de la madre cuando el huevo tiene el suyo comprendido entre 25 *y* 27

Se trata de la distribución de la variable X condicionada a que Y tome valores en el intervalo $[25,27)$; para la primera fila:

$$f_{1|1} = \frac{n_{11}}{n_{\cdot 1}} \equiv f_{[500,550)|[25,27)} = \frac{n\big(x \in [500,550), y \in [25,27)\big)}{n\big(y \in [25,27)\big)} = \frac{15}{27} = \frac{5}{9} = 0{,}56$$

n_{ij}: Frecuencias absolutas del par (x, y)
$n_{\cdot j}$: Frecuencias marginales absolutas de Y

De manera análoga para las otras dos filas, se tiene:

$$f_{2|1} = f_{[550,600)|[25,27)} = \frac{12}{27} = \frac{4}{9} = 0{,}44$$

$$f_{3|1} = f_{[600,650)|[25,27)} = \frac{0}{27} = 0$$

Así, para la categoría inferior de peso de los huevos, el $56\,\%$ de las madres se encuentran a su vez en la categoría inferior respectiva y el $44\,\%$, en la intermedia.

c) Media, mediana y moda del peso de los huevos

La media de Y se podría calcular de la siguiente forma:

$$\bar{y} = \frac{1}{n} \cdot \sum_{j=1}^{s=4} y_j \cdot n_{\cdot j} = \frac{y_1 \cdot n_{\cdot 1} + \cdots + y_4 \cdot n_{\cdot 4}}{n} = \frac{26 \cdot 27 + 28 \cdot 28 + 30 \cdot 25 + 32 \cdot 30}{110} = 29{,}05$$

n: Número total de observaciones
y_j: Marcas de clase
$n_{\cdot j}$: Frecuencias marginales absolutas de Y

Para determinar la mediana, que sería el segundo cuartil, hay que encontrar el intervalo de Y cuya frecuencia marginal absoluta acumulada es inmediatamente superior a:

$$N_{\cdot j} = \frac{r}{k} \cdot n = \frac{2}{4} \cdot 110 = 55$$

$N_{\cdot j}$: Frecuencia marginal absoluta acumulada de Y
r: Posición del cuantil
k: Orden del cuantil
n: Número total de observaciones

Se comprueba fácilmente que se trata del tercer intervalo, $[29, 31)$:

$$N_{[29,31)} = N_{\cdot 3} = \sum_{j=1}^{s-1=3} n_{\cdot j} = n_{\cdot 1} + n_{\cdot 2} + n_{\cdot 3} = 27 + 28 + 25 = 80$$

$n_{\cdot j}$: Frecuencias marginales absolutas de Y

La mediana en sí se calcularía como sigue:

$$Me = L_{3i} + A \cdot \frac{\frac{n}{2} - N_{\cdot 2}}{n_{\cdot 3}} = 29 + (31 - 29) \cdot \frac{\frac{110}{2} - 55}{25} = 29$$

L_{3i}: Límite inferior del tercer intervalo
A: Amplitud del intervalo
n: Número total de observaciones
$N_{\cdot 2}$: Frecuencia absoluta acumulada marginal del segundo intervalo
$n_{\cdot 3}$: Frecuencia absoluta marginal del tercer intervalo

Por otro lado, la moda se encontraría dentro del intervalo $[31, 33)$, puesto que es el que mayor frecuencia marginal tiene.

d) Nivel de representatividad de la media del peso de la madre cuando el huevo está comprendido entre 25 *y* 27

Se puede estimar a través del coeficiente de variación:

$$CV = \frac{S_X}{|\bar{x}|}$$

S_X: Desviación típica
$|\bar{x}|$: Media aritmética

La media aritmética de X para ese intervalo de Y se calcula como sigue:

$$\bar{x}|Y \in [25, 27) = \sum_{i=1}^{r=3} x_i \cdot f_{i|1} = 525 \cdot \frac{5}{9} + 575 \cdot \frac{4}{9} + 625 \cdot 0 = 547{,}22$$

x_i: Marcas de clase
$f_{i|j}$: Frecuencias condicionadas de X a Y

Para calcular la desviación típica, primero se calcula la varianza:

$$S^2_{X|Y\in[25,27)} = \sum_{i=1}^{r=3} x_i^2 \cdot f_{i|1} - \left(\bar{x}|Y \in [25,27)\right)^2 =$$

$$= 525^2 \cdot \frac{5}{9} + 575^2 \cdot \frac{4}{9} + 625^2 \cdot 0 - (547{,}22)^2 = 617{,}28$$

Así:

$$CV = \frac{S_X}{|\bar{x}|} = \frac{\sqrt{617{,}28}}{547{,}22} = 0{,}045$$

Dado que $CV < 0{,}1$, se puede decir que la concentración de los datos es alta y, por tanto, la media es representativa de la distribución para ese intervalo de Y.

e) Independencia de las variables

Las variables son independientes si la frecuencia condicionada coincide con la marginal, lo que también se puede expresar de la siguiente manera:

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j} \quad \forall_{i,j}$$

f_{ij}: Frecuencias relativas del par (x, y)
$f_{i\cdot}$: Frecuencias marginales relativas de X
$f_{\cdot j}$: Frecuencias marginales relativas de Y

Así, para el par $(x, y) \in [500, 550) \times [25, 27)$, se tiene:

$$f_{11} = \frac{n_{11}}{n} = \frac{15}{110}$$
$$f_{1\cdot} = \frac{n_{1\cdot}}{n} = \frac{44}{110}$$
$$f_{\cdot 1} = \frac{n_{\cdot 1}}{n} = \frac{27}{110}$$

n_{ij}: Frecuencias absolutas del par (x, y)
n: Número total de observaciones
$n_{i\cdot}$, $n_{\cdot j}$: Frecuencias marginales absolutas

Dado que:

$$f_{11} = \frac{15}{110} \neq \frac{44}{110} \cdot \frac{27}{110} = f_{1\cdot} \cdot f_{\cdot 1}$$

Las variables no son independientes.

f) Dependencia lineal de las variables

El grado de dependencia lineal se puede cuantificar a través del coeficiente de correlación de Pearson:

$$r = \frac{S_{XY}}{S_X \cdot S_Y}$$

S_{XY}: Covarianza de X e Y
S_X: Desviación típica de X
S_Y: Desviación típica de Y

En primer lugar, se calculan las varianzas:

$$S_X^2 = \sum_{i=1}^{r=3} x_i^2 \cdot f_{i\cdot} - \bar{x}^2 = 1.583{,}47$$

$$S_Y^2 = \sum_{j=1}^{s=4} y_j^2 \cdot f_{\cdot j} - \bar{y}^2 = 5{,}14$$

x_i, y_j: Marcas de clase
$f_{i\cdot}$, $f_{\cdot j}$: Frecuencias marginales relativas
$\bar{x}$, $\bar{y}$: Medias aritméticas

Y a continuación, la covarianza, que se podría expresar como:

$$S_{XY} = \sum_{i=1}^{r=3} \sum_{j=1}^{s=4} x_i \cdot y_j \cdot f_{ij} - \bar{x} \cdot \bar{y} = 44{,}03$$

f_{ij}: Frecuencias relativas del par (x, y)

Así:

$$r = \frac{44{,}03}{\sqrt{1.583{,}47 \cdot 5{,}14}} = 0{,}49$$

Que indica que hay una relación lineal positiva pero no muy acentuada, puesto que $0{,}49$ está considerablemente lejos de 1.

III. AJUSTE Y REGRESIÓN BIDIMENSIONAL

Ejercicio III.1.

Dada la distribución bidimensional de la tabla adjunta, determinar la recta de ajuste de Y en función de X y la bondad de dicho ajuste.

X	1	2	3	4	5	6
Y	2	5	9	13	17	21

En primer lugar, se determina la media aritmética de cada variable:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1 + \cdots + 6}{6} = 3{,}50$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i = \frac{2 + \cdots + 21}{6} = 11{,}17$$

x_i, y_i: Valores
n: Número total de observaciones

A continuación, la varianza de X:

$$S_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{(1 - 3{,}50)^2 + \cdots + (6 - 3{,}50)^2}{6} = 2{,}92$$

A partir de las medias aritméticas, se puede estimar la covarianza de ambas variables:

$$S_{XY} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n} - \bar{x} \cdot \bar{y} = \frac{1 \cdot 2 + \cdots + 6 \cdot 21}{6} - 3{,}50 \cdot 11{,}17 = 11{,}25$$

Ahora se puede hacer el cálculo de los parámetros de la recta de ajuste:

$$b = \frac{S_{xy}}{S_x^2} = \frac{11{,}25}{2{,}92} = 3{,}86$$

$$a = \bar{y} - b \cdot \bar{x} = 11{,}17 - 3{,}86 \cdot 3{,}50 = -2{,}33$$

a: Punto de corte de ordenadas
b: Pendiente

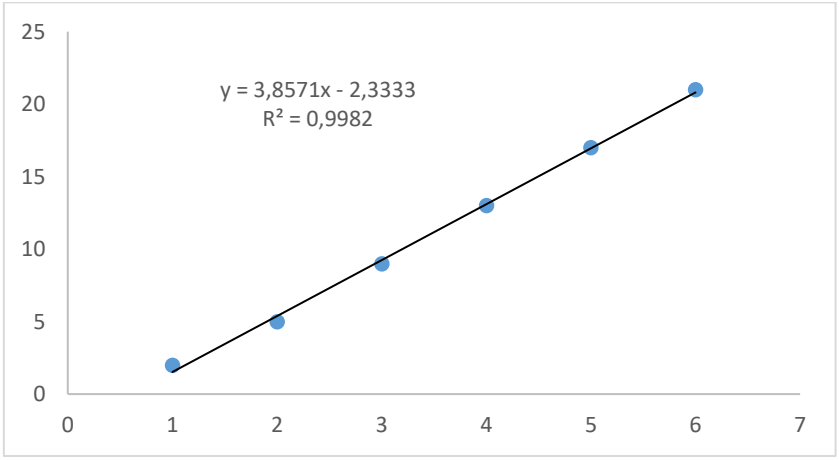

Figura III.1: Recta de ajuste de Y en función de X

Por tanto, la recta de ajuste de Y en función de X:

$$y = 3{,}86 \cdot x - 2{,}33$$

Y la bondad del ajuste:

$$R^2 = \frac{S_{XY}^2}{S_X^2 \cdot S_Y^2} = \frac{11{,}25^2}{2{,}92 \cdot 43{,}47} = 0{,}9982$$

El $99{,}82\ \%$ de la variabilidad de Y queda explicada por X a través de la recta de ajuste; por tanto, el modelo propuesto explica casi perfectamente las variaciones de la variable Y.

IV. COMBINATORIA

Ejercicio IV.1.

En un instituto, los alumnos de segundo de bachillerato deciden realizar un sorteo para el viaje de fin de curso; para numerar las papeletas, utilizan únicamente los dígitos 1, 2, 3, 4 y 5. Cuántas papeletas distintas de cuatro dígitos podrán vender si:

a) Los cuatros dígitos son distintos

Se trata de una variación sin repetición (dado que los dígitos han de ser distintos) de cinco elementos (los cinco dígitos posibles) tomados de cuatro en cuatro:

$$V_{5,4} = \frac{5!}{(5-4)!} = \frac{5!}{1!} = 5 \cdot 4 \cdot 3 \cdot 2 = 120$$

b) Pueden aparecer dígitos repetidos

También es una variación, pero en este caso con repetición:

$$VR_{5,4} = 5^4 = 625$$

c) Aparecen tres unos y un cinco

Dado que se forman dos grupos, uno de tres unos y otro con un único cinco, a posicionar en cuatro lugares, todos los dígitos propuestos han de intervenir en cada papeleta, por lo que se trata de una permutación con repetición:

$$PR_4^{3,1} = \frac{4!}{3!\,1!} = 4$$

d) Sólo se utilizan los dígitos 2, 3, 4 y 5 sin repetir ninguno

Se trata de una permutación sin repetición de cuatro elementos (en este caso no se dice que sean cuatro dígitos cualesquiera de los cinco posibles, sino cuatro en concreto) tomados de cuatro en cuatro:

$$P_4 = V_{4,4} = 4! = 24$$

e) Sólo se utilizan los dígitos 2, 3, 4 y 5 pero se pueden repetir

Es un caso análogo al de *b)*, pero al igual que en *d)* se indica exactamente qué dígitos se han de usar, y no cualquiera de los cinco, por lo que:

$$VR_{4,4} = 4^4 = 256$$

f) No se tiene en cuenta el orden, pero los dígitos son distintos

Como se pretende elegir cuatro elementos de un total de cinco y no importa el lugar que ocupen, se trata de una combinación sin repetición:

$$C_{5,4} = \frac{V_{5,4}}{P_4} = \frac{5!}{(5-4)!\,4!} = \frac{5!}{1!\,4!} = 5$$

g) No se tiene en cuenta el orden, pero los dígitos pueden estar repetidos

En este caso, se trata de una combinación con repetición:

$$CR_{5,4} = \binom{5+4-1}{4} = \binom{8}{4} = \frac{(5+4-1)!}{4!\,(5-1)!} = \frac{8!}{4!\,4!} = 70$$

V. PROBABILIDAD

Ejercicio V.1.

Sabiendo que $P(A \cap B) = 0,6$ y que $P(A \cap \bar{B}) = 0,2$, se pide calcular la probabilidad de A.

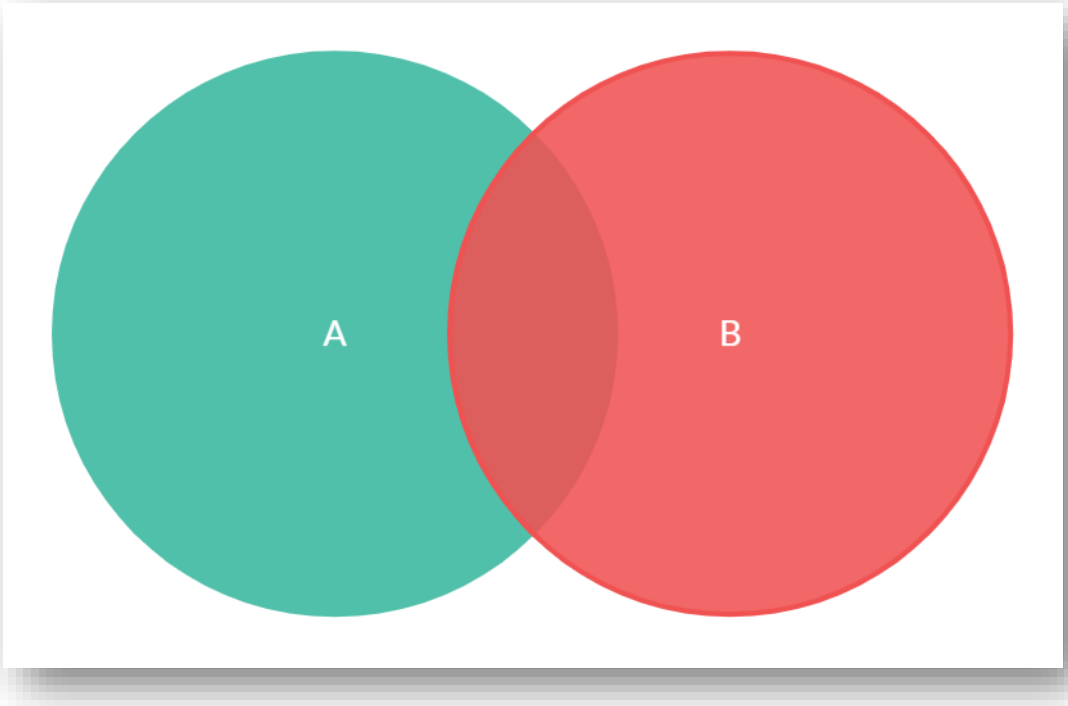

Figura V.1: Intersección de dos conjuntos

Se sabe que la probabilidad de la zona de intersección de ambos conjuntos, en rojo oscuro, es:

$$P(A \cap B) = 0{,}6$$

Por otro lado, el conjunto complementario de B, $\bar{B}$, equivaldría a todo el espacio ajeno a B, tanto la parte coloreada en verde de A —con la que no intersecta— como la zona en blanco; por tanto, $A \cap \bar{B}$ equivale precisamente a dicha parte de A coloreada sólo en verde, y a partir del enunciado se sabe que:

$$P(A \cap \bar{B}) = 0{,}2$$

Por tanto, la probabilidad de A equivaldría a la probabilidad de la zona coloreada en rojo oscuro más la probabilidad de la zona coloreada en verde, es decir:

$$P(A) = P[(A \cap B) \cup (A \cap \bar{B})]$$

Dada la siguiente notación:

$$X_1 = (A \cap B)$$
$$X_2 = (A \cap \bar{B})$$

Se tiene:

$$P(X_1 \cup X_2) = P(X_1) + P(X_2) - P(X_1 \cap X_2)$$

Si X_1 y X_2 fueran conjuntos disjuntos —o sucesos incompatibles—, mutuamente excluyentes, $P(X_1 \cap X_2) = P(\emptyset) = 0$; es lo que se pretende comprobar, puesto que ésa es la única probabilidad desconocida:

$$P(X_1 \cap X_2) = P[(A \cap B) \cap (A \cap \bar{B})]$$

Merced a la propiedad distributiva:

$$P[(A \cap B) \cap (A \cap \bar{B})] = P(A \cap B \cap \bar{B})$$

Dado que $B \cap \bar{B} = \emptyset$, $P(A \cap \emptyset) = 0$ y, por tanto y en efecto, $P(X_1 \cap X_2) = 0$.

En conclusión:

$$P(A) = P(X_1) + P(X_2) - P(X_1 \cap X_2) = 0{,}6 + 0{,}2 - 0 = 0{,}8$$

Ejercicio V.2.

Sea una urna con dos bolas blancas, tres negras y cuatro rojas, denotada como $\boldsymbol{U(2B, 3N, 4R)}$. Se extraen tres bolas de forma sucesiva: la primera es negra, la segunda se oculta y la tercera es blanca. Hallar la probabilidad de que la segunda sea roja.

La extracción de una bola negra en primer lugar deja la urna como sigue:

$$U(2B, 2N, 4R)$$

En segundo lugar, existen tres posibilidades, ya que la bola extraída podría ser de cualquier color, blanca, negra o roja:

$$U_1(1B, 2N, 4R)$$
$$U_2(2B, 1N, 4R)$$
$$U_3(2B, 2N, 3R)$$

En tercer y último lugar, se sabe a ciencia cierta que se ha extraído una bola blanca.

Por tanto, se conoce el resultado final del experimento y se desea determinar la probabilidad de ocurrencia de la etapa previa; es decir, la probabilidad de obtener una bola roja en la segunda etapa, $P(U_3)$, condicionada a que se haya extraído una blanca en la tercera, $P(B)$, lo que implica aplicar el teorema de Bayes, que a su vez hace uso del de la probabilidad total:

$$P(U_3|B) = \frac{P(B|U_3) \cdot P(U_3)}{P(B)} = \frac{P(B|U_3) \cdot P(U_3)}{\sum_{i=1}^{3} P(B|U_i) \cdot P(U_i)}$$

Dado que en la primera etapa quedan ocho bolas, las probabilidades de extraer una blanca (dos casos favorables), una negra (dos casos favorables) o una roja (cuatro casos favorables) en la segunda son las siguientes:

$$P(U_1) = P(U_2) = \frac{2}{8} = \frac{1}{4}$$
$$P(U_3) = \frac{4}{8} = \frac{1}{2}$$

Por otro lado, la probabilidad de extraer una bola blanca en la tercera fase estará condicionada por lo ocurrido en la segunda, que es algo incierto, pero se tiene por seguro que al llegar a la última etapa sólo quedan siete bolas (o casos posibles); así, caben tres posibilidades, teniendo en cuenta que en el supuesto U_1 sólo habría una bola blanca (un caso favorable) y tanto en U_2 como en U_3, dos:

$$P(B|U_1) = \frac{1}{7}$$

$$P(B|U_2) = P(B|U_3) = \frac{2}{7}$$

Por tanto, la probabilidad de que la segunda bola sea roja es:

$$P(U_3|B) = \frac{\frac{2}{7}\cdot\frac{1}{2}}{\frac{1}{7}\cdot\frac{1}{4}+\frac{2}{7}\cdot\frac{1}{4}+\frac{2}{7}\cdot\frac{1}{2}} = \frac{4}{7}$$

Ejercicio V.3.

Para un sistema de alarma, la probabilidad de que funcione en caso de peligro es de $0,95$, mientras que la de que funcione por error —sin que haya peligro— es de $0,03$. Si la probabilidad de que haya peligro es de $0,1$:

a) *Calcular el porcentaje de veces en las que, habiendo funcionado la alarma, no hubiese realmente peligro*

Se puede definir un suceso P, que haya peligro, y su complementario, $\bar{P}$, que no haya peligro, de manera que:

$$P(P) = 0{,}10 \Longrightarrow P(\bar{P}) = 1 - 0{,}10 = 0{,}90$$

También se puede definir un suceso F, que la alarma funcione, y su complementario, $\bar{F}$, que la alarma no funcione, de manera que se tienen las siguientes probabilidades condicionadas:

$$P(F|P) = 0{,}95 \Longrightarrow P(\bar{F}|P) = 1 - 0{,}95 = 0{,}05$$
$$P(F|\bar{P}) = 0{,}03 \Longrightarrow P(\bar{F}|\bar{P}) = 1 - 0{,}03 = 0{,}97$$

Por otro lado, lo que se pide en este apartado es también una probabilidad condicionada, pero de que no haya peligro si ha funcionado la alarma, la cual se puede calcular a través del teorema de Bayes:

$$P(\bar{P}|F) = \frac{P(F|\bar{P})\cdot P(\bar{P})}{P(F)}$$

El numerador, a su vez, se puede calcular según el teorema de la probabilidad total:

$$P(F) = \sum_{i=1}^{2} P(F \mid P_i)\cdot P(P_i) = P(F|P)\cdot P(P) + P(F|\bar{P})\cdot P(\bar{P}) =$$
$$= 0{,}95\cdot 0{,}10 + 0{,}03\cdot 0{,}90 = 0{,}122$$

Obviamente, no se refiere a la probabilidad de que la alarma sea capaz de funcionar en un sentido técnico —se asume que la alarma no está defectuosa— sino de que se active o no.

Por tanto:

$$P(\bar{P}|F) = \frac{0{,}03\cdot 0{,}90}{0{,}122} = 0{,}2213 \equiv 22{,}13\,\%$$

b) *Hallar la probabilidad de que haya peligro y la alarma no funcione*

Se trata de la intersección entre ambos conjuntos:

$$P(P \cap \bar{F})$$

Por la definición de probabilidad condicionada se tiene:

$$P(P|\bar{F}) = \frac{P(P \cap \bar{F})}{P(P)} \Leftrightarrow P(P|\bar{F}) \cdot P(P) = P(P \cap \bar{F}) \equiv P(\bar{F} \cap P) = P(\bar{F}|P) \cdot P(P)$$

Por tanto, se tiene:

$$P(P \cap \bar{F}) = P(\bar{F}|P) \cdot P(P) = 0{,}05 \cdot 0{,}10 = 0{,}005 \equiv 0{,}50\ \%$$

c) *Calcular la probabilidad de que, no habiendo funcionado la alarma, haya peligro*

De nuevo, por el teorema de Bayes:

$$P(P|\bar{F}) = \frac{P(\bar{F}|P) \cdot P(P)}{P(\bar{F})} = \frac{P(\bar{F}|P) \cdot P(P)}{1 - P(F)} = \frac{0{,}05 \cdot 0{,}10}{1 - 0{,}122} = 0{,}0057 \equiv 0{,}57\ \%$$

VI. VARIABLE ALEATORIA

Ejercicio VI.1:

La función de distribución asociada a la producción de una máquina, en miles de unidades, es la siguiente:

$$F(x) = \begin{cases} 0, & x < 0 \\ x \cdot (2 - x), & 0 \leq x \leq k \\ 1, & x > k \end{cases}$$

a) Determinar k para que F sea efectivamente una función de distribución

Cuando se trata de una VA continua, su función de distribución, a diferencia de para el caso discreto, no presenta discontinuidades de salto sino que es continua en todo $\mathbb{R}$; por tanto, $F(x)$ ha de serlo también en el punto de división k, para lo cual han de coincidir los límites laterales:

$$\lim_{x \to k^+} F(x) = \lim_{x \to k^-} F(x)$$

$$\lim_{x \to k^+} 1 = 1$$

$$\lim_{x \to k^-} x \cdot (2 - x) = k \cdot (2 - k)$$

$$1 = k \cdot (2 - k) \Longrightarrow k^2 - 2 \cdot k + 1 = 0 \Longrightarrow k = 1$$

Así, para $k = 1$:

$$F(x) = \begin{cases} 0, & x < 0 \\ x \cdot (2 - x), & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

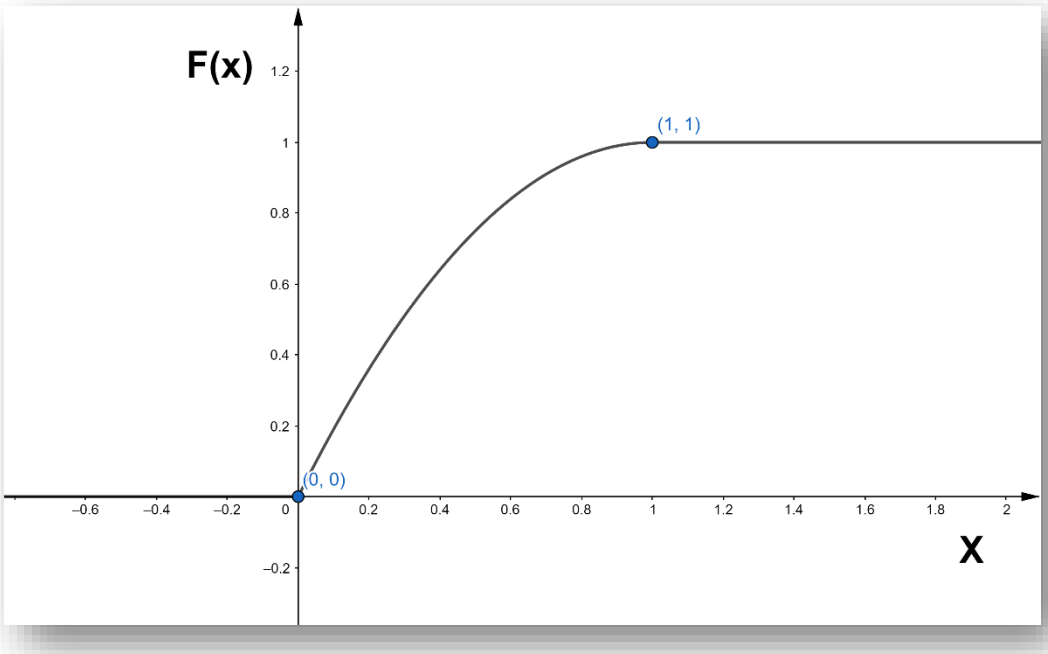

Figura VI.1: Función de distribución

b) Calcular la función de densidad

Se obtiene por derivación de la función de distribución:

$$f(x) = \begin{cases} 2 - 2 \cdot x, & 0 \leq x \leq 1 \\ 0, & en\ el\ resto \end{cases}$$

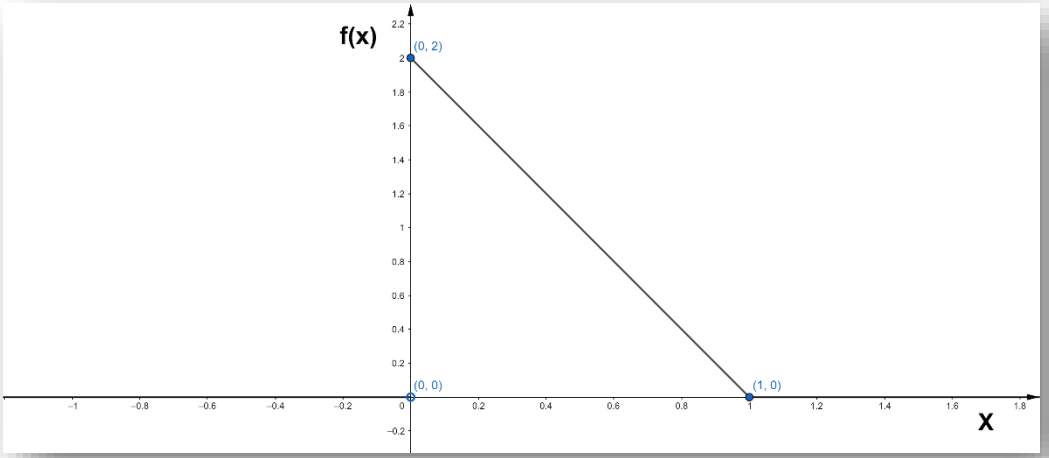

Figura VI.2: Función de densidad

Se observa que la probabilidad de producir menos de 0 unidades o más de 1.000 es nula, y, dada la pendiente de la recta, la probabilidad de producir pocas unidades es mayor que la de producir muchas; por ejemplo, a pesar de que Δ tiene la misma amplitud tanto para la producción de menos de 200 unidades como de más de 800, el área bajo la curva y, por tanto, la probabilidad es mucho mayor en el primer caso que en el segundo.

c) Obtener la media y la varianza de la producción

La media viene dada por la esperanza matemática:

$$E[X] = \int_{-\infty}^{+\infty} x \cdot f(x) \cdot dx$$

Dado que la función vale 0 fuera del intervalo $[0, 1]$:

$$\mu = \int_0^1 x \cdot (2 - 2 \cdot x) \cdot dx = \left(x^2 - \frac{2 \cdot x^3}{3} \right) \Big|_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

Es decir, la producción media sería de $333{,}33$ unidades.

Y la varianza se puede calcular a partir de la media:

$$V[X] = \sigma^2 = E[X^2] - (E[X])^2$$

$$E[X^2] = \int_{-\infty}^{+\infty} x^2 \cdot f(x) \cdot dx = \int_0^1 x^2 \cdot (2 - 2 \cdot x) \cdot dx = \left(\frac{2}{3} \cdot x^3 - \frac{1}{2} \cdot x^4 \right) \Big|_0^1 = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$$

$$\sigma^2 = \frac{1}{6} - \left(\frac{1}{3} \right)^2 = \frac{1}{18}$$

Es decir, la varianza de la producción sería de $55{,}55$ unidades al cuadrado, lo que supone una desviación típica de $7{,}45$ unidades.

d) ¿Cuál es la probabilidad de que la producción sea inferior a 500 unidades? ¿Y superior a 250?

A partir de la función de densidad:

$$P(X \leq 0{,}5) = \int_{-\infty}^{0{,}5} f(x) \cdot dx = \int_{0}^{0{,}5} (2 - 2 \cdot x) \cdot dx = (2 \cdot x - x^2)|_0^{0{,}5} = 0{,}75$$

$$P(X > 0{,}25) = 1 - P(X \leq 0{,}25) = 1 - \int_{0}^{0{,}25} (2 - 2 \cdot x) \cdot dx = 1 - (2 \cdot x - x^2)|_0^{0{,}25} = 0{,}56$$

VII. MODELOS PROBABILÍSTICOS

Ejercicio VII.1:

En una población de animales, se sabe que el 60 % son machos. Si se toma una muestra de 10 individuos, ¿cuál es la probabilidad de que 7 sean hembras?

Si se tomase un animal al azar de la población con el objetivo de determinar su sexo, se estaría ante un experimento de Bernoulli, puesto que sólo hay dos posibles resultados: macho y hembra; a través de la VA discreta "sexo del animal", ambos sucesos se podrían asociar a los valores 1 y 0, con probabilidades invariantes p y q, respectivamente.

Dado que el 60 % de la población son machos, se tiene:

$$P(Y = 1) = 0{,}6$$
$$P(Y = 0) = 1 - 0{,}6 = 0{,}4$$

Por otro lado, la muestra tomada, también al azar, equivale a realizar 10 experimentos de Bernoulli, por lo que se puede definir otra VA discreta "número de hembras", sumatorio de las 10 VVAA de Bernoulli, que sigue una distribución binomial:

$$X \sim B(10, 0{,}4)$$

Por tanto, la probabilidad de que haya 7 hembras en la muestra es:

$$P(X = 7) = \binom{10}{7} \cdot 0{,}4^7 \cdot (1 - 0{,}4)^{10-7} = \frac{10!}{7! \cdot (10-7)!} \cdot 0{,}4^7 \cdot 0{,}6^3 = 0{,}042$$

Es decir, un 4,2 %.

Si se construye la función de probabilidad con todos los resultados posibles:

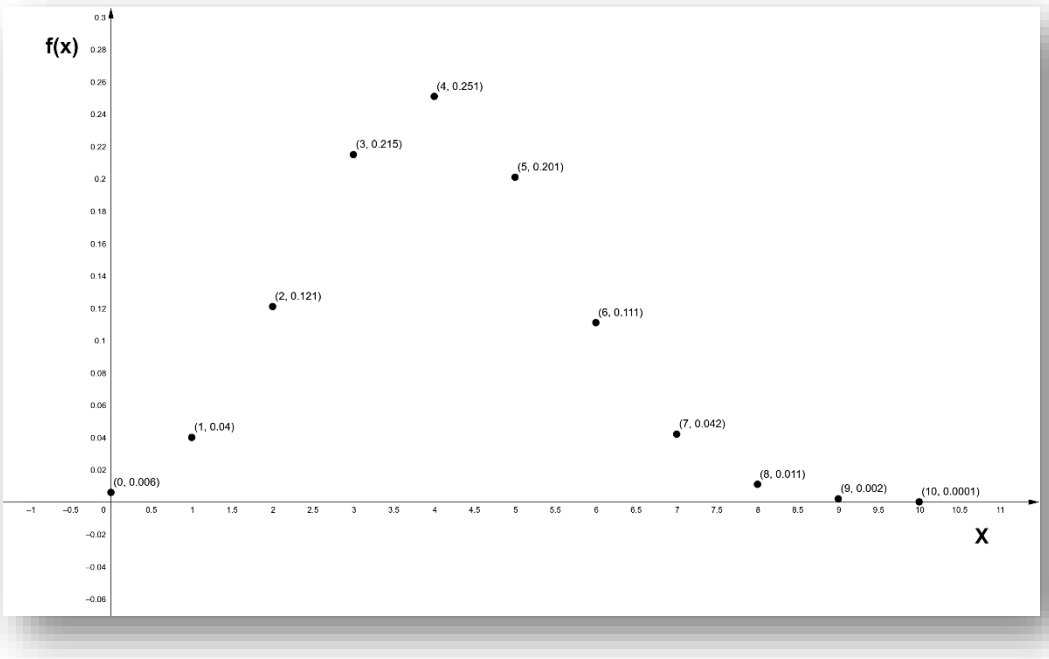

Figura VII.1: Función de cuantía o probabilidad de la VA "número de hembras"

Se observa que el valor que mayor masa de probabilidad acumula, el $25{,}1\ \%$, es el 4 —de un total de 10—.

Ejercicio VII.2:

El número medio de vehículos por minuto que llegan a una gasolinera es igual a dos.

a) ¿Cuál es la probabilidad de que en un minuto lleguen cinco vehículos?

Se trata de un proceso de Poisson en el que el soporte es temporal, la unidad prefijada de dicho soporte es un minuto y el promedio de éxitos, $\lambda = n \cdot p$, es, precisamente, el número medio de vehículos; por tanto:

$$X \sim P(2)$$

$$P(X = 5) = \frac{2^5}{5!} \cdot e^{-2} = 0{,}036$$

La probabilidad es del $3{,}6\ \%$.

b) ¿Cuál es la probabilidad de que en cinco minutos no llegue ninguno?

La unidad prefijada cambia de uno a cinco minutos, por lo que si en el primer caso la media es 2, en el segundo será 10:

$$X \sim P(10)$$

$$P(Y = 0) = \frac{10^0}{0!} \cdot e^{-10} = 0{,}000045$$

La probabilidad es residual, del $0{,}0045\ \%$.

Ejercicio VII.3:

Con los datos del anterior ejercicio, y suponiendo que acaba de llegar un vehículo, calcular la probabilidad de que transcurran más de cinco minutos hasta que aparezca el siguiente.

Desde el punto de vista de un proceso de Poisson, estaríamos en el mismo supuesto que el del segundo apartado del anterior ejercicio, puesto que la probabilidad de que transcurran más de cinco minutos hasta que aparezca el siguiente vehículo es equivalente a que en cinco minutos no llegue ninguno.

Por otro lado, si se define la VA Z continua "tiempo transcurrido entre la llegada de dos coches", y sabiendo del anterior ejercicio que $\lambda = 2$, ésta seguiría una distribución de tipo exponencial tal que:

$$Z \sim E(2)$$

De su definición de función de distribución se tiene:

$$P(Z \leq 5) = 1 - e^{-2 \cdot 5}$$

Pero como se pide que el tiempo transcurrido sea superior a 5:

$$P(Z > 5) = 1 - P(Z \leq 5) = 1 - (1 - e^{-2 \cdot 5}) = e^{-10} = 0{,}000045$$

Y, por tanto, se comprueba:

$$P(Z > 5) = P(Y = 0)$$

Ejercicio VII.4:

La resistencia de una muestra de un determinado material viene dada por una VA X con la siguiente función de densidad:

$$f(x) = \begin{cases} x, & 0 \le x < 1 \\ \dfrac{2 \cdot x + 1}{8}, & 1 \le x \le 2 \\ 0, & en\ el\ resto \end{cases}$$

a) Calcular su función de distribución

Si $x < 0$:

$$F(x) = 0$$

Si $0 \le x < 1$:

$$F(x) = \int_{-\infty}^{x} f(x) \cdot dx = 0 + \int_{0}^{x} x \cdot dx = \frac{x^2}{2}$$

Si $1 \le x \le 2$:

$$F(x) = \int_{-\infty}^{x} f(x) \cdot dx = 0 + \int_{0}^{1} x \cdot dx + \int_{1}^{x} \frac{2 \cdot x + 1}{8} \cdot dx =$$

$$= \left(\frac{x^2}{2}\right)\Big|_0^1 + \left(\frac{x^2 + x}{8}\right)\Big|_1^x = \frac{1}{2} + \frac{x^2 + x - 2}{8}$$

Si $x > 2$:

$$F(x) = 1$$

Por tanto:

$$F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{x^2}{2}, & 0 \le x < 1 \\ \dfrac{1}{2} + \dfrac{x^2 + x - 2}{8}, & 1 \le x \le 2 \\ 1, & x > 2 \end{cases}$$

b) Calcular $P(0{,}5 \le X \le 1{,}5)$

Dado que la función de distribución acumula la probabilidad entre $-\infty$ y $+\infty$, la probabilidad de que la VA tome valores entre $0{,}5$ y $1{,}5$ equivale a determinar el valor de $F(x)$ en $x = 1{,}5$ y restarle el que adquiere en $x = 0{,}5$:

$$P(0{,}5 \le X \le 1{,}5) = F(1{,}5) - F(0{,}5) = \frac{1}{2} + \frac{1{,}5^2 + 1{,}5 - 2}{8} - \frac{0{,}5^2}{2} = 0{,}59$$

También se podría calcular a partir de la función de densidad como sigue:

$$P(0{,}5 \le X \le 1{,}5) = \int_{0{,}5}^{1} x \cdot dx + \int_{1}^{1{,}5} \frac{2 \cdot x + 1}{8} \cdot dx = \left(\frac{x^2}{2}\right)\Big|_{0{,}5}^{1} + \left(\frac{x^2 + x}{8}\right)\Big|_{1}^{1{,}5} =$$

$$= \left(\frac{1}{2} - \frac{0{,}5^2}{2}\right) + \left(\frac{1{,}5^2 + 1{,}5}{8} - \frac{1 + 1}{8}\right) = 0{,}59$$

c) Una muestra de material se encuentra en estado ideal de resistencia entre $0{,}5$ *y* $1{,}5$*; si se consideran* 10 *muestras, ¿cuál es la probabilidad de que al menos el* $70\ \%$ *de ellas tenga resistencia ideal?*

Del anterior apartado se sabe que:

$$P(0{,}5 \leq X \leq 1{,}5) = 0{,}59$$

Por tanto, la probabilidad de que la muestra de material se encuentre en un estado ideal de resistencia es:

$$P(I) = 0{,}59$$

Y la probabilidad de que no se encuentre en un estado ideal es:

$$P(\bar{I}) = 1 - P(I) = 1 - 0{,}59 = 0{,}41$$

Así, el experimento consistente en determinar si una muestra es ideal o no es un experimento de Bernoulli con probabilidad de éxito $p = 0{,}59$; en consecuencia, y dado que se consideran $n = 10$ muestras, se puede definir una VA Y "número de muestras con resistencia ideal" que sigue una distribución binomial:

$$Y \sim B(10, 0{,}59)$$

Como se pide la probabilidad de que al menos siete de las muestras tengan resistencia ideal, y teniendo en cuenta que la binomial es una distribución discreta:

$$P(Y \geq 7) = \sum_{i=7}^{10} P(Y = i) = P(Y = 7) + P(Y = 8) + P(Y = 9) + P(Y = 10) =$$

$$= \binom{10}{7} \cdot 0{,}59^7 \cdot 0{,}41^3 + \cdots + \binom{10}{10} \cdot 0{,}59^{10} \cdot 0{,}41^0 = 0{,}36$$

Es decir, la probabilidad es del $36\ \%$.

VIII. INFERENCIA ESTADÍSTICA

Ejercicio VIII.1:

En una urna se tienen 100 bolas: 60 rojas, 25 blancas y 15 amarillas. ¿Cuál es la probabilidad de que, al extraer dos de ellas, la primera sea blanca y la segunda roja?

a) Si se devuelve la primera bola a la urna

Se trata de un muestreo aleatorio simple con reposición, lo que garantiza la independencia de las VA que compondrían una muestra obtenida de este modo.

Si se definen los sucesos B_1 como "sacar la primera bola blanca" y R_2 como "sacar la segunda bola roja", dado que ambos son independientes:

$$P(B_1 \cap R_2) = P(B_1) \cdot P(R_2) = \frac{25}{100} \cdot \frac{60}{100} = 0{,}150$$

b) Si no se devuelve

El muestreo pasa de ser con a ser sin reposición, por lo que deja de cumplirse la premisa de la independencia: la extracción de la segunda bola está condicionada a la de la primera. Por tanto:

$$P(B_1 \cap R_2) = P(B_1) \cdot P(R_2|B_1) = \frac{25}{100} \cdot \frac{60}{99} = 0{,}152$$

Ejercicio VIII.2:

Se quiere realizar un estudio sobre el tiempo semanal dedicado a la lectura en una población de 1.000 habitantes, caracterizada por la siguiente tabla:

i	*Edades*	f_i	σ_i
1	< 18	0,25	0,10
2	$19 - 35$	0,40	0,30
3	$36 - 55$	0,20	0,50
4	< 55	0,15	0,10

i: Grupos de edad
f_i: Frecuencias relativas
σ_i: Desviaciones típicas

Dado que la población está dividida —o se ha dividido— en subpoblaciones homogéneas, se va a proceder a realizar un muestreo estratificado; por otro lado, ya que se conoce tanto el tamaño de los estratos (f_i) como su variabilidad interna (σ_i), la asignación de unidades se puede hacer mediante afijación óptima, de manera que los estratos más heterogéneos contribuyan más a la composición final de la muestra:

$$n_i = n \cdot \frac{\sigma_i \cdot N_i}{\sum_{j=1}^{k} \sigma_j \cdot N_j}$$

n_i: Unidades muestrales para el estrato i-ésimo
n: Tamaño muestral total
σ_i: Desviación típica del estrato i-ésimo
N_i: Tamaño del estrato i-ésimo
k: Número de estratos

σ_j: Desviación típica del estrato j-ésimo
N_j: Tamaño del estrato j-ésimo

Así, para una muestra de 100 individuos, se tiene:

$$n_1 = n \cdot \frac{\sigma_1 \cdot N_1}{\sum_{j=1}^{4} \sigma_j \cdot N_j} = n \cdot \frac{\cancel{N} \cdot \sigma_1 \cdot f_1}{\cancel{N} \cdot \sum_{j=1}^{k} \sigma_j \cdot f_j} = 100 \cdot \frac{0{,}1 \cdot 0{,}25}{(0{,}1 \cdot 0{,}25 + \cdots + 0{,}1 \cdot 0{,}15)} \cong 10$$

N: Tamaño muestral total

Y de manera análoga para el resto de estratos:

i	n_i
1	10
2	46
3	38
4	6

IX. ESTIMACIÓN PUNTUAL

Ejercicio IX.1:

Al lanzar un dado se obtiene el espacio equiprobable $S = \{1, 2, 3, 4, 5, 6\}$. Sea X la VA "doble del número obtenido".

a) Determinar la distribución de probabilidad

Dado que los valores de la distribución son el doble de los obtenidos y que todos ellos son equiprobables:

x_i	$p(x_i)$
2	1/6
4	1/6
6	1/6
8	1/6
10	1/6
12	1/6

De tal manera que:

$$\sum_{i=1}^{n} p(x_i) = 1$$

b) Calcular la media aritmética

$$\mu_X = E(X) = \sum_{i=1}^{n=6} x_i \cdot p(x_i) = 2 \cdot \left(\frac{1}{6}\right) + \cdots + 12 \cdot \left(\frac{1}{6}\right) = \frac{42}{6} = 7$$

c) Calcular la varianza

$$\sigma_X^2 = Var(X) = E(X^2) - \big(E(X)\big)^2 =$$

$$= \sum_{i=1}^{n=6} {x_i}^2 \cdot p(x_i) - \big(E(X)\big)^2 = \left(2^2 \cdot \left(\frac{1}{6}\right) + \cdots + 12^2 \cdot \left(\frac{1}{6}\right)\right) - 7^2 = 60{,}7 - 49 = 11{,}7$$

Ejercicio IX.2:

En una población $N(\mu, \sigma)$ se calcula μ a través de una MAS de tamaño 2 mediante dos estimadores:

$$\mu_1 = \frac{x_1 + \cdots + x_n}{n - 1}$$
$$\mu_2 = \frac{x_1 + \cdots + x_n}{n + 1}$$

a) *Determinar si son insesgados*

Dado que la muestra sólo tiene dos elementos, los estimadores quedan como sigue:

$$\mu_1 = \frac{x_1 + \cdots + x_n}{n - 1} = \frac{x_1 + x_2}{2 - 1} = x_1 + x_2$$

$$\mu_2 = \frac{x_1 + \cdots + x_n}{n+1} = \frac{x_1 + x_2}{2+1} = \frac{x_1 + x_2}{3}$$

Para que un estimador sea insesgado, su valor esperado ha de ser equivalente al parámetro que pretende estimar; para el primer caso:

$$E(\mu_1) = \mu$$

Sin embargo:

$$E(\mu_1) = E(x_1 + x_2) = E(x_1) + E(x_2) = \mu + \mu = 2 \cdot \mu$$

Los valores esperados de x_1 y x_2 coinciden con la esperanza de la VA, que es equivalente a la media poblacional, μ; es decir, el parámetro que se quiere estimar. Se observa que no se cumple la igualdad propuesta.

Por añadidura:

$$Sesgo(\mu_1) = E(\mu_1) - \mu = 2 \cdot \mu - \mu = \mu$$

El sesgo es distinto de cero, lo que confirma que el estimador es sesgado.

Para el segundo estimador:

$$E(\mu_1) = E\left(\frac{x_1 + x_2}{3}\right) = \frac{1}{3} \cdot \big(E(x_1) + E(x_2)\big) = \frac{1}{3} \cdot (\mu + \mu) = \frac{2}{3} \cdot \mu$$

Tampoco se cumple que $E(\mu_2) = \mu$.

Por otro lado:

$$Sesgo(\mu_2) = E(\mu_2) - \mu = \frac{2}{3} \cdot \mu - \mu = \frac{1}{3} \cdot \mu$$

Con lo cual, el segundo estimador tampoco es insesgado, pero su sesgo es menor que el del primero.

b) *Determinar si son eficientes*

Se determina la varianza del primer estimador:

$$V(\mu_1) = V(x_1 + x_2) = V(x_1) + V(x_2) = \sigma^2 + \sigma^2 = 2 \cdot \sigma^2$$

De manera análoga a lo que ocurría con la esperanza, las varianzas de x_1 y x_2 coinciden con la varianza de la VA.

Y la del segundo:

$$V(\mu_2) = V\left(\frac{x_1 + x_2}{3}\right) = \frac{1}{3^2} \cdot \big(V(x_1) + V(x_2)\big) = \frac{1}{9} \cdot (\sigma^2 + \sigma^2) = \frac{1}{9} \cdot \sigma^2$$

El segundo estimador es más eficiente que el primero dado que su varianza es menor.

c) *Calcular el ECM*

Para calcular el error cuadrático medio se utilizan las propiedades determinadas en los dos apartados previos:

$$ECM(\mu_1) = V(\mu_1) + \big(Sesgo(\mu_1)\big)^2 = 2 \cdot \sigma^2 + \mu^2$$

$$ECM(\mu_2) = V(\mu_2) + \big(Sesgo(\mu_2)\big)^2 = \frac{1}{9} \cdot \sigma^2 + \frac{1}{9} \cdot \mu^2$$

A mayor ECM, menor fiabilidad del estimador; por tanto, el segundo es mejor estimador —como ya se había comprobado al determinar tanto el sesgo como la eficiencia—.

X. INTERVALOS DE CONFIANZA

Ejercicio X.1:

Una empresa quiere determinar su gasto promedio en transporte; para ello, toma 100 desplazamientos al azar realizados por sus empleados, a partir de los cuales estima que, de media, se ha realizado un desembolso de 625 €. Gracias a estudios previos realizados, la empresa sabe que el gasto en transporte se distribuye de manera normal y que su variabilidad, estimada como desviación típica, es de 300 €. Para un nivel de confianza del 95 %, ¿a qué conclusión se puede llegar?

En primer lugar, se puede definir una VA X "gasto en transporte de la empresa" con la siguiente distribución:

$$X \sim N(\mu, 300)$$

Se quiere calcular el gasto promedio, es decir, μ, para un nivel de confianza, γ, del 95 %, lo que implica determinar unas cotas tales que la probabilidad de que la media poblacional se halle entre ellas sea al menos de $0{,}95$. Dichas cotas definirán un intervalo de confianza de nivel $1-\alpha$, donde α es el nivel de significación.

μ es desconocida, pero se dispone de un estimador puntual, $\bar{X} = 625$, calculado a partir de una MAS con $n = 100$, que sirve como aproximación inicial al parámetro deseado.

Por definición, la media muestral es a su vez una VA con la siguiente distribución:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(\mu, \frac{300}{\sqrt{100}}\right)$$

Así:

$$\bar{X} \sim N(\mu, 30)$$

Dado que la distribución del estimador depende también del parámetro, es necesario calcular una función pivote de $\bar{X}$ cuya distribución sea conocida pero no se vea afectada por μ.

Para ello, la opción más sencilla es la normalización o tipificación de $\bar{X}$:

$$h(\bar{X}) = \frac{\bar{X} - \bar{x}}{S} = \frac{\bar{X} - \mu}{30} \sim N(0, 1)$$

Si se denotan las cotas a determinar, que están en función del nivel de significación, como $\lambda_1(\alpha) \equiv \lambda_1$ y $\lambda_2(\alpha) \equiv \lambda_2$, se tiene:

$$P\left(\lambda_1 \leq \frac{\bar{X} - \mu}{30} \leq \lambda_2\right) = \gamma = 1 - \alpha = 0{,}95$$

Dada la relación existente entre el nivel de confianza y el de significación, se puede concluir que $\alpha = 1 - \gamma = 1 - 0{,}95 = 0{,}5$. Al ser la normal tipo una distribución simétrica, se opta por repartir equitativamente el nivel de significación entre ambas colas —esto es, $0{,}25$— para obtener así el intervalo de confianza de longitud mínima.

Como se observa en la figura X.1, que muestra un detalle de la distribución normal tipo correspondiente al semieje positivo de abscisas, se ha de encontrar el valor, que equivaldría a λ_2, que deja por debajo de sí una probabilidad de $0{,}975$.

Si consultamos la tabla de la figura X.2, este valor es $1{,}96$; dada la simetría con respecto al eje de ordenadas de la normal tipo, $\lambda_1 = -1{,}96$, que sería el valor que deja por encima de sí una probabilidad de $0{,}975$ —es decir, entre colas habría un área equivalente a $0{,}95$, el nivel de confianza—.

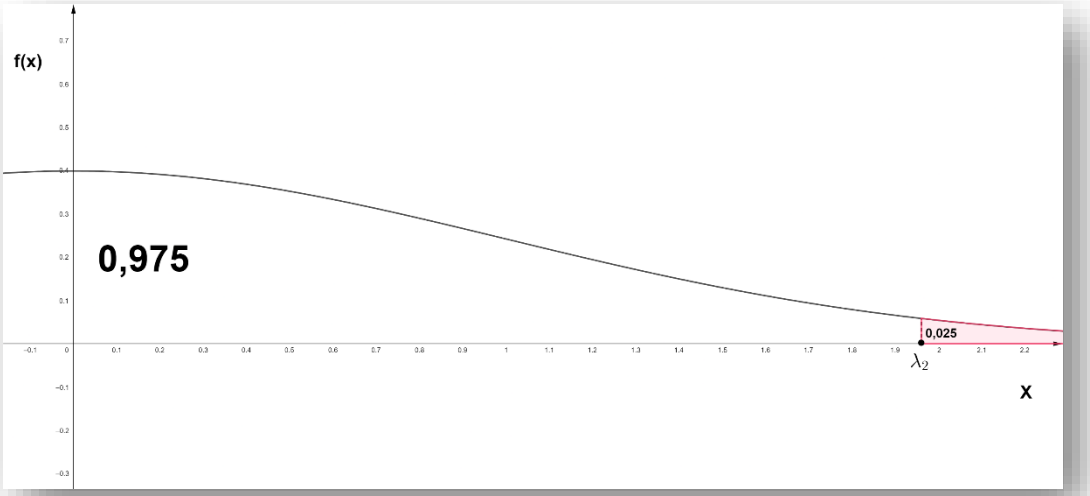

Figura X.1: Detalle de la distribución normal tipo

Despejando:

$$P\left(\lambda_1 \leq \frac{\bar{X}-\mu}{30} \leq \lambda_2\right) = P(\bar{X} + 30 \cdot \lambda_1 \leq \mu \leq \bar{X} + 30 \cdot \lambda_2) =$$

$$= P(\bar{X} - 30 \cdot 1{,}96 \leq \mu \leq \bar{X} + 30 \cdot 1{,}96) = P(\bar{X} - 58{,}8 \leq \mu \leq \bar{X} + 58{,}8) = 0{,}95$$

	0	0'01	0'02	0'03	0'04	0'05	0'06	0'07	0'08	0'09
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	0'5753
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	0'7190	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	0'7823	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	0'8413	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	0'9772	0'9778	0'9783	0'9788	0'9793	0'9798	0'9803	0'9808	0'9812	0'9817

Figura X.2: Tabla de la distribución normal tipo para un nivel de confianza del 95 %

Una vez se sustituye el estimador obtenido a partir de la realización muestral ya no se puede hablar de probabilidad, puesto que el parámetro a determinar estará ($p = 1$) o no ($p = 0$) incluido en el intervalo resultante, por lo que la notación pasa a ser:

$$I_{95\%}(\mu) = [625 - 58{,}8; 625 + 58{,}8] = [566{,}2; 683{,}8]$$

Así, de cada 100 intervalos construidos siguiendo esta metodología a partir de muestras de este mismo tamaño, 95 de ellos contendrán a la media poblacional.

Ejercicio X.2:

A partir de una muestra de 20 linternas cuyos periodos de duración en horas han sido 503, 480, 345, 427, 386, 432, 429, 378, 440, 434, 429, 436, 451, 466, 394, 422, 412, 507, 433 y 480, se quiere obtener un intervalo de confianza al 95 % para la vida media de una población de linternas que se distribuye normalmente.

Sea X una VA "duración en horas de una linterna" con la siguiente distribución:

$$X \sim N(\mu, \sigma)$$

Se pide determinar la vida media poblacional, es decir, μ, pero además se desconoce σ. Por tanto, no se puede utilizar la función pivote del anterior ejercicio, puesto que habría dos incógnitas:

$$\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1)$$

En consecuencia, mediante la combinación de dos VVAA, la media muestral $\bar{X}$ y la cuasivarianza muestral S_c^2, se puede construir la siguiente función pivote, que se distribuye según una t de Student con $n - 1$ grados de libertad:

$$\frac{\sqrt{n} \cdot (\bar{X} - \mu)}{S_C} \sim t_{n-1}$$

	0'9995	0'995	0'9875	0'975	0'95	0'875	0'85	0'8	0'75	0'7	0'65	0'6	0'55
1	636'58	63'656	25'452	12'706	6'3137	2'4142	1'9626	1'3764	1'0000	0'7265	0'5095	0'3249	0'1584
2	31'600	9'9250	6'2054	4'3027	2'9200	1'6036	1'3862	1'0607	0'8165	0'6172	0'4447	0'2887	0'1421
3	12'924	5'8408	4'1765	3'1824	2'3534	1'4226	1'2498	0'9785	0'7649	0'5844	0'4242	0'2767	0'1366
4	8'6101	4'6041	3'4954	2'7765	2'1318	1'3444	1'1896	0'9410	0'7407	0'5686	0'4142	0'2707	0'1338
5	6'8685	4'0321	3'1634	2'5706	2'0150	1'3009	1'1558	0'9195	0'7267	0'5594	0'4082	0'2672	0'1322
6	5'9587	3'7074	2'9687	2'4469	1'9432	1'2733	1'1342	0'9057	0'7176	0'5534	0'4043	0'2648	0'1311
7	5'4081	3'4995	2'8412	2'3646	1'8946	1'2543	1'1192	0'8960	0'7111	0'5491	0'4015	0'2632	0'1303
8	5'0414	3'3554	2'7515	2'3060	1'8595	1'2403	1'1081	0'8889	0'7064	0'5459	0'3995	0'2619	0'1297
9	4'7809	3'2498	2'6850	2'2622	1'8331	1'2297	1'0997	0'8834	0'7027	0'5435	0'3979	0'2610	0'1293
10	4'5868	3'1693	2'6338	2'2281	1'8125	1'2213	1'0931	0'8791	0'6998	0'5415	0'3966	0'2602	0'1289
11	4'4369	3'1058	2'5931	2'2010	1'7959	1'2145	1'0877	0'8755	0'6974	0'5399	0'3956	0'2596	0'1286
12	4'3178	3'0545	2'5600	2'1788	1'7823	1'2089	1'0832	0'8726	0'6955	0'5386	0'3947	0'2590	0'1283
13	4'2209	3'0123	2'5326	2'1604	1'7709	1'2041	1'0795	0'8702	0'6938	0'5375	0'3940	0'2586	0'1281
14	4'1403	2'9768	2'5096	2'1448	1'7613	1'2001	1'0763	0'8681	0'6924	0'5366	0'3933	0'2582	0'1280
15	4'0728	2'9467	2'4899	2'1315	1'7531	1'1967	1'0735	0'8662	0'6912	0'5357	0'3928	0'2579	0'1278
16	4'0149	2'9208	2'4729	2'1199	1'7459	1'1937	1'0711	0'8647	0'6901	0'5350	0'3923	0'2576	0'1277
17	3'9651	2'8982	2'4581	2'1098	1'7396	1'1910	1'0690	0'8633	0'6892	0'5344	0'3919	0'2573	0'1276
18	3'9217	2'8784	2'4450	2'1009	1'7341	1'1887	1'0672	0'8620	0'6884	0'5338	0'3915	0'2571	0'1274
19	3'8833	2'8609	2'4334	2'0930	1'7291	1'1866	1'0655	0'8610	0'6876	0'5333	0'3912	0'2569	0'1274
20	3'8496	2'8453	2'4231	2'0860	1'7247	1'1848	1'0640	0'8600	0'6870	0'5329	0'3909	0'2567	0'1273

Figura X.3: Tabla de la t de Student para 19 grados de libertad y un nivel de confianza del 95 %

A partir de la MAS con $n = 20$ se puede calcular $\bar{X}$:

$$\bar{X} = \frac{1}{20} \cdot \sum_{i=1}^{20} X_i = \frac{X_1 + \cdots + X_{20}}{20} = \frac{503 + \cdots + 480}{20} = 434{,}2$$

Y S_c^2:

$$S_c^2 = \frac{\sum_{i=1}^{r}(X_i - \bar{X})^2}{n-1} = \frac{(X_1 - \bar{X})^2 + \cdots + (X_{20} - \bar{X})^2}{19} =$$
$$= \frac{(503 - 434{,}2)^2 + \cdots + (480 - 434{,}2)^2}{19} = 1.650{,}9$$
$$S_C = \sqrt{S_c^2} = 40{,}6$$

Dado que $\gamma = 0{,}95$, $\alpha = 0{,}05$; al ser también la t de Student una función simétrica con respecto al eje de ordenadas, el nivel de significación se reparte equitativamente entre las colas, con lo cual $\alpha/2 = 0{,}025$.

Así, el intervalo de confianza se construye de la siguiente manera:

$$I_{95\%}(\mu) = \left[\bar{X} - t_{n-1;\,1-\frac{\alpha}{2}} \cdot \frac{S_C}{\sqrt{n}};\ \bar{X} + t_{n-1;\,1-\frac{\alpha}{2}} \cdot \frac{S_C}{\sqrt{n}}\right] =$$
$$= \left[434{,}2 - t_{19;\,0{,}975} \cdot \frac{40{,}6}{\sqrt{20}};\ 434{,}2 + t_{19;\,0{,}975} \cdot \frac{40{,}6}{\sqrt{20}}\right]$$

Consultando en la tabla de la t de Student, como se puede ver en la figura X.3, el valor que para 19 grados de libertad deja por debajo de sí el 97,5 % del área es $\lambda_2 = t_{19,0,975} = 2{,}0930$.

En conclusión:

$$I_{95\%}(\mu) = \left[434{,}2 - 2{,}0930 \cdot \frac{40{,}6}{\sqrt{20}};\ 434{,}2 + 2{,}0930 \cdot \frac{40{,}6}{\sqrt{20}}\right] = [415{,}2;\ 453{,}2]$$

Ejercicio X.3:

Se sabe que el peso por comprimido de un cierto preparado farmacéutico se distribuye según una gausiana. Con el objeto de estudiar la varianza de la distribución, se extrae una MAS de 6 elementos. Sabiendo que la varianza muestral es igual a 40, se pretende estimar la varianza poblacional mediante un intervalo de confianza al 90 %.

Sea X una VA "peso por comprimido de un fármaco" con la siguiente distribución:

$$X \sim N(\mu, \sigma)$$

Se pretende determinar la varianza poblacional, σ^2, pero hay que tener en cuenta que tampoco se conoce la media poblacional, μ; en esta situación, mediante el teorema de Fisher se puede construir la siguiente función pivote, que se distribuye según una chi-cuadrado con $n-1$ grados de libertad:

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Dado que se dispone de una MAS con $n = 6$, la única incógnita de la función pivote sería, precisamente, σ^2; a partir de ella se puede llegar al siguiente intervalo de confianza, que dependería de la varianza muestral, S^2, cuyo valor es conocido:

$$I_{1-\alpha}(\sigma^2) = \left[\frac{n \cdot S^2}{\chi^2_{n-1;\,1-\frac{\alpha}{2}}};\ \frac{n \cdot S^2}{\chi^2_{n-1;\,\frac{\alpha}{2}}}\right] = \left[\frac{n \cdot S^2}{\lambda_2};\ \frac{n \cdot S^2}{\lambda_1}\right]$$

Dado que se pide que el nivel de confianza sea del $90\,\%$, el de significación será del $10\,\%$, a repartir entre ambas colas de la chi-cuadrado, que no es simétrica, por lo que es necesario estimar ambos valores críticos, $\lambda_1 < \lambda_2$, por separado; así:

- $\lambda_1 = \chi^2_{6-1;\,0{,}1/2} = \chi^2_{5;\,0{,}05}$ es el valor que deja un $5\,\%$ del área de la distribución por debajo de él.
- $\lambda_2 = \chi^2_{6-1;\,1-0{,}1/2} = \chi^2_{5;\,0{,}95}$ es el valor que deja un $95\,\%$ del área de la distribución por debajo de él.

Recurriendo a la tabla de la chi-cuadrado, tenemos que $\lambda_1 = 1{,}1455$:

	0,5	0,45	0,4	0,35	0,3	0,25	0,2	0,15	0,125	0,1	0,05	0,025	0,01	0,005
1	0,4549	0,3573	0,2750	0,2059	0,1485	0,1015	0,0642	0,0358	0,0247	0,0158	0,0039	0,0010	0,0002	0,0000
2	1,3863	1,1957	1,0217	0,8616	0,7133	0,5754	0,4463	0,3250	0,2671	0,2107	0,1026	0,0506	0,0201	0,0100
3	2,3660	2,1095	1,8692	1,6416	1,4237	1,2125	1,0052	0,7978	0,6924	0,5844	0,3518	0,2158	0,1148	0,0717
4	3,3567	3,0469	2,7528	2,4701	2,1947	1,9226	1,6488	1,3665	1,2188	1,0636	0,7107	0,4844	0,2971	0,2070
5	~~4,3515~~	~~3,9959~~	~~3,6555~~	~~3,3251~~	~~2,9999~~	~~2,6746~~	~~2,3425~~	~~1,9938~~	~~1,8082~~	~~1,6103~~	1,1455	0,8312	0,5543	0,4118
6	5,3481	4,9519	4,5702	4,1973	3,8276	3,4546	3,0701	2,6613	2,4411	2,2041	1,6354	1,2373	0,8721	0,6757

Figura X.4: Tabla I de la chi-cuadrado para 5 grados de libertad y un nivel de confianza del 90 %

Y $\lambda_2 = 11{,}070$:

	0'9995	0'995	0'9875	0'975	0'95	0'875	0'85	0'8	0'75	0'7	0'65	0'6	0'55
1	12'115	7'8794	6'2385	5'0239	3'8415	2'3535	2'0722	1'6424	1'3233	1'0742	0'8735	0'7083	0'5707
2	15'201	10'597	8'7641	7'3778	5'9915	4'1589	3'7942	3'2189	2'7726	2'4079	2'0996	1'8326	1'5970
3	17'731	12'838	10'861	9'3484	7'8147	5'7394	5'3170	4'6416	4'1083	3'6649	3'2831	2'9462	2'6430
4	19'998	14'860	12'762	11'143	9'4877	7'2140	6'7449	5'9886	5'3853	4'8784	4'4377	4'0446	3'6871
5	~~22'106~~	~~16'750~~	~~14'544~~	~~12'832~~	11'070	8'6248	8'1152	7'2893	6'6257	6'0644	5'5731	5'1319	4'7278
6	24'102	18'548	16'244	14'449	12'592	9'9917	9'4461	8'5581	7'8408	7'2311	6'6948	6'2108	5'7652

Figura X.5: Tabla II de la chi-cuadrado para 5 grados de libertad y un nivel de confianza del 90 %

En consecuencia:

$$I_{90\%}(\sigma^2) = \left[\frac{6 \cdot 40}{11{,}070};\ \frac{6 \cdot 40}{1{,}1455}\right] = [21{,}7;\ 209{,}6]$$

Ejercicio X.4:

En unas elecciones, uno de los candidatos desea estimar, al $95\,\%$ de confianza, la proporción de votantes que están a su favor. Con este fin, toma una muestra aleatoria de 100 votantes y observa que el $55\,\%$ son partidarios suyos; ¿qué conclusiones puede obtener para el conjunto del electorado?

Sea X una VA "votantes favorables" con la siguiente distribución:

$$X \sim B(n, p)$$

Se quiere determinar la proporción poblacional, p, que es desconocida, y para ello se dispone de un estimador puntual, $\hat{p} = 0{,}55$, obtenido a partir de una MAS con $n = 100$.

La proporción muestral es una VA que, dado que $n \geq 30$, por el TCL en su forma LL se puede convertir en la siguiente función pivote, distribuida aproximadamente según una normal tipo:

$$\frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \xrightarrow{d} N(0,1)$$

Y a partir de la cual se puede llegar al siguiente intervalo de confianza:

$$I_{1-\alpha}(p) = \left[\hat{p} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}\right]$$

Dado que $\gamma = 0{,}95$, $\alpha = 0{,}5$; al tratarse de una normal tipo, se reparte el nivel de significación entre ambas colas, por lo que el valor a localizar, $Z_{1-\alpha/2} = Z_{1-0{,}5/2} = Z_{0{,}975}$, será aquel que deje el 97,5 % del área bajo la curva por debajo (o por encima, dado la simetría de la distribución con respecto al eje de ordenadas) de sí:

	0	0'01	0'02	0'03	0'04	0'05	0'06	0'07	0'08	0'09
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	0'5753
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	0'7190	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	0'7823	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	0'8413	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	0'9772	0'9778	0'9783	0'9788	0'9793	0'9798	0'9803	0'9808	0'9812	0'9817

Figura X.6: Tabla de la distribución normal tipo para un nivel de confianza del 95 %

En consecuencia:

$$I_{95\%}(p) = \left[0{,}55 \pm 1{,}96 \cdot \sqrt{\frac{0{,}55 \cdot (1-0{,}55)}{100}}\right] \equiv [0{,}45; 0{,}65]$$

XI. CONTRASTE DE HIPÓTESIS

Ejercicio XI.1:

En una región se cree que la altura media de sus habitantes mayores de 18 años es 175 cm y se acepta un modelo normal para la distribución de alturas con desviación típica poblacional de 9 cm. Para un nivel de significación del 5 % y a partir de una MAS de tamaño 25 cuya media arroja un valor de 172 cm, ¿qué se puede decir sobre la suposición inicial?

Sea X una VA "altura de los habitantes mayores de edad" con la siguiente distribución:

$$X \sim N(175, 9)$$

La hipótesis nula, H_0, es precisamente que la media poblacional, μ, es 175 cm, un valor concreto, por lo que se trata de un contraste bilateral, en el que la hipótesis alternativa, H_1, sería la negación de la nula:

$$\begin{cases} H_0: \mu = 175 \\ H_1: \mu \neq 175 \end{cases}$$

Para decidir entre ambas hipótesis, se dispone de una MAS con $n = 25$. Puesto que el parámetro bajo estudio es la media poblacional, lo lógico es utilizar en el proceso inferencial su estimador ideal, esto es, la media muestral, $\bar{X}$, que es 172 cm; el criterio de decisión consiste en que si la diferencia entre ambas medias, la poblacional y la muestral, es grande, se rechazará H_0 y se aceptará H_1, mientras que de lo contrario no habrá evidencia suficiente para poder decir que la hipótesis nula es falsa.

Como se puede deducir, no es necesario que al finalizar el contraste se verifique con total exactitud la hipótesis nula, sino que es suficiente con que el valor obtenido se encuentre en el interior de una región de aceptación centrada en la media poblacional y limitada por dos valores críticos —esto es, un intervalo de confianza—, los cuales a su vez marcarán el final y el principio de las dos regiones de rechazo, localizadas en las colas de la distribución, como se puede ver en la figura XI.1.

Por definición, la media muestral sigue la siguiente distribución:

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

$\bar{X}$ se puede tipificar, de manera que, bajo el supuesto de que H_0 no sea falsa, d sigue una distribución normal tipo:

$$d = \frac{\bar{X} - \bar{x}}{S} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Este estadístico ofrece una medida de la discrepancia entre la evidencia muestral, representada por $\bar{X}$, y la hipótesis nula, representada por μ, por lo que puede ser usado para validar el contraste.

Dado que la curva de la normal tipo es simétrica con respecto de la media —esto es, el eje de ordenadas, ya que la media poblacional tipificada es cero—, si se reparte el nivel de significación —la probabilidad de cometer un error de tipo I, es decir, rechazar H_0 cuando es cierta—, α, entre ambas colas de la normal, se pueden

obtener sendos valores críticos que separan la región de aceptación (RA) central de las de rechazo (RR) periféricas:

- $d_{c_1} = -Z_{\alpha/2} = -Z_{0,05/2} = -Z_{0,025}$ es el valor que deja un 2,5 % del área de la distribución por debajo de él (o el 97,5 % por encima).
- $d_{c_2} = Z_{\alpha/2} = Z_{0,05/2} = Z_{0,025}$ es el valor que deja un 2,5 % del área de la distribución por encima de él (o el 97,5 % por debajo).

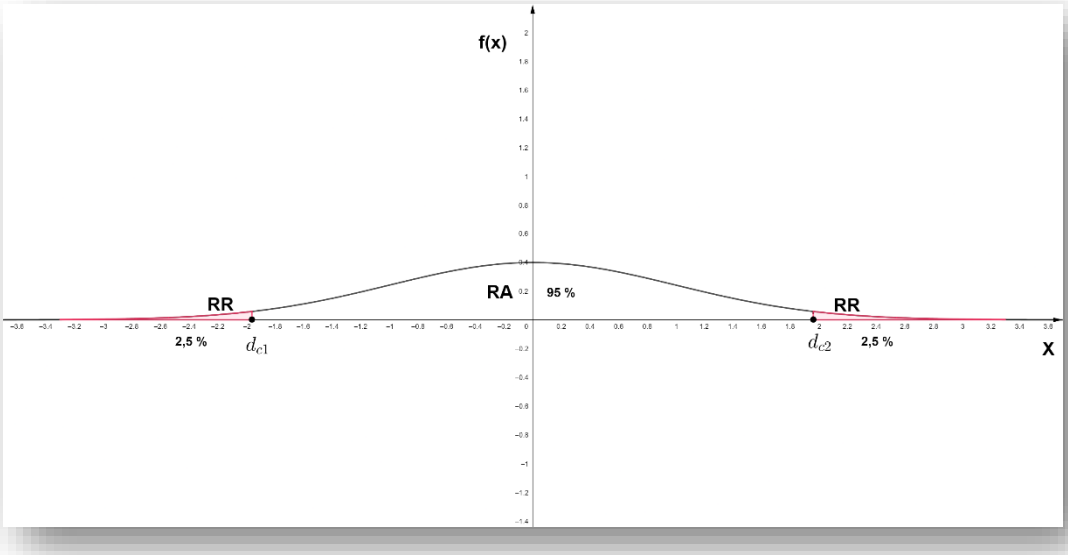

Figura XI.1: Regiones de aceptación y rechazo para una normal tipo al 5 % de significación

A partir de la tabla de la normal tipo, se obtiene el valor para λ_2 (se busca aquel que deja una probabilidad de 0,975 por debajo de él):

	0	0'01	0'02	0'03	0'04	0'05	0'06	0'07	0'08	0'09
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	0'5753
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	0'7190	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	0'7823	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	0'8413	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	0'9772	0'9778	0'9783	0'9788	0'9793	0'9798	0'9803	0'9808	0'9812	0'9817

Figura XI.2: Tabla de la distribución normal tipo para un nivel de significación del 5 %

En consecuencia, y dado que λ_1 es el valor especular con respecto al eje de ordenadas de λ_2 , la región de aceptación es $[-1{,}96; 1{,}96]$.

Por otro lado, el valor observado para el estadístico es:

$$d_0 = \frac{172 - 175}{\frac{9}{\sqrt{25}}} = -1{,}67$$

Que se encuentra dentro de la región de aceptación, por lo que no se puede rechazar H_0.

Ejercicio XI.2:

El volumen (en miles de litros) que diariamente venía envasando una planta embotelladora se distribuye según una normal con media 150 y desviación típica 5. Sin embargo, hace tres meses que se han introducido cambios en el proceso productivo y se quiere comprobar si suponen una ventaja comparativa antes de convertirlos en definitivos. Para ello, el jefe de planta ha tomado una MAS de 25 días y ha concluido que el volumen medio embotellado es 153. Suponiendo que la varianza no ha cambiado, se pide plantear un contraste de hipótesis con una significación del $2\,\%$ para tomar una decisión.

Sea X una VA "volumen diario en miles de litros" con la siguiente distribución:

$$X \sim N(150, 5)$$

La hipótesis nula es que la producción se mantiene en el valor promedio habitual, mientras que la alternativa es que con el nuevo sistema la producción se ve alterada:

$$\begin{cases} H_0: \mu = 150 \\ H_1: \mu \neq 150 \end{cases}$$

A pesar de que parece tratarse de nuevo de un contraste bilateral, la MAS, con $n = 25$, arroja un valor promedio de 153, por lo que la hipótesis alternativa habría de ser no sólo que la producción se ve alterada sino que se incrementa:

$$\begin{cases} H_0: \mu = 150 \\ H_1: \mu > 150 \end{cases}$$

Y, por tanto, se trata de un contraste unilateral por la derecha, por lo que sólo habría una región crítica —obviamente, a la derecha de la región de aceptación—.

Dado el estadístico anteriormente utilizado:

$$d = \frac{\bar{X} - \bar{x}}{S} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

A partir de la MAS, el valor observado sería:

$$d_0 = \frac{153 - 150}{\frac{5}{\sqrt{25}}} = 3$$

Para resolver el contraste, en este caso, se recurre al concepto de p-valor:

$$p = P(d > d_0 | H_0) = P(d > 3)$$

Así, para un contraste unilateral por la derecha, el p-valor sería la probabilidad de que, dada H_0 —es decir, bajo el supuesto de que la hipótesis nula no es falsa—, el estadístico de prueba sea superior al valor observado.

Haciendo uso de la tabla de la normal tipo, habría que determinar qué probabilidad deja por encima de sí un valor de 3; como se aprecia en la figura XI.3,

por debajo de ese valor se encuentra prácticamente toda el área bajo la curva. En consecuencia:

$$p = P(d > 3) = 1 - P(d < 3) = 1 - 0{,}9987 = 0{,}0013$$

Teniendo en cuenta que el nivel de significación es del $2\,\%$:

$$p = 0{,}0013 < 0{,}02$$

Lo que implica rechazar H_0, dado que $p \ll \alpha$.

	0	0'01	0'02	0'03	0'04	0'05	0'06	0'07	0'08	0'09
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	0'5753
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	0'7190	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	0'7823	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	0'8413	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	0'9772	0'9778	0'9783	0'9788	0'9793	0'9798	0'9803	0'9808	0'9812	0'9817
2'1	0'9821	0'9826	0'9830	0'9834	0'9838	0'9842	0'9846	0'9850	0'9854	0'9857
2'2	0'9861	0'9864	0'9868	0'9871	0'9875	0'9878	0'9881	0'9884	0'9887	0'9890
2'3	0'9893	0'9896	0'9898	0'9901	0'9904	0'9906	0'9909	0'9911	0'9913	0'9916
2'4	0'9918	0'9920	0'9922	0'9925	0'9927	0'9929	0'9931	0'9932	0'9934	0'9936
2'5	0'9938	0'9940	0'9941	0'9943	0'9945	0'9946	0'9948	0'9949	0'9951	0'9952
2'6	0'9953	0'9955	0'9956	0'9957	0'9959	0'9960	0'9961	0'9962	0'9963	0'9964
2'7	0'9965	0'9966	0'9967	0'9968	0'9969	0'9970	0'9971	0'9972	0'9973	0'9974
2'8	0'9974	0'9975	0'9976	0'9977	0'9977	0'9978	0'9979	0'9979	0'9980	0'9981
2'9	0'9981	0'9982	0'9982	0'9983	0'9984	0'9984	0'9985	0'9985	0'9986	0'9986
3	0'9987	0'9987	0'9987	0'9988	0'9988	0'9989	0'9989	0'9989	0'9990	0'9990
3'1	0'9990	0'9991	0'9991	0'9991	0'9992	0'9992	0'9992	0'9992	0'9993	0'9993

Figura XI.3: Masa de probabilidad que queda por debajo de 3 para una normal tipo

A modo de comprobación, se puede resolver el ejercicio de modo análogo al anterior:

Al tratarse de un contraste unilateral por la derecha, la probabilidad correspondiente al nivel de significación se concentrará por completo en el extremo derecho; así, y según la tabla de la normal tipo, el valor crítico d_c que deja por encima de sí un $2\,\%$ del área de la curva (o una masa de probabilidad de $0{,}98$ por debajo) es $2{,}05$. Dado que el valor observado es 3, resulta obvio que se encuentra en la zona de rechazo.

Ejercicio XI.3:

Los rodamientos esféricos que fabrica una máquina han de tener un diámetro uniforme para ser aptos para su uso. El responsable del taller asegura que la varianza es de $0,0250$ cm^2, pero al tomar una muestra de 50 elementos se obtiene un valor de $0,0272$ cm^2; ¿qué se puede concluir?

Sea X una VA "diámetro de un rodamiento esférico" con la siguiente distribución:

$$X \sim N(\mu, 0{,}025)$$

Como hipótesis nula se asume que la varianza poblacional, σ^2, es la que asegura el responsable; sin embargo, la MAS, con $n = 50$, arroja un valor de $S^2 = 0{,}0272\ cm^2$, por lo que como hipótesis alternativa se establece que la varianza ha de ser superior a la esperada:

$$\begin{cases} H_0: \sigma^2 = 0{,}025 \\ H_1: \sigma^2 > 0{,}025 \end{cases}$$

Se trataría de un contraste unilateral a la derecha y, dado que no se conoce la media poblacional, se recurriría al siguiente estadístico:

$$d = \frac{n \cdot S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Si se asume un nivel de significación del $5\ \%$, acumulando toda la probabilidad de cometer un error de tipo I en la cola derecha, habría que localizar aquel valor que, para $n - 1 = 50 - 1 = 49$ grados de libertad, deja un $95\ \%$ de la curva de la chi-cuadrado por debajo de él:

	0'9995	0'995	0'9875	0'975	0'95	0'875	0'85	0'8	0'75	0'7	0'65	0'6	0'55
1	12'115	7'8794	6'2385	5'0239	3'8415	2'3535	2'0722	1'6424	1'3233	1'0742	0'8735	0'7083	0'5707
2	15'201	10'597	8'7641	7'3778	5'9915	4'1589	3'7942	3'2189	2'7726	2'4079	2'0996	1'8326	1'5970
3	17'731	12'838	10'861	9'3484	7'8147	5'7394	5'3170	4'6416	4'1083	3'6649	3'2831	2'9462	2'6430
4	19'998	14'860	12'762	11'143	9'4877	7'2140	6'7449	5'9886	5'3853	4'8784	4'4377	4'0446	3'6871
5	22'106	16'750	14'544	12'832	11'070	8'6248	8'1152	7'2893	6'6257	6'0644	5'5731	5'1319	4'7278
6	24'102	18'548	16'244	14'449	12'592	9'9917	9'4461	8'5581	7'8408	7'2311	6'6948	6'2108	5'7652
7	26'018	20'278	17'885	16'013	14'067	11'326	10'748	9'8032	9'0371	8'3834	7'8061	7'2832	6'8000
8	27'867	21'955	19'478	17'535	15'507	12'636	12'027	11'030	10'219	9'5245	8'9094	8'3505	7'8325
9	29'667	23'589	21'034	19'023	16'919	13'926	13'288	12'242	11'389	10'656	10'006	9'4136	8'8632
10	31'419	25'188	22'558	20'483	18'307	15'198	14'534	13'442	12'549	11'781	11'097	10'473	9'8922
11	33'138	26'757	24'056	21'920	19'675	16'457	15'767	14'631	13'701	12'899	12'184	11'530	10'920
12	34'821	28'300	25'530	23'337	21'026	17'703	16'989	15'812	14'845	14'011	13'266	12'584	11'946
13	36'477	29'819	26'985	24'736	22'362	18'939	18'202	16'985	15'984	15'119	14'345	13'636	12'972
14	38'109	31'319	28'422	26'119	23'685	20'166	19'406	18'151	17'117	16'222	15'421	14'685	13'996
15	39'717	32'801	29'843	27'488	24'996	21'384	20'603	19'311	18'245	17'322	16'494	15'733	15'020
16	41'308	34'267	31'250	28'845	26'296	22'595	21'793	20'465	19'369	18'418	17'565	16'780	16'042
17	42'881	35'718	32'644	30'191	27'587	23'799	22'977	21'615	20'489	19'511	18'633	17'824	17'065
18	44'434	37'156	34'027	31'526	28'869	24'997	24'155	22'760	21'605	20'601	19'699	18'868	18'086
19	45'974	38'582	35'399	32'852	30'144	26'189	25'329	23'900	22'718	21'689	20'764	19'910	19'107
20	47'498	39'997	36'760	34'170	31'410	27'376	26'498	25'038	23'828	22'775	21'826	20'951	20'127
21	49'010	41'401	38'113	35'479	32'671	28'559	27'662	26'171	24'935	23'858	22'888	21'992	21'147
22	50'510	42'796	39'458	36'781	33'924	29'737	28'822	27'301	26'039	24'939	23'947	23'031	22'166
23	51'999	44'181	40'794	38'076	35'172	30'911	29'979	28'429	27'141	26'018	25'006	24'069	23'185
24	53'478	45'558	42'124	39'364	36'415	32'081	31'132	29'553	28'241	27'096	26'063	25'106	24'204
25	54'948	46'928	43'446	40'646	37'652	33'247	32'282	30'675	29'339	28'172	27'118	26'143	25'222
26	56'407	48'290	44'762	41'923	38'885	34'410	33'429	31'795	30'435	29'246	28'173	27'179	26'240
27	57'856	49'645	46'071	43'195	40'113	35'570	34'574	32'912	31'528	30'319	29'227	28'214	27'257
28	59'299	50'994	47'375	44'461	41'337	36'727	35'715	34'027	32'620	31'391	30'279	29'249	28'274
29	60'734	52'335	48'674	45'722	42'557	37'881	36'854	35'139	33'711	32'461	31'331	30'283	29'291
30	62'160	53'672	49'967	46'979	43'773	39'033	37'990	36'250	34'800	33'530	32'382	31'316	30'307
35	69'197	60'275	56'365	53'203	49'802	44'753	43'640	41'778	40'223	38'859	37'623	36'475	35'386
40	76'096	66'766	62'665	59'342	55'758	50'424	49'244	47'269	45'616	44'165	42'848	41'622	40'459
50	~~89'560~~	~~79'490~~	~~75'039~~	~~71'420~~	67'505	61'647	60'346	58'164	56'334	54'723	53'258	51'892	50'592

Figura XI.4: Tabla II de la chi-cuadrado para 50 grados de libertad y un nivel de confianza del 95 %

La tabla no es lo suficientemente específica, pero si interpolamos entre 40 y 50 grados de libertad, se obtiene el siguiente valor crítico:

$$d_c = \chi^2_{49;\,0,95} = 66{,}33$$

Bajo la hipótesis nula, el valor observado resulta ser:

$$d = \frac{n \cdot S^2}{\sigma^2} = \frac{50 \cdot 0{,}0272}{0{,}0250} = 54{,}40$$

Que cae claramente dentro de la región de aceptación, por lo que no se rechazaría la hipótesis nula.

XII. CONTRASTES NO PARAMÉTRICOS

Ejercicio XII.1:

Se desea contrastar la hipótesis de normalidad de una población de la cual se ha extraído la siguiente muestra: $12,3$, $11,0$, $10,7$, $12,4$, $11,7$, $13,1$, $9,9$, $12,6$, $11,8$, $10,2$ y $10,5$.

Se va a proceder a realizar un contraste de Shapiro-Wilk; para ello, en primer lugar, se han de ordenar los elementos de la muestra de mayor a menor:

$$13{,}1\ 12{,}6\ 12{,}4\ 12{,}3\ 11{,}8\ 11{,}7\ 11{,}0\ 10{,}7\ 10{,}5\ 10{,}2\ 9{,}9$$

Para $n = 11$, los valores tabulados son los que se muestran en la siguiente figura:

	n									
i	11	12	13	14	15	16	17	18	19	20
1	0'5601	0'5475	0'5359	0'5251	0'5150	0'5056	0'4968	0'4886	0'4808	0'4734
2	0'3315	0'3325	0'3325	0'3318	0'3306	0'3290	0'3273	0'3253	0'3232	0'3211
3	0'2260	0'2347	0'2412	0'2495	0'2495	0'2521	0'2540	0'2553	0'2561	0'2565
4	0'1429	0'1586	0'1707	0'1802	0'1878	0'1988	0'1988	0'2027	0'2059	0'20S5
5	0'0695	0'0922	0'1099	0'1240	0'1353	0'1447	0'1524	0'1587	0'1641	0'1686
6	0'0000	0'0303	0'0539	0'0727	0'0880	0'1005	0'1109	0'1197	0'1271	0'1334
7			0'0000	0'0240	0'0433	0'0593	0'0725	0'0837	0'0932	0'1013
8					0'0000	0'0196	0'0359	0'0496	0'0612	0'0711
9							0'0000	0'0163	0'0303	0'0422
10									0'0000	0'0140

Figura XII.1: Coeficientes de Shapiro-Wilk para un tamaño muestral de 11

Se calculan las diferencias entre los pares de valores que equidistan del centro y se multiplican por los coeficientes tabulados:

$x_{(n-i+1)}$	$x_{(i)}$	$x_{(n-i+1)} - x_{(i)}$	a_{n-i+1}	$(x_{(n-i+1)} - x_{(i)}) \cdot (a_{n-i+1})$
13,1	**9,9**	3,2	0,5601	1,7923
12,6	**10,2**	2,4	0,3315	0,7956
12,4	**10,5**	1,9	0,2260	0,4294
12,3	**10,7**	1,6	0,1429	0,2286
11,8	**11,0**	0,8	0,0695	0,0556

$x_{(n-i+1)}$: Valores por encima de la mediana
$x_{(i)}$: Valores por debajo de la mediana
a_{n-i+1}: Coeficientes tabulados en función del tamaño muestral

Así:

$$b = 1{,}7923 + 0{,}7956 + 0{,}4294 + 0{,}2286 + 0{,}0556 = 3{,}3015$$

Dado el estadístico experimental:

$$W_{exp} = \frac{b^2}{(n-1)\cdot S_C^2} = \frac{3{,}3015^2}{(11-1)\cdot 1{,}1482} = 0{,}9493$$

Para un valor de 1, estaríamos antes una réplica de una normal.

El valor crítico, para un nivel de significación del 5 %, resulta ser:

	α								
n	0'01	0'02	0'05	0'10	0'50	0'90	0'95	0'98	0'99
3	0'753	0'756	0'767	0'789	0'959	0'998	0'999	1'000	1'000
4	0'687	0'707	0'748	0'792	0'935	0'987	0'992	0'996	0'997
5	0'686	0'715	0'762	0'806	0'927	0'979	0'986	0'991	0'993
6	0'713	0'743	0'788	0'826	0'927	0'974	0'981	0'986	0'989
7	0'730	0'760	0'803	0'838	0'928	0'972	0'979	0'985	0'988
8	0'749	0'778	0'818	0'851	0'932	0'972	0'978	0'984	0'987
9	0'764	0'791	0'829	0'859	0'935	0'972	0'978	0'984	0'986
10	0'781	0'806	0'842	0'869	0'938	0'972	0'978	0'983	0'986
11	0'792	0'817	0'850	0'876	0'940	0'973	0'979	0'984	0'986
12	0'805	0'828	0'859	0'883	0'943	0'973	0'979	0'984	0'986

Figura XII.2: Valores críticos del test de Shapiro-Wilk

$$W_{11;\,0,05} = 0{,}85$$

Dado que $W_{exp} > W_{n,\alpha}$, no se cae en la región crítica y se puede asumir la hipótesis de que la muestra tiene una estructura normal.

Ejercicio XII.2:

Se desea contrastar si la mediana de la población de la cual se ha extraído la siguiente MAS vale 5: 4, 5, 6, 5, 3, 4, 2, 7, 6, 5, 4, 3, 8, 8, 9, 4, 6, 7, 2, 5, 6.

Se plantea el siguiente contraste T de Wilcoxon:

$$\begin{cases} H_0: Me = 5 \\ H_1: Me \neq 5 \end{cases}$$

En primer lugar, se calcula la diferencia en valor absoluto entre cada elemento de la muestra y la mediana hipotética; a continuación, para aquellos valores en que la diferencia no es nula se asignan rangos de manera ordenada (primero aquellos con una diferencia de 1, luego de 2, etc.) y se promedian:

x_i	Me_0	$x_i - Me_0$	*Signo*	$\lvert x_i - Me_0 \rvert$	*Rangos*	*Promedios*
4	5	−1	−	1	1	4,5
5		0				
6		1	+	1	2	4,5
5		0				
3		−2	−	2	9	10,5
4		−1	−	1	3	4,5
2		−3	−	3	13	14,5
7		2	+	2	10	10,5
6		1	+	1	4	4,5
5		0				
4		−1	−	1	5	4,5
3		−2	−	2	11	10,5
8		3	+	3	14	14,5
8		3	+	3	15	14,5

9		4	+	4	17	17,0
4		−1	−	1	6	4,5
6		1	+	1	7	4,5
7		2	+	2	12	10,5
2		−3	−	3	16	14,5
5		0				
6		1	+	1	8	4,5

x_i: Valores
Me_o: Mediana bajo la hipótesis nula

El número de diferencias no nulas resulta ser $n' = 17$.

Se procede a hacer los sumatorios de rangos promedio, agrupados en función del signo de la diferencia:

$$T^- = 4{,}5 + 10{,}5 + 4{,}5 + 14{,}5 + 4{,}5 + 10{,}5 + 4{,}5 + 14{,}5 = 68{,}0$$
$$T^+ = 4{,}5 + 10{,}5 + 4{,}5 + 14{,}5 + 14{,}5 + 17{,}0 + 4{,}5 + 10{,}5 + 4{,}5 = 85{,}0$$

Por tanto, el estadístico experimental:

$$T_{exp} = mín(T^-, T^+) = mín(68{,}0; 85{,}0) = 68{,}0$$

Para $n' = 17$ y $\alpha = 0{,}05$, se tiene:

n	α			
	0'005	0'01	0'025	0'05
5	0	0	0	0
6	0	0	2	3
7	0	0	2	3
8	0	1	3	5
9	1	3	5	8
10	3	5	8	10
11	5	7	10	13
12	7	9	13	17
13	9	12	17	21
14	12	15	21	25
15	15	19	25	30
16	19	23	29	35
17	23	27	34	41
18	27	32	40	47

Figura XII.3: Tabla para el contraste T de Wilcoxon

Dado que $T_{exp} = 68 > T_{17;\,0{,}05} = 41$, no se cae en la región crítica y por tanto no se puede rechazar H_0.