

Problemas de Bioestadística



Alfredo Sánchez Alberca
asalber@ceu.es
<https://aprendeconalf.es>

Tabla de contenidos

Prefacio	3
Licencia	3
1 Estadística descriptiva	4
2 Regresión y correlación	76
3 Probabilidad	114
4 Variables aleatorias discretas	140
5 Variables aleatorias continuas	159
6 Intervalos de confianza	177
7 Contrastes de hipótesis	193
8 Análisis de la varianza de un factor	205
9 Contrastes de la chi-cuadrado	209

Prefacio

Colección de problemas de Bioestadística para los grados de Ciencias de la Salud (Farmacia, Medicina, Fisioterapia, Óptica, Psicología, Biotecnología, etc.).

Los problemas están agrupados por temas y la mayoría incluye la solución, que puede desplegarse pulsando sobre el bloque *Solución* que aparece bajo el enunciado. Se recomienda intentar resolver cada problema antes de consultar la solución.

Los problemas marcados con una estrella ★ han aparecido en exámenes de años anteriores.

Licencia

Esta obra está bajo una licencia [Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional](#).

1 Estadística descriptiva

Ejercicio 1.1. Se realizó una encuesta a 40 personas de más de 70 años sobre el número de medicamentos distintos que tomaban habitualmente. El resultado de dicha encuesta fue el siguiente:

3 1 2 2 0 1 4 2 3 5 1 3 2 3 1 4 2 4 3 2
3 5 0 1 2 0 2 3 0 1 1 5 3 4 2 3 0 1 2 3

Se pide:

1. Construir la tabla de distribución de frecuencias de la muestra.
2. Dibujar el diagrama de barras de frecuencias relativas y el polígono de frecuencias asociado.
3. Dibujar el diagrama de barras de frecuencias relativas acumuladas y el polígono de frecuencias asociado.

Ejercicio 1.2. Se realizó una encuesta a 40 personas de más de 70 años sobre el número de medicamentos distintos que tomaban habitualmente. El resultado de dicha encuesta fue el siguiente:

3 1 2 2 0 1 4 2 3 5 1 3 2 3 1 4 2 4 3 2
3 5 0 1 2 0 2 3 0 1 1 5 3 4 2 3 0 1 2 3

Calcular los siguientes estadísticos:

1. Media.
2. Mediana.
3. Moda.
4. Varianza.
5. Desviación típica.
6. Coeficiente de variación.

Solución

1. $\bar{x} = 2.225$ medicamentos.
2. $Med = 2$ medicamentos.
3. $Mod = 2$ y 3 medicamentos.

4. $s^2 = 1.974$ medicamentos².
5. $s = 1.405$ medicamentos.
6. $cv = 0.632$, lo que indica que la hay una dispersión moderada-alta.

Ejercicio 1.3. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias en el mes de noviembre. Los datos observados fueron:

15 23 12 10 28 7 12 17 20 21 18 13 11 12 26
30 6 16 19 22 14 17 21 28 9 16 13 11 16 20

Se pide:

1. Construir la tabla de distribución de frecuencias agrupando en clases.
2. Dibujar el diagrama más adecuado para representar la distribución de frecuencias.
3. Dibujar el diagrama más apropiado para representar la distribución de frecuencias acumuladas.

Ejercicio 1.4. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias en el mes de noviembre. Los datos observados fueron:

15 23 12 10 28 7 12 17 20 21 18 13 11 12 26
30 6 16 19 22 14 17 21 28 9 16 13 11 16 20

Calcular los siguientes estadísticos:

1. Media aritmética
2. Mediana.
3. Moda.
4. Tercer decil.
5. Percentil 62.

Solución

Agrupando en las clases $(5, 10]$, $(10, 15]$, $(15, 20]$, $(20, 25]$, $(25, 30]$.

1. $\bar{x} = 16.67$ ingresos.
2. $Me = 16.11$ ingresos.
3. $Mo = (10, 15]$ y $(15, 20]$ ingresos.
4. $D_3 = 12.78$ ingresos.
5. $P_{62} = 18.1$ ingresos.

Ejercicio 1.5. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias en el mes de noviembre. Los datos observados fueron:

115 23 12 10 28 7 12 17 20 21 18 13 11 12 26
30 6 16 19 22 14 17 21 28 9 16 13 11 16 20

Calcular los siguientes estadísticos:

1. Varianza
2. Desviación típica.
3. Coeficiente de variación de Pearson.

Solución

Agrupando en las clases $(5, 10]$, $(10, 15]$, $(15, 20]$, $(20, 25]$, $(25, 30]$.

1. $s^2 = 36.75$ ingresos².
2. $s = 6.06$ ingresos.
3. $cv = 0.365$, lo que indica que hay poca dispersión.

Ejercicio 1.6. ★ El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

0 1 2 1 3 0 1 0 1 2 0 1
1 1 2 0 1 3 2 1 2 1 0 1

Se pide:

1. Construir la tabla de distribución de frecuencias.
2. Dibujar el diagrama de barras y el polígono de frecuencias asociado.
3. Dibujar el diagrama de barras de frecuencias acumuladas el polígono de frecuencias asociado.

Ejercicio 1.7. ★ El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

0 1 2 1 3 0 1 0 1 2 0 1
1 1 2 0 1 3 2 1 2 1 0 1

Se pide:

1. Media.
2. Mediana.

3. Moda.
4. Cuartiles.
5. Percentil 32.

💡 Solución

1. $\bar{x} = 1.125$ lesiones.
2. $Me = 1$ lesiones.
3. $Mo = 1$ lesiones.
4. $C_1 = 0.5$ lesión, $C_2 = 1$ lesiones y $C_3 = 2$ lesiones.
5. $P_{32} = 1$ lesión.

Ejercicio 1.8. ★ El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

0	1	2	1	3	0	1	0	1	2	0	1
1	1	2	0	1	3	2	1	2	1	0	1

Se pide:

1. rango intercuartílico e interpretarlo.
2. Calcular el coeficiente de asimetría e interpretarlo.

💡 Solución

1. $C_1 = 0.5$ lesión, $C_2 = 1$ lesiones y $C_3 = 2$ lesiones. $RI = 1.5$ lesiones, lo que indica que hay bastante dispersión central.
2. $\bar{x} = 1.125$ lesiones, $s^2 = 0.776$ lesiones², $s = 0.88$ lesiones y $g_1 = 0.49$ lo que indica que la distribución es un poco asimétrica a la derecha.

Ejercicio 1.9. La siguiente tabla expresa la distribución de las puntuaciones obtenidas por un grupo de alumnos:

0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
7	8	13	6	7	6	6	5	6	2

Se pide:

1. Dibujar el histograma y polígono de frecuencias.
2. Calcular la media aritmética, la mediana y la moda.
3. Calcular el percentil 92.
4. Calcular la desviación típica.

5. Calcular el coeficiente de asimetría.
6. Calcular del coeficiente de curtosis.

 Solución

1. $\bar{x} = 42.42$ puntos, $Med = 38.3$ puntos y $Mod = (20, 30]$ puntos.
2. $P_{92} = 84.53$ puntos.
3. $s^2 = 695$ puntos², $s = 26.36$ puntos.
4. $g_1 = 0.335$, que indica que la distribución es un poco asimétrica hacia la derecha.
5. $g_2 = -1.06$, que indica que la distribución es bastante platicúrtica.

Ejercicio 1.10. Con el fin de realizar un estudio sobre el aprovechamiento de la energía solar, se han contabilizado las horas de sol registradas durante el mes de enero en las estaciones meteorológicas españolas. Los datos obtenidos son los siguientes:

Horas de Sol	Nº de estaciones
De 50 a 70	2
De 70 a 90	6
De 90 a 110	12
De 110 a 130	12
De 130 a 150	16
De 150 a 170	18
De 170 a 190	10
De 190 a 210	2
De 210 a 230	2
De 230 a 250	2

Hallar la media de horas de sol habidas en dicho mes, la desviación típica y el coeficiente de asimetría e interpretarlos.

 Solución

$\bar{x} = 140$ horas, $s^2 = 1480$ horas² y $s = 38.47$ horas, lo que indica que hay poca dispersión y la media es bastante representativa. $g_1 = 0.226$, lo que indica que la distribución es un poco asimétrica hacia la derecha.

Ejercicio 1.11. En un hospital se ha tomado nota de la concentración de anticuerpos de inmunoglobulina M en el suero sanguíneo de personas sanas, y han resultado los siguientes datos por litro. Entre paréntesis figura el sexo de la persona (H para hombre y M para mujer).

(H) 1.071	(H) 0.955	(H) 0.730	(M) 0.908	(M) 0.859
(H) 0.927	(M) 0.962	(M) 1.543	(H) 1.094	(M) 0.847
(H) 1.214	(M) 1.456	(M) 1.516	(M) 1.002	(M) 0.799
(M) 0.881	(M) 1.096	(M) 0.964	(H) 0.973	(H) 1.222
(H) 0.887	(H) 1.022	(M) 0.881	(M) 1.420	(M) 1.205
(M) 0.822	(M) 0.920	(M) 0.544	(H) 1.254	(H) 2.048
(M) 1.053	(M) 0.673	(M) 1.454	(H) 1.160	(H) 1.327
(M) 1.005	(H) 1.017	(M) 0.806	(H) 1.337	(H) 0.926
(M) 1.029	(H) 1.516	(M) 1.231	(H) 1.249	(M) 1.627
(M) 1.081	(H) 1.416	(M) 1.033	(M) 1.417	(M) 1.031

Se pide

1. Construir la tabla de frecuencias desde 0.5 en clases de amplitud 0.1.
2. Dibujar el histograma de frecuencias relativas y frecuencias relativas acumuladas.
3. ¿En qué población es más representativa la media, en la de hombres o en la de mujeres?

Solución

Llamando X a la concentración de inmunoglobulina en hombres e Y a la concentración en mujeres:

1. $\bar{x} = 1.175$, $s_x = 0.284$, $cv_x = 0.24$, $\bar{y} = 1.077$, $s_y = 0.279$, $cv_y = 0.26$, lo que indica que hay una poca menos dispersión en los hombres y su media es un poco más representativa.

Ejercicio 1.12. ★ En un estudio sobre el crecimiento se tomaron dos muestras, una de niños recién nacidos y otra de niños con un año de edad. Las estaturas observadas en cada muestra fueron:

Recién nacidos: 51, 50, 51, 53, 49, 50, 53, 50, 47, 50

Un año: 62, 65, 69, 71, 65, 66, 68, 69

¿En cuál de las dos muestras es más representativa la media? Justificar la respuesta.

Solución

Llamemos X a las estaturas de los niños recién nacidos e Y a las estaturas de los niños de 1 año. La media muestral es tanto más representativa cuanto más pequeño es el coeficiente de variación. Por tanto, para ver en qué muestra es más representativa la media, tenemos que comparar los coeficientes de correlación de ambas muestras. El coeficiente de correlación está definido por la fórmula

$$cv = \frac{s}{\bar{x}},$$

de modo que, para calcular los coeficientes de variación tenemos que calcular la media y la desviación típica de cada muestra. Llamando X_1 a la talla de los recién nacidos y X_2 a la talla de los niños de un año, tenemos

$$\begin{aligned}\bar{x} &= \frac{\sum_i x_i}{n_x} = \frac{51 + 50 + \dots + 50}{10} = \frac{504}{10} = 50.4\text{cm}, \\ s_x^2 &= \frac{\sum_i x_i^2}{n_x} - \bar{x}^2 = \frac{51^2 + 50^2 + \dots + 50^2}{10} - 50.4^2 = 2543 - 2540.16 = 2.84\text{cm}^2, \\ s_x &= \sqrt{2.84} = 1.69\text{cm}, \\ \bar{y} &= \frac{\sum_i y_i}{n_y} = \frac{62 + 65 + \dots + 69}{8} = \frac{535}{8} = 66.875\text{cm}, \\ s_y^2 &= \frac{\sum_i y_i^2}{n_y} - \bar{y}^2 = \frac{62^2 + 65^2 + \dots + 69^2}{8} - 66.875^2 = 4479.625 - 4472.266 = 7.36\text{cm}^2, \\ s_y &= \sqrt{7.36} = 2.713\text{cm}.\end{aligned}$$

Sustituyendo ahora en la fórmula del coeficiente de variación obtenemos

$$cv_x = \frac{s_x}{\bar{x}} = \frac{1.69}{50.4} = 0.034 \quad cv_y = \frac{s_y}{\bar{y}} = \frac{2.713}{66.875} = 0.041.$$

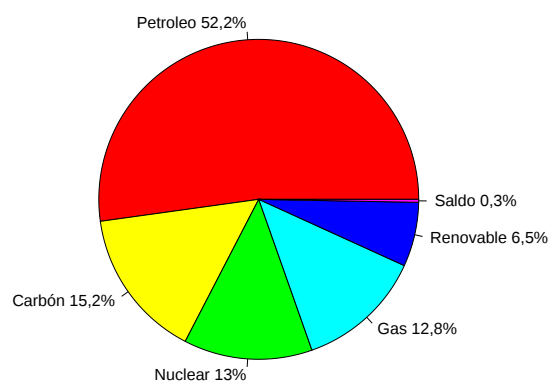
En consecuencia, como $cv_x < cv_y$, la media es más representativa en la muestra de niños recién nacidos.

Ejercicio 1.13. Las temperaturas medias mensuales (en °C) durante el año 2001 en Madrid y Sevilla fueron:

Ciudad	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Madrid	7.2	8.4	12.2	13.7	16.7	23.3	24.2	25.5	20.4	16.2	8.1	4.2
Sevilla	12.1	13.6	16.8	19.0	20.9	27.0	26.6	28.2	24.7	21.3	13.8	11.5

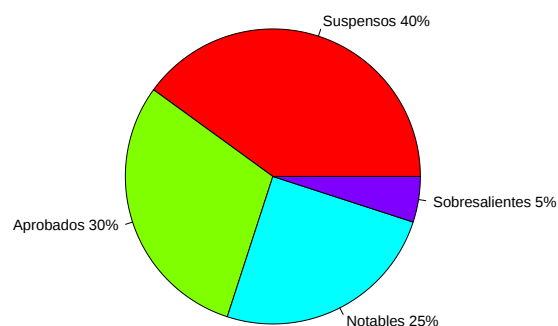
¿En cuál de la dos ciudades ha habido una mayor variación de temperatura?

Ejercicio 1.14. El siguiente diagrama muestra los datos de una muestra de consumo de energía primaria en la industria española durante el año 2002.



Realizar un estudio descriptivo de los datos que incluya todos los estadísticos posibles.

Ejercicio 1.15. ★ El siguiente diagrama refleja el porcentaje de calificaciones obtenidas en un examen realizado a 80 alumnos:



Se pide:

1. Construir la tabla de frecuencias para las calificaciones.
2. Dibujar el polígono de frecuencias acumuladas.
3. Calcular todos los estadísticos de tendencia central que sean posibles.
4. A partir de la variable calificación, construir la variable nota con los siguientes intervalos: Suspenso $[0, 5)$, Aprobado $[5, 7)$, Notable $[7, 9)$ y Sobresaliente $[9, 10]$, y calcular la nota media y estudiar su representatividad.

Nota: En los tres primeros apartados se debe trabajar con la variable calificación, mientras que en el último debe utilizarse la variable nota.

💡 Solución

1. $Me = \text{Aprobado}$ y $Mo = \text{Suspenso}$.
2. $\bar{x} = 5.275$ puntos, $s = 2.447$ puntos y $cv = 0.464$, de manera que la media es moderadamente representativa.

Ejercicio 1.16. ★ Sea la variable estadística agrupada en intervalos cuya distribución de frecuencias viene dada por la siguiente tabla:

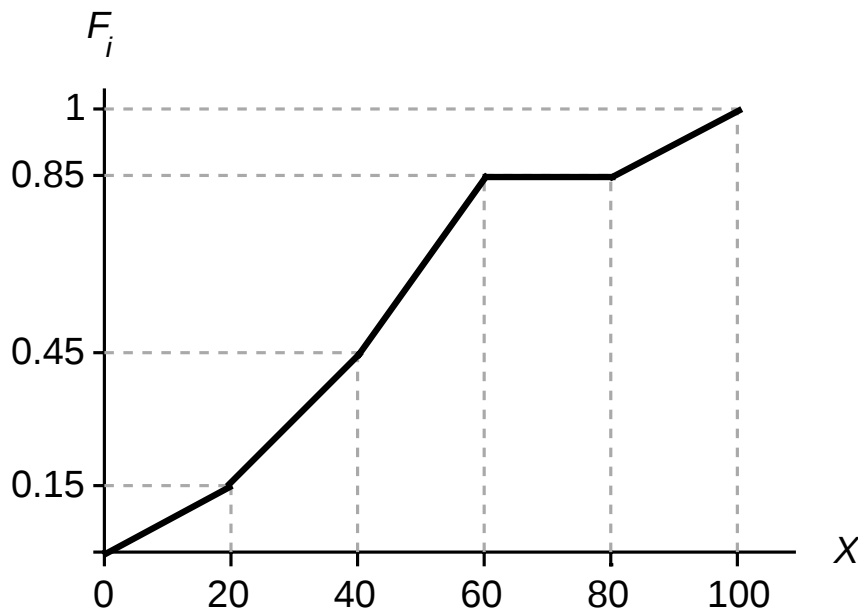
Intervalos	n_i	f_i	N_i	F_i
[0, 10)	10	0.25		
[10, 20)			22	
[20, 30)		0.30		
[30, 40)				

1. Completar la tabla y hallar la desviación típica.
2. Calcular la mediana y el rango intercuartílico e interpretarlos.

 Solución

1. $\bar{x} = 18.5$, $s^2 = 102.75$ y $s = 10.14$.
2. $Med = 18.33$, $C_1 = 10$, $C_3 = 26.27$ y $RI = 16.17$ lo que indica, teniendo en cuenta que el rango de toda la distribución es 40, que la dispersión es moderada.

Ejercicio 1.17. Dada la gráfica correspondiente a un polígono acumulativo de frecuencias relativas de una variable estadística agrupada en intervalos de una muestra de tamaño 20:



se pide:

1. Construir la tabla de frecuencias.

2. Dibujar el histograma correspondiente.
3. Calcular la mediana y la moda.
4. Calcular la media aritmética y la desviación típica.

 Solución

1. $Med = 42.5$ y $Mod = (40, 60)$.
2. $\bar{x} = 44$, $s^2 = 564$, $s = 23.75$.

Ejercicio 1.18. ★ La siguiente tabla representa la distribución de frecuencias del tiempo de atención (en minutos) en una consulta médica:

Intervalos	n_i	f_i	N_i	F_i
[0, 5)	2			
[5, 10)			8	
[10, 15)				0.7
[15, 20)	6			

1. Completar la tabla.
2. Dibujar el polígono de frecuencias acumuladas.

Ejercicio 1.19. ★ La siguiente tabla representa la distribución de frecuencias del tiempo de atención (en minutos) en una consulta médica:

Intervalos	n_i	f_i	N_i	F_i
[0, 5)	2			
[5, 10)			8	
[10, 15)				0.7
[15, 20)	6			

1. Completar la tabla.
2. Calcular el coeficiente de variación y el rango intercuartílico e interpretar los resultados.

 Solución

$\bar{x} = 11.5$, $s^2 = 24$, $s = 4.9$, $cv = 0.426$, lo que indica que hay una dispersión moderada. $C_1 = 7.5$, $C_3 = 15.83$, $RI = 8.33$, lo que indica que la dispersión central también es moderada.

Ejercicio 1.20. ★ Para obtener información acerca del número de consultas al médico, X , que los abonados de una compañía de seguro médico realizan cada mes, trabajando con una muestra de 200 abonados, la distribución de frecuencias fue:

x_i	n_i
0	83
1	51
2	26
3	15
4	9
5	6
6	5
7	3
8	1
9	1

Y se pide calcular:

1. Media, desviación típica y coeficiente de variación del número de visitas al médico. Interpretar el coeficiente de variación.
2. Coeficiente de asimetría de la distribución. Interpretarlo.
3. Percentiles 10 y 90. Interpretarlos.

Solución

1. $\bar{x} = 1.41$ consultas, $s^2 = 3.2919$ consultas², $s = 1.8144$ consultas y $cv = 1.29$, lo que indica que hay muchas dispersión y la media no es muy representativa.
2. $g_1 = 1.66$ lo cual indica que la distribución es bastante asimétrica hacia la derecha, pero no lo suficiente para considerarla anormal.
3. $P_{10} = 0$ lo que indica que el 10% de las personas de la muestra que menos consultas realizaron, realizaron 0 consultas, y $P_{90} = 4$ lo que indica que el 10% de las personas de la muestra que más consultas realizaron, realizaron 4 o más consultas.

Ejercicio 1.21. ★ Los siguientes datos corresponden a 49 mediciones del colesterol HDL (en mg/dl) tomadas en personas de edades y hábitos similares:

69.4	68.3	60.4	70.1	93.1	71.1	58.2
73.6	70.9	71.5	57.1	58.7	56.9	74.8
55.6	66.6	63.7	74.5	77.3	88.0	70.3
75.9	67.5	70.0	67.3	67.6	77.8	70.6
64.9	68.0	85.1	72.8	61.6	69.2	69.6
68.1	50.1	60.2	52.0	76.4	77.2	82.4
74.6	71.9	73.8	82.2	71.9	73.6	69.1

1. Agrupar los datos en 6 clases comenzando en 50 y terminando en 93,2.
2. Calcular el rango intercuartílico e interpretarlo.
3. Razonar si la media de estos datos es representativa de los mismos.

Solución

1. $C_1 = 64.874$ mg/dl, $C_3 = 75.071$ mg/dl y $RI = 10.197$ mg/dl. Esto es bastante menos de la mitad del recorrido de la variable, lo que indica que los datos centrales están bastante concentrados.
2. $\bar{x} = 69.469$ mg/dl, $s^2 = 71.952$ (mg/dl)², $s = 8.482$ mg/dl y $cv = 0.122$, lo que indica que hay poca dispersión y por tanto la media es muy representativa.

Ejercicio 1.22. Se ha llevado a cabo un estudio sobre el número de radiografías realizadas durante el último año a un grupo de 200 personas, y la información se presenta en la siguiente tabla incompleta:

Radiografías	Personas	f_i	F_i
0	84	0.20	0.72
1			
2			
3	24		0.02
4			
5			

1. Completar tabla.
2. Calcular media, mediana, desviación típica y coeficiente de variación e interpretar los resultados.

Solución

1. $\bar{x} = 1.62$ radiografías, $s^2 = 1.875$ radiografías², $s = 1.37$ radiografías y $cv = 0.845$, lo que indica que hay bastante dispersión y la media no es muy representativa.

Ejercicio 1.23. ★ En un estudio diseñado para investigar la efectividad de un nuevo producto anestésico local, la misma cantidad de producto fue suministrada a 20 pacientes, y se midió el tiempo transcurrido hasta lograr cierto grado de sensibilidad. Los resultados, en minutos, son los siguientes:

38, 43, 52, 64, 39, 54, 51, 47, 42, 58, 63, 36, 39, 47, 49, 46, 52, 44, 38, 57

1. Agrupar los datos desde 35 a 65 en 6 clases diferentes.
2. Una vez agrupados, calcular: Media, Desviación Típica y Coeficiente de Asimetría.

3. Teniendo en cuenta la distribución agrupada y suponiendo que todos aquellos datos que se encuentren por arriba del percentil 95 tienen un comportamiento anormal, ¿cuáles de los pacientes se puede considerar que han tenido un tiempo de insensibilidad anormal?

 Solución

1. $\bar{x} = 47.75$ minutos, $s^2 = 66.188$ minutos², $s = 8.136$ minutos. $g_1 = 0.268$ lo que indica que la distribución es ligeramente asimétrica hacia la derecha.
2. $P_{95} = 62.5$ minutos.

Ejercicio 1.24. ★ Si a todos los datos de una muestra se les suma una misma cantidad positiva, ¿cómo se ve afectada la representatividad de la media? ¿Y si se multiplican por un mismo número distinto de 0? Razonar la respuesta.

 Solución

Sea X una variable cualquiera y consideremos la variable $Y = X + c$ resultante de sumar una constante $c > 0$ a los valores de X . Entonces, aplicando las propiedades de las transformaciones lineales tenemos

$$cv_y = \frac{s_y}{|\bar{y}|} = \frac{s_x}{|\bar{x} + c|} < \frac{s_x}{|\bar{x}|} = cv_x,$$

luego al ser menor el coeficiente de variación Y hay menos dispersión relativa y su media es más representativa.

Por otro lado si ahora tomamos $Y = cX$ como la variable resultante de multiplicar los datos de X por una constante $c \neq 0$, tenemos, de nuevo aplicando las propiedades de las transformaciones lineales,

$$cv_y = \frac{s_y}{|\bar{y}|} = \frac{cs_x}{|c\bar{x}|} = \frac{s_x}{|\bar{x}|} = cv_x,$$

de modo que el coeficiente de variación no se altera y por tanto la representatividad de la media es la misma.

Ejercicio 1.25. ★ Como parte de un proyecto de investigación, los investigadores obtuvieron los siguientes datos respecto a los niveles de peróxido lípido (en nmol/ml) en el suero de 30 individuos adultos bajo tratamiento de Diabetes Mellitus:

3.09	6.06	7.34	5.32	4.29	5.36	6.01	7.84	3.87	5.23
4.67	7.89	5.16	6.32	6.45	3.21	5.98	6.45	7.12	4.13
5.16	3.04	4.56	5.67	5.98	6.23	7.34	5.32	4.21	7.13

Agrupar los datos en 5 clases de amplitud unidad, comenzando en 3, y sobre la distribución obtenida:

1. Calcular media, desviación típica y coeficiente de variación de los niveles de peróxido lípido. Interpretar el coeficiente de variación.
2. Calcular cuartiles de la distribución e interpretarlos.
3. Dibujar el diagrama de caja y bigotes y comprobar si hay o no datos atípicos.

Solución

1. $\bar{x} = 5.667$ nmol/ml, $s^2 = 1.668$ (nmol/ml)², $s = 1.292$ nmol/ml y $cv = 0.228$, lo que indica que hay poca dispersión y la media es bastante representativa.
2. $C_1 = 4.7$ nmol/ml, $C_2 = 5.667$ nmol/ml y $C_3 = 6.75$ nmol/ml.
3. Las vallas son $v_1 = 1.625$ y $v_2 = 9.825$. Todos los datos están entre las vallas y no hay datos atípicos. Los bigotes son $b_1 = 3.04$ nmol/ml y $b_2 = 7.89$ nmol/ml.

Ejercicio 1.26. ★ A continuación figura la distribución de edades de una muestra de 65 individuos sujetos a rehabilitación tras un infarto de miocardio:

Edad	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)
Personas	6	12	23	19	5

¿Se puede asumir que la muestra proviene de una población normal?

Utilizar las siguientes sumas: $\sum x_i = 4275$ años, $\sum (x_i - \bar{x})^2 = 7462$ años², $\sum (x_i - \bar{x})^3 = -18249$ años³, $\sum (x_i - \bar{x})^4 = 2099636$ años⁴.

Solución

$\bar{x} = 65.769$ años, $s^2 = 114.823$ años², $s = 10.716$ años. $g_1 = -0.228$ y $g_2 = -0.55$, como tanto el coeficiente de asimetría como el de apuntamiento están entre -2 y 2, podemos suponer que los datos provienen de una población normal.

Ejercicio 1.27. ★ Para obtener información acerca del porcentaje de albúmina en el suero proteico de personas adultas, se analizaron muestras de 32 personas, con los siguientes resultados:

70.2	63.5	65.8	67.9	60.1	69.7	64.2	65.3
62.8	68.4	65.2	66.3	70.7	71.8	68.7	71.9
64.4	62.4	60.4	67.0	62.9	65.9	67.5	66.6
67.8	70.5	63.1	65.3	69.5	71.4	61.0	64.3

1. Agrupar la distribución de porcentajes de albúmina en 6 clases de igual amplitud, desde 60 hasta 72.
2. En la distribución agrupada calcular media, desviación típica, y cuartiles.
3. ¿Es representativa la media de la muestra de porcentajes de albúmina?
4. Dibujar el diagrama de caja y bigotes de la distribución y determinar si hay o no algún dato atípico.

Solución

1. $\bar{x} = 66.3125\%$, $s^2 = 9.9023\%^2$, $s = 3.1468\%$. $C_1 = 64\%$, $C_2 = 66\%$, $C_3 = 69\%$.
2. $cv = 0.05$ lo que indica que hay muy poca dispersión y la media es muy representativa.

Ejercicio 1.28. Para determinar la validez de un terreno para un determinado cultivo, se realizó un muestreo sistemático del terreno obteniendo 18 valoraciones del PH del suelo-agua (en relación 1:2.5). Los valores obtenidos fueron:

5.82 6.23 6.17 7.11 6.44 6.08 6.03 5.91 6.83
6.55 6.24 6.12 6.32 5.86 6.64 6.73 7.24 6.02

Se pide:

1. Calcular el coeficiente de asimetría e interpretarlo.
2. Aplicar una transformación a la variable para obtener otra variable con una simetría más normal.

Ejercicio 1.29. Para determinar la eficacia de un nuevo método para la medición del hematocrito en sangre, se repitió la medida 8 veces sobre una misma muestra de sangre, obteniéndose los siguientes resultados (en porcentaje de hematocrito sobre volumen de plasma sanguíneo):

42.2 42.1 41.9 41.8 42 42.1 41.9 42

¿Se puede afirmar que se trata de un buen método de medición?

Solución

Para ver si se trata de un buen método de medición, tenemos que comprobar que entre las sucesivas mediciones no haya grandes diferencias. Se trata, por tanto, de medir la variabilidad de la muestra y para ello calculamos el coeficiente de variación:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{42.2 + \dots + 42}{8} = \frac{336}{8} = 42\%, \\ s^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{42.2^2 + \dots + 42^2}{8} - 42^2 = \frac{14112.12}{8} - 42^2 = 0.015\%^2, \\ s &= \sqrt{0.015} = 0.1225\%, \\ cv &= \frac{s}{|\bar{x}|} = \frac{0.1225}{42} = 0.003.\end{aligned}$$

Como el coeficiente de variación es muy pequeño, podemos concluir que la variabilidad entre las mediciones es ínfima y por tanto se trata de un buen método de medición.

Ejercicio 1.30. En cuestionario sobre la dependencia de las personas mayores de 75 años se preguntaba sobre la necesidad de ayuda en el desarrollo normal de su vida. Las posibles respuestas eran:

1. Ninguna ayuda.
2. Ayuda al subir las escaleras.
3. Ayuda al subir las escaleras y al incorporarse de una posición sentada o tumbada.
4. Ayuda al subir las escaleras, al incorporarse, y al vestirse.
5. Ayuda para prácticamente todo.

El cuestionario lo respondieron 20 personas, y los resultado obtenidos fueron

B D A B C C B C D E A B C E A B C D B B A A B

Construir la tabla de distribución de frecuencias y el diagrama correspondiente.

Ejercicio 1.31. En cuestionario sobre la dependencia de las personas mayores de 75 años se preguntaba sobre la necesidad de ayuda en el desarrollo normal de su vida. Las posibles respuestas eran:

1. Ninguna ayuda.
2. Ayuda al subir las escaleras.
3. Ayuda al subir las escaleras y al incorporarse de una posición sentada o tumbada.
4. Ayuda al subir las escaleras, al incorporarse, y al vestirse.
5. Ayuda para prácticamente todo.

El cuestionario lo respondieron 20 personas, y los resultado obtenidos fueron

B D A B C C B C D E A B C E A B C D B B A A B

Se pide:

1. Calcular las medidas de tendencia central.
2. Calcular los cuartiles y el decil 8.
3. ¿Qué se puede decir sobre la dispersión?

Solución

1. Me entre b y c y $Mo = b$.
2. $C_1 = b$, C_2 entre b y c , C_3 entre c y d , $D_8 = d$.
3. Suponiendo que hay la misma distancia entre categorías y asignando rangos a cada valor ($a = 1, b = 2, c = 3, d = 4, e = 5$), tenemos $\bar{x} = 2.7$, $s^2 = 1.41$, $s = 1.187$ y $cv = 0.44$, lo que indica una dispersión moderada.

Ejercicio 1.32. ★ Como parte de un proyecto de investigación, los investigadores obtuvieron los siguientes datos respecto a los niveles (en partes por millón) de cierto contaminante químico en 30 suelos diferentes de la Comunidad de Madrid:

3.09	6.06	7.34	5.32	4.29	5.36	6.01	7.84	3.87	5.23
4.67	7.89	5.16	6.32	6.45	3.21	5.98	6.45	7.12	4.13
5.16	3.04	4.56	5.67	5.98	6.23	7.34	5.32	4.21	7.13

Agrupar los datos en 5 clases de amplitud unidad, comenzando en 3, y sobre la distribución obtenida:

1. Calcular media, desviación típica y coeficiente de variación de los niveles de contaminante. Interpretar el coeficiente de variación.
2. Calcular cuartiles de la distribución e interpretarlos.
3. Dibujar el diagrama de caja y bigotes y comprobar si hay o no datos atípicos.

Solución

1. $\bar{x} = 5.667$ ppm, $s^2 = 1.668$ ppm², $s = 1.292$ ppm y $cv = 0.228$, lo que indica que hay poca dispersión y la media es bastante representativa.
2. $C_1 = 4.7$ ppm, $C_2 = 5.667$ ppm y $C_3 = 6.75$ ppm.
3. Las vallas son $v_1 = 1.625$ y $v_2 = 9.825$. Todos los datos están entre las vallas y no hay datos atípicos. Los bigotes son $b_1 = 3.04$ ppm y $b_2 = 7.89$ ppm.

Ejercicio 1.33. ★ Se desea realizar un estudio sobre los días necesarios para tratar una determinada lesión deportiva. Se utilizaron para ello dos tratamientos diferentes, y se observaron 50 pacientes con cada uno de los tratamientos, obteniendo los siguientes resultados:

Días	A	B
------	---	---

20-40	5	8
40-60	20	15
60-80	18	20
80-100	7	7

1. ¿En cuál de los dos tratamientos es más representativa la media del número de sesiones necesarias?
2. ¿Qué tratamiento presenta una distribución más asimétrica?
3. ¿Qué tratamiento presenta una distribución más apuntada?

Usar las siguientes sumas:

A: $\sum x_i = 3040$ días, $\sum (x_i - \bar{x})^2 = 14568$ días², $\sum (x_i - \bar{x})^3 = 17011.2$ días³, $\sum (x_i - \bar{x})^4 = 9989603$ días⁴

B: $\sum x_i = 3020$ días, $\sum (x_i - \bar{x})^2 = 16992$ días², $\sum (x_i - \bar{x})^3 = -42393.6$ días³, $\sum (x_i - \bar{x})^4 = 12551516$ días⁴

Solución

1. $\bar{x}_A = 60.8$ sesiones, $s_A^2 = 291.36$ sesiones², $s_A = 17.0693$ sesiones y $cv_A = 0.28$.
 $\bar{x}_B = 60.4$ sesiones, $s_B^2 = 339.84$ sesiones², $s_B = 18.4348$ sesiones y $cv_B = 0.31$.
Así pues, como $cv_A < cv_B$ la media es un poco más representativa para el tratamiento A.
2. $g_{1_A} = 0.068$ y $g_{1_B} = -0.14$ lo que indica que la distribución para el tratamiento A es ligeramente asimétrica hacia la derecha y la del tratamiento B ligeramente asimétrica hacia la izquierda, aunque ambas son casi simétricas.
3. $g_{2_A} = -0.65$ y $g_{2_B} = -0.83$.

Ejercicio 1.34. ★ El número de muertos en accidentes de carretera durante el 2005 en España fue el siguiente

Mes	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Muertos	272	269	293	274	308	314	374	354	310	318	269	297

Se pide:

1. Calcular el coeficiente de variación del número mensual de muertos e interpretarlo.
2. Calcular la mediana del número mensual de muertos.

Solución

1. Para calcular el coeficiente de variación necesitamos tanto la media como la desviación típica de la variable. Si llamamos X al número mensual de muertos, teniendo en cuenta que los valores se repiten una única vez, y trabajando con ellos sin agrupar, ya que sólo tenemos 12 valores diferentes correspondientes a los 12 meses del año, podemos completar fácilmente la tabla necesaria para el cálculo:

x_i	x_i^2
272	73984
269	72361
293	85849
274	75076
308	94864
314	98596
374	139876
354	125316
310	96100
318	101124
269	72361
297	88209
$\sum x_i = 2652$	$\sum x_i^2 = 1123716$

Con ello:

$$\bar{x} = \frac{\sum x_i}{12} = \frac{3651}{12} = 304.333$$

$$s^2 = \frac{\sum x_i^2}{12} - \bar{x}^2 = \frac{1123716}{12} - 304.333^2 = 1024.425 \Rightarrow s = +\sqrt{1024.425} = 32.007$$

Y el coeficiente de variación vale:

$$Cv = \frac{s}{|\bar{x}|} = \frac{32.007}{304.333} = 0.105 = 10.5\%$$

Por lo tanto, la dispersión relativa de la muestra es bastante pequeña (10.5%) y podemos afirmar que la media de la distribución es bastante representativa.

2. Ordenando los datos de menor a mayor obtenemos la siguiente secuencia:

269, 269, 272, 274, 293, 297, 308, 310, 314, 318, 354, 374

Y para calcular la mediana, una vez ordenados los 12 datos de mayor a menor, nos debemos fijar en los valores de la variable que ocupan las dos posiciones centrales, sexta y séptima, ya que tenemos un número par de datos. En la sexta posición encontramos 297 muertos, y en la séptima 308. Por lo tanto:

$$Me = \frac{297 + 308}{2} = 302.5$$

Ejercicio 1.35. ★ Un médico de familia analiza el número de recetas que ha expedido

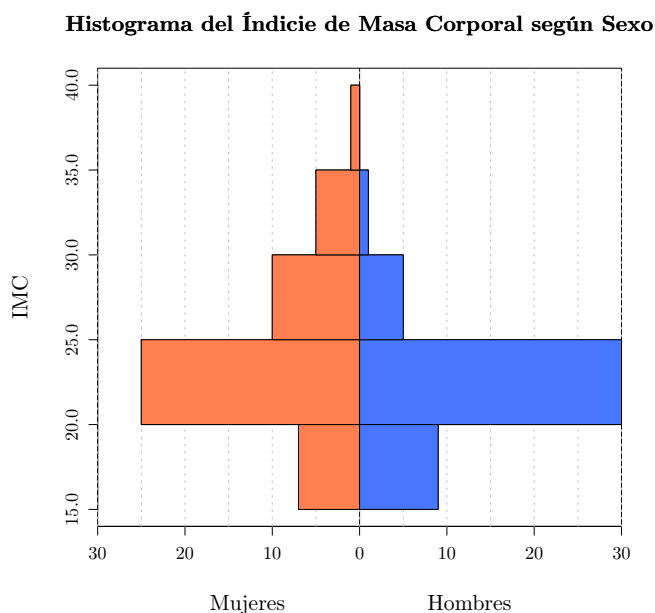
entre sus abonados en los dos últimos meses, obteniendo la siguiente distribución:

Recetas	0	1	2	3	4	6	7	9	12
Abonados	401	203	95	150	166	40	10	12	3

1. Calcular la media, desviación típica y coeficiente de variación del número de recetas e interpretar este último.
2. Si un número medio de recetas superior a 2 puede suponer un fraude, ¿existen pruebas significativas para afirmar que no hay fraude? Justificar la respuesta.
3. Teniendo en cuenta la definición dada más abajo, calcular la Media Recortada 10% ($MR_{0.1}$) del número de recetas. ¿Cuándo crees que será conveniente la utilización de la media recortada en lugar de la media aritmética?

Definición: La *media recortada* $p\%$ es la media aritmética de la muestra que queda al quitar el $p\%$ de valores menores y el $p\%$ de valores mayores de la muestra original.

Ejercicio 1.36. ★ El siguiente histograma refleja la distribución del índice de masa corporal en una muestra de hombres y mujeres.



Se pide:

1. Dibujar el diagrama de sectores para el sexo.
2. ¿En qué grupo es más representativa la media? Justificar la respuesta.
3. ¿Cómo calcularías la media de toda la muestra a partir de las medias de hombres y mujeres? ¿Cuánto vale?

Solución

Llamemos H a la variable que mide el índice de masa corporal para los hombres y M para las mujeres.

1. A partir del histograma de frecuencias absolutas de las mujeres podemos obtener fácilmente las frecuencias absolutas de los intervalos mirando la altura de las barras, y a partir de las frecuencias absolutas podemos calcular el resto de frecuencias. La tabla de frecuencias para las mujeres es:

M	n_i	f_i	N_i	F_i
15-20	7	0.15	7	0.15
20-25	25	0.52	32	0.67
25-30	10	0.21	42	0.88
30-35	5	0.10	47	0.98
35-40	1	0.02	48	1.00
Suma	48	1.00		

Del mismo modo, a partir del histograma de frecuencias absolutas de los hombres obtenemos la tabla de frecuencias para hombres:

H	n_i	f_i	N_i	F_i
15-20	9	0.20	9	0.20
20-25	30	0.67	39	0.87
25-30	5	0.11	44	0.98
30-35	1	0.02	45	1.00
Suma	45	1.00		

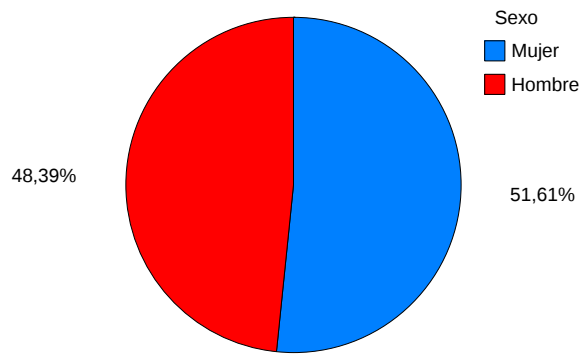
2. La variable Sexo sólo tiene dos categorías que son Hombre y Mujer, de modo que el diagrama de sectores asociado sólo tendrá dos sectores. La tabla de frecuencias de la variable sexo es:

Sexo	n_i	f_i
Mujer	48	0.516
Hombre	45	0.484
Suma	93	1.000

Así pues, para calcular el ángulo correspondiente a cada sector basta con multiplicar 360° (el ángulo de la circunferencia completa) por la frecuencia relativa de cada categoría:

$$\alpha_m = 360^\circ 0.516 = 186^\circ \quad \alpha_h = 360^\circ 0.484 = 174^\circ.$$

Y el diagrama de sectores asociados es:



3. La media será más representativa en el grupo que tenga menos dispersión, y para comparar la dispersión de ambos grupos necesitamos el coeficiente de variación. Calculamos para ello los estadísticos necesarios a partir de las tablas de frecuencias. Para las mujeres tenemos:

M	m_i	n_i	$m_i n_i$	$m_i^2 n_i$
15-20	17.5	7	122.5	2143.75
20-25	22.5	25	562.5	12656.25
25-30	27.5	10	275.0	7562.50
30-35	32.5	5	162.5	5281.25
35-40	37.5	1	37.5	1406.25
Suma		48	1160.0	29050.00

$$\begin{aligned}\bar{m} &= \frac{\sum m_i}{n_m} = \frac{1160}{48} = 24.17, \\ s_m^2 &= \frac{\sum m_i^2}{n_m} - \bar{m}^2 = \frac{29050}{48} - 24.17^2 = 21.1806, \\ s_m &= \sqrt{21.1806} = 4.6, \\ cv_m &= \frac{s_m}{|\bar{m}|} = \frac{4.6}{24.17} = 0.19.\end{aligned}$$

Y para los hombres:

H	h_i	n_i	$h_i n_i$	$h_i^2 n_i$
15-20	17.5	9	157.5	2756.25
20-25	22.5	30	675.0	15187.50
25-30	27.5	5	137.5	3781.25
30-35	32.5	1	32.5	1056.25
Suma		45	1002.5	22781.25

$$\begin{aligned}\bar{h} &= \frac{\sum h_i}{n_h} = \frac{1002.5}{45} = 22.28, \\ s_h^2 &= \frac{\sum h_i^2}{n_h} - \bar{h}^2 = \frac{22781.25}{45} - 22.28^2 = 9.9506, \\ s_h &= \sqrt{9.9506} = 3.15, \\ cv_h &= \frac{s_h}{|\bar{h}|} = \frac{3.15}{22.28} = 0.14.\end{aligned}$$

Así pues, ambas muestras tienen un coeficiente de variación bajo y por tanto las medias son muy representativas, pero lo es un poco más la de los hombres ya que tienen una poca menos dispersión.

- Para calcular la media de toda la muestra podemos hacer una media ponderada de las medias de las mujeres y los hombres donde los pesos son los tamaños muestrales de cada grupo.

$$\bar{x} = \frac{p_m \bar{m} + p_h \bar{h}}{p_m + p_h} = \frac{45 \cdot 22.28 + 48 \cdot 24.17}{45 + 48} = 23.25.$$

Ejercicio 1.37. ★ Se ha medido el tiempo necesario, en días, para curar un tipo concreto de infección (tiempo transcurrido desde el comienzo del tratamiento hasta el alta) en dos grupos de pacientes a los que se han aplicado antibióticos diferentes. En un grupo de 80 pacientes tratados con el antibiótico A, los resultados obtenidos fueron:

Tiempo (días)	n_i
[0, 30)	50
[30, 60)	20
[60, 90)	10

Mientras que en un grupo de 100 pacientes tratados con el antibiótico B se obtuvo:

Tiempo (días)	n_i
[0, 30)	40
[30, 60)	50
[60, 90)	10

1. Calcular la media y la desviación típica del número de días hasta el alta, para cada uno de los antibióticos.
2. ¿En qué antibiótico es más representativa la media?. Justificar la respuesta numéricamente.
3. ¿Cuánto vale el percentil 90 del tiempo de curación para los pacientes tratados con B?
4. ¿Qué tiempo de recuperación ha sido relativamente más alto, uno de 40 días de un paciente tratado con A, o uno de 45 días de otro tratado con B? Justificar la respuesta numéricamente.

Ejercicio 1.38. ★ Para estudiar la eficacia de un nuevo fármaco en el tratamiento de la hipertensión arterial, se tomaron los valores de la tensión arterial diastólica (TAD) en 8 pacientes antes del tratamiento (X) y después de un mes de tratamiento (Y), obteniéndose los siguientes resultados:

X	100	104	96	98	105	106	110	100
Y	92	97	94	100	94	92	95	104

1. ¿En qué variable es más representativa la media?
2. Hallar el coeficiente de asimetría de la variable Y e interpretarlo.

Solución

1. Para ver en qué muestra es más representativa la media tenemos que comparar las dispersiones de ambas muestras mediante el coeficiente de variación. Para ello, procedemos al cálculo de los estadísticos necesarios:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{100 + \dots + 100}{8} = \frac{819}{8} = 102.375, \\ s_x^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{100^2 + \dots + 100^2}{8} - 102.375^2 = \frac{83997}{8} - 10480,6406 = 18.9844, \\ s_x &= \sqrt{18.9844} = 4.3571, \\ cv_x &= \frac{s_x}{|\bar{x}|} = \frac{4.3571}{102.375} = 0.0426, \\ \bar{y} &= \frac{\sum y_j}{n} = \frac{92 + \dots + 104}{8} = \frac{768}{8} = 96, \\ s_y^2 &= \frac{\sum y_j^2}{n} - \bar{y}^2 = \frac{92^2 + \dots + 104^2}{8} - 96^2 = \frac{73850}{8} - 9216 = 15.25, \\ s_y &= \sqrt{15.25} = 3.9051, \\ cv_y &= \frac{s_y}{|\bar{y}|} = \frac{3.9051}{96} = 0.0407,\end{aligned}$$

Ambos coeficientes de variación son muy pequeños, lo cual indica que hay muy poca dispersión en las muestras y las medias son muy representativas, aunque un poco más la de Y por tener un coeficiente de variación más pequeño.

2. El coeficiente de asimetría de Y es

$$g_1 = \frac{\sum (y_j - \bar{y})^3 / n}{s_y^3} = \frac{((92 - 96)^3 + \dots + (104 - 96)^3) / 8}{3.9051^3} = \frac{432/8}{59,552} = 0.9068,$$

lo que indica que la distribución es asimétrica a la derecha aunque no lo suficiente como para rechazar la hipótesis de que los datos provienen de una población normal.

Ejercicio 1.39. ★ Una piscifactoría realiza un estudio para determinar si una especie de pez está en peligro extinción en los ríos de una región. Para ello se midió el número de peces en 10 sitios diferentes obteniendo los siguientes resultados en unidades por hectómetro cúbico:

36 – 11 – 7 – 18 – 35 – 27 – 21 – 15 – 25 – 12

Se pide:

1. Calcular las medidas de tendencia central.
2. Calcular el coeficiente de apuntamiento e interpretarlo.

Solución

Llamemos X al número de peces por hectómetro cúbico.

1. Las medidas de tendencia central son la media aritmética, la mediana y la moda. Calculamos primero la media:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{36 + \dots + 12}{10} = \frac{207}{10} = 20.7 \text{ peces,}$$

Para calcular la mediana ordenamos los valores de la muestra:

7 – 11 – 12 – 15 – 18 – 21 – 25 – 27 – 35 – 36

Como el tamaño de la muestra es par, habrá dos individuos que ocupen las posiciones centrales de la muestra, y estos serán los que ocupen las posiciones $n/2 = 10/2 = 5$ y $n/2 + 1 = 6$, es decir, los valores 18 y 21. Así pues la mediana es

$$Me = \frac{18 + 21}{2} = 19.5 \text{ peces.}$$

Por último, no podemos decir que haya moda porque todos los valores tienen frecuencia 1.

2. La fórmula del coeficiente de apuntamiento es

$$g_2 = \frac{\sum (x_i - \bar{x})^4}{ns^4} - 3,$$

así que necesitamos calcular previamente la desviación típica:

$$s^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{36^2 + \dots + 12^2}{10} - 20.7^2 = \frac{5179}{10} - 428.49 = 89.41,$$
$$s = \sqrt{89.41} = 9.4557.$$

Así pues, tenemos

$$g_2 = \frac{(36 - 20.7)^4 + \dots + (12 - 20.7)^4}{10 \cdot 9.4557^4} - 3 = \frac{149449.657}{79941.9504} - 3 = -1.13,$$

lo que indica que la distribución es bastante platicúrtica, es decir, con menos apuntamiento que una distribución normal.

Ejercicio 1.40. ★ Las temperaturas (en °C) observadas en Barcelona y La Coruña en 12 días elegidos aleatoriamente el último año fueron:

Barcelona	13	14	16	18	21	25	28	28	25	21	17	13
La Coruña	13	13	15	16	17	20	22	23	22	19	15	13

Se pide:

1. ¿En cuál de las dos ciudades hay mayor variación de temperaturas?
2. ¿Por debajo de qué valor estarán las temperaturas de la mitad de los días en ambas ciudades?

Solución

Llamemos B a la variable que mide las temperaturas en Barcelona, y C a la que mide las temperaturas en La Coruña.

1. Para ver en qué ciudad hay mayor variación de temperatura tenemos que calcular los coeficientes de variación de ambas ciudades y compararlos.

$$\begin{aligned}\bar{b} &= \frac{\sum b_i}{n} = \frac{13 + 14 + \dots + 13}{12} = \frac{239}{12} = 19.9167, \\ s_b^2 &= \frac{\sum b_i^2}{n} - \bar{b}^2 = \frac{13^2 + 14^2 + \dots + 13^2}{12} - 19.9167^2 = \frac{5103}{12} - 396.6749 = 28.5764, \\ s_b &= \sqrt{28.5764} = 5.3457, \\ cv_b &= \frac{s_b}{|\bar{b}|} = \frac{5.3457}{19.9167} = 0.2684, \\ \bar{c} &= \frac{\sum c_i}{n} = \frac{13 + 13 + \dots + 13}{12} = \frac{208}{12} = 17.3333, \\ s_c^2 &= \frac{\sum c_i^2}{n} - \bar{c}^2 = \frac{13^2 + 13^2 + \dots + 13^2}{12} - 17.3333^2 = \frac{3760}{12} - 300.4433 = 12.8889, \\ s_c &= \sqrt{12.8889} = 3.5901, \\ cv_c &= \frac{s_c}{|\bar{c}|} = \frac{3.5901}{17.3333} = 0.2071.\end{aligned}$$

Por consiguiente, como el coeficiente de variación es mayor en Barcelona, es en esta ciudad donde hay mayor variación de temperaturas.

2. El valor que cumple que por debajo del mismo están la mitad de los valores de la muestra es la mediana. Para calcular la mediana en ambas ciudades ordenamos los valores de la muestra de menor a mayor, y puesto que tenemos

un tamaño muestral par $n = 12$, busquemos los dos valores centrales, que ocuparan las posiciones $n/2 = 6$ y $n/2 + 1 = 7$. En el caso de Barcelona tenemos que dichos valores son

Posición	1	2	3	4	5	6	7	8	9	10	11	12
B	13	13	14	16	17	18	21	21	25	25	28	28

Como son valores diferentes, la mediana es $Me_b = \frac{18 + 21}{2} = 19,5$.

En el caso de La Coruña tenemos

Posición	1	2	3	4	5	6	7	8	9	10	11	12
C	13	13	13	15	15	16	17	19	20	22	22	23

Y al igual que antes, la mediana es $Me_c = \frac{16 + 17}{2} = 16,5$.

Ejercicio 1.41. ★ Las pruebas de determinación del cociente intelectual realizadas en un colectivo de alumnos universitarios reflejan los siguientes resultados:

C.I.	Nº de alumnos
[80,90)	3
[90,100)	12
[100,110)	21
[110,120)	24
[120,130)	13
[130,140)	2

1. Calcular el coeficiente de variación del cociente intelectual. ¿Es representativa la media?
2. Si se considera *muy inteligente* a una persona cuyo cociente intelectual se encuentra en el 10% de los más inteligentes, ¿cuál será el cociente que delimita la categoría *muy inteligente* en el colectivo de alumnos?

Solución

1. Para calcular el coeficiente de variación necesitamos la media y la desviación típica. Para facilitar los cálculos, añadimos dos nuevas columnas a la tabla de distribución de frecuencias:

C.I.	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
[80,90)	85	3	255	21675
[90,100)	95	12	1140	108300
[100,110)	105	21	2205	231525
[110,120)	115	24	2760	317400
[120,130)	125	13	1625	203125
[130,140)	135	2	270	36450
Sumas		75	8255	918475

A partir de aquí calculamos los estadísticos que necesitamos.

$$\begin{aligned}\bar{x} &= \frac{\sum x_i n_i}{N} = \frac{8255}{75} = 110.07, \\ s^2 &= \frac{\sum x_i^2 n_i}{N} - \bar{x}^2 = \frac{918475}{75} - 110.07^2 = 131.6622, \\ s &= \sqrt{131.6622} = 11.47.\end{aligned}$$

Finalmente calculamos el coeficiente de variación:

$$cv = \frac{s}{|\bar{x}|} = \frac{11.47}{110.07} = 0.1.$$

Puesto que el coeficiente de variación es pequeño, podemos concluir que hay poca dispersión en la muestra y por tanto, la media es representativa.

2. Buscamos el coeficiente intelectual por encima del cual está el 10% más inteligente de la clase, o lo que es lo mismo, por debajo del cual está el 90% menos inteligente de la clase. En definitiva se trata de calcular el percentil 90. Para ello necesitamos la columna de frecuencias acumuladas en la tabla

C.I.	n_i	N_i
[80,90)	3	3
[90,100)	12	15
[100,110)	21	36
[110,120)	24	60
[120,130)	13	73
[130,140)	2	75

La frecuencia acumulada correspondiente al percentil 90 es $N_{P_{90}} = 90 \cdot 75/100 = 67.5$. Mirando en la columna de frecuencias acumuladas de la tabla de frecuencias comprobamos que dicha frecuencia se alcanza dentro del intervalo [120,130). Interpolando en este intervalo obtenemos el percentil que buscamos

$$P_{90} = 120 + \frac{67.5 - 60}{13} 10 = 125.77.$$

Ejercicio 1.42. ★ En un laboratorio farmacéutico se van a fabricar dos tipos de cápsulas A y B, cuyos contenidos en principio activo debe ser 250 y 500 mg. respectivamente. Para ello se prueban dos máquinas dosificadoras, una para cada tipo de cápsulas, siendo los contenidos en principio activo de las cápsulas fabricadas con ellas los siguientes:

A	B
238	488
245	494
236	476
248	483
241	492
243	

¿En cuál de las dos máquinas es más representativa la media? Razonar la respuesta.

Solución

Llamaremos A a la variable que mide la cantidad de principio activo que pone la máquina A en cada cápsula y B a la variable que mide la cantidad de principio activo que pone la máquina B en cada cápsula. Para ver en cual de las dos muestras es más representativa la media, tenemos que calcular los coeficientes de variación en cada caso.

Calculamos primero los estadísticos necesarios para la máquina A :

$$\begin{aligned}\bar{A} &= \frac{\sum A_i}{n_A} = \frac{1451}{6} = 241.833, \\ s_A^2 &= \frac{\sum A_i^2 n_A}{n_A} - \bar{A}^2 = \frac{350999}{6} - 241.833^2 = 16.472, \\ s_A &= \sqrt{16.472} = 4.058, \\ cv_A &= \frac{s_A}{\bar{A}} = \frac{4.058}{241.833} = 0.0167,\end{aligned}$$

En el caso de la máquina B tenemos:

$$\begin{aligned}\bar{B} &= \frac{\sum B_i}{n_B} = \frac{2433}{5} = 486.6, \\ s_B^2 &= \frac{\sum B_i^2 n_B}{n_B} - \bar{B}^2 = \frac{1184109}{5} - 486.6^2 = 42.24, \\ s_B &= \sqrt{42.24} = 6.499, \\ cv_B &= \frac{s_B}{\bar{B}} = \frac{6.499}{486.6} = 0.0133.\end{aligned}$$

Como el coeficiente de variación de la muestra de B es un poco menor que el de A , la dispersión será menor en la muestra de B y $\bar{B}$ será un poco más representativa que $\bar{A}$.

Ejercicio 1.43. ★ En un centro de reclutamiento se han medido las alturas de 150 jóvenes, obteniéndose la siguiente tabla de frecuencias:

Altura:	Número de jóvenes:
1.50-1.60	12
1.60-1.70	30
1.70-1.80	52
1.80-1.90	42
1.90-2.00	11
2.00-2.10	3

1. Calcular el coeficiente de variación. ¿Es representativa la media?
2. Si se considera alto a un joven cuya altura corresponde como mínimo a la del percentil 85, ¿cuál es la mínima altura que tendrá un alto?

Solución

1. El coeficiente de variación se define como

$$cv = \frac{s}{|\bar{x}|},$$

de modo que necesitamos calcular la media y la desviación típica, pero antes de nada, completamos la tabla de frecuencias.

x_i	n_i	N_i	$x_i n_i$	$x_i^2 n_i$
1.50-1.60	12	12	18.6	28.83
1.60-1.70	30	42	49.5	81.68
1.70-1.80	52	94	91	159.25
1.80-1.90	42	136	77.7	143.75
1.90-2.00	11	147	21.45	41.83
2.00-2.10	3	150	6.15	12.61
Sumas	150		264.4	467.94

Ahora calculamos la media y la desviación típica:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i n_i}{N} = \frac{264.4}{150} = 1.7627. \\ s^2 &= \frac{\sum x_i^2 n_i}{N} - \bar{x}^2 = \frac{467.94}{150} - 1.7627^2 = 0.013. \\ s &= \sqrt{0.02} = 0.114.\end{aligned}$$

Por consiguiente, el coeficiente de variación es

$$cv = \frac{0.114}{1.7627} = 0.065,$$

lo que indica que hay muy poca dispersión y la media es muy representativa.

2. Tenemos que calcular el percentil 85 para ver a partir de qué estatura un joven es considerado alto. La frecuencia absoluta acumulada correspondiente al percentil 85 es $85 \cdot N/100 = 85 \cdot 150/100 = 127.5$. Mirando la tabla de frecuencias en la columna de las frecuencias absolutas acumuladas podemos comprobar que la frecuencia 127.5 se alcanza en algún punto del intervalo 1.80-1.90. Para obtener el percentil 85 interpolamos en dicho intervalo:

$$P_{85} = 1.80 + \frac{127.5 - 94}{42}(1.90 - 1.80) = 1.88.$$

Ejercicio 1.44. ★ Se ha realizado un estudio sobre la tensión arterial en dos ciudades A y B. Se tomó una muestra de 20 individuos, que arrojó los siguientes valores:

Tensión	135	128	137	110	154	142	121	127	114	103
Ciudad	A	B	A	B	A	A	A	B	B	B

Tensión	98	96	114	123	132	141	132	121	98	136
Ciudad	A	B	A	B	A	B	A	B	B	A

Se pide:

1. Construir la tabla de frecuencias para la tensión, agrupando en clases de amplitud 10, entre 95 y 155.
2. Dibujar el polígono de frecuencias acumuladas correspondiente a la tabla anterior.
3. Calcular el coeficiente de asimetría e interpretarlo.
4. Calcular el tercer decil e interpretarlo.
5. Considerando los datos sin agrupar, ¿en qué ciudad son más homogéneas las tensiones?

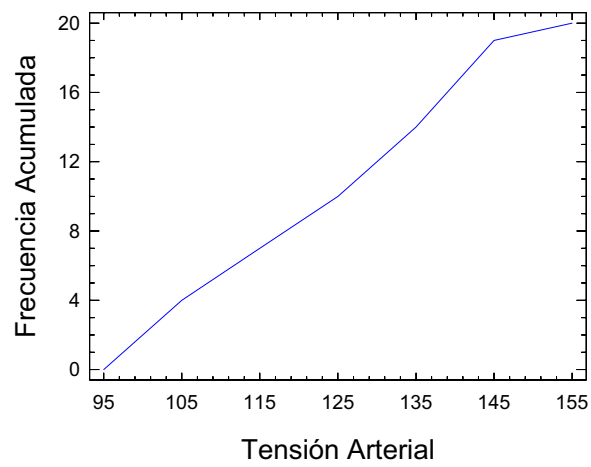
Solución

Llamemos X a la variable que mide la tensión arterial.

1. Agrupando en clases de amplitud 10, obtenemos 6 clases entre 95 y 155. La tabla de frecuencias correspondiente es

X	x_i	n_i	f_i	N_i	F_i
[95,105)	100	4	0.2	4	0.2
[105,115)	110	3	0.15	7	0.35
[115,125)	120	3	0.15	10	0.5
[125,135)	130	4	0.2	14	0.7
[135,145)	140	5	0.25	19	0.95
[145,155)	150	1	0.05	10	1

2. El polígono de frecuencias absolutas acumuladas correspondiente a esta tabla es el siguiente



3. El coeficiente de asimetría se calcula mediante la fórmula

$$g_1 = \frac{\sum (x - \bar{x})^3 n_i / n}{s^3}.$$

A partir de la tabla anterior, calculamos los estadísticos necesarios:

X	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
[95,105)	100	4	400	40000
[105,115)	110	3	330	36300
[115,125)	120	3	360	43200
[125,135)	130	4	520	67600
[135,145)	140	5	700	98000
[145,155)	150	1	150	22500
Sumas		20	2460	307600

$$\begin{aligned}\bar{x} &= \frac{\sum x_i n_i}{n} = \frac{2460}{20} = 123, \\ s^2 &= \frac{\sum x_i^2 n_i}{N} - \bar{x}^2 = \frac{307600}{20} - 123^2 = 251, \\ s &= \sqrt{251} = 15.84.\end{aligned}$$

Por último calculamos el coeficiente de asimetría

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
[95,105)	100	4	-23	-48668
[105,115)	110	3	-13	-6591
[115,125)	120	3	-3	-81
[125,135)	130	4	7	1372
[135,145)	140	5	17	24565
[145,155)	150	1	27	19683
Sumas		20		-9720

$$g_1 = \frac{\sum (x - \bar{x})^3 n_i / n}{s^3} = \frac{-9720/20}{15.84^3} = -0.12.$$

Esto indica que la distribución es ligeramente asimétrica a la izquierda.

4. El tercer decil D_3 tiene frecuencia absoluta acumulada $N_{D_3} = 3n/10 = 3 \cdot 20/10 = 6$, y mirando en la columna de frecuencias absolutas acumuladas comprobamos que corresponderá a un individuo de la clase $[105, 115)$. Interpolando en dicho intervalo tenemos que el tercer decil es

$$D_3 = 105 + \frac{6 - 4}{3} 10 = 111,67,$$

lo cual indica que el 30% de la población tiene un tensión inferior a 111.67.

5. Llamemos a y b a las variables que miden las tensiones de las ciudades A y B respectivamente. Las tensiones serán más homogéneas en la ciudad que haya menos dispersión. Para comparar las dispersiones de ambas poblaciones utilizamos el coeficiente de variación que no depende de las unidades. Calculamos los estadísticos necesarios trabajando directamente desde la muestra sin agrupar:

$$\begin{aligned}\bar{a} &= \frac{\sum a_i}{n} = \frac{135 + \dots + 136}{10} = \frac{1301}{10} = 130.1, \\ s_a^2 &= \frac{\sum a_i^2}{n} - \bar{a}^2 = \frac{135^2 + \dots + 136^2}{10} - 130.1^2 = \frac{171459}{10} - 16926.01 = 219.89, \\ s_a &= \sqrt{219.89} = 14.83, \\ cv_a &= \frac{s_a}{|\bar{a}|} = \frac{14.83}{130.1} = 0.1139, \\ \bar{b} &= \frac{\sum b_j}{n} = \frac{128 + \dots + 98}{10} = \frac{1161}{10} = 116.1, \\ s_b^2 &= \frac{\sum b_j^2}{n} - \bar{b}^2 = \frac{128^2 + \dots + 98^2}{10} - 116.1^2 = \frac{136689}{10} - 13479.21 = 189.69, \\ s_b &= \sqrt{189.69} = 13.77, \\ cv_b &= \frac{s_b}{|\bar{b}|} = \frac{13.77}{116.1} = 0.1186.\end{aligned}$$

Así pues, comparando los coeficientes de variación en ambas ciudades comprobamos que son prácticamente iguales aunque es un poco menor la dispersión en la ciudad A por lo que las tensiones serán un poco más homogéneas.

Ejercicio 1.45. ★ En un grupo de alumnos se ha medido el número de días que faltaron

a clase una muestra de alumnos de primer año y otra de alumnos repetidores, obteniendo:

Primer año:	8	2	3	2	1	4	0	3
Repetidores:	4	6	20	12	16	8	5	

Se pide:

1. Calcular la mediana en ambos casos.
2. ¿En cuál de las dos muestras hay menos dispersión?
3. ¿Comparar el apuntamiento de ambas muestras?

Solución

Llamemos X a la variable que mide el número de días que faltaron los alumnos de primer año, e Y al de los alumnos repetidores

1. Para calcular la mediana primero ordenamos los datos de ambas muestras:

Primer año:	0	1	2	2	3	3	4	8
Repetidores:	4	5	6	8	12	16	20	

En el primer caso, como el tamaño de la muestra es par, habrá dos valores que estarán en el centro de la distribución, que serán los que ocupen la posiciones $n/2 = 8/2 = 4$ y $n/2 + 1 = 5$, es decir, los valores 2 y 3, de modo que la mediana es su media:

$$Me_x = \frac{2+3}{2} = 2.5.$$

En el segundo caso, como el tamaño muestral es impar, sólo habrá un valor central que será el que ocupe la posición $(n+1)/2 = 8/2 = 4$, es decir, el valor 8, y por tanto $Me_y = 8$.

2. Para compara la dispersión de ambas muestras tenemos que calcular el coeficiente de variación, y para ello, previamente hay que calcular la media y la desviación típica de cada muestra:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{8 + \dots + 3}{8} = \frac{23}{8} = 2.875, \\ s_x^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{8^2 + \dots + 3^2}{8} - 2.875^2 = \frac{107}{8} - 8,2656 = 5.1094, \\ s_x &= \sqrt{5.1094} = 2.2604, \\ cv_x &= \frac{s_x}{|\bar{x}|} = \frac{2.2604}{2.875} = 0.7862, \\ \bar{y} &= \frac{\sum y_j}{n} = \frac{4 + \dots + 5}{7} = \frac{71}{7} = 10.143, \\ s_y^2 &= \frac{\sum y_j^2}{n} - \bar{y}^2 = \frac{4^2 + \dots + 5^2}{7} - 10.143^2 = \frac{941}{7} - 102.8804 = 31.5510, \\ s_y &= \sqrt{31.5510} = 5.617, \\ cv_y &= \frac{s_y}{|\bar{y}|} = \frac{5.617}{10.143} = 0.5538.\end{aligned}$$

Así pues, como el coeficiente de variación es menor en Y , hay menos dispersión el grupo de los repetidores.

3. El coeficiente de apuntamiento en ambas muestras es

$$\begin{aligned}g_{2_x} &= \frac{\sum (x_i - \bar{x})^4}{ns_x^4} - 3 = \frac{(8 - 2.875)^4 + \dots + (3 - 2.875)^4}{8 \cdot 2.2604^4} - 3 = \frac{773.3379}{208.8484} - 3 = 0.703, \\ g_{2_y} &= \frac{\sum (y_j - \bar{y})^4}{ns_y^4} - 3 = \frac{(4 - 10.143)^4 + \dots + (5 - 10.143)^4}{7 \cdot 5.617^4} - 3 = \frac{13068.6239}{6968.1218} - 3 = -1.125.\end{aligned}$$

Por tanto, la primera muestra tiene un apuntamiento mayor de lo normal (leptocúrtica) mientras que la segunda muestra tiene un apuntamiento bastante menor de lo normal (platicúrtica).

Ejercicio 1.46. ★ Al realizar un estudio sobre el peso de las mujeres mayores de 30 años en una determinada población, se obtuvieron los siguientes datos:

$$\begin{array}{cccccccccccc}72 & - & 66 & - & 51 & - & 87 & - & 65 & - & 57 & - & 73 & - & 84 & - & 67 & - & 78 \\58 & - & 62 & - & 75 & - & 56 & - & 68 & - & 74 & - & 57 & - & 65 & - & 73 & - & 67\end{array}$$

Realizar un estudio descriptivo agrupando los datos en 4 clases de amplitud 10 comenzando en el 50, que incluya:

1. Histograma de frecuencias absolutas y frecuencias absolutas acumuladas y los correspondientes polígonos.

2. Rango intercuartílico e interpretación.
3. Estudiar la representatividad de la media.

Ejercicio 1.47. ★ Se realiza un estudio sobre el nivel de colesterol en sangre de los trabajadores de una empresa que utilizan habitualmente el comedor de la misma. Los resultados fueron los siguientes:

214 – 205 – 197 – 204 – 186 – 192 – 216 – 176 – 192 – 196
 207 – 182 – 218 – 203 – 194 – 176 – 203 – 214 – 188 – 198

Se pide:

1. Obtener la distribución de frecuencias agrupadas en clases de amplitud 10, comenzando en 170.
2. Dibujar el histograma y el polígono de frecuencias.
3. Calcular el coeficiente de variación de Pearson e interpretar los resultados.
4. Calcular el rango intercuartílico.

Ejercicio 1.48. En un grupo de alumnos universitarios se realiza un estudio sobre la velocidad de lectura. Para ellos se les dio a leer un texto de un periódico y se contó el número de palabras que leyeron en un minuto, obteniendo los siguientes resultados:

286 – –242 – –185 – –244 – –296 – –211 – –233 – –179 – –190 – –274
 255 – –222 – –260 – –216 – –204 – –247 – –232 – –257 – –196 – –261

Se pide:

1. Construir la tabla de frecuencias agrupando en intervalos de amplitud 25 comenzando en 175.
2. Calcular las medidas de tendencia central. ¿Es representativa la media?
3. Calcular el coeficiente de apuntamiento e interpretarlo.

Solución

1. $\bar{x} = 233.75$, $Med = 235$, $Mod = [225, 250)$ y $[250 - 275]$. $s^2 = 1017.19$, $s = 31.89$ y $cv = 0.14$ lo que indica que hay poca dispersión y la media es muy representativa.
2. $g_2 = -1.1$ de manera que la distribución es bastante platicúrtica.

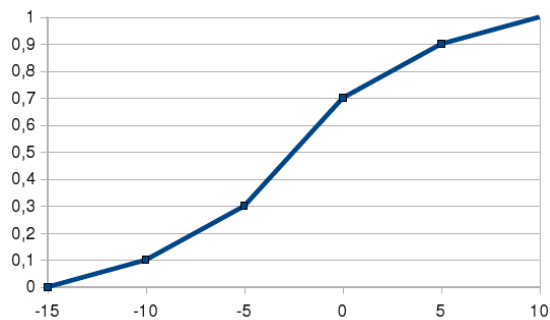
Ejercicio 1.49. ★ Una piscifactoría realiza un estudio para determinar si una especie de pez está en peligro extinción en los ríos de una región. Para ello se midió el número de peces en 12 sitios diferentes obteniendo los siguientes resultados en unidades por hectómetro cúbico:

16 – 11 – 14 – 18 – 21 – 27 – 21 – 15 – 22 – 12 – 2 – 19

Se pide:

1. Dibujar el diagrama de cajas y eliminar de la muestra los datos atípicos si existen.
2. Calcular las medidas de tendencia central. ¿Cuál de ellas es más representativa?
3. Calcular el coeficiente de apuntamiento e interpretarlo.

Ejercicio 1.50. ★ Dado el siguiente polígono correspondiente a una muestra de 60 individuos



Calcular:

1. Tabla de frecuencias
2. Coeficiente de variación. ¿Se puede decir que hay mucha dispersión?
3. Decil 7.
4. Coeficiente de asimetría e interpretarlo.

Ejercicio 1.51. ★ La siguiente tabla muestra los datos de emisiones de CO_2 y CH_4 (en Kg/hab) y el producto interior bruto per cápita (en miles US\$) de varios países en el último año:

País	CO_2	CH_4	PIB
Austria	7.60	0.97	38.40
España	6.73	0.81	30.12
Francia	5.71	0.94	33.19
EEUU	19.40	1.72	45.84
Alemania	9.80	0.83	34.18
Canadá	15.60	3.08	38.43
Italia	7.29	0.58	30.44
Japón	9.44	0.16	33.58

País	CO ₂	CH ₄	PIB
Australia	17.48	6.36	36.26
Reino Unido	8.99	0.76	35.13

País	CO ₂	CH ₄	PIB
Bolivia	1.05	3.44	40.13
Niger	0.1	0.12	0.67
Senegal	0.35	0.76	1.69
Pakistán	0.65	0.59	2.59
Filipinas	0.83	0.46	3.38
Perú	0.94	0.75	7.80
Túnez	2.17	0.48	7.47
Nepal	0.13	0.90	1.21
Nicaragua	0.7	0.32	2.62
Mauritania	0.97	0.85	2.01

Se pide: Agrupar los datos de CH₄ en 5 clases desde el 0 hasta el 2 y calcular:

1. Media, varianza y coeficiente de variación. ¿Existe mucha dispersión? Justificar la respuesta.
2. Calcular el Rango Intercuartílico e interpretarlo.
3. Tipificar los datos y calcular el coeficiente de asimetría e interpretarlo.

Ejercicio 1.52. Se realizó una encuesta a 40 personas de más de 70 años sobre el número de medicamentos distintos que tomaban habitualmente. El resultado de dicha encuesta fue el siguiente:

3 - -1 - -2 - -2 - -0 - -1 - -4 - -2 - -3 - -5 - -1 - -3 - -2 - -3 - -1 - -4 - -2 - -4 - -3 -
3 - -5 - -0 - -1 - -2 - -0 - -2 - -3 - -0 - -1 - -1 - -5 - -3 - -4 - -2 - -3 - -0 - -1 - -2 -

Se pide:

1. Obtener la distribución de frecuencias de la muestra.
2. Dibujar el diagrama de barras y el polígono de frecuencias asociados.
3. Dibujar el diagrama de frecuencias acumuladas.
4. Calcular la media aritmética, la mediana y la moda.
5. Calcular la varianza y la desviación típica.
6. Calcular el coeficiente de variación de Pearson.

 Solución

1. $\bar{x} = 2.225$, $Med = 2$ y $Mod = 2$.
2. $s^2 = 1.974$, $s = 1.405$.
3. $cv = 0.632$.

Ejercicio 1.53. ★ Se considera la variable estadística agrupada en clases cuya distribución de frecuencias viene dada por la siguiente tabla:

X	n_i	f_i	N_i	F_i
[0,20)		0.25		
[20,40)		0.3		
[40,60)				0.9
[60,80)	6			

1. Completar razonadamente la tabla.
2. Hallar el rango intercuartílico e interpretarlo.
3. Hallar el coeficiente de apuntamiento e interpretarlo.

Ejercicio 1.54. ★ Se han medido las estaturas, X , en metros de 15 individuos, obteniéndose los siguientes sumatorios:

$$\sum_{i=1}^{15} x_i = 26.40 \text{ m}, \sum_{i=1}^{15} x_i^2 = 52.45 \text{ m}^2, \sum_{i=1}^{15} (x_i - \bar{x})^3 = -8.42 \text{ m}^3$$

Se pide:

1. Calcular media, desviación típica, coeficiente de variación y coeficiente de asimetría de la variable. Interpretar los dos últimos.
2. Si trabajamos con una nueva variable $Y = 1.1X - 0.2$, ¿cuánto valdrán los estadísticos anteriores? Justificar adecuadamente la respuesta.

 Solución

1. Teniendo en cuenta los sumatorios que nos dan como datos, y el número total de datos en la muestra ($n = 15$), obtenemos:

$$\begin{aligned}
\bar{x} &= \frac{\sum x_i}{n} = \frac{26.40}{15} = 1.76 \text{ m}, \\
s^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{52.45}{15} - 1.76^2 = 0.399 \text{ m}^2, \\
s &= +\sqrt{s^2} = +\sqrt{0.399} = 0.631 \text{ m}, \\
CV &= \frac{s}{|\bar{x}|} = \frac{0.631}{1.76} = 0.359 = 35.9, \\
g_1 &= \frac{\frac{\sum (x_i - \bar{x})^3}{n}}{s^3} = \frac{\frac{-8.42}{15}}{0.631^3} = -2.234.
\end{aligned}$$

En cuanto a la interpretación del coeficiente de variación, su valor es igual a 0.359, lo cual indica que la dispersión relativa en torno a la media es considerable (supera el 30%), por lo que la media no es demasiado representativa de la distribución. Para el coeficiente de asimetría hemos obtenido un valor de -2.234 , negativo lo cual indica que la distribución presenta cola a la izquierda.

2. Si tenemos una nueva variable obtenida mediante la transformación lineal: $Y = 1.1X - 0.2$, le podemos aplicar el teorema que nos dice cuáles serán la nueva media y desviación típica de una variable Y obtenida mediante una transformación lineal aplicada a otra X . Este teorema dice que si $Y = aX + b$, entonces $\bar{y} = a\bar{x} + b$, y $s_y = |a| s_x$. Por lo tanto:

$$\bar{y} = 1.1 \cdot \bar{x} - 0.2 = 1.736, \quad s_y = 1.1 \cdot s_x = 0.694.$$

Con los resultados anteriores, el coeficiente de variación de la nueva variable resulta muy sencillo de calcular:

$$CV_y = \frac{s_y}{|\bar{y}|} = \frac{0.694}{1.736} = 0.400 = 40.0\%$$

Y para el coeficiente de asimetría utilizamos su definición basada en sumatorios:

$$\begin{aligned}
g_{1y} &= \frac{\frac{\sum (y_i - \bar{y})^3}{n}}{s_y^3} = \frac{\frac{\sum (1.1 \cdot x_i - 0.2 - (1.1 \cdot \bar{x} - 0.2))^3}{15}}{(1.1 \cdot s_x)^3} = \\
&= \frac{\frac{\sum (1.1 \cdot x_i - 1.1 \cdot \bar{x})^3}{15}}{(1.1 \cdot s_x)^3} = \frac{1.1^3 \sum (x_i - \bar{x})^3}{1.1^3 (s_x)^3} = \frac{\sum (x_i - \bar{x})^3}{(s_x)^3} = g_{1x} = -2.234.
\end{aligned}$$

que como se puede comprobar no cambia.

Ejercicio 1.55. Dada la siguiente muestra del número de ingresos en urgencias en un determinado hospital,

5 – 7 – 11 – 7 – 4 – 7 – 24 – 7 – 8 – 8 – 9 – 8 – 13 – 7 – 8 – 17 – 7 – 6

estudiar la asimetría de la muestra y transformar los datos para conseguir una simetría más normal.

 Solución

$\bar{x} = 9.055$ ingresos, $s^2 = 21.4969$ ingresos², $s = 4.6365$ ingresos y $g_1 = 2.02$. Para corregir asimetría tomamos $Y = \ln(X)$ y resulta $\bar{y} = 2.108$, $s_y^2 = 0.1668$, $s_y = 0.4084$ y $g_1 = 0.95$.

Ejercicio 1.56. En un experimento se ha medido el tiempo de respuesta a un determinado estímulo visual en grupo de 16 individuos, obteniendo los siguientes resultados (en centésimas de segundo):

73, 42, 67, 55, 62, 52, 59, 82, 46, 61, 72, 51, 64, 55, 61.

Comprobar si hay datos atípicos en la muestra. En tal caso, sustituirlos por máximos o mínimos valores normales y hacer un resumen descriptivo de la muestra.

 Solución

Los cuartiles son $C_1 = 52$ cs, $C_3 = 67$ cs y $RI = 15$ cs. Las vallas son $v_1 = 29.5$ y $v_2 = 89.5$, luego no hay datos atípicos. $\bar{x} = 60.13$ cs, $s^2 = 105.58$ cs², $s = 10.28$ cs, $cv = 0.17$, $g_1 = 0.26$ y $g_2 = -0.37$.

Ejercicio 1.57. En un grupo de operarios se ha medido porcentaje de tareas bien realizadas de dos tipos A y B , obteniendo los siguientes resultados:

% A	58	55	36	67	60	31	53	42
% B	76	81	82	84	80	75	87	79

Se pide:

1. Calcular las puntuaciones típicas de cada individuo en ambos tipos de tareas.
2. Teniendo en cuenta la distribuciones de puntuaciones de las tareas, ¿en qué tipo de tareas funciona mejor el primer individuo? ¿Y el último? Justificar la respuesta.

 Solución

$\bar{x}_A = 50.25\%$, $\bar{x}_B = 80.5\%$, $s_A^2 = 138.44\%^2$, $s_B^2 = 13.75\%^2$, $s_A = 11.7\%$ y $s_B = 3.71\%$. Las puntuaciones típicas son

% A	0.66	0.4	-1.21	1.42	0.83	-1.64	0.23	-0.7
% B	-1.21	0.13	0.4	0.94	-0.13	-1.48	1.75	-0.4

Ejercicio 1.58. En un cuestionario se ha preguntado a un grupo de individuos por su nivel de estudios (SE: sin estudios, EB: estudios básicos, ES: estudios secundarios, EU: estudios universitarios, ED: estudios de doctorado) obteniendo los siguientes resultados:

EU, ES, ES, EU, EB, EB, ED, SE, EB, SE, EU, ES, ES, ES, EB, EB, SE, ED, EB, ES, ES, SE, EU, EU, EB, ES, EU, ES, EB, ES

Se pide:

1. Construir la tabla de frecuencias y los diagramas asociados.
2. Dar una medida de representatividad.
3. Calcular los cuartiles y el percentil 90.

 Solución

1. $Me = ES$ y $Mo = ES$.
2. $C_1 = EB$, $C_2 = ES$, $C_3 = EU$ y $P_{90} = EU$.

Ejercicio 1.59. ★ El tiempo de recuperación, medido en días, de 10 pacientes con una determinada lesión de rodilla ha sido

46, 54, 48, 62, 51, 50, 54, 50, 49, 52.

Calcular el coeficiente de asimetría y de apuntamiento e interpretarlos.

 Solución

Llamemos X a la variable que mide el tiempo de recuperación de la lesión de rodilla. Para calcular el coeficiente de asimetría y de apuntamiento necesitamos calcular primero la media y la desviación típica de la variable. Puesto que son pocos datos ($n = 10$), trabajaremos con los datos muestrales sin agrupar.

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{46 + 54 + \dots + 52}{10} = \frac{516}{10} = 51.6, \\ s_x^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{46^2 + 54^2 + \dots + 52^2}{10} - 51.6^2 = \frac{26802}{10} - 2662.56 = 17.64, \\ s_x &= \sqrt{17.64} = 4.2\end{aligned}$$

El coeficiente de asimetría es

$$g_1 = \frac{\sum (x_i - \bar{x})^3 / n}{s^3} = \frac{((46 - 51.6)^3 + \dots + (52 - 51.6)^3) / 10}{4.2^3} = \frac{904.32 / 10}{74.088} = 1.22,$$

lo que indica que la distribución es bastante asimétrica a la derecha.

El coeficiente de apuntamiento es

$$g_2 = \frac{\sum (x_i - \bar{x})^4 / n}{s^4} - 3 = \frac{((46 - 51.6)^4 + \dots + (52 - 51.6)^4) / 10}{4.2^4} - 3 = \frac{12975.312 / 10}{311.1696} - 3 = 1.17,$$

lo que indica que la distribución es bastante más apuntada de lo normal, es decir, es leptocúrtica.

Ejercicio 1.60. ★ La ingesta calórica durante la comida para un individuo al que se ha hecho un seguimiento estricto durante 15 días, expresada en kilocalorías, ha sido:

1121, 1138, 1101, 1025, 1323, 1450, 1005, 1214, 1040, 1321
1225, 1428, 1105, 1202, 1483, 1465, 1362, 1325, 1010, 1311

Agrupar los datos en 5 clases de amplitud 100, comenzando en 1000 kilocalorías, y utilizando dicha agrupación calcular:

1. Coeficiente de variación.
2. Rango intercuartílico.
3. Coeficiente de asimetría.

Interpretar los resultados.

Solución

Llamemos X al número de kilocalorías ingeridas al día.

1. Teniendo en cuenta la agrupación propuesta y lo que se nos pide en los apartados siguientes, la tabla que necesitamos es:

Intervalos	x_i	n_i	N_i	$x_i n_i$	$x_i^2 n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
[1000, 1100)	1050	4	4	4200	4410000	-205	-34460500
[1100, 1200)	1150	4	8	4600	5290000	-105	-4630500
[1200, 1300)	1250	3	11	3750	4687500	-5	-375
[1300, 1400)	1350	5	16	6750	9112500	95	4286875
[1400, 1500)	1450	4	20	5800	8410000	195	29659500
Sumas		20		25100	31910000		-5145000

De todo ello:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i n_i}{\sum n_i} = \frac{25100}{20} = 1255 \text{ kcal}, \\ s_x^2 &= \frac{\sum x_i^2 n_i}{\sum n_i} - \bar{x}^2 = \frac{31910000}{20} - 1255^2 = 20475 \text{ kcal}^2, \\ s_x &= +\sqrt{s_x^2} = 143.091 \text{ kcal},\end{aligned}$$

y el coeficiente de variación:

$$cv_x = \frac{s_x}{|\bar{x}|} = 0.114$$

Para su interpretación, teniendo en cuenta que 0.114 está cercano a 0, la desviación típica será muy pequeña en comparación con la media, lo cual implica que la media muestral será bastante representativa.

2. El rango intercuartílico es la diferencia entre el tercer cuartil y el primero, así que se hay que calcular el primer y el tercer cuartil. Para el primero de los cuartiles, teniendo en cuenta que le corresponde una frecuencia acumulada $n/4$, es decir 5, luego se encuentra el intervalo $[1100, 1200)$, y aplicando semejanza de triángulos:

$$\frac{8 - 4}{1200 - 1100} = \frac{5 - 4}{C_1 - 1100} \Leftrightarrow C_1 = 1125 \text{ kcal}.$$

De la misma forma pero procediendo con una frecuencia acumulada $3n/4 = 15$ para C_3 , lo cual indica que se encuentra en el intervalo $[1300, 1400)$:

$$\frac{16 - 11}{1400 - 1300} = \frac{15 - 11}{C_3 - 1300} \Leftrightarrow C_3 = 1380 \text{ kcal}.$$

Por lo tanto el rango intercuartílico vale $RI = C_3 - C_1 = 1380 - 1125 = 255 \text{ kcal}$,] cuya interpretación es que entre 1125 kcal y 1380 kcal se encuentran el 50% de los individuos de la muestra que comen una cantidad media de kilocalorías (lejos de los excesos'' del 25% que más kilocalorías toman, o de los defectos'' del 25% que menos kilocalorías ingieren). Se puede observar, así mismo, que estos datos centrales no están muy dispersos.

3. El coeficiente de asimetría es

$$g_1 = \frac{\frac{\sum (x_i - \bar{x})^3 n_i}{N}}{s_x^3} = \frac{-5145000}{143.091^3} = -0.0878,$$

cuya interpretación es que la distribución es casi simétrica, con una ligera asimetría a la izquierda.

Ejercicio 1.61. ★ En un laboratorio se ha medido el número de hijos que tuvieron unas ratas que habían seguido un tratamiento de fertilidad, obteniendo los siguientes resultados:

7 – 5 – 6 – 6 – 8 – 9 – 10 – 8 – 7 – 6 – 8 – 7 – 9 – 11 – 8 – 7 – 7 – 7 – 6 – 8

1. Existen datos atípicos en la muestra? Justificar la respuesta.
Nota: En caso de detectarse algún dato atípico, no quitarlo.
2. Calcular los coeficientes de asimetría y apuntamiento e interpretarlos. Se puede concluir que la muestra proviene de una población normal?

 Solución

1. El 11 es un dato atípico.
2. $g_1 = 0.61$ lo que indica que la distribución es asimétrica hacia la derecha, y $g_2 = 0.084$, lo que indica que la distribución es un poco platicúrtica. Como ambos valores están en el intervalo $[-2, 2]$ se puede suponer que la muestra viene de una población normal.

Ejercicio 1.62. ★ En un estudio sobre el número de kilocalorías diarias en la dieta de los españoles mayores de edad se han obtenido los siguientes datos, en una muestra de tamaño 200, separada por sexos:

Kilocaloras	Hombres	Mujeres
[1000, 1400)	1	15
[1400, 1800)	10	24
[1800, 2200)	25	28
[2200, 2600)	34	15
[2600, 3000)	26	10
[3000, 3400)	12	0

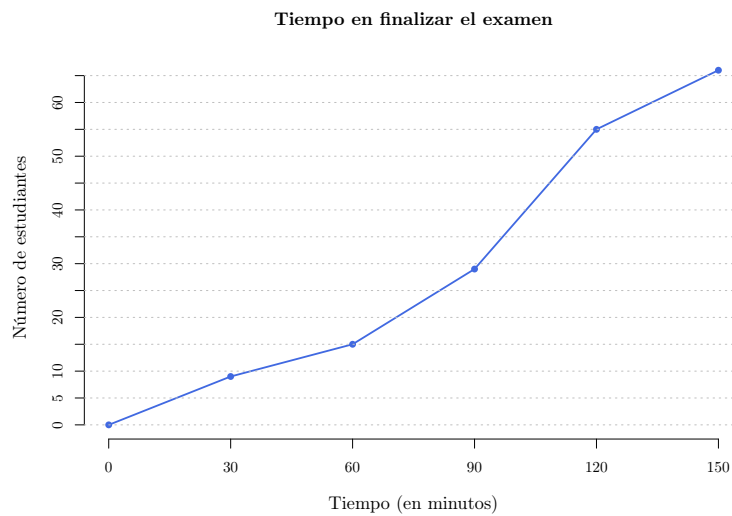
1. Qué media de consumo de kilocaloras es más representativa de su distribución, la de hombres o la de mujeres? Justificar adecuadamente la respuesta.
2. Cuánto vale el percentil 90 del consumo de kilocaloras en los hombres?
3. Cuánto vale el coeficiente de apuntamiento de la distribución global, considerando hombres y mujeres? Interpretarlo.

Solución

Llamando h y m al número de kilocalorías en hombres y mujeres respectivamente:

1. $\bar{h} = 2407.407$ kcal, $s_h = 468.1927$ kcal y $cv_h = 0.1945$.
 $\bar{m} = 1917.391$ kcal $s_m = 484.6802$ kcal y $cv_m = 0.2528$.
2. El percentil 90 en los hombres vale 3040 kcal.
3. $g_2 = -0.73$, lo que indica que la distribución es platicúrtica.

Ejercicio 1.63. ★ En un examen de estadística al que se han presentado 66 alumnos se ha contado el número de exámenes finalizados cada media hora, obteniendo el siguiente polígono:



Se pide:

1. ¿Cuánto tiempo tiene que pasar para que hayan finalizado el examen la mitad de los alumnos? ¿Y el 90% de los estudiantes?
2. ¿Qué porcentaje de alumnos habrá terminado a los 100 minutos de examen?
3. ¿Cuál es el tiempo de duración del examen que mejor representa a los estudiantes de la muestra? ¿Es un valor representativo?

Solución

1. El tiempo que transcurra hasta que hayan finalizado el examen la mitad de los alumnos es precisamente la mediana. La frecuencia absoluta acumulada correspondiente a la mediana es $66/2 = 33$ de modo que la mediana cae en

la clase $[90, 120)$, e interpolando en esta clase se obtiene

$$Me = 90 + \frac{33 - 29}{26}30 = 94.6154 \text{ min.}$$

2. Para ver si la muestra tienen mucha dispersión hay que calcular la desviación típica o el coeficiente de variación, pero antes necesitamos conocer las frecuencias absolutas de cada clase. En la tabla que nos dan aparece el número de exámenes finalizados para cada instante, que es, en realidad, la frecuencia absoluta acumulada. A partir de ella se puede obtener fácilmente la frecuencia absoluta de cada clase simplemente restando a la frecuencia absoluta acumulada de la clase siguiente la de la clase actual. Así se obtiene la siguiente tabla de frecuencias.

	x_i	n_i	N_i
$(0,30]$	15	9	9
$(30,60]$	45	6	15
$(60,90]$	75	14	29
$(90,120]$	105	26	55
$(120,150]$	135	11	66

Para calcular la desviación típica se añaden nuevas columnas a la tabla de frecuencias con los cálculos necesarios:

	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
(0,30]	15	9	135	2025
(30,60]	45	6	270	12150
(60,90]	75	14	1050	78750
(90,120]	105	26	2730	286650
(120,150]	135	11	1485	200475
Σ		66	5670	580050

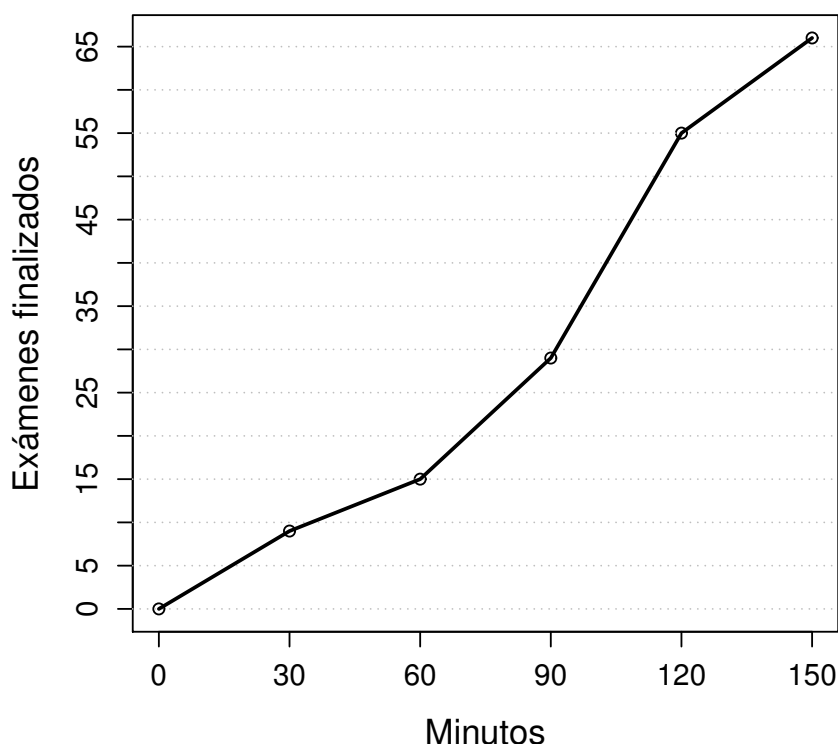
A partir de la tabla calculamos los estadísticos que se piden:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i n_i}{n_x} = \frac{5670}{66} = 85.9091, \\ s^2 &= \frac{\sum x_i^2 n_i}{n_x} - \bar{x}^2 = \frac{580050}{66} - 7380.3735 = 1408.2645, \\ s &= \sqrt{1408.2645} = 37.5268, \\ cv &= \frac{s}{|\bar{x}|} = \frac{37.5268}{|85.9091|} = 0.4368.\end{aligned}$$

Lo que indica que hay una dispersión moderada.

Ejercicio 1.64. ★ En un examen de estadística al que se han presentado 66 alumnos se ha contado el número de exámenes finalizados cada media hora, obteniendo el siguiente polígono:

Histograma de exámenes finalizados



Se pide:

1. ¿Existe mucha dispersión en la muestra? Justificar la respuesta.
2. ¿Cuánto tiempo tiene que pasar para que hayan finalizado el examen la mitad de los alumnos?
3. Estudiar la asimetría de la distribución.

💡 Solución

1. Para ver si la muestra tienen mucha dispersión hay que calcular la desviación típica o el coeficiente de variación, pero antes necesitamos conocer las frecuencias absolutas de cada clase. En la tabla que nos dan aparece el número de exámenes finalizados para cada instante, que es, en realidad, la frecuencia absoluta acumulada. A partir de ella se puede obtener fácilmente la frecuencia absoluta de cada clase simplemente restando a la frecuencia absoluta acumulada de la clase siguiente la de la clase actual. Así se obtiene la siguiente tabla de frecuencias.

	x_i	n_i	N_i
(0,30]	15	9	9
(30,60]	45	6	15
(60,90]	75	14	29
(90,120]	105	26	55
(120,150]	135	11	66

Para calcular la desviación típica se añaden nuevas columnas a la tabla de frecuencias con los cálculos necesarios:

	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
(0,30]	15	9	135	2025
(30,60]	45	6	270	12150
(60,90]	75	14	1050	78750
(90,120]	105	26	2730	286650
(120,150]	135	11	1485	200475
Σ		66	5670	580050

A partir de la tabla calculamos los estadísticos que se piden:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i n_i}{n_x} = \frac{5670}{66} = 85.9091, \\ s^2 &= \frac{\sum x_i^2 n_i}{n_x} - \bar{x}^2 = \frac{580050}{66} - 7380.3735 = 1408.2645, \\ s &= \sqrt{1408.2645} = 37.5268, \\ cv &= \frac{s}{|\bar{x}|} = \frac{37.5268}{|85.9091|} = 0.4368.\end{aligned}$$

Lo que indica que hay una dispersión moderada.

- El tiempo que transcurra hasta que hayan finalizado el examen la mitad de los alumnos es precisamente la mediana. La frecuencia absoluta acumulada correspondiente a la mediana es $66/2 = 33$ de modo que la mediana cae en la clase $[90, 120)$, e interpolando en esta clase se obtiene

$$Me = 90 + \frac{33 - 29}{26} 30 = 94.6154 \text{ min.}$$

- Para estudiar la asimetría de la muestra se calcula el coeficiente de asimetría. Para ello se añade una nueva columna a la tabla de frecuencias con las desviaciones a la media al cubo:

	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
(0,30]	15	9	-70.9091	-3208841.4726
(30,60]	45	6	-40.9091	-410781.3674
(60,90]	75	14	-10.9091	-18175.8077
(90,120]	105	26	19.0909	180906.0856
(120,150]	135	11	49.0909	1301355.3719
Σ		66		-2155537.1901

A partir de la tabla calculamos los estadísticos que se piden:

$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n_x}{s_x^3} = \frac{-2155537.1901/66}{37.5268^3} = -0.618.$$

Lo que indica que la distribución es asimétrica hacia la izquierda.

Ejercicio 1.65. ★ Clasificar las siguientes variables:

- Horas diarias de ejercicio.
- Nacionalidad.
- Tensión arterial.

4. Gravedad de una enfermedad.
5. Número de lesiones en una temporada.
6. Ingesta diaria de calorías.
7. Talla de ropa.
8. Asignaturas aprobadas en un curso.

Solución

1. Variable cuantitativa continua.
2. Atributo nominal.
3. Variable cuantitativa continua.
4. Atributo ordinal.
5. Variable cuantitativa discreta.
6. Variable cuantitativa continua.
7. Atributo ordinal.
8. Variable cuantitativa discreta.

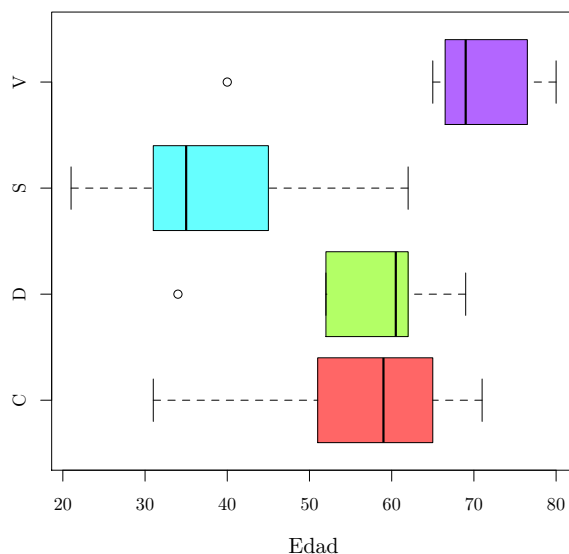
Ejercicio 1.66. ★ La siguiente tabla representa la distribución de frecuencias del número de usos del seguro de salud en un año de una muestra de clientes de una compañía de seguros.

Usos:	0	1	2	3	4	5	7
Clientes:	4	8	6	3	2	1	1

Dibujar el diagrama de caja. ¿Cómo es la simetría de la distribución?

Ejercicio 1.67. ★ El diagrama de caja siguiente corresponde a la distribución de edades según el estado civil de una muestra de personas.

Diagrama de caja de edades según estado civil



1. ¿Qué grupo tiene edades mayores?
2. ¿Qué grupo tiene una menor dispersión central en las edades?
3. ¿Qué grupos tienen datos atípicos?
4. ¿Qué grupo tienes una distribución de edad más asimétrica?

💡 Solución

1. Viudos.
2. Divorciados.
3. Viudos y divorciados.
4. Divorciados.

Ejercicio 1.68. ★ La presión arterial sistólica (en mmHg) de una muestra de personas es

135 128 137 110 154 142 121 127 114 103

1. Calcular los estadísticos de tendencia central.
2. ¿Cómo es la dispersión relativa con respecto a la media?
3. ¿Cómo es la asimetría de la distribución muestral?
4. ¿Cómo es la curtosis de la distribución muestral?
5. Si se sabe que el método utilizado para medir la presión está sesgado, y, para obtener los valores correctos, hay que aplicar la transformación $y = 1.2x - 5$,

¿cuáles serán los estadísticos anteriores para los valores correctos de la presión arterial sistólica?

Usar las siguientes sumas: $\sum x_i = 1271$ mmHg, $\sum (x_i - \bar{x})^2 = 2188.9$ mmHg², $\sum (x_i - \bar{x})^3 = 2764.32$ mmHg³, $\sum (x_i - \bar{x})^4 = 1040080$ mmHg⁴.

Solución

Llamando x a la presión arterial sistólica.

1. $\bar{x} = 127.1$ mmHg, $Me = 127.5$ mmHg, Mo todos los valores.
2. $s = 14.7949$ mmHg y $cv = 0.1164$.
3. $g_1 = 0.0854$ luego la distribución es casi simétrica.
4. $g_2 = -0.8292$ luego la distribución es menos apuntada de lo normal (platicúrtica).
5. $\bar{y} = 147.52$ mmHg, $Me = 148$ mmHg, $Mo = 157$ mmHg, $s = 17.7539$ mmHg, $cv = 0.1203$, $g_1 = 0.0854$ y $g_2 = -0.8292$.

Ejercicio 1.69. ★ La siguiente tabla contiene las frecuencias de embarazos, abortos y nacimientos de una muestra de 999 mujeres de una ciudad.

Num	Embarazos	Abortos	Nacimientos
0	61	751	67
1	64	183	80
2	328	51	400
3	301	10	300
4	122	2	90
5	81	2	62
6	29	0	0
7	11	0	0
8	2	0	0

1. ¿Cuántos datos atípicos hay en el número de nacimientos de la muestra?
2. ¿Qué variable tiene menor dispersión con respecto a la media?
3. ¿Qué valor es relativamente mayor 7 embarazos o 4 abortos? Justificar la respuesta.

Usar las siguientes sumas:

Embarazos: $\sum x_i = 2783$, $\sum x_i^2 = 9773$.

Abortos: $\sum x_i = 333$, $\sum x_i^2 = 559$.

Nacimientos: $\sum x_i = 2450$, $\sum x_i^2 = 7370$.

Solución

Llamando x al número de embarazos, y al número de abortos y z al número de nacimientos.

1. 129 datos atípicos.
2. Embarazos: $\bar{x} = 2.7858$, $s_x = 1.422$ y $cv_x = 0.5105$.
Abortos: $\bar{y} = 0.3333$, $s_y = 0.6697$ y $cv_y = 2.009$.
Nacimientos: $\bar{z} = 2.4525$, $s_z = 1.1674$ y $cv_z = 0.476$.
3. La puntuación típica de 7 embarazos es 2.9635, y la puntuación típica de 4 abortos es 5.4754, luego es relativamente mayor el número de abortos.

Ejercicio 1.70. ★ Se ha modificado un gen de una especie de rata para facilitar la metabolización del colesterol en sangre. Para comprobar la efectividad de la modificación genética se tomaron dos grupos de 20 ratas, unas con el gen modificado y las otras no, y se sometieron a una misma dieta con distintas concentraciones de aceite de palma durante un mes, obteniendo los siguientes resultados:

Cantidad de aceite de palma en gr (igual en las dos muestras)

$$\sum x_i = 640.6467, \sum x_i^2 = 23508.6387, \sum (x_i - \bar{x})^3 = -5527.08, \sum (x_i - \bar{x})^4 = 792910$$

Colesterol en sangre en mg/dl de ratas no modificadas genéticamente

$$\sum y_j = 2945.8545, \sum y_j^2 = 439517.5975, \sum (y_j - \bar{y})^3 = 604.08, \sum (y_j - \bar{y})^4 = 3717331.07$$

Colesterol en sangre en mg/dl de ratas modificadas genéticamente

$$\sum y_j = 2126.5899, \sum y_j^2 = 226824.5373, \sum (y_j - \bar{y})^3 = -629.4, \sum (y_j - \bar{y})^4 = 48248.29$$

Se pide:

1. ¿En qué muestra es más representativa la media del colesterol, en las ratas modificadas genéticamente o en las no modificadas?
2. ¿En qué muestra es más asimétrica la distribución del colesterol?
3. ¿En qué muestra es menos normal el apuntamiento del colesterol?
4. ¿Qué nivel de colesterol es relativamente mayor, el de una rata modificada genéticamente con nivel 130 mg/dl, o el de una rata no modificada con nivel 145 mg/dl?

Solución

1. Ratas no modificadas genéticamente: $\bar{y} = 147.2927$ mg/dl, $s_y^2 = 280.7332$ (mg/dl)², $s = 16.7551$ mg/dl y $cv_y = 0.1138$.
Ratas modificadas genéticamente: $\bar{y} = 106.3295$ mg/dl, $s_y^2 = 35.265$ (mg/dl)², $s = 5.9384$ mg/dl y $cv_y = 0.0558$.
Así, la media de las ratas modificadas genéticamente es más representativa ya que su coeficiente de variación es menor.
2. Ratas no modificadas genéticamente: $g_1 = 0.0064$.
Ratas modificadas genéticamente: $g_1 = 0.1503$.

Por tanto, la distribución de ratas modificadas genéticamente es más asimétrica ya que el coeficiente de asimetría está más alejado de 0.

3. Ratas no modificadas genéticamente: $g_2 = -0.6416$.

Ratas modificadas genéticamente: $g_2 = 1.0602$.

Por tanto, el apuntamiento de la distribución de ratas modificadas genéticamente es menos normal ya que el coeficiente de apuntamiento está más alejado de 0.

4. Ratas no modificadas genéticamente: $z(145) = -0.1368$.

Ratas modificadas genéticamente: $z(130) = 3.986$.

Por tanto, un nivel de colesterol de 130 mg/dl en ratas modificadas genéticamente es relativamente mayor que 145 mg/dl en ratas no modificadas genéticamente.

2 Regresión y correlación

Ejercicio 2.1. Dada la siguiente tabla de correlación:

$X \setminus Y$	1	2	3
$[-2, 2)$	3	6	1
$[2, 6)$	4	7	3
$[6, 10)$	5	3	0

Determinar:

1. Las distribuciones marginales. Media, Moda y Mediana.
2. Rectas de Regresión.
3. Coeficiente de correlación lineal. Interpretar el resultado.

Ejercicio 2.2. Una compañía de asistencia sanitaria hace un estudio del número de veces que, durante el último trimestre, han acudido sus asegurados a consultas de especialistas, en función de su edad. En la siguiente tabla se reflejan los resultados obtenidos:

Edad \ Cons.	0	1	2	3	4
$[30, 40)$	6	2	2	0	0
$[40, 50)$	4	3	6	4	1
$[50, 60)$	0	2	4	5	3
$[60, 70)$	0	0	3	4	5
$[70, 80)$	0	0	0	4	6

Se pide:

1. Recta de regresión del número de consultas sobre la edad.
2. Coeficiente de correlación e interpretarlo.
3. ¿Cuántas consultas se espera que realice una persona de 52 años? ¿Es fiable esta predicción?

Ejercicio 2.3. Se determina la pérdida de actividad que experimenta un medicamento desde el momento de su fabricación a lo largo del tiempo, obteniéndose el siguiente resultado:

Tiempo (en años)	1	2	3	4	5
Actividad restante (%)	96	84	70	58	52

Se pide:

1. Calcular la recta de regresión de la actividad sobre el tiempo transcurrido.
2. Según el modelo lineal, ¿cuánto tiempo debe pasar para que la actividad del fármaco sea del 80%? ¿Cuándo será nula la actividad?

Solución

Llamando T al tiempo y A a la actividad del fármaco:

1. $\bar{t} = 3$ años, $\bar{a} = 72\%$, $s_t^2 = 2$ años², $s_a^2 = 264\%^2$, $s_{ta} = -22.8$ años·%.
Recta de regresión de actividad sobre tiempo: $a = -11.4t + 106.2$.
2. Recta de regresión de tiempo sobre actividad: $t = -0.086a + 9.2182$.
 $t(80) = 2.3091$ años y $t(0) = 9.2182$ años.

Ejercicio 2.4. Las temperaturas medias mensuales (en °C) y las precipitaciones totales mensuales (en mm) durante el año 2001 en Madrid fueron:

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Temp.	7.2	8.4	12.2	13.7	16.7	23.3	24.2	25.5	20.4	16.2	8.1	4.2
Prec.	73.6	31.7	72.1	20.7	37.1	3.8	3.3	1.5	23.1	67.0	12.4	18.0

¿Existe relación lineal entre las precipitaciones y la temperatura? De acuerdo a esta relación, ¿qué cantidad de precipitaciones se espera que haya un mes con una temperatura media de 15°C? ¿Es fiable esta predicción?

Ejercicio 2.5. Se ha realizado un estudio comparativo de las puntuaciones obtenidas por los alumnos en un test de ingreso en la universidad (X), y el número de asignaturas aprobadas en el primer curso (Y). Los resultados obtenidos se expresan en la siguiente tabla:

$X \setminus Y$	0	1	2	3	4
$[0, 10)$	2	2	1	0	0
$[10, 20)$	1	1	2	2	0
$[20, 30)$	0	1	3	4	1
$[30, 40)$	0	0	2	2	6

Se desea calcular:

1. Recta de regresión de X sobre Y .
2. Coeficiente de correlación e interpretación del mismo.
3. Si la universidad en cuestión sólo contara con alumnos que al menos logren aprobar dos asignaturas, ¿qué número de preguntas respondidas correctamente exigirá en el test?

Solución

1. $\bar{x} = 23$ puntos, $\bar{y} = 2.4$ asignaturas, $s_x^2 = 116$ puntos², $s_y^2 = 1.5733$ asignaturas², $s_x = 10.7703$ puntos, $s_y = 1.2453$ asignaturas y $s_{xy} = 9.8$ puntos·asignaturas.
Recta de regresión de X sobre Y : $x = 6.2288y + 8.0508$.
2. $r = 0.73$, lo que quiere decir que hay buena relación lineal entre las puntuaciones y las asignaturas aprobadas y además es creciente (a mayor puntuación en el test, más asignaturas aprobadas).

Ejercicio 2.6. ★ La tabla siguiente representa la distribución bidimensional de frecuencias de una muestra de 80 personas en un estudio sobre la relación entre el nivel de colesterol en sangre (X) en mg/dl y la tensión arterial máxima (Y) en mmHg.

$X \setminus Y$	[110,130)	[130,150)	[150,170)	n_x
[170,190)		4		12
[190,210)	10	12	4	
[210,230)	7		8	
[230,250)	1			18
n_y		30	24	

Se pide:

1. Completar la tabla.
2. Calcular la recta de regresión del nivel de colesterol sobre la tensión.
3. Usar el modelo lineal para predecir el colesterol esperado para una persona con una tensión arterial de 160 mmHg.
4. Según el modelo lineal, ¿cuál es la tensión arterial máxima esperada para una persona cuyo nivel de colesterol es 270 mg/dl.

Usar las siguientes sumas: $\sum x_i n_i = 16960$ mg/dl, $\sum y_j n_j = 11160$ mmHg, $\sum x_i^2 n_i = 3627200$ (mg/dl)², $\sum y_j^2 n_j = 1576800$ mmHg² y $\sum x_i y_j n_{ij} = 2378800$ mg/dl·mmHg.

💡 Solución

1. Tabla de frecuencias

$X \setminus Y$	[110,130)	[130,150)	[150,170)	n_x
[170,190)	8	4	0	12
[190,210)	10	12	4	26
[210,230)	7	9	8	24
[230,250)	1	5	12	18
n_y	26	30	24	80

2. $\bar{x} = 212$ mg/dl, $\bar{y} = 139.5$ mmHg, $s_x^2 = 396$ (mg/dl)², $s_y^2 = 249.75$ mmHg² y $s_{xy} = 161$ mg/dl·mmHg. Recta de regresión del nivel de colesterol sobre la tensión arterial: $x = 122.0721 + 0.6446y$.

3. $x(160) = 225.2152$ mg/dl.

4. Recta de regresión de la tensión arterial sobre el colesterol: $y = 0.4066x + 53.3081$.

$y(270) = 163.0808$ mmHg.

Ejercicio 2.7. ★ En un centro dietético se está probando una nueva dieta de adelgazamiento en una muestra de 12 individuos. Para cada uno de ellos se ha medido el número de días que lleva con la dieta y el número de kilos perdidos desde entonces, obteniéndose los siguientes resultados:

(33, 3.9), (51, 5.9), (30, 3.2), (55, 6.0), (38, 4.9), (62, 6.2),
(35, 4.5), (60, 6.1), (44, 5.6), (69, 6.2), (47, 5.8), (40, 5.3)

Se pide:

1. Dibujar el diagrama de dispersión. Según la nube de puntos, ¿qué tipo de modelo explicaría mejor la relación entre los días de dieta y los kilos perdidos?
2. Calcular el modelo lineal y el logarítmico de los kilos perdidos con respecto a los días de dieta.
3. Utilizar el mejor de los modelos anteriores para predecir en número de kilos perdidos tras 40 días de dieta y tras 100 días. ¿Son fiables estas predicciones?

Usar las siguientes sumas (X =Días de dieta e Y =Peso perdido): $\sum x_i = 564$ días, $\sum \log(x_i) = 45.8086$ log(días), $\sum y_j = 63.6$ kg, $\sum x_i^2 = 28234$ días², $\sum \log(x_i)^2 = 175.6603$ log(días)², $\sum y_j^2 = 347.7$ kg², $\sum x_i y_j = 3108.5$ días·kg, $\sum \log(x_i) y_j = 245.4738$ log(días)·kg.

 Solución

Llamando X a los días de dieta, Y a los kg perdidos y $Z = \log X$.

1. $\bar{x} = 47$ días, $\bar{y} = 5.3$ kg, $s_x^2 = 143.833$ días², $s_y^2 = 0.885$ kg², $s_{xy} = 9.942$ días·kg. Modelo lineal: $y = 0.069x + 2.051$.
 $\bar{z} = 3.82 \log(\text{días})$, $s_z^2 = 0.07 \log^2(\text{días})$, $s_{yz} = 0.22 \log(\text{días}) \cdot \text{kg}$.
Modelo logarítmico: $y = 3.4 \log y - 7.67$.
2. Modelo lineal: $r^2 = 0.78$, modelo logarítmico: $r^2 = 0.86$.
Predicciones con el modelo logarítmico: $y(40) = 4.86$ kg y $y(100) = 7.98$ kg.
Las predicciones son fiables ya que el coeficiente de determinación es alto, aunque la de 100 días no lo es tanto por estar fuera del rango de valores observados en la muestra.

Ejercicio 2.8. ★ Al realizar un estudio sobre la dosificación de un cierto medicamento, se trataron 6 pacientes con dosis diarias de 2 mg, 7 pacientes con 3 mg y otros 7 pacientes con 4 mg. De los pacientes tratados con 2 mg, 2 curaron al cabo de 5 días, y 4 al cabo de 6 días. De los pacientes tratados con 3 mg diarios, 2 curaron al cabo de 3 días, 4 al cabo de 5 días y 1 al cabo de 6 días. Y de los pacientes tratados con 4 mg diarios, 5 curaron al cabo de 3 días y 2 al cabo de 5 días.

Se pide:

1. Construir la tabla de la distribución conjunta de frecuencias.
2. Obtener las distribuciones de frecuencias marginales y calcular los principales estadísticos para cada variable.
3. Calcular la covarianza e interpretarla.

 Solución

Llamando X a la dosis e Y al tiempo de curación:

1. $\bar{x} = 3.05$ mg, $\bar{y} = 4.55$ días, $s_x^2 = 0.648$ mg², $s_y^2 = 1.448$ días², $s_x = 0.805$ mg, $s_y = 1.203$ días y $s_{xy} = -0.678$ mg·días, lo que indica que hay una relación lineal decreciente.

Ejercicio 2.9. ★ Al realizar un estudio sobre la dosificación de un cierto medicamento, se trataron 6 pacientes con dosis diarias de 2 mg, 7 pacientes con 3 mg y otros 7 pacientes con 4 mg. De los pacientes tratados con 2 mg, 2 curaron al cabo de 5 días, y 4 al cabo de 6 días. De los pacientes tratados con 3 mg diarios, 2 curaron al cabo de 3 días, 4 al cabo de 5 días y 1 al cabo de 6 días. Y de los pacientes tratados con 4 mg diarios, 5 curaron al cabo de 3 días y 2 al cabo de 5 días.

Se pide:

1. Dar el coeficiente de correlación e interpretación.
2. Determinar el tiempo esperado de curación para una dosis de 5 mg diarios.

Solución

Llamando X a la dosis e Y al tiempo de curación:

1. $\bar{x} = 3.05$ mg, $\bar{y} = 4.55$ días, $s_x^2 = 0.648$ mg², $s_y^2 = 1.448$ días², $s_x = 0.805$ mg, $s_y = 1.203$ días y $s_{xy} = -0.678$ mg·días.
 $r = -0.7$, que quiere decir que hay buena relación lineal entre la dosis y el tiempo de curación, y además es decreciente (a mayor dosis, menor tiempo de curación).
2. Recta de regresión del tiempo de curación sobre la dosis: $y = -1.046x + 7.741$.
 $y(5) = 2.511$ días.

Ejercicio 2.10. ★ Después de tomar un litro de vino se ha medido la concentración de alcohol en la sangre en distintos instantes, obteniendo:

Tiempo después (minutos)	30	60	90	120	150	180
Concentración (gramos/litro)	1.6	1.7	1.5	1.1	0.7	0.2

Se pide:

1. Calcular la recta de regresión de la concentración en función del tiempo.
2. ¿Qué concentración de alcohol habrá a los 100 minutos?
3. Si la concentración máxima de alcohol en la sangre que permite la ley para poder conducir es 0.8 g/l, ¿cuánto tiempo habrá que esperar después de tomarse un litro de vino para poder conducir sin infringir la ley?

Ejercicio 2.11. Se consideran dos variables aleatorias X e Y tales que:

- La recta de regresión de Y sobre X viene dada por la ecuación: $y - x - 2 = 0$.
- La recta de regresión de X sobre Y viene dada por la ecuación: $y - 4x + 22 = 0$.

Calcular:

1. Valores de $\bar{x}$ e $\bar{y}$.
2. Coeficiente de correlación lineal.

Solución

1. $\bar{x} = 8$ y $\bar{y} = 10$.
2. $r = 0.5$.

Ejercicio 2.12. En el ajuste rectilíneo a una distribución bidimensional se sabe que $\bar{x} = 2$, $\bar{y} = 1$, y el coeficiente de correlación lineal es 0 ($r = 0$).

1. Si $x = 10$, ¿cuál será el valor interpolado para y ?
2. Si $y = 5$, ¿cuál será el valor interpolado para x ?
3. Dibuja las rectas de regresión de Y sobre X , y la de X sobre Y .

 Solución

1. $y(10) = 1$.
2. $x(5) = 2$.

Ejercicio 2.13. ★ En un estudio para relacionar la longitud de la línea de la vida de la mano izquierda y la duración de la vida de una persona se han obtenido datos de 50 personas con los siguientes resultados (X =longitud de la línea en cm, Y =edad al morir en años):

$$\sum y = 3333 \quad \sum y^2 = 231933 \quad \sum x = 459.9 \quad \sum x^2 = 4308.57 \quad \sum xy = 30949.$$

A la vista de estos resultados, ¿cuanto vivirá, por termino medio, una persona con una línea de longitud 7.5 cm? ¿Es fiable esta estimación?

 Solución

$\bar{x} = 9.198$ cm, $\bar{y} = 66.66$ años, $s_x^2 = 1.568$ cm², $s_y^2 = 195.104$ años² y $s_{xy} = 6.393$ cm·años.

Recta de regresión de la edad al morir sobre la longitud de la línea de la vida:
 $y = 4.077x + 29.158$.

$y(7.5) = 59.736$ años.

$r^2 = 0.13$, lo que quiere decir que casi no hay relación lineal entre las variables y la predicción anterior no es fiable.

Ejercicio 2.14. ★ En el estudio de regresión lineal con dos variables X e Y se sabe que $\bar{x} = 30$, $\bar{y} = 70$ y el coeficiente de correlación lineal es 0.8. También se sabe que para $x = 42$ el valor que predice la recta de regresión para y es 78.

Se pide:

1. Calcular el valor de x que se predice cuando $y = 74$.
2. Explicar razonadamente en cuál de las dos variables es más representativa la media.

 Solución

1. Recta de regresión de X sobre Y : $x = 0.96x - 37.2$.
 $x(74) = 33.84$.
2. $cv_x = 0.0408\sqrt{s_{xy}} > cv_y = 0.0146\sqrt{s_{xy}}$ y por tanto es más representativa la media de Y pues tiene menor dispersión relativa.

Ejercicio 2.15. ★ Se han medido dos variables S y T en 10 individuos, obteniéndose los siguientes resultados:

$$(-1.5, 2.25), (0.8, 0.64), (-0.2, 0.04), (-0.8, 0.64), (0.4, 0.16), \\ (0.2, 0.04), (-2.1, 4.41), (-0.4, 0.16), (1.5, 2.25), (2.1, 4.41).$$

Se pide:

1. Calcular la covarianza de S y T .
2. ¿Se puede afirmar que S y T son independientes? Justificar la respuesta.
3. ¿Qué valor predice la correspondiente recta de regresión para $t = 2$?

 Solución

1. $\bar{s} = 0$, $\bar{t} = 1.5$ y $s_{st} = 0$.
2. No podemos afirmar que S y T son independientes, sólo se puede afirmar que no hay relación lineal.
3. $s(2) = 0$.

Ejercicio 2.16. En un experimento se ha medido el número de bacterias por unidad de volumen en un cultivo, cada hora transcurrida, obteniendo los siguientes resultados:

Horas	0	1	2	3	4	5	6	7	8
Nº Bacterias	25	28	47	65	86	121	190	290	362

Se pide:

1. Dibujar el diagrama de dispersión. Según este diagrama, ¿qué tipo de modelo explicaría mejor la relación entre el número de bacterias y las horas transcurridas?
2. Dibujar el diagrama de dispersión tomando una escala logarítmica para el número de bacterias.
3. Según el modelo anterior, ¿Cuántas bacterias tendríamos al cabo de 3 horas y media? ¿Y al cabo de 10 horas? ¿Son fiables estas predicciones?
4. ¿Cuánto tiempo tendría que transcurrir para que en el cultivo hubiese 100 bacterias?

Solución

Llamando X a las horas, Y a las bacterias y Z al logaritmo neperiano de las bacterias:

1. $\bar{x} = 4$ horas, $\bar{z} = 4.5149 \log(\text{bacterias})$, $s_x^2 = 6.6667 \text{ horas}^2$, $s_z^2 = 0.8361 \log^2(\text{bacterias})$ y $s_{xz} = 2.3466 \text{ horas} \cdot \log(\text{bacterias})$.

Modelo lineal del logaritmo de las bacterias sobre las horas: $z = 0.3520x + 3.1070$.

Modelo exponencial de las bacterias sobre las horas: $y = e^{0.3520x+3.1070}$.

$y(3.5) = 76.6254$ bacterias y $y(10) = 755.0986$ bacterias.

2. Modelo lineal de las horas sobre el logaritmo de las bacterias: $x = 2.8218z - 8.7403$.

Modelo logarítmico de las horas sobre las bacterias: $x = 2.8218 \log y - 8.7403$.

$x(100) = 4.25$ horas.

Ejercicio 2.17. Para evaluar la percepción de los ciudadanos sobre la contaminación atmosférica, se ha realizado un estudio en el que se ha medido en 12 ciudades la concentración media de CO (en mg/m^3 diarios), y la percepción mediana en la calidad el aire (en una muestra de individuos de tamaño fijo), medida en la escala MM=Muy Mala, M=Mala, A=Aceptable, B=Buena y MB=Muy Buena. Los resultados obtenidos fueron:

$(12.8, A), (11.6, A), (9.8, B), (10.3, MB), (15.7, MM), (18.2, M),$
 $(11.8, B), (16.7, M), (14.5, M), (12.1, A), (19.4, MM), (7.9, MB)$

¿Existe relación entre la percepción de los habitantes de estas ciudades y la concentración de monóxido de carbono en la atmósfera de las mismas?

Ejercicio 2.18. En un estudio sobre la influencia del tabaco en los embarazos se ha medido en una muestra de 20 madres el número medio de cigarrillos diarios que fumaban las madres y el peso del recién nacido, obteniendo los siguientes resultados

Cigarrillos	2	3	10	8	12	6	6	5	4	9	14	3	7	8	2
Peso (kg)	3.1	3.3	2.5	3.3	2.6	3.1	3.0	3.4	3.4	2.7	2.5	3.7	3.1	3	3.6

Se pide:

1. Construir el modelo de regresión logarítmico del peso sobre el número de cigarrillos.
2. Según este modelo, ¿cuanto pesará el recién nacido si la madre fumaba 15 cigarrillos diarios? Es fiable esta predicción.
3. ¿Es mejor el modelo lineal a la hora de hacer predicciones?

Ejercicio 2.19. ★ La tabla siguiente contiene los datos de las presiones sistólicas de 15 individuos en función de la edad de estos.

Edad(x)	20	30	40	50	60
	121	131	132	136	134
Sistólica(y)	130	125	129	128	142
	125	128	131	134	137

1. ¿Qué porcentaje de la varianza de la presión sistólica se explica, mediante un modelo de regresión lineal, por la varianza de la edad?
2. ¿Qué edad le correspondería a un individuo que presenta una presión sistólica de 133? ¿Es fiable esta predicción? Razona la respuesta.

Solución

1. $\bar{x} = 40$ años, $\bar{y} = 130.867$ mmHg, $s_x^2 = 200$ años², $s_y^2 = 26.295$ mmHg² y $s_{xy} = 58.667$ años·mmHg.
 $r^2 = 0.654$, luego el modelo lineal explica el 65.4% de la varianza de la presión sistólica.
2. Recta de regresión de la edad sobre la presión sistólica: $x = 2.231y - 251.978$.
 $x(133) = 44.745$ años. La predicción es bastante fiable pues el coeficiente de determinación es alto.

Ejercicio 2.20. ★ Se ha realizado un estudio de regresión para ver la relación que existe entre la velocidad de transformación de una determinada sustancia química en una reacción y la temperatura a la que se realiza dicha reacción (manteniendo las cantidades de reactivos constantes). Según una recta de regresión, a 10 °C le correspondería una velocidad de 5 gr/min, y a 30 °C le correspondería una velocidad de 15 gr/min. Y según la otra recta, a una velocidad de 8 gr/min le correspondería una temperatura de 17 °C, y a una velocidad de 16 gr/min le correspondería una temperatura de 31 °C. Se pide:

1. Calcular las ecuaciones de las rectas de regresión.
2. Calcular las medias de ambas variables.
3. Calcular el coeficiente de determinación. ¿Podemos decir que las predicciones del enunciado son fiables? Justificar la respuesta.

Ejercicio 2.21. ★ En un estudio ambiental de una comunidad autónoma se afirma que el número de hectáreas quemadas en los últimos 6 años está relacionado con la cantidad de precipitación media caída en la comunidad, en litros por metro cuadrado. Los datos que han manejado son:

Año	Hectáreas quemadas	Precipitación (l/m ²)
2000	1250	420
2001	1400	380
2002	850	460
2003	1650	370
2004	900	410
2005	1700	310

1. Calcular la recta de regresión del número de hectáreas quemadas en función de la precipitación media anual.
2. ¿Es el modelo lineal un buen modelo de ajuste para la nube de puntos? Justificar la respuesta.

Ejercicio 2.22. Se desea comprobar si el número de ventas de un fármaco depende del descuento que se aplique sobre él. Para ello se ha medido el número de ventas en farmacias que aplican distintos descuentos obteniendo la siguiente muestra:

Descuento (%)	20	16	15	10	12	11	16	8	18	12	12	10	15	14
Ventas	98	46	40	15	21	19	50	8	71	24	21	16	39	32

Se pide:

1. Construir los modelos exponencial y logarítmico.
2. ¿Cuál de ellos expresa mejor la relación entre el descuento y las ventas?
3. ¿Qué descuento tendremos que aplicar si queremos vender al menos 50 fármacos?

Ejercicio 2.23. ★ Para ver si un aditivo para la gasolina mejora la combustión aumentando la emisión de dióxido de carbono, se ha hecho un estudio en el que se ha medido la cantidad de aditivo añadida a cada litro de gasolina y el porcentaje de CO₂ emitido por un mismo motor, obteniendo la siguiente muestra:

Aditivo (cl/l)	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
CO ₂ (%)	11.2	12.0	12.7	13.3	13.5	13.7	13.8	13.9

Se pide:

1. Calcular el modelo de regresión lineal y logarítmico del CO₂ sobre el aditivo. ¿Cuál de los dos modelos es mejor?
2. Según el mejor de los modelos anteriores, ¿cuánto CO₂ se producirá para 0.5 cl de aditivo? ¿y para 2 cl? ¿Son fiables estas predicciones?

3. La normativa sobre emisión de gases exige que el porcentaje mínimo de CO_2 en la combustión debe superar al menos el 12.5%. ¿Cuánto aditivo es necesario para garantizar esto?

Ejercicio 2.24. ★ La siguiente tabla muestra los datos de emisiones de CO_2 y CH_4 (en kg/hab) y el producto interior bruto per cápita (en miles US\$) de varios países en el último año:

País	CO_2	CH_4	PIB
Austria	7.60	0.97	38.40
España	6.73	0.81	30.12
Francia	5.71	0.94	33.19
EEUU	19.40	1.72	45.84
Alemania	9.80	0.83	34.18
Canadá	15.60	3.08	38.43
Italia	7.29	0.58	30.44
Japón	9.44	0.16	33.58
Australia	17.48	6.36	36.26
Reino Unido	8.99	0.76	35.13

País	CO_2	CH_4	PIB
Bolivia	1.05	3.44	40.13
Niger	0.1	0.12	0.67
Senegal	0.35	0.76	1.69
Pakistán	0.65	0.59	2.59
Filipinas	0.83	0.46	3.38
Perú	0.94	0.75	7.80
Túnez	2.17	0.48	7.47
Nepal	0.13	0.90	1.21
Nicaragua	0.7	0.32	2.62
Mauritania	0.97	0.85	2.01

Utilizando los datos sin agrupar, calcular el modelo de regresión logarítmico que explique las emisiones de CO_2 en función del PIB y utilizarlo para predecir las emisiones de un país con 10 mil US\$ de PIB. ¿Es fiable la predicción?

Ejercicio 2.25. ★ En un banco de sangre se mantiene el plasma a 0°F . Cuando se necesita para una transfusión se calienta en un horno a una temperatura constante de 120°F . En un experimento se ha medido la temperatura del plasma a distintos instantes desde el comienzo del calentamiento. Los resultados son:

Tiempo (min)	5	8	15	25	30	37	45	60
Temperatura ($^\circ\text{F}$)	25	50	86	102	110	114	118	120

Se pide:

1. Dibujar el diagrama de dispersión. ¿Qué modelo explicaría la relación entre la temperatura y el tiempo?
2. ¿Qué transformación de escala tendríamos que realizar en las variables para tener una nube de puntos con una tendencia lineal? Hacer la representación gráfica.
3. Construir el modelo de regresión logarítmico de la temperatura sobre el tiempo.
4. Según el modelo, ¿qué temperatura habrá a los 15 minutos? ¿Es fiable la predicción? Justificar la respuesta.

Usar las siguientes sumas ($X=\text{Tiempo}$ e $Y=\text{Temperatura}$): $\sum x_i = 225 \text{ min}$, $\sum \log(x_i) = 24.5289 \log(\text{min})$, $\sum y_j = 725 \text{ }^\circ\text{F}$, $\sum \log(y_j) = 35.2051 \log(^\circ\text{F})$, $\sum x_i^2 = 8833 \text{ min}^2$, $\sum \log(x_i)^2 = 80.4703 \log(\text{min})^2$, $\sum y_j^2 = 74345 \text{ }^\circ\text{F}^2$, $\sum \log(y_j)^2 = 157.1023 \log(^\circ\text{F})^2$, $\sum x_i y_j = 24393 \text{ min} \cdot ^\circ\text{F}$, $\sum x_i \log(y_j) = 1048.0142 \text{ min} \cdot \log(^\circ\text{F})$, $\sum \log(x_i) y_j = 2431.7096 \log(\text{min}) \cdot ^\circ\text{F}$, $\sum \log(x_i) \log(y_j) = 111.1165 \log(\text{min}) \log(^\circ\text{F})$.

Solución

1. Un modelo logarítmico.
2. Aplicar una transformación logarítmica al tiempo, $z = \log(x)$.
3. $\bar{z} = 28.125 \log(\text{min})$, $s_z^2 = 0.6577 \log^2(\text{min})$, $\bar{y} = 90.625 \text{ }^\circ\text{F}$, $s_y^2 = 1080.2344 \text{ }^\circ\text{F}^2$ y $s_{zy} = 26.0969 \log(\text{min}) \cdot ^\circ\text{F}$.
Modelo logarítmico de la temperatura sobre el tiempo: $y = -31.0325 + 39.6781 \log(x)$.
4. $y(15) = 76.4176 \text{ }^\circ\text{F}$. $r^2 = 0.9586$, que está muy cerca de 1, por lo que la predicción es fiable.

Ejercicio 2.26. ★ La concentración de un fármaco en sangre, C en mg/dl, es función del tiempo, t en horas, y viene dada por la siguiente tabla:

Tiempo (horas)	2	3	4	5	6	7	8
Concentración de fármaco en sangre	25	36	48	64	86	114	168

Se pide:

1. Construir el modelo de regresión lineal de la concentración del fármaco sobre el tiempo.
2. Construir el modelo de regresión exponencial de la concentración del fármaco sobre el tiempo.
3. Usar el mejor de los dos modelos anteriores para predecir la concentración de fármaco en sangre que habrá a las 4.8 horas. ¿Es fiable la predicción?

Usar las siguientes sumas (C =Concentración del fármaco y T =tiempo): $\sum t_i = 35$ h, $\sum \log(t_i) = 10.6046 \log(\text{h})$, $\sum c_j = 541$ mg/dl, $\sum \log(c_j) = 29.147 \log(\text{mg/dl})$, $\sum t_i^2 = 203 \text{ h}^2$, $\sum \log(t_i)^2 = 17.5206 \log(\text{h})^2$, $\sum c_j^2 = 56937 (\text{mg/dl})^2$, $\sum \log(c_j)^2 = 124.0131 \log(\text{mg/dl})^2$, $\sum t_i c_j = 3328 \text{ h} \cdot \text{mg/dl}$, $\sum t_i \log(c_j) = 154.3387 \text{ h} \cdot \log(\text{mg/dl})$, $\sum \log(t_i) c_j = 951.6961 \log(\text{h}) \cdot \text{mg/dl}$, $\sum \log(t_i) \log(c_j) = 46.08046 \log(\text{h}) \cdot \log(\text{mg/dl})$.

Solución

En el primer apartado de este problema debemos trabajar con el modelo exponencial de la concentración en función del tiempo, por lo que vamos a tener que calcular la recta de regresión de $z = \ln C$ en función de t . Además, en el segundo apartado debemos trabajar con el modelo lineal de t en función de C . Por lo tanto, la tabla con los sumatorios precisos es:

t_i	c_i	t_i^2	c_i^2	$t_i \cdot c_i$	$z_i = \ln_i$	z_i^2	$t_i \cdot z_i$
2	25	4	625	50	3.219	10.362	6.438
3	36	9	1296	108	3.584	12.845	10.752
4	48	16	2304	192	3.871	14.985	15.484
5	64	25	4096	320	4.159	17.297	20.795
6	86	36	7396	516	4.454	19.838	26.724
7	114	49	12996	798	4.736	22.430	33.152
8	168	64	28224	1344	5.124	26.255	40.992
$\sum =$ 35	541	203	56937	3328	29.147	124.012	154.337

1. Para el modelo exponencial de la concentración en función del tiempo tenemos en cuenta que:

$$C = a \cdot e^{bt} \Leftrightarrow \ln C = \ln(a \cdot e^{bt}) = \ln a + bt$$

Por lo tanto, si $z = \ln C$, entonces:

$$z = \ln a + bt$$

Y el modelo exponencial se transforma en un modelo lineal de z en función de t .

Por otra parte, sabemos que la recta de regresión de z en función de t viene dada por:

$$z - \bar{z} = \frac{s_{tz}}{s_t^2}(t - \bar{t})$$

Y teniendo en cuenta los sumatorios obtenidos:

$$\begin{aligned}\bar{t} &= \frac{\sum t_i}{n} = \frac{35}{7} = 5, \\ \bar{z} &= \frac{\sum z_i}{n} = \frac{29.147}{7} = 4.164, \\ s_t^2 &= \frac{\sum t_i^2}{n} - \bar{t}^2 = \frac{203}{7} - 5^2 = 4, \\ s_z^2 &= \frac{\sum z_i^2}{n} - \bar{z}^2 = \frac{124.012}{7} - 4.164^2 = 0.38, \\ s_{tz} &= \frac{\sum t_i z_i}{n} - \bar{t} \cdot \bar{z} = \frac{154.337}{7} - 5 \cdot 4.164 = 1.228.\end{aligned}$$

Donde la media de t viene dada en horas, su varianza en horas al cuadrado, la media de z no tiene unidades (z es un logaritmo neperiano), tampoco las tiene su varianza, y la covarianza tiene las unidades de t , es decir horas.

Con todo ello, la ecuación de la recta de regresión de z en función t vale:

$$z - 4.164 = \frac{1.228}{4}(t - 5) \Leftrightarrow z = 2.629 + 0.307 \cdot t$$

Por lo tanto, teniendo en cuenta que:

$$z = \ln a + b \cdot t = 2.629 + 0.307 \cdot t$$

obtenemos fácilmente que $b = 0.307$, y para a despejamos tomando exponenciales:

$$\ln a = 2.629 \Leftrightarrow a = e^{2.629} = 13.860.$$

Con todo ello, cuando $t_0 = 4.8$ horas, el valor obtenido para C_0 (en mg/dl) vale:

$$C(4.8) = 13.860e^{0.307 \cdot 4.8} = 60.498 \text{ mg/dl}.$$

Para ver si es fiable o no la predicción, calculamos el coeficiente de determinación (o el coeficiente de correlación):

$$r^2 = \frac{s_{tz}^2}{s_t^2 s_z^2} = \frac{1.228^2}{4 \cdot 0.377} = 0.999,$$

luego, mediante el modelo exponencial estamos explicando un 99.9% de la variabilidad de la nube de puntos, y el modelo exponencial es muy bueno. Por lo tanto, si el modelo es muy bueno y además la predicción la realizamos en $t_0 = 4.8$, que está dentro del rango en el que hemos calculado el modelo, sin duda la predicción también será muy fiable.

2. Para este nuevo apartado debemos predecir el tiempo que debe transcurrir para que la concentración sea de 100 mg/dl mediante un modelo lineal. Por lo tanto necesitamos la recta de regresión del tiempo en función de la concentración:

$$t - \bar{t} = \frac{s_{tC}}{s_C^2} (C - \bar{C})$$

Mediante los sumatorios obtenidos en la tabla del comienzo, calculamos:

$$\begin{aligned}\bar{C} &= \frac{\sum C_i}{n} = \frac{541}{7} = 77.286, \\ s_C^2 &= \frac{\sum z_i^2}{n} - \bar{z}^2 = \frac{56937}{7} - 77.286^2 = 2160.731, \\ s_{tC} &= \frac{\sum t_i C_i}{n} - \bar{t} \cdot \bar{C} = \frac{3328}{7} - 5 \cdot 77.286 = 90.000.\end{aligned}$$

Donde la media de C viene dada en mg/dl, su varianza en $(\text{mg/dl})^2$, y la covarianza en horas $\cdot$ (mg/dl).

Sustituyendo todo en la ecuación de la recta obtenemos:

$$t - 5 = \frac{90.000}{2160.731} (C - 77.286) \Leftrightarrow t = 0.0417 \cdot C + 1.781$$

Por lo tanto, si $C_0 = 100$ entonces $t_0 = 5.951$ horas. Para ver si la predicción es adecuada, de nuevo calculamos el coeficiente de determinación:

$$r^2 = \frac{s_{tC}^2}{s_t^2 s_C^2} = \frac{90.000^2}{4 \cdot 2160.731} = 0.937.$$

Lo cual nos confirma que el modelo lineal, aunque peor que el exponencial, sigue siendo un muy buen modelo. Si a eso unimos que estamos realizando la predicción dentro del rango de concentraciones en las que lo hemos calculado, concluimos que sí que será fiable.

Ejercicio 2.27. La actividad de una sustancia radiactiva en función del tiempo (en número de desintegraciones por segundo) viene dada por la siguiente tabla:

t (horas)	0	10	20	30	40	50	60	70
A (10^7 desintegraciones/s)	25.9	8.16	2.57	0.81	0.25	0.08	0.03	0.01

1. Representar los datos de la actividad en función del tiempo. A la vista de la representación, ¿qué modelo de regresión explicaría mejor la relación entre la actividad y el tiempo transcurrido?
2. Representar los datos de la actividad en función del tiempo en papel semilogarítmico (con escala logarítmica en el eje de ordenadas).
3. Calcular la ecuación de la recta de regresión del logaritmo neperiano de la actividad en función del tiempo.
4. Teniendo en cuenta que, en teoría, la actividad de una sustancia radiactiva en función del tiempo viene dada por la ecuación:

$$A(t) = A_0 e^{-\lambda t}$$

donde A_0 es la actividad inicial y λ es la llamada constante de desintegración, propia de cada sustancia radiactiva, utilizar la pendiente de la ecuación de la recta obtenida en el apartado anterior para calcular la constante de desintegración radiactiva de la sustancia con la que se han generado los datos.

Solución

Llamando X al tiempo e Y al logaritmo de la actividad

1. $\bar{x} = 35$, $\bar{y} = -0.7421$, $s_x^2 = 525$, $s_y^2 = 6.6664$ y $s_{xy} = -59.1434$.
Recta de regresión de Y sobre X : $y = -0.1127x + 3.2008$.
2. $\lambda = 0.1127$.

Ejercicio 2.28. Para oscilaciones de pequeña amplitud, el periodo T de oscilación de un péndulo simple viene dado por:

$$T = 2\pi\sqrt{\frac{L}{g}}$$

donde L es la longitud del péndulo y g la aceleración de la gravedad. Para comprobar que dicha ley es cierta, se mide T para varias longitudes del péndulo, obteniéndose la siguiente tabla:

L (cm)	52.5	68.0	99.0	116.0	146.0
T (seg)	1.449	1.639	1.999	2.153	2.408

Se pide:

1. Representar los datos del periodo de oscilación frente a la longitud del péndulo. ¿Sería adecuado un modelo lineal para ajustar la nube de puntos?
2. Representar los datos del periodo de oscilación frente a la longitud en papel logarítmico (con escala logarítmica tanto en el eje de abscisas como en el de ordenadas). ¿Qué modelo de regresión sería adecuado para ajustar la nube de puntos obtenida?
3. Tomar logaritmos neperianos tanto del periodo de oscilación como de la longitud y representar en una gráfica los logaritmos obtenidos. ¿Qué modelo de regresión sería adecuado para ajustar la nube de puntos obtenida?
4. Calcular la ecuación de la recta de regresión que mejor ajusta la nube de puntos del apartado anterior.
5. Teniendo en cuenta el valor del término independiente de la recta obtenida en el apartado anterior, calcular el valor de g .

Solución

Llamando X al logaritmo de la longitud e Y al logaritmo del periodo:

1. $\bar{x} = 4.5025 \log \text{cm}$, $\bar{y} = 0.6407 \log \text{s}$, $s_x^2 = 0.1353 \log^2 \text{cm}$, $s_y^2 = 0.0339 \log^2 \text{s}$, $s_{xy} = 0.0677 \log \text{cm} \log \text{s}$.
Recta de regresión de Y sobre X : $y = 0.5006x - 1.6132$.
2. $g = 994,4145 \text{ cm/s}^2$.

Ejercicio 2.29. En análisis colorimétrico, es frecuente utilizar la fracción de luz que absorbe una determinada sustancia disuelta como una medida de la concentración con la que dicha sustancia está presente en la disolución, siempre y cuando se utilice luz monocromática y la misma longitud recorrida por la luz en cada una de las mediciones. Si llamamos I_0 a la intensidad de luz incidente, I a la intensidad de luz transmitida y C

a la concentración de la sustancia analizada, en un experimento de análisis colorimétrico realizado con Mn y una longitud de onda de 525 nm, se han obtenido los siguientes datos, donde la concentración de Mn viene dada en mg por cada 100 ml de disolución:

C	1.00	2.00	3.00	4.00
I/I_0	0.418	0.149	0.058	0.026

Se pide:

1. Representar los datos considerando I/I_0 en función de C . A la vista de la nube de puntos, ¿qué modelo de regresión sería el más adecuado para expresar la relación entre las variables?
2. Representar los datos pero en papel semilogarítmico.
3. Calcular la ecuación de la recta de regresión del logaritmo neperiano de I/I_0 frente a C .

Solución

Llamando C a la concentración e Z al logaritmo neperiano de I/I_0 .

1. $\bar{c} = 2.5$ mg/100ml, $\bar{z} = -2.3183$, $s_c^2 = 1.25$ (mg/100ml)², $s_z^2 = 1.0788$ y $s_{cz} = -1.1595$.
Recta de regresión de Z sobre C : $z = -0.9276c + 0.0007$.

Ejercicio 2.30. Se han recogido por medio de unos cuestionarios los niveles de estrés y energía de 14 mujeres durante un año. A partir de las respuestas del cuestionario se han asignado puntuaciones a cada una de ellas de manera que a mayor puntuación mayor grado de estrés y energía. Los datos recogidos son:

Edad	21	31	19	21	30	20	22	23	45	24	26	19	25	21
Estrés	25	19	20	19	24	6	29	25	49	0	10	25	13	23
Energía	25	20	45	60	50	50	10	60	40	60	50	60	85	50

Se pide:

1. Dibujar un diagrama de dispersión que refleje la relación entre el estrés y la energía.
2. ¿Existe relación lineal entre el estrés y la energía? ¿Y entre el estrés y la edad? Justificar la respuesta.
3. ¿Qué efecto tendría sobre el coeficiente de correlación lineal de la edad y el estrés la eliminación del individuo de 45 años? Justificar la respuesta.
4. Calcular el coeficiente de correlación de Spearman entre estrés y energía e interpretarlo. ¿Coinciden las conclusiones con las que se deducen del coeficiente de correlación lineal?

Solución

Llamando X a la edad, Y al estrés y Z a la energía:

1. $r_{yz}^2 = 0.14$, lo que indica que casi no hay relación entre el estrés y la energía y $r_{xy}^2 = 0.31$ lo que indica que hay una ligera relación entre el estrés y la edad.
2. El coeficiente de correlación lineal disminuye hasta valer casi 0, lo que indica que la relación lineal entre el estrés y la edad del apartado anterior se debe a este dato atípico, así que, realmente no hay relación entre estrés y edad.
3. $r_s = -0.41$ lo que indica que hay una ligera relación decreciente entre energía y estrés.

Ejercicio 2.31. Para comprobar el efecto de la herencia genética sobre la inteligencia se desarrolló un estudio en el que se midió el coeficiente intelectual de varias parejas de gemelos, obteniendo los siguientes resultados:

(128, 132) (116, 112) (86, 98) (65, 81) (104, 96) (111, 111) (101, 105) (72, 75)

Calcular el coeficiente de determinación lineal e interpretarlo. ¿Tiene sentido calcular el coeficiente de correlación?

Solución

Llamando X al coeficiente intelectual del primer hermano e Y al del segundo: $\bar{x} = 97.875$, $\bar{y} = 101.25$, $s_x^2 = 418.3594$, $s_y^2 = 288.4375$, $s_{xy} = 326.5313$ y $r^2 = 0.8836$, lo que indica que existe bastante relación lineal entre el coeficiente intelectual de los gemelos. No tiene sentido el coeficiente de correlación lineal porque es indiferente el orden en que tomemos a los gemelos.

Ejercicio 2.32. En un estudio sobre la búsqueda visual se realiza un prueba que consiste en presentarle a un sujeto una matriz de n símbolos y pedirle que pulse rápidamente un botón si entre los símbolos se encuentra uno concreto, u otro botón diferente si no aparece dicho símbolo. El tiempo de respuesta de cada participante (en centésimas de segundo) y el número de símbolos de cada matriz aparecen en la siguiente tabla:

Matrices con	n	4	5	6	7	8	9	10	11	12
el símbolo	T	22	24	23	31	33	45	42	46	50
Matrices sin	n	4	5	6	7	8	9	10	11	
el símbolo	T	25	24	32	35	43	49	52	56	

Se pide:

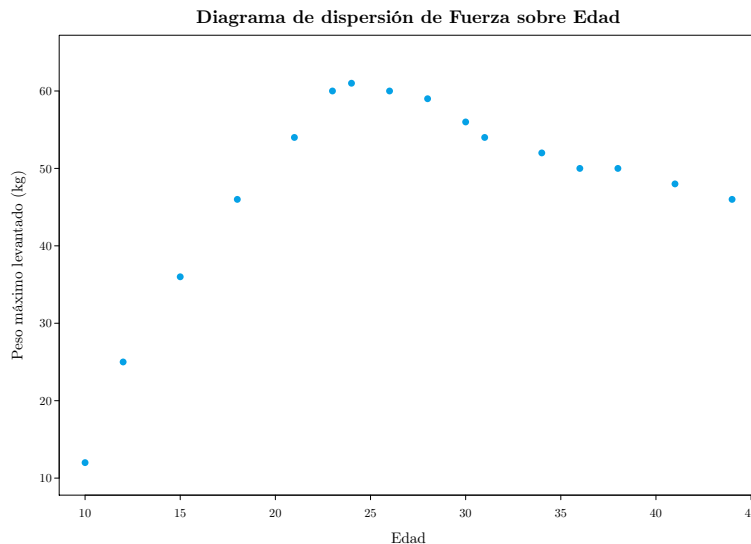
1. Construir la recta de regresión del tiempo de respuesta sobre el número de símbolos para las matrices con el símbolo y también para las matrices sin el símbolo.
2. ¿En qué matrices, las que tienen el símbolo o las que no, explica mejor el tiempo de respuesta el número de símbolos? Justificar la respuesta.
3. Según los modelos anteriores, ¿cuánto tiempo tardará en responder una persona elegida al azar en una matriz de 20 símbolos que contenga al símbolo? ¿Y si no lo contuviese?

Solución

Llamando X al número de símbolos e Y al tiempo de respuesta:

1. Matrices con el símbolo: $\bar{x} = 8$ símbolos, $\bar{y} = 35.1111$ seg, $s_x^2 = 6.6667$ símbolos², $s_y^2 = 104.4321$ seg², $s_{xy} = 25.4446$ símbolos·seg.
Recta de regresión del tiempo sobre el número de símbolos: $y = 3.8333x + 4.4444$. Matrices sin el símbolo: $\bar{x} = 7.5$ símbolos, $\bar{y} = 39.5$ seg, $s_x^2 = 5.25$ símbolos², $s_y^2 = 132.25$ seg², $s_{xy} = 26$ símbolos·seg.
Recta de regresión del tiempo sobre el número de símbolos: $y = 4.9525x + 2.3571$.
2. $r^2 = 0.9292$ en las matrices con el símbolo y $r^2 = 0.9736$ en las matrices sin el símbolo, así que el número de símbolos explica un poco mejor el tiempo de respuesta en las matrices sin el símbolo.
3. $y(20) = 81.11$ seg si la matriz contiene el símbolo y $y(20) = 101.4$ seg si la matriz no contiene el símbolo.

Ejercicio 2.33. Se ha realizado un estudio para averiguar la relación entre la edad y la fuerza física. Para ello se ha medido la edad de 16 participantes y el máximo peso (en kg) que eran capaces de levantar. Los resultados obtenidos fueron:



Se pide:

1. Calcular el coeficiente de determinación lineal de la muestra completa.
2. Calcular el coeficiente de determinación lineal de la muestra de personas menores de 25 años.
3. Calcular el coeficiente de determinación lineal de la muestra de personas mayores de 25 años.
4. ¿Para qué grupo de edades es más fuerte la relación lineal entre la fuerza física y la edad?

Usar las siguientes sumas (X =Edad e Y =Peso máximo levantado).

- Muestra total: $\sum x_i = 431$ años, $\sum y_j = 769$ kg, $\sum x_i^2 = 13173$ años², $\sum y_j^2 = 39675$ kg² y $\sum x_i y_j = 21792$ años·kg.
- Muestra menores de 25 años: $\sum x_i = 123$ años, $\sum y_j = 294$ kg, $\sum x_i^2 = 2339$ años², $\sum y_j^2 = 14418$ kg² and $\sum x_i y_j = 5766$ años·kg.
- Muestra mayores de 25 años: $\sum x_i = 308$ años, $\sum y_j = 475$ kg, $\sum x_i^2 = 10834$ años², $\sum y_j^2 = 25257$ kg² y $\sum x_i y_j = 16026$ años·kg.

Solución

Llamando X a la edad e Y al peso levantado.

1. $\bar{x} = 26.9375$ años, $s_x^2 = 97.6836$ años², $\bar{y} = 48.0625$ kg, $s_y^2 = 169.6836$ kg², $s_{xy} = 67.3164$ años·kg y $r^2 = 0.2734$.
2. $\bar{x} = 15.5714$ años, $\bar{y} = 42$ kg, $s_x^2 = 25.3878$ años², $s_y^2 = 295.7143$ kg², $s_{xy} = 85.7143$ años·kg y $r^2 = 0.9786$.
3. $\bar{x} = 35.2222$ años, $\bar{y} = 52.7778$ kg, $s_x^2 = 32.6173$ años², $s_y^2 = 20.8395$ kg², $s_{xy} = -25.5062$ años·kg y $r^2 = 0.9571$.
4. La relación lineal entre la edad y la fuerza física es un poco más fuerte en los menores de 25 años.

Ejercicio 2.34. Para evaluar la capacidad de aprendizaje en la realización de una tarea, se ha medido el tiempo que tarda en realizarse una tarea en sucesivas repeticiones de la misma. Los resultados obtenidos son:

Repetición	1	2	3	4	5	6	7	8	9	10
Tiempo (min)	80	65	56	50	48	43	41	38	37	35

Se pide:

1. Dibujar el diagrama de dispersión.
2. En vista del diagrama de dispersión, construir el modelo de regresión más adecuado del tiempo en función de las repeticiones.

3. ¿Qué porcentaje de la variabilidad del tiempo explican las repeticiones?
4. ¿Cuanto tiempo tardará por término medio en la 5 repetición de la tarea?

Solución

Llamando X a las repeticiones, Y al tiempo y Z al logaritmo neperiano del tiempo, se tienen:

1. $\bar{x} = 5.5$ repeticiones, $\bar{z} = 3.8644 \ln(\text{min})$, $s_x^2 = 8.25$ repeticiones², $s_z^2 = 0.0637 \ln^2(\text{min})$ y $s_{xz} = -0.7014$ repeticiones $\cdot\ln(\text{min})$.
 Recta de regresión del logaritmo del tiempo sobre las repeticiones: $z = -0.085x + 4.3320$.
 Modelo exponencial del tiempo sobre las repeticiones: $y = e^{-0.085x+4.3320}$.
2. $R^2 = 0.9364$, es decir, un 93.64%.
3. $y(5) = 49.74$ min.

Ejercicio 2.35. En un estudio se ha preguntado a un grupo de personas sobre su ideología política X (izquierda, centro o derecha) y su opinión sobre la subida o bajada de impuestos Y , obteniendo la siguiente tabla de frecuencias:

$X \backslash Y$	Bajada	Mantenimiento	Subida
Izquierda	2	6	8
Centro	3	4	3
Derecha	6	5	3

¿Se puede concluir que existe relación entre la ideología y la opinión sobre la subida o bajada de impuestos? Justificar la respuesta.

Solución

$\chi^2 = 4.4$ y $C = 0.49$ lo que indica que existe bastante relación entre las variables.

Ejercicio 2.36. Un estudio sobre 100 personas concluye que 26 personas son fumadores y bebedores habituales, 12 son bebedores pero no fumadores, 18 son fumadores pero no bebedores y 44 no beben ni fuman habitualmente. Según estos datos, ¿podemos decir que existe relación entre el tabaco y la bebida? Justificar la respuesta.

Solución

$\chi^2 = 14.83$ y $C = 0.36$ lo que indica que hay una relación moderada entre los hábitos de fumar y beber.

Ejercicio 2.37. En un estudio en el que participaron las 8 universidades de una región se ha valorado la excelencia docente e investigadora, estableciendo los siguientes rankings (de mejor a peor):

Ranking docencia	3	4	8	5	2	1	6	7
Ranking investigación	6	5	4	3	7	8	1	2

¿Se puede decir que existe relación entre la excelencia docente e investigadora? Justificar la respuesta.

 Solución

$r_s = -0.83$, lo que indica una fuerte relación inversa entre la excelencia docente y la excelencia investigadora.

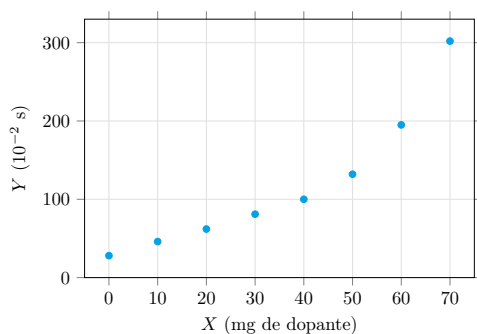
Ejercicio 2.38. ★ En un grupo de pacientes se analiza el efecto de una sustancia dopante en el tiempo de respuesta a un determinado estímulo. Para ello, se suministra en sucesivas dosis, de 0 a 70 mg, la misma cantidad de dopante a todos los miembros del grupo, y se anota el tiempo medio de respuesta al estímulo, expresado en centésimas de segundo.

X (mg)	0	10	20	30	40	50	60	70
Y (10^{-2} s)	28	46	62	81	100	132	195	302

1. Representar la nube de puntos. A la vista de la representación, ¿crees que el modelo lineal es el que mejor explica el tipo de relación entre las variables?
2. Calcular la recta de regresión del tiempo en función de la cantidad de dopante, y utilizarla para predecir el tiempo de reacción medio para una cantidad de dopante de 25 mg.
3. Hacer la misma predicción del apartado anterior con el modelo exponencial. ¿Qué predicción es mejor?
4. Si para el estímulo estudiado los tiempos de reacción superiores a un segundo se consideran peligrosos para la salud, ¿a partir de qué nivel debería regularse, e incluso prohibirse, la administración de la sustancia dopante?

 Solución

1. El diagrama de dispersión de Y sobre X es el siguiente



A la vista del diagrama se puede decir que el modelo lineal no es el que mejor se ajustaría a la nube de puntos, sino posiblemente el exponencial.

2. La recta de regresión de Y sobre X , tiene ecuación

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

Calculamos primero los estadísticos que necesitamos en la ecuación:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{0 + 10 + \dots + 70}{8} = \frac{280}{8} = 35, \\ s_x^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{0^2 + 10^2 + \dots + 70^2}{8} - 35^2 = \frac{14000}{8} - 35^2 = 7261.6875, \\ s_x &= \sqrt{7261.6875} = 22.91, \\ \bar{y} &= \frac{\sum y_j}{n} = \frac{28 + 46 + \dots + 302}{8} = \frac{946}{8} = 118.25, \\ s_y^2 &= \frac{\sum y_j^2}{n} - \bar{y}^2 = \frac{28^2 + 46^2 + \dots + 302^2}{8} - 118.25^2 = \frac{169958}{8} - 13983.0625 = 7261.6875, \\ s_y &= \sqrt{7261.6875} = 85.22, \\ s_{xy} &= \frac{\sum x_i y_j}{N} - \bar{x} \bar{y} = \frac{0 \cdot 28 + 10 \cdot 46 + \dots + 70 \cdot 302}{8} - 35 \cdot 118.25 = \\ &= \frac{47570}{8} - 4138.75 = 1807.5.\end{aligned}$$

Sustituyendo en la ecuación anterior estos estadísticos calculados obtenemos la recta de regresión de Y sobre X .

$$y = 118.25 + \frac{1807.5}{525}(x - 35) = 3.44x - 2.25.$$

Según esta recta, el tiempo de reacción medio para una cantidad dopante de 25 mg sería

$$y(25) = 3.44 \cdot 25 - 2.25 = 83.82 \text{ centésimas de segundo.}$$

3. Para ver la cantidad dopante que le corresponde 1 segundo de tiempo de reacción, necesitamos utilizar la recta de regresión de X sobre Y . La ecuación de esta recta es

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 35 + \frac{1807.5}{7261.6875}(y - 118.25) = 0.25y + 5.57.$$

Y una vez que tenemos la recta de regresión, para estimar la dosis correspondiente a 1 segundo, hacemos la predicción para $y = 100$ centésimas de segundo (ya que las unidades de X son centésimas de segundo 10^{-2} s). Sustituyendo en la ecuación anterior tenemos

$$x(100) = 0.25 \cdot 100 + 5.57 = 30.46 \text{ mg.}$$

Como la relación entre X e Y es creciente ($s_{xy} > 0$), a partir de 30.46 mg el tiempo de reacción sería superior a un segundo, y por tanto, peligroso para la salud.

Ejercicio 2.39. ★ La artrosis reumatoide es una enfermedad reumática que aparece con frecuencia en las personas mayores. Uno de los índices más utilizados para ver el grado de actividad de la enfermedad es el RADAI (Rheumatoid Arthritis Disease Activity Index), que mide el grado de actividad en una escala de 0 (mínima actividad) a 3 (máxima actividad). Para ver de qué manera influye la edad en el grado de actividad de la enfermedad se ha seleccionado un grupo de personas mayores y se ha medido el índice RADAI en ellos, obteniendo la siguiente tabla de frecuencias:

RADAI\Edad	40-50	50-60	60-70	70-80
0-1	8	6	2	1
1-2	4	7	5	2
2-3	0	2	6	7

Se pide:

1. Estudiar si existe relación lineal entre la edad y el RADAI.
2. Calcular la recta de regresión del RADAI sobre la edad. Según la recta, ¿cuánto aumentaría el grado de actividad de la enfermedad por cada año que pasa?
3. Si se considera que los pacientes con un RADAI de 2 o superior necesitan ayuda en sus actividades diarias, ¿a qué edad se empezaría a necesitar esta ayuda?

Solución

Llamemos X a la variable que mide la edad e Y a la que mide el RADAI.

1. Para ver si existe relación lineal entre X e Y , podemos calcular el coeficiente de correlación lineal, pero para ello necesitamos la media y desviación típica de cada variable y la covarianza. Antes de calcular estos estadísticos, obtenemos las distribuciones marginales de cada variable a partir de la tabla:

$Y \backslash X$	40-50	50-60	60-70	70-80	n_y
0-1	8	6	2	1	17
1-2	4	7	5	2	18
2-3	0	2	6	7	15
n_x	12	15	13	10	50

A partir de aquí calculamos los estadísticos anteriores:

$$\begin{aligned}\bar{x} &= \frac{\sum_i x_i n_i}{n} = \frac{45 \cdot 12 + 55 \cdot 15 + 65 \cdot 13 + 75 \cdot 10}{50} = \frac{2960}{50} = 59.2, \\ \bar{y} &= \frac{\sum_j y_j n_j}{n} = \frac{0.5 \cdot 17 + 1.5 \cdot 18 + 2.5 \cdot 15}{50} = \frac{73}{50} = 1.46, \\ s_x^2 &= \frac{\sum_i x_i^2 n_i}{n} - \bar{x}^2 = \frac{45^2 \cdot 12 + 55^2 \cdot 15 + 65^2 \cdot 13 + 75^2 \cdot 10}{50} - 59.2^2 = \\ &= \frac{180850}{50} - 3504.64 = 112.36, \\ s_x &= \sqrt{112.36} = 10.6, \\ s_y^2 &= \frac{\sum_j y_j^2 n_j}{n} - \bar{y}^2 = \frac{0.5^2 \cdot 17 + 1.5^2 \cdot 18 + 2.5^2 \cdot 15}{50} - 1.46^2 = \\ &= \frac{138.5}{50} - 2.1316 = 0.6384, \\ s_y &= \sqrt{0.6384} = 0.8, \\ s_{xy} &= \frac{\sum_{ij} x_i y_j n_{ij}}{n} - \bar{x} \bar{y} = \frac{45 \cdot 0.5 \cdot 8 + 55 \cdot 1.5 \cdot 4 + \dots + 75 \cdot 2.5 \cdot 7}{50} - 59.2 \cdot 1.46 = \\ &= \frac{4570}{50} - 86.432 = 4.968.\end{aligned}$$

Con estos datos, el coeficiente de correlación lineal es

$$r = \frac{s_{xy}}{s_x s_y} = \frac{4.968}{10.6 \cdot 0.8} = 0.59,$$

que indica que existe relación lineal aunque no demasiado fuerte, sino más bien moderada.

2. La recta de regresión de Y sobre X es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}) = 1.46 + \frac{4.968}{112.36}(x - 59.2) = 0.0442x - 1.1575.$$

El aumento del grado de actividad del RADAI por cada año que pasa nos lo da el coeficiente de regresión de Y sobre X , que es la pendiente de la recta de regresión que hemos calculado, es decir, 0.0442 por cada año.

3. Para predecir a qué edad se empezaría a necesitar ayuda, necesitamos calcular la recta de regresión de X sobre Y , que tiene ecuación

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 5.92 + \frac{4.968}{0.6384}(y - 1.46) = 7.782y + 47.83.$$

Sustituyendo y por 2 en esta ecuación tenemos

$$x(2) = 7.782 \cdot 2 + 47.83 = 63.4 \text{ años.}$$

Ejercicio 2.40. ★ En un equipo de baloncesto se ha introducido un programa de estiramientos para ver si se consigue reducir el número de lesiones. Durante toda una temporada cada jugador realizó ejercicios de estiramiento durante un número fijo de minutos en cada entrenamiento. Al finalizar la temporada se midió el número de lesiones y se obtuvieron los resultados de la siguiente tabla:

Minutos de estiramiento	0	30	10	15	5	25	35	40
Lesiones	4	1	2	2	3	1	0	1

Se pide:

1. Calcular la recta de regresión del número de lesiones con respecto al tiempo de estiramiento.
2. ¿Cual es la disminución de lesiones esperada por cada minuto de estiramiento?
3. ¿Cuántos minutos de estiramiento debe realizar un jugador para no tener ninguna lesión? ¿Es fiable esta predicción?

Usar las siguientes sumas (X =Número de minutos de estiramiento, e Y =Número de lesiones): $\sum x_i = 160$ minutos, $\sum y_j = 14$ lesiones, $\sum x_i^2 = 4700$ minutos², $\sum y_j^2 = 36$ lesiones² y $\sum x_i y_j = 160$ minutos·lesiones.

Solución

Llamemos X a la variable que mide el tiempo de estiramiento, e Y a la que mide el número de lesiones en cada jugador.

1. La recta de regresión de Y sobre X , tiene ecuación

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

Calculamos primero los estadísticos que necesitamos en la ecuación:

$$\begin{aligned}
\bar{x} &= \frac{\sum x_i}{N} = \frac{0 + 30 + \dots + 40}{8} = \frac{160}{8} = 20, \\
s_x^2 &= \frac{\sum x_i^2}{N} - \bar{x}^2 = \frac{0^2 + 30^2 + \dots + 40^2}{8} - 20^2 = \frac{4700}{8} - 20^2 = 187.5, \\
s_x &= \sqrt{187.5} = 13.69, \\
\bar{y} &= \frac{\sum y_j}{N} = \frac{4 + 1 + \dots + 1}{8} = \frac{14}{8} = 1.75, \\
s_y^2 &= \frac{\sum y_j^2}{N} - \bar{y}^2 = \frac{4^2 + 1^2 + \dots + 1^2}{8} - 1.75^2 = \frac{36}{8} - 1.75^2 = 1.4375, \\
s_y &= \sqrt{1.4375} = 1.2, \\
s_{xy} &= \frac{\sum x_i y_j}{N} - \bar{x} \bar{y} = \frac{0 \cdot 4 + 30 \cdot 1 + \dots + 40 \cdot 1}{8} - 20 \cdot 1.75 = \\
&= \frac{160}{8} - 20 \cdot 1.75 = -15.
\end{aligned}$$

Sustituyendo en la ecuación anterior estos estadísticos calculados obtenemos la recta de regresión de Y sobre X .

$$y = 1.75 - \frac{15}{187.5}(x - 20) = -0.08x + 3.35.$$

El incremento que experimenta la variable Y por cada unidad que se incrementa la variable X según la recta de regresión, es su pendiente o coeficiente de regresión de Y sobre X , que en este caso es -0.08 . Así, pues, por cada minuto más de estiramiento se espera tener 0.08 lesiones menos.

2. Para predecir el número de minutos que debería estirar un jugador que quiere tener 0 lesiones, debemos calcular antes la recta de regresión de tiempo de estiramiento sobre número de lesiones. La ecuación de esta recta es

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 20 - \frac{15}{1.4375}(y - 1.75) = -10.43y + 38.26.$$

Una vez que tenemos la recta de regresión, para estimar el valor de X para $y = 0$, basta con sustituir y por 0 en esta ecuación y obtenemos

$$x(0) = -10.43 \cdot 0 + 38.26 = 38.26.$$

Por último, para ver si esta estimación es fiable, calculamos el coeficiente de correlación lineal

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-15}{13.69 \cdot 1.2} = -0.91.$$

Como el coeficiente de correlación lineal está próximo a -1, el modelo lineal es un buen modelo y por tanto sus predicciones serán fiables.

Ejercicio 2.41. ★ Un profesor está interesado en analizar la relación existente entre la nota que esperan obtener los alumnos en los exámenes de su asignatura con la que de verdad obtienen una vez corregidos dichos exámenes. La tabla muestra la nota esperada y la obtenida para 10 alumnos diferentes:

Alumno	Nota esperada	Nota obtenida
1	3.0	5.1
2	6.0	4.8
3	7.0	6.0
4	8.0	4.2
5	3.0	5.2
6	9.0	7.5
7	2.0	3.6
8	5.0	3.0
9	8.0	6.5
10	2.0	0.8

1. Calcular la recta de regresión de la nota obtenida en función de la nota esperada.
2. Calcular el coeficiente de correlación lineal e interpretarlo.
3. ¿Cuál es la nota que esperaba obtener un alumno que en realidad saca un 4.0?

Solución

Llamemos X a la nota esperada e Y a la nota real obtenida.

1. La ecuación de la recta de regresión de Y sobre X es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

Calculamos primero los estadísticos que necesitamos en la ecuación:

$$\begin{aligned}
\bar{x} &= \frac{\sum x_i}{n} = \frac{53}{10} = 5.3, \\
s_x^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{345}{10} - 5.3^2 = 6.41, \\
s_x &= \sqrt{6.41} = 2.5318, \\
\bar{y} &= \frac{\sum y_j}{n} = \frac{46.7}{10} = 4.67, \\
s_y^2 &= \frac{\sum y_j^2}{n} - \bar{y}^2 = \frac{250.83}{10} - 4.67^2 = 3.2741, \\
s_y &= \sqrt{3.2741} = 1.8094, \\
s_{xy} &= \frac{\sum x_i y_j}{n} - \bar{x} \bar{y} = \frac{278.6}{10} - 5.3 \cdot 4.67 = 3.109.
\end{aligned}$$

Sustituyendo en la ecuación anterior estos estadísticos calculados obtenemos la recta de regresión de Y sobre X .

$$y = 4.67 + \frac{3.109}{6.41}(x - 5.3) = 0.485x + 2.0994.$$

2. El coeficiente de correlación lineal es

$$r = \frac{s_{xy}}{s_x s_y} = \frac{3.109}{2.5318 \cdot 1.8094} = 0.6786,$$

y según este valor, podemos decir que existe una dependencia creciente moderada.

3. Para predecir la nota esperada por un alumno que saca un 4, necesitamos la recta de regresión de X sobre Y . La ecuación de esta recta es

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 5.3 + \frac{3.109}{3.2741}(y - 4.67) = 0.9496y + 0.8655.$$

Sustituyendo y por 4 en esta ecuación obtenemos la predicción deseada

$$x(4) = 0.9496 \cdot 4 + 0.8655 = 4.6639.$$

Ejercicio 2.42. ★ Los investigadores están estudiando la correlación entre obesidad y la respuesta individual al dolor. La obesidad se mide como porcentaje sobre el peso ideal (x). La respuesta al dolor se mide utilizando el umbral de reflejo de flexión nociceptiva (y), que es una medida de sensación de punzada. Se obtuvieron los siguientes resultados:

x	89	90	75	30	51	75	62	45	90	20
y	10	12	4	4.5	5.5	7	9	8	15	3

1. Dibujar el diagrama de dispersión. Según la nube de puntos, ¿qué modelo explicaría mejor la relación entre el umbral de reflejo y el porcentaje sobre el peso ideal?
2. Según el modelo de regresión más adecuado, ¿qué respuesta al dolor se espera que tenga una persona con una obesidad del 50%? ¿Es esta predicción fiable?
3. Según el modelo de regresión más adecuado, ¿cuál es la obesidad esperada de una persona con un umbral reflejo de flexión nociceptiva de 10? ¿Es esta predicción fiable?

Usar las siguientes sumas (X =Obesidad e Y =Respuesta al dolor): $\sum x_i = 629$, $\sum \log(x_i) = 40.4121$, $\sum y_j = 92.2$, $\sum \log(y_j) = 21.339$, $\sum x_i^2 = 45445$, $\sum \log(x_i)^2 = 165.6795$, $\sum y_j^2 = 960.14$, $\sum \log(y_j)^2 = 47.6231$, $\sum x_i y_j = 6537.7$, $\sum x_i \log(y_j) = 1443.1275$, $\sum \log(x_i) y_j = 387.5728$, $\sum \log(x_i) \log(y_j) = 88.3696$.

Solución

1. $\bar{x} = 62.9$, $s_x^2 = 588.09$, $\bar{y} = 9.22$, $s_y^2 = 11.0056$ and $s_{xy} = 82.0356$.
Recta de regresión de la respuesta al dolor sobre la obesidad: $y = 1.3232 + 0.1255x$. $r^2 = 0.8422$.
 $\overline{\log(x)} = 4.0412$, $s_{\log(x)}^2 = 0.2366$ and $s_{\log(x)y} = 1.4973$. Modelo logarítmico de la respuesta al dolor sobre la obesidad: $y = -16.3578 + 6.3293 \log(x)$.
 $r^2 = 0.8611$.
 $y(50) = 8.4023$.
2. Modelo exponencial de la obesidad sobre la respuesta al dolor: $x = e^{2.7868+0.1361y}$.
 $x(10) = 63.2648$.

Ejercicio 2.43. ★ Se realiza un estudio para establecer una ecuación mediante la cual se pueda utilizar la concentración de estrona en saliva para predecir la concentración del esteroide en plasma libre. Se extrajeron los siguientes datos de 10 varones sanos:

Estrona	1.4	7.5	8.5	9.0	9.0	11	13	14	14.5	16
Esteroides	30.0	25.0	31.5	27.5	39.5	38.0	43.0	49.0	55.0	48.5

1. Comprobar la idoneidad del modelo lineal de regresión. Si el modelo es apropiado, hallar la recta de regresión de la concentración de estrona en función de la concentración de esteroide.
2. Si un individuo presenta una concentración de estrona en saliva de 10, ¿qué concentración de esteroide en plasma libre predeciría el modelo de regresión lineal?

3. Para los dos primeros individuos, calcular los r^2 que se comenten al utilizar el modelo de regresión lineal para predecir la concentración de estrona. Razonar a que se deben estos r^2 .

Ejercicio 2.44. ★ En una análisis de niños sanos se deseaba establecer si existía relación lineal entre la edad (en años) del niño y el ángulo de Clarke (en grados), obteniéndose en una muestra de 7 niños los valores que aparecen a continuación:

Edad	3	4	5	6	7	8	9
Ángulo de Clarke	24	26	30	31	34	32	33

1. Calcular la ecuación de la recta de regresión del Ángulo de Clarke en función de la edad.
2. ¿Qué tanto por ciento de la variabilidad de la nube de puntos explicamos con el modelo lineal? ¿Se puede considerar un modelo bueno?
3. Calcular el coeficiente de correlación de Spearman e interpretarlo. ¿Está en consonancia con el coeficiente de correlación lineal?

Ejercicio 2.45. ★ Se quiere estudiar la relación entre las concentraciones de dos sustancias X e Y en la sangre. Para ello se han medido las concentraciones de estas sustancias en siete individuos, ambas en microgramos por decilitro de sangre, obteniendo los siguientes resultados

X	2.1	4.9	9.8	11.7	5.9	8.4	9.2
Y	1.3	1.5	1.7	1.8	1.5	1.7	1.7

Se pide:

1. ¿Existe relación lineal entre Y y X ?
2. ¿Existe relación potencial entre Y y X ?
3. Utilizar el mejor de los modelos anteriores para predecir la concentración de Y para $x = 8 \mu\text{gr/dl}$. ¿Es fiable la predicción? Justificar la respuesta.

Solución

Para el modelo lineal se tiene

1. Para ver si existe relación lineal entre Y y X se calcula el coeficiente de determinación lineal:

$$\begin{aligned}
\bar{x} &= \frac{\sum x_i}{n} = \frac{2.1 + \dots + 9.2}{7} = \frac{52}{7} = 7.4286 \text{ } \mu\text{gr/dl}, \\
s_x^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{2.1^2 + \dots + 9.2^2}{7} - 7.4286^2 = \frac{451.36}{7} - 55.1841 = 9.2963 \text{ } (\mu\text{gr/dl})^2, \\
\bar{y} &= \frac{\sum y_j}{n} = \frac{1.3 + \dots + 1.7}{7} = \frac{11.2}{7} = 1.6 \text{ } \mu\text{gr/dl}, \\
s_y^2 &= \frac{\sum y_j^2}{n} - \bar{y}^2 = \frac{1.3^2 + \dots + 1.7^2}{7} - 1.6^2 = \frac{18.1}{7} - 2.56 = 0.0257 \text{ } (\mu\text{gr/dl})^2, \\
s_{xy} &= \frac{\sum x_i y_j}{n} - \bar{x} \bar{y} = \frac{2.1 \cdot 1.3 + \dots + 9.2 \cdot 1.7}{7} - 7.4286 \cdot 1.6 = \frac{86.57}{7} - 11.8858 = 0.4814 \text{ } (\mu\text{gr/dl})^2, \\
r^2 &= \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{0.4814^2}{9.2963 \cdot 0.0257} = 0.9696.
\end{aligned}$$

Como el coeficiente de determinación lineal está muy próximo a 1, podemos concluir que existe una relación lineal muy fuerte entre X e Y .

2. Del mismo modo, para ver si existe relación potencial entre Y y X se calcula el coeficiente de determinación potencial. Teniendo en cuenta que la ecuación del modelo potencial $y = ax^b$ se puede convertir en lineal aplicando el logaritmo tanto a X como a Y , $\ln y = \ln a + b \ln x$, el coeficiente de determinación potencial entre X y Y es el mismo que el coeficiente de determinación lineal entre $\ln(X)$ y $\ln(Y)$. Así pues, para calcularlo primero construimos las variables $U = \ln(X)$ y $V = \ln(Y)$:

$U =$	0.7419	1.5892	2.2824	2.4596	1.7750	2.1282	2.2192
$\ln X$							
$V =$	0.2624	0.4055	0.5306	0.5878	0.4055	0.5306	0.5306
$\ln Y$							

Y el coeficiente de determinación potencial vale:

$$\begin{aligned}\bar{u} &= \frac{\sum u_i}{n} = \frac{0.7419 + \dots + 2.2192}{7} = \frac{13.1955}{7} = 1.8851 \ln(\mu\text{gr/dl}), \\ s_u^2 &= \frac{\sum u_i^2}{n} - \bar{u}^2 = \frac{0.7419^2 + \dots + 2.2192^2}{7} - 1.8851^2 = \frac{26.9397}{7} - 3.5536 = 0.295 \ln^2(\mu\text{gr/dl}), \\ \bar{v} &= \frac{\sum v_j}{n} = \frac{0.2624 + \dots + 0.5306}{7} = \frac{3.253}{7} = 0.4647 \ln(\mu\text{gr/dl}), \\ s_v^2 &= \frac{\sum v_j^2}{n} - \bar{v}^2 = \frac{0.2624^2 + \dots + 0.5306^2}{7} - 0.4647^2 = \frac{1.5878}{7} - 0.2159 = 0.0109 \ln^2(\mu\text{gr/dl}), \\ s_{uv} &= \frac{\sum u_i v_j}{n} - \bar{u}\bar{v} = \frac{0.7419 \cdot 0.2624 + \dots + 2.2192 \cdot 0.5306}{7} - 1.8851 \cdot 0.4647 = \\ &= \frac{6.5224}{7} - 0.876 = 0.0557 \ln^2(\mu\text{gr/dl}), \\ r^2 &= \frac{s_{uv}^2}{s_u^2 s_v^2} = \frac{0.0557^2}{0.295 \cdot 0.0109} = 0.9688.\end{aligned}$$

Así pues, el modelo potencial también es muy buen modelo para explicar la relación entre Y y X aunque es un poco mejor el lineal.

3. Como el modelo lineal es un poco mejor que el potencial, hay que hacer la predicción con el modelo lineal. Para ello se calcula la recta de regresión de Y sobre X , que tiene ecuación

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}) = 1.6 + \frac{0.4814}{9.2963}(x - 7.4286) = 0.0518x + 1.2153.$$

Finalmente, la concentración de Y para $x = 8 \mu\text{gr/dl}$ será

$$y = 0.0518 \cdot 8 + 1.2153 = 1.6296 \mu\text{gr/dl}.$$

Ejercicio 2.46. Dar ejemplos de:

1. Dos variables no relacionadas.
2. Dos variables relacionadas de manera creciente.
3. Dos variables relacionadas de manera decreciente.

Ejercicio 2.47. ★ Se ha modificado un gen de una especie de rata para facilitar la metabolización del colesterol en sangre. Para comprobar la efectividad de la modificación genética se tomaron dos grupos de 20 ratas, unas con el gen modificado y las otras no,

y se sometieron a una misma dieta con distintas concentraciones de aceite de palma durante un mes, obteniendo los siguientes resultados:

Cantidad de aceite de palma en gr (igual en las dos muestras)

$$\sum x_i = 640.6467, \sum x_i^2 = 23508.6387.$$

Colesterol en sangre en mg/dl de ratas no modificadas genéticamente

$$\sum y_j = 2945.8545, \sum y_j^2 = 439517.5975, \sum x_i y_j = 98156.0658.$$

Colesterol en sangre en mg/dl de ratas modificadas genéticamente

$$\sum y_j = 2126.5899, \sum y_j^2 = 226824.5373, \sum x_i y_j = 69517.3648.$$

Se pide:

1. ¿En qué grupo se ajusta mejor la recta de regresión del nivel de colesterol sobre la cantidad de aceite de palma?
2. Según el modelo de regresión lineal, ¿qué cantidad de colesterol se espera que tenga una rata modificada genéticamente que toma 25 gr de aceite de palma en la dieta? ¿Y una no modificada?
3. ¿Qué cantidad máxima de aceite de palma hay que suministrar a una rata no modificada genéticamente para que su colesterol no pase de 150 mg/dl? ¿Es fiable la predicción?

Solución

1. Cantidad de aceite de palma: $\bar{x} = 32.0323$ gr, $s_x^2 = 149.3614$ gr².
Ratas no modificadas genéticamente: $\bar{y} = 147.2927$ mg/dl, $s_y^2 = 280.7332$ (mg/dl)², $s_{xy} = 189.6733$ gr·mg/dl y $r^2 = 0.858$.
Ratas modificadas genéticamente: $\bar{y} = 106.3295$ mg/dl, $s_y^2 = 35.265$ (mg/dl)², $s_{xy} = 69.8861$ gr·mg/dl y $r^2 = 0.9273$.
Por tanto, la recta de regresión se ajusta mejor en las ratas modificadas genéticamente ya que el coeficiente de determinación es mayor.
2. Recta de regresión de Y sobre X en Ratas no modificadas genéticamente: $y = 106.615 + 1.2699x$. Predicción: $y(25) = 138.3624$ mg/dl.
Recta de regresión de Y sobre X en ratas modificadas genéticamente: $y = 91.3416 + 0.4679x$. Predicción: $y(25) = 103.0391$ mg/dl.
3. Recta de regresión de X sobre Y en ratas no modificadas genéticamente: $x = -67.4838 + 0.6756y$. Predicción: $x(150) = 33.8615$ gr. La predicción es muy fiable ya que el coeficiente de determinación está próximo a 1.

3 Probabilidad

Ejercicio 3.1. En una estantería en la que hay 3 cajas de un medicamento A y 2 de un medicamento B , se eligen 3 al azar. ¿Cuál es la probabilidad de que se hayan elegido 2 cajas del medicamento A y 1 del B ?

 Solución

36/60.

Ejercicio 3.2. En un laboratorio hay 4 frascos de ácido sulfúrico y 2 de ácido nítrico, y en otro hay 1 frascos de ácido sulfúrico y 3 de ácido nítrico. Se saca al azar un frasco de cada laboratorio. Hallar la probabilidad de que:

1. Los dos frascos sean de ácido sulfúrico.
2. Los dos sean de ácido nítrico.
3. Uno sea de ácido sulfúrico y otro de ácido nítrico.
4. Calcular la probabilidad de estos mismos sucesos si el frasco elegido en el primer laboratorio se introduce en el segundo antes de sacar el frasco de este.

 Solución

1. 4/24.
2. 6/24.
3. 14/24.
4. 8/30, 8/30 y 14/30 respectivamente.

Ejercicio 3.3. Sean A y B sucesos de un mismo espacio muestral tales que: $P(A) = 3/8$, $P(B) = 1/2$, $P(A \cap B) = 1/4$.

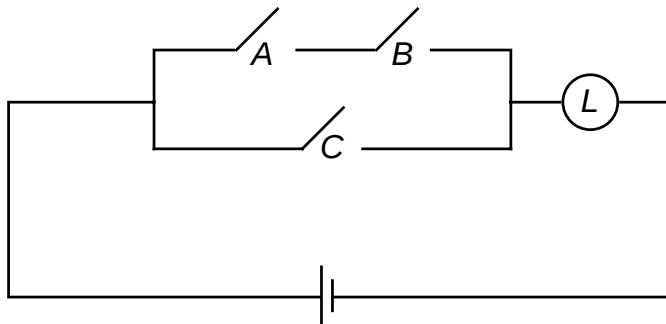
Calcular:

1. $P(A \cup B)$.
2. $P(\overline{A})$ y $P(\overline{B})$.
3. $P(\overline{A} \cap \overline{B})$.
4. $P(A \cap \overline{B})$.
5. $P(A/B)$.
6. $P(A/\overline{B})$.

 Solución

1. $P(A \cup B) = 5/8$.
2. $P(\overline{A}) = 5/8$ y $P(\overline{B}) = 1/2$.
3. $P(\overline{A} \cap \overline{B}) = 3/8$.
4. $P(A \cap \overline{B}) = 1/8$.
5. $P(A/B) = 1/2$.
6. $P(A/\overline{B}) = 1/4$.

Ejercicio 3.4. Dado el siguiente circuito



si la probabilidad de estar cerrado el interruptor A es 0.8, el B 0.9 y el C 0.7, ¿cuál es la probabilidad de que esté encendida la lámpara L ?

Ejercicio 3.5. La probabilidad de contraer hepatitis a partir de una unidad de sangre es 0.01. Un paciente recibe dos unidades de sangre durante su estancia en el hospital. ¿Cuál es la probabilidad de que contraiga hepatitis como consecuencia de ello?

 Solución

0.0199.

Ejercicio 3.6. En un lapso de tiempo una ameba puede morir con probabilidad $1/4$ y dividirse en dos con probabilidad $1/2$. Durante el siguiente lapso de igual duración con cada ameba, independientemente de su origen, ocurre lo mismo. ¿Cuántas amebas y con qué probabilidad pueden existir al final del segundo intervalo de tiempo?

Ejercicio 3.7. ★ Supongamos los sucesos independientes A_1 y A_2 con probabilidades p_1 y p_2 respectivamente, e incompatibles ambos con el suceso A_3 con probabilidad p_3 (A_1 , A_2 y A_3 son sucesos de un mismo espacio muestral). Calcular en función de p_1 , p_2 y p_3 las probabilidades de los siguientes sucesos:

1. $A_1 - (A_2 \cup A_3)$.

2. $\overline{A_1 \cup A_2 \cup A_3}$.
3. $A_1/(A_2 \cup A_3)$.

Ejercicio 3.8. Sean A y B sucesos de un mismo espacio muestral, tales que $P(A) = 0.6$ y $P(A \cup B) = 0.9$. Calcular $P(B)$ si:

1. A y B son incompatibles.
2. A y B son independientes.

 Solución

1. $P(B) = 0.3$.
2. $P(B) = 0.75$.

Ejercicio 3.9. ★ En un estudio sobre el tabaco, se informa que el 40% de los fumadores tiene un padre fumador, el 25% tiene una madre fumadora, y el 52% tiene al menos uno de los dos padres fumadores. Se elige una persona fumadora al azar. Calcular:

1. Probabilidad de que la madre sea fumadora si lo es el padre.
2. Probabilidad de que la madre sea fumadora si no lo es el padre.
3. ¿Son independientes el tener padre fumador y el tener madre fumadora?

 Solución

Llamando PF al suceso que consiste en tener un padre fumador y MF a tener una madre fumadora, del enunciado se tiene que $P(PF) = 0.4$, $P(MF) = 0.25$ y $P(PF \cup MF) = 0.52$.

1. Nos piden la probabilidad de tener madre fumadora condicionada por tener padre fumador. Según la definición de probabilidad condicionada se tiene

$$P(MF/PF) = \frac{P(MF \cap PF)}{P(PF)}.$$

A su vez, la probabilidad de la intersección de tener madre y padre fumadores puede calcularse a partir de la fórmula de la unión:

$$P(MF \cup PF) = P(MF) + P(PF) - P(MF \cap PF) \Leftrightarrow P(MF \cap PF) = P(MF) + P(PF) - P(MF \cup PF)$$

de modo que la probabilidad condicionada queda

$$P(MF/PF) = \frac{P(MF \cap PF)}{P(PF)} = \frac{0.13}{0.4} = 0.33.$$

2. Ahora nos piden la probabilidad de tener madre fumadora condicionada por tener padre no fumador. De nuevo, según la definición de probabilidad condicionada se tiene

$$P(MF/\overline{PF}) = \frac{P(MF \cap \overline{PF})}{P(\overline{PF})}.$$

Como el suceso $MF \cap \overline{PF}$ es la diferencia del suceso MF y el suceso PF , su probabilidad se puede calcular como

$$P(MF \cap \overline{PF}) = P(MF) - P(MF \cap PF) = 0.25 - 0.13 = 0.12,$$

de modo que la probabilidad condicionada queda

$$P(MF/\overline{PF}) = \frac{P(MF \cap \overline{PF})}{P(\overline{PF})} = \frac{0.12}{1 - 0.4} = 0.2.$$

3. Como $P(MF) = 0.25 \neq 0.33 = P(MF/PF)$ se puede concluir que los sucesos no son independientes.

Ejercicio 3.10. El tétanos es mortal en el 70% de los casos. Si tres personas contraen el tétanos, ¿Cuál es la probabilidad de que mueran al menos dos de los tres?

 Solución

0.784.

Ejercicio 3.11. ★ Un equipo de atención primaria de salud realiza un estudio de la población, para evaluar la incidencia de hipertensión e hipercolesterolemia. Para ello analizan a 1000 personas de dicha población, seleccionadas aleatoriamente, encontrándose que 180 presentan hipertensión, 140 hipercolesterolemia y 800 ninguna de ambas. Se pide calcular la probabilidad de que una persona tomada al azar.

1. Presente ambas enfermedades.
2. Presente hipertensión si no presenta hipercolesterolemia.

 Solución

Llamando HT a tener hipertensión y HC a tener hipercolesterolemia:

1. $P(HT \cap HC) = 0.12$.
2. $P(HT/\overline{HC}) = 0.0698$.

Ejercicio 3.12. ★ Tras observar los resultados de la prueba de selectividad se sabe que el 40% de los alumnos aprueba el examen de Matemáticas, el 30% el examen de Física y el 55% suspenden los dos. Si se elige un alumno al azar, calcular:

1. Probabilidad de que haya aprobado al menos uno de los dos exámenes.
2. Probabilidad de que haya aprobado Matemáticas si ha aprobado Física.
3. Probabilidad de que haya aprobado Física si ha suspendido Matemáticas.
4. ¿Son independientes aprobar Matemáticas y aprobar Física?

 Solución

Llamando M al suceso correspondiente a aprobar Matemáticas y F a aprobar Física:

1. $P(M \cup F) = 0.45$.
2. $P(M/F) = 0.83$.
3. $P(F/\bar{M}) = 0.08$.
4. No son independientes.

Ejercicio 3.13. ★ La probabilidad de que una lesión A se reproduzca es $4/5$, la de que se reproduzca otra lesión B es $1/2$, y la de que ambas se reproduzcan $1/3$. Hallar la probabilidad de que:

1. Sólo se reproduzca la lesión B .
2. Al menos una se reproduzca.
3. Se reproduzca la lesión B si se ha reproducido la A .
4. Se reproduzca la lesión B si no se reproduce la lesión A .

 Solución

Llamemos A y B a los sucesos consistentes en que se reproduzcan las respectivas lesiones. Según el enunciado tenemos $P(A) = 4/5$, $P(B) = 1/2$ y $P(A \cap B) = 1/3$.

1. Para que sólo se reproduzca la lesión B debe cumplirse el suceso $B - A$, es decir $B \cap \bar{A}$.

$$P(B \cap \bar{A}) = P(B) - P(A \cap B) = 1/2 - 1/3 = 1/6.$$

2. Para que se reproduzca al menos una lesión debe cumplirse el suceso $A \cup B$.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 4/5 + 1/2 - 1/3 = 29/30.$$

3. El suceso consistente en que se reproduzca la lesión B si se ha reproducido la A es B/A .

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{4/5} = 5/12.$$

4. El suceso consistente en que se reproduzca la lesión B si no se ha reproducido la A es $B/\bar{A}$.

$$P(B/\bar{A}) = \frac{P(\bar{A} \cap B)}{P(\bar{A})} = \frac{1/6}{1 - 4/5} = 5/6.$$

Ejercicio 3.14. A partir de una investigación realizada, se sabe que el 10% de las personas de 50 años sufren un tipo particular de artritis. Se ha desarrollado un procedimiento para detectar esta enfermedad, y por las pruebas realizadas se observa que si se aplica el procedimiento a individuos que padecen la enfermedad, da positivo en el 85% de los casos, mientras que si se aplica a individuos sanos, da positivo en el 4% de los casos. Se pide:

1. Calcular la probabilidad de que realizado el procedimiento a una persona, el resultado sea positivo.
2. Si el resultado de aplicar el procedimiento a una persona ha sido positivo, ¿Cuál es la probabilidad de que padezca la enfermedad?

Ejercicio 3.15. ★ En un experimento aleatorio se pueden dar dos sucesos A y B , y se sabe que $P(B) = 0.4$, $P(A/B) = 0.3$ y $P(A/\bar{B}) = 0.2$. Calcular las siguientes probabilidades:

1. $P(A)$.
2. $P(\bar{A} \cap \bar{B})$.
3. $P(\bar{A} \cup \bar{B})$.

Solución

1. $P(A) = 0.24$.
2. $P(\bar{A} \cap \bar{B}) = 0.52$.
3. $P(\bar{A} \cup \bar{B}) = 0.88$.

Ejercicio 3.16. ★ En un servicio clínico digestivo se sabe que, de cada 1000 pacientes con dolor de estómago, 700 presentan gastritis, 200 presentan úlcera y 100 presentan cáncer. En el análisis de la sintomatología gástrica, se ha comprobado que las probabilidades de presentar vómitos son 0.3 en el caso de gastritis, 0.6 en el caso de úlcera y 0.9 en el caso de cáncer. Llega un nuevo paciente con dolor de estómago que, además, presenta vómitos. ¿Qué diagnosticaríamos?

 Solución

Llamando G a tener gastritis, U a tener úlcera, C a tener cáncer y V a tener vómitos, $P(G/V) = 0.5$, $P(U/V) = 0.286$ y $P(C/V) = 0.214$, de modo que se diagnosticaría gastritis.

Ejercicio 3.17. ★ Un estudiante se somete a un examen de tipo test en el que cada pregunta tiene 3 respuestas posibles. El estudiante se sabe el 40% de las preguntas, y el resto las contesta al azar. Se elige al azar una pregunta. ¿Qué probabilidad hay de que no la supiera si la contestó correctamente?

 Solución

1/3.

Ejercicio 3.18. ★ Un test diseñado para diagnosticar el cáncer de cuello uterino da resultado positivo en el 10% de los casos en los que no existe la enfermedad, y da negativo en el 5% de los casos en los que sí que existe la enfermedad.

Se sabe que en una cierta población de mujeres, el 4% padece dicha enfermedad. Si una mujer elegida aleatoriamente se somete al test, y da positivo, ¿qué probabilidad hay de que padezca la enfermedad?

Ejercicio 3.19. ★ Se ha desarrollado un nuevo test diagnóstico para detectar el síndrome de Down en niños recién nacidos, con un sensibilidad del 80% y una especificidad del 90%. Si en una determinada población en la que hay un 1% de recién nacidos con el síndrome, al aplicarle el test a un niño, da positivo, ¿cuál es la probabilidad de que tenga el síndrome? ¿le diagnosticarías la enfermedad? ¿Cuál debería ser la especificidad mínima del test para diagnosticar el síndrome en el caso de dar positivo?

Nota: La *sensibilidad* de un test diagnóstico es la proporción de personas con la enfermedad que tienen un resultado positivo en el test, mientras que la *especificidad* del test es la proporción de personas sin la enfermedad que tienen un resultado negativo en el test.

 Solución

Llamando S a tener el síndrome de Down y $+$ a que el test de positivo, $P(S/+) = 0.0748$ y $P(\bar{S}/+) = 0.9252$, de modo que no se diagnosticaría el síndrome al ser más probable que no lo tenga.

La especificidad mínima para que el test diagnostique el síndrome es $P(-/\bar{S}) = 0.9919$.

Ejercicio 3.20. ★ En un estudio se han probado tres tipos de tratamientos A , B y C contra una determinada enfermedad. De los pacientes participantes en el estudio, el 50% fueron tratados con el tratamiento A , el 30% con el B y el 20% con el C . Posteriormente se observaron los pacientes que sanaron y los que tuvieron algún efecto secundario, según se muestra en la siguiente tabla:

Tratamiento	Sanados	Con efectos secundarios
A	86%	12%
B	92%	14%
C	81%	6%

Se pide:

1. Si se selecciona un enfermo al azar, ¿cuál es la probabilidad de que haya sanado? ¿Y de que haya tenido algún efecto secundario?
2. Si un enfermo ha sanado, ¿qué tratamiento es más probable que haya recibido? ¿Y si en vez de decirnos que ha sanado nos dicen que no ha tenido efectos secundarios?
3. Si en total hay un 8% pacientes que no sanaron pero que tampoco tuvieron efectos secundarios, ¿cuál es la probabilidad de que un enfermo se haya curado sin tener efectos secundarios?

Solución

Llamado S a sanar y E a tener efectos secundarios:

1. $P(S) = 0.868$ y $P(E) = 0.114$.
2. $P(A/S) = 0.495$, $P(B/S) = 0.318$ y $P(C/S) = 0.187$.
 $P(A/\bar{E}) = 0.497$, $P(B/\bar{E}) = 0.291$ y $P(C/\bar{E}) = 0.212$.
 En ambos casos el tratamiento más probable es el A .
3. $P(S \cap \bar{E}) = 0.806$.

Ejercicio 3.21. En una plantación hay tres variedades de uvas A , B y C en un 45%, 35% y 20% respectivamente. Se sabe que el porcentaje de vides infectadas por una bacteria es del 5% en la variedad A , del 12% en la variedad B y del 18% en la C . Si se toman uvas de una vid elegida al azar, ¿cuál es la probabilidad de que estén infectadas? Si las uvas cogidas de una vid están infectadas, ¿de qué variedad es más probable que sean?

Ejercicio 3.22. Se sabe que el 25% de las plantas afectadas por un vertido tóxico están contaminadas. Se ha desarrollado un procedimiento para detectar la contaminación, y por las pruebas realizadas se observa que si se aplica el procedimiento a plantas contaminadas, da positivo en el 96% de los casos, mientras que si se aplica plantas no contaminadas, da positivo en el 8% de los casos. Se pide:

1. Calcular la probabilidad de que al aplicar el procedimiento a una planta elegida al azar, el resultado sea positivo.
2. Si el resultado de aplicar el procedimiento a una planta ha sido positivo, ¿Cuál es la probabilidad de que esté contaminada?

Ejercicio 3.23. Un test para detectar la contaminación del agua tiene una sensibilidad del 97.5% y una especificidad del 95%. Se pide:

1. ¿Qué relación existe entre la probabilidad de que el agua esté contaminada y la de que el test de positivo?
2. Si se aplica el test a muestras de 20 embalses y resulta positivo en 5 de las muestras, según la relación anterior, ¿qué porcentaje de embalses se estima que están contaminados?
3. Suponiendo que el porcentaje real de embalses contaminados es del 12%, ¿cuál es la probabilidad de que un embalse esté contaminado si el resultado del test es negativo?

Ejercicio 3.24. ★ El dolor intenso sin derrame en una zona concreta de la articulación de la rodilla es síntoma de esguince en el Ligamento Lateral Externo de la misma (L.L.E.). Si los esguinces en dicho ligamento se clasifican como: de grado 1, cuando hay simple distensión, que se presenta en un 60% de los casos; de grado 2, cuando hay ruptura parcial, que se presenta en un 30% de los casos; y de grado 3, cuando hay ruptura total, que se presenta en un 10%. Y teniendo en cuenta que el síntoma se presenta en un 80% de los que tienen el esguince de grado 1, en un 90% de los de grado 2, y en un 100% de los de grado 3:

1. Si una persona se produce un esguince de L.L.E., ¿cuál es la probabilidad total de que padezca dolor intenso sin derrame?
2. Si una persona llega a una consulta con dolor intenso sin derrame en la zona adecuada de la rodilla, ¿cuál sería el diagnóstico?
3. De un total de 10000 personas analizadas con dolor intenso sin derrame en la zona adecuada de la rodilla, ¿cuántas se espera que hayan sufrido un esguince de grado 1? ¿Y de grado 2? ¿Y de grado 3?
4. Si mantenemos iguales el resto de probabilidades dadas como dato, ¿cuáles deben ser las probabilidades de esguince de grado 2 y de grado 3 para que la probabilidad de esguince de grado 2 si se padece el síntoma sea igual a la de grado 3 si se padece el síntoma?

Ejercicio 3.25. ★ En una población se ha vacunado a la tercera parte de los individuos contra la gripe. Trascurrido el invierno, se comprueba que la probabilidad de estar vacunado si se tiene la gripe es 0.2, y que el 10% de los vacunados tuvieron gripe.

1. ¿Cuál fue la incidencia de la epidemia de gripe?
(Nota: La incidencia de una epidemia es la probabilidad de personas infectadas).

2. ¿Cuál es la probabilidad de que una persona no vacunada contraiga la gripe?
3. ¿Se puede afirmar que la vacuna tiene alguna eficacia?

Solución

Llamando G al suceso consistente en tener la gripe y V a estar vacunado:

1. $P(G) = 1/6$.
2. $P(G/\bar{V}) = 0.2$.
3. Si resulta eficaz, aunque poco.

Ejercicio 3.26. ★ Una zona de un Parque Nacional se ha repoblado con 10000 árboles, de los cuales 5000 han sido pinos, 4000 encinas y 1000 fresnos. Al cabo de un año se hace recuento y se observa que han sobrevivido 4500 pinos, 2000 encinas y 900 fresnos. Suponiendo que los resultados obtenidos son aplicables a una nueva zona del Parque Nacional que se va a repoblar:

1. ¿Cuál es la probabilidad de que un árbol que se va a plantar muera durante el primer año?
2. ¿Cuál es la probabilidad de que muera y sea pino?
3. Si sabemos que el árbol ha vivido, ¿cuál es la probabilidad de que sea un fresno?
4. Si sabemos que el árbol ha muerto, ¿cuál es la probabilidad de que sea una encina?

Ejercicio 3.27. Sean A , B y C sucesos arbitrarios de un experimento aleatorio. Se consideran los siguientes sucesos:

$$\begin{aligned} E_1 &= \{\text{al menos dos de los sucesos ocurren}\}, \\ E_2 &= \{\text{exactamente dos de los sucesos ocurren}\}, \\ E_3 &= \{\text{al menos uno de los sucesos ocurre}\}, \\ E_4 &= \{\text{exactamente uno de los sucesos ocurre}\}, \\ E_5 &= \{\text{no m' a s de dos sucesos ocurren}\}. \end{aligned}$$

Expresar E_1 , E_2 , E_3 , E_4 y E_5 en función de A, B y C .

Ejercicio 3.28. La probabilidad de que un hombre siga vivo dentro de 25 años es $3/5$, y la de que su esposa lo esté es $2/3$. Si ambos sucesos son independientes, hallar la probabilidad de que en ese momento:

1. Ambos estén vivos.
2. Sólo el hombre viva.
3. Sólo viva la esposa.
4. Al menos uno esté vivo.
5. Viva el hombre si vive la esposa.
6. Viva el hombre si no vive la esposa.

Ejercicio 3.29. La probabilidad de que un hombre siga vivo dentro de 25 años es $3/5$, la de que su esposa lo esté es $2/3$ y la de que ambos vivan es $1/2$. Hallar la probabilidad de que en ese momento:

1. Sólo el hombre viva.
2. Sólo viva la esposa.
3. Al menos uno esté vivo.
4. Viva el hombre si vive la esposa.
5. Viva el hombre si no vive la esposa.

Ejercicio 3.30. El 40% de los alumnos que han suspendido matemáticas deciden preparar el examen extraordinario en una academia, el 35% decide asistir a las clases que organiza la propia universidad y el resto lo prepara por su cuenta. De experiencias previas, se sabe que el 30% de los alumnos de academia consigue aprobar el examen, de los alumnos que asisten a la universidad aprueban el 45%, y del resto de los alumnos aprueban el 25%. Se pide:

1. ¿Cuál es la probabilidad de que un alumno elegido al azar apruebe el examen?
2. Si un alumno ha suspendido, ¿cuál es la probabilidad de que haya asistido a una academia?

Ejercicio 3.31. Se dispone de dos urnas, la primera con 10 bolas blancas y 6 bolas negras, y la segunda con 5 bolas rojas, 8 bolas azules y 3 bolas verdes. Construir el espacio muestral del experimento que consiste en sacar una bola de cada urna, y del experimento que consistiría en sacar dos bolas de cada urna.

Ejercicio 3.32. Una ciudad se abastece de agua de tres pantanos: A en un 50%, B en un 40% y C en un 10%. El agua del pantano A produce diarrea en un 3% de los casos, la del B en un 1% y la del C en un 7%. ¿Cual es la probabilidad de que una persona que beba agua en la ciudad tenga diarrea? Si una persona presenta diarrea por causa del agua, ¿de qué pantano es más probable que provenga dicha agua?

Ejercicio 3.33. Construir el espacio muestral de los siguientes experimentos aleatorios:

1. Seleccionar una persona al azar y medir su sexo y si es fumadora o no.
2. Seleccionar una persona al azar y medir su grupo sanguíneo y si es fumadora o no.
3. Seleccionar una persona al azar y medir su sexo, grupo sanguíneo y se es fumadora o no.

Ejercicio 3.34. ★ Un test diagnóstico permite detectar la presencia de un tóxico en el pescado con una sensibilidad del 96% y una especificidad del 92%. Si un vertido ha afectado a las tres quintas partes del pescado de un lago, se pide:

1. ¿Cuál es la probabilidad de que al aplicar el test a un pescado elegido al azar de negativo?

2. Si el test da negativo en un pescado, ¿es más probable que esté contaminado o que no lo esté?
3. ¿Podríamos decir lo mismo en un lago donde el 90% de los pescados hayan sido contaminados?

Ejercicio 3.35. Se han realizado tres campañas A , B y C de sensibilización para el reciclaje de vidrio y papel. La campaña A se ha aplicado al 50% de la población, la B al 20%, y al resto se la ha aplicado la C . Posteriormente se observaron las personas que acabaron reciclando habitualmente el vidrio y el papel, obteniendo los siguientes resultados

A	56%	42%
B	52%	46%
C	71%	38%

Se pide:

1. Si se selecciona un individuo al azar, ¿cuál es la probabilidad de que recicle vidrio? ¿Y de que recicle papel?
2. Si un individuo recicla vidrio, ¿qué campaña es más probable que haya recibido? ¿Y si en vez de decirnos que recicla vidrio nos dicen que recicla papel?
3. Si en total hay un 18% de personas que no reciclan ni vidrio ni papel, ¿cuál es la probabilidad de que una persona recicle vidrio si no recicla papel?

Ejercicio 3.36. ★ Una enfermedad se trata con 3 fármacos diferentes, A , B y C . El fármaco A se aplica a un 30% de los hombres y a un 50% de las mujeres; el fármaco B se aplica a un 60% de los hombres y a un 20% de las mujeres; y el fármaco C a un 10% de los hombres y al 30% de las mujeres. Además, se sabe que del total de afectados por la enfermedad el 40% son hombres y el 60% mujeres.

Teniendo en cuenta que cada persona enferma es tratada únicamente con un fármaco, se pide:

1. Si tenemos un paciente con esa enfermedad, ¿cuál es el fármaco que con más probabilidad se le aplicaría?
2. Si tenemos un paciente tratado con B , ¿cuál es la probabilidad de que sea hombre?
3. Si tenemos un paciente que no ha sido tratado con C , ¿cuál es la probabilidad de que se mujer?

Ejercicio 3.37. ★ Una enfermedad se trata con 3 medicamentos diferentes: A en un 50% de los casos, B en un 30% y C en un 20%, todo ello independientemente de si se es hombre o mujer. Si sabemos que el medicamento A produce efectos secundarios en un 5% de los hombres y en un 10% de las mujeres, el B en un 15% de los hombres y en un 5% de las mujeres, y el C en un 8% de los hombres y en un 13% de las mujeres, se pide:

1. ¿En qué colectivo resulta más probable que haya efectos secundarios, en hombres o en mujeres? Justificar adecuadamente la respuesta.
2. Calcular la probabilidad de que un hombre que presenta efectos secundarios haya sido tratado con C, y la de que una mujer que no los presenta haya sido tratada con A.
3. Si en total de enfermos hay un 65% de hombres y un 35% de mujeres, ¿qué probabilidad hay de que un enfermo que no presenta efectos secundarios sea mujer?

Solución

Según el enunciado, si llamamos A al suceso haber sido tratado con el medicamento A, y de forma similar a los sucesos B y C , entonces sus respectivas probabilidades son: $P(A) = 0.5$, $P(B) = 0.3$ y $P(C) = 0.2$, y todo ello tanto para hombres como para mujeres.

Además, también se nos dan como datos las sucesivas probabilidades condicionadas de que se produzcan efectos secundarios al tomar cada uno de los fármacos, tanto en hombres como en mujeres. Si llamamos EH al suceso efectos secundarios en hombres y EM al suceso efectos secundarios en mujeres, entonces los datos del problema son: $P(EH/A) = 0.05$, $P(EM/A) = 0.10$, $P(EH/B) = 0.15$, $P(EM/B) = 0.05$, $P(EH/C) = 0.08$, y $P(EM/C) = 0.13$. De lo anterior pueden deducirse fácilmente las probabilidades de sus complementarios: $P(\overline{EH}/A) = 0.95$, $P(\overline{EM}/A) = 0.90$, $P(\overline{EH}/B) = 0.85$, $P(\overline{EM}/B) = 0.95$, $P(\overline{EH}/C) = 0.92$, y $P(\overline{EM}/C) = 0.87$.

1. Teniendo en cuenta las probabilidades anteriores y aplicando el teorema de probabilidad total:

$$\begin{aligned} P(EH) &= P(A)P(EH/A) + P(B)P(EH/B) + P(C)P(EH/C) = \\ &= 0.5 \cdot 0.05 + 0.3 \cdot 0.15 + 0.2 \cdot 0.08 = 0.086, \end{aligned}$$

$$\begin{aligned} P(EM) &= P(A)P(EM/A) + P(B)P(EM/B) + P(C)P(EM/C) = \\ &= 0.5 \cdot 0.1 + 0.3 \cdot 0.05 + 0.2 \cdot 0.13 = 0.091. \end{aligned}$$

Por lo tanto, resulta más probable que sean las mujeres las que tengan efectos secundarios.

2. Según el teorema de Bayes, la probabilidad de que un hombre que presenta efectos secundarios haya sido tratado con C es:

$$P(C/EH) = \frac{P(C)P(EH/C)}{P(EH)} = \frac{0.2 \cdot 0.08}{0.086} = 0.186.$$

De manera muy parecida, la probabilidad de que una mujer que no presenta efectos secundarios haya sido tratada con A vale:

$$P(A/\overline{EM}) = \frac{P(A)P(\overline{EM}/A)}{P(\overline{EM})} = \frac{0.5 \cdot 0.9}{1 - 0.091} = 0.495.$$

3. Si en el total de enfermos hay un 65% de hombres ($P(H) = 0.65$) y un 35% de mujeres ($P(M) = 0.35$), y teniendo en cuenta que ya hemos calculado en el primer apartado las probabilidades de padecer efectos secundarios si se es hombre ($P(EH) = 0.086$) y de efectos secundarios si se es mujer ($P(EM) = 0.091$), aplicando el teorema de Bayes:

$$P(M/\overline{E}) = \frac{P(M)P(\overline{EM})}{P(H)P(\overline{EH}) + P(M)P(\overline{EM})} = \frac{0.35 \cdot 0.909}{0.65 \cdot 0.914 + 0.35 \cdot 0.909} = 0.349.$$

Ejercicio 3.38. Se sabe que los grupos sanguíneos en una determinada población se distribuyen con las siguientes frecuencias:

$$0 : 30\% \quad A : 45\% \quad B : 18\% \quad AB : 7\%$$

Por otro lado, también se sabe que la octava parte de los individuos del grupo 0 tienen RH negativo, así como la cuarta parte del grupo A, la mitad del grupo B, y la tercera parte del grupo AB.

Se pide:

1. ¿Cuál es la probabilidad de que un individuo elegido al azar sea del tipo A y tenga RH positivo?
2. ¿Cuál es la probabilidad de que un individuo elegido al azar tenga RH negativo o sea del grupo universal?
3. Si un individuo elegido al azar tiene RH positivo, ¿cuál es la probabilidad de que pertenezca al grupo B?

Solución

Llamando 0 a tener grupo 0, A a tener grupo A, B a tener grupo B, AB a tener grupo AB, + a tener RH positivo y - a tener RH negativo, del enunciado se tiene:

$$\begin{array}{lll} P(0) = 0.3 & P(-/0) = 1/8 & P(+/0) = 1 - 1/8 = 7/8, \\ P(A) = 0.45 & P(-/A) = 1/4 & P(+/A) = 1 - 1/4 = 3/4, \\ P(B) = 0.18 & P(-/B) = 1/2 & P(+/B) = 1 - 1/2 = 1/2, \\ P(AB) = 0.07 & P(-/AB) = 1/3 & P(+/AB) = 1 - 1/3 = 2/3. \end{array}$$

1. La probabilidad que nos piden es la de la intersección del suceso A y el suceso +, que vale

$$P(A \cap +) = P(A)P(+/A) = 0.45 \frac{3}{4} = 0.34.$$

2. La probabilidad que nos piden es la de la unión del suceso $-$ con el suceso 0 , que vale

$$P(- \cup 0) = P(-) + P(0) - P(- \cap 0).$$

A su vez, para calcular la probabilidad de tener RH negativo podemos utilizar el teorema de la probabilidad total, partiendo de que los grupos sanguíneos forman un sistema completo de suceso, de manera que

$$P(-) = P(0)P(-/0) + P(A)P(-/A) + P(B)P(-/B) + P(AB)P(-/AB) = 0.3 \frac{1}{8} + 0.45 \frac{1}{4} + 0.18 \frac{1}{2} + 0.07$$

Por otro lado,

$$P(- \cap 0) = P(0)P(-/0) = 0.3 \frac{1}{8} = 0.04.$$

Así que, finalmente la probabilidad que nos piden es

$$P(- \cup 0) = P(-) + P(0) - P(- \cap 0) = 0.26 + 0.3 - 0.04 = 0.53.$$

3. Ahora nos piden la probabilidad del suceso B condicionado por el suceso $+$, que puede calcularse utilizando el teorema de Bayes:

$$P(B/+)=\frac{P(B)P(+/B)}{P(+)}=\frac{P(B)P(+/B)}{1-P(-)}=\frac{0.18\cdot 0.5}{1-0.26}=0.12.$$

Ejercicio 3.39. Sabemos que el test de Bender para detectar alteraciones cerebrales tiene una sensibilidad del 88% y una especificidad del 96%. Por otro lado, sabemos que en una población la probabilidad de que un individuo elegido al azar tenga alteraciones cerebrales y además de positivo en el test es 0.08. Se pide:

1. Calcular el porcentaje de personas que tienen alteraciones cerebrales en la población.
2. ¿Es efectivo el test en esta población para detectar la ausencia de alteraciones cerebrales?

Solución

Llamando E a presentar alteraciones cerebrales y $+$ y $-$ a que el test de positivo y negativo respectivamente:

1. $P(E) = 0.0901$, es decir, un 9.09%.
2. $P(\bar{E}/-) = 0.99$, lo cual indica que es muy efectivo para detectar la ausencia de alteraciones cerebrales.

Ejercicio 3.40. Los estudios epidemiológicos indican que el 20% de las personas mayores sufren un deterioro neuropsicológico. También se sabe que la tomografía axial computerizada (TAC) puede detectar ese trastorno en el 90% de los que lo padecen, pero que también puede diagnosticarlo en el 5% de las personas que no lo tienen. Si se toma una persona al azar y el TAC da positivo, ¿cuál es la probabilidad de que realmente esté enfermo?

Solución

La tomografía axial computerizada es un test diagnóstico que se utiliza para detectar el deterioro neuropsicológico, así que, si llamamos E a sufrir el deterioro neuropsicológico, entonces $P(E) = 0.2$ ya que el 20% de la población tienen esta enfermedad, y si llamamos $+$ al que el TAC da positivo, entonces $P(+|E) = 0.9$ pues el test detecta la enfermedad en el 90% de los pacientes enfermos, y $P(+|\bar{E}) = 0.05$ pues el test también la diagnostica en el 5% de los pacientes sin esta enfermedad. Para saber cuál es la probabilidad de que una persona en la que el test da positivo tenga la enfermedad, necesitamos calcular la probabilidad de estar enfermo, condicionado por que el test da positivo, y para calcularla podemos utilizar el teorema de Bayes, ya que E y $\bar{E}$ forman un sistema completo de sucesos:

$$P(E|+) = \frac{P(E)P(+|E)}{P(+)} = \frac{P(E)P(+|E)}{P(E)P(+|E) + P(\bar{E})P(+|\bar{E})} = \frac{0.2 \cdot 0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.05} = 0.82.$$

Ejercicio 3.41. ★ Un fisioterapeuta dispone de dos técnicas A y B para rehabilitar una determinada lesión. Se sabe que dicha lesión es 3 veces más frecuentes en personas mayores de 30 años que en personas jóvenes. También se sabe que en las personas jóvenes, la técnica A es efectiva en el 30% de los casos y la técnica B en el 60%, mientras que en personas mayores, la técnica A es efectiva en el 50% de los casos y la técnica B en el 70%. Si, independiente de la edad, aplica la técnica A el mismo número de veces que la técnica B , se pide:

1. ¿Cuál es la probabilidad de que una persona joven elegida al azar se cure? ¿y la de que se cure una persona mayor?

- ¿Cuál es la probabilidad de que una persona cualquiera se cure?
- Si una persona mayor no se ha curado, ¿cuál es la probabilidad de que fuese tratado con la técnica A ?

Solución

Llamemos J al suceso consistente en que una persona con la lesión sea joven. Según el enunciado $P(\bar{J}) = 3P(J)$. Como además, se cumple que $P(J) + P(\bar{J}) = 1$, tenemos que $P(J) + 3P(J) = 4P(J) = 1$, con lo que $P(J) = 1/4 = 0.25$ y $P(\bar{J}) = 0.75$. Por otro lado, llamemos C al suceso consistente en curarse, y A y B a los sucesos consistentes en aplicar las respectivas técnicas de rehabilitación. Según el enunciado, la técnica A cura al 30% de los jóvenes con lesión, lo que, traducido a probabilidad, puede escribirse como $P(C/J \cap A) = 0.3$. Del mismo modo, del resto del enunciado se deduce $P(C/J \cap B) = 0.6$, $P(C/\bar{J} \cap A) = 0.5$ y $P(C/\bar{J} \cap B) = 0.7$. Con estos datos, pasamos a contestar a las preguntas.

- El suceso consistente en que una persona joven se cure puede expresarse como C/J . Para calcular su probabilidad podemos utilizar el teorema de la probabilidad total aprovechando que A/J y B/J forma un sistema completo de sucesos.

$$P(C/J) = P(A/J)P(C/J \cap A) + P(B/J)P(C/J \cap B).$$

Como ambas técnicas se aplican el mismo número de veces independientemente de la edad, tenemos que $P(A/J) = 0.5$ y $P(B/J) = 0.5$, y por tanto,

$$P(C/J) = 0.5 \cdot 0.3 + 0.5 \cdot 0.6 = 0.45.$$

Del mismo modo, la probabilidad de que se cure una persona mayor es

$$P(C/\bar{J}) = P(A/\bar{J})P(C/\bar{J} \cap A) + P(B/\bar{J})P(C/\bar{J} \cap B) = 0.5 \cdot 0.5 + 0.5 \cdot 0.7 = 0.6.$$

- De nuevo, para calcular la probabilidad de que una persona lesionada se cure, podemos utilizar el teorema de la probabilidad total, aprovechando que J y $\bar{J}$ también forman un sistema completo de sucesos. Así, tenemos

$$P(C) = P(J)P(C/J) + P(\bar{J})P(C/\bar{J}) = 0.25 \cdot 0.45 + 0.75 \cdot 0.6 = 0.5625.$$

- Por último, para calcular la probabilidad de que una persona mayor fuese tratada con la técnica A si se ha curado, podemos utilizar el teorema de Bayes:

$$P(A/\bar{J} \cap \bar{C}) = \frac{P(A/\bar{J})P(\bar{C}/\bar{J} \cap A)}{P(\bar{C}/\bar{J})} = \frac{P(A/\bar{J})(1 - P(C/\bar{J} \cap A))}{1 - P(C/\bar{J})} = \frac{0.5 \cdot (1 - 0.5)}{1 - 0.6} = 0.625.$$

Ejercicio 3.42. ★ Para comprobar la eficacia de un test diagnóstico se lleva a cabo una experiencia cuyos resultados se recogen en la siguiente tabla:

	Test +	Test -
Enfermos	4680	120
No Enfermos	80	2020

Calcular para dicho test:

1. Las probabilidades de verdadero negativo, verdadero positivo, falso negativo y falso positivo.
2. Los valores predictivos, tanto el positivo como el negativo.
3. La probabilidad de diagnóstico acertado.

Ejercicio 3.43. ★ Para diagnosticar una misma enfermedad se utilizan dos test A y B completamente independientes. Si la prevalencia de la enfermedad en una población es de un 2%, la sensibilidad de A es de un 95%, la sensibilidad de B es de un 97%, la especificidad de A es de un 90%, y la de B de un 85%, calcular:

1. El valor predictivo positivo del test A .
2. La probabilidad de que, aplicados ambos a un individuo cualquiera de la población, alguno de los test dé positivo.
3. La probabilidad de que, aplicados ambos a un individuo cualquiera de la población, los dos den diagnóstico erróneo.

Ejercicio 3.44. ★ Al aplicar un test diagnóstico a una población se obtuvieron un 1% de personas enfermas en las que el test dio negativo, un 2% de personas sanas en las que el test dio positivo, y un 90% de personas sanas en las que el test dio negativo. Se pide:

1. Calcular la prevalencia de la enfermedad.
2. Calcular la sensibilidad del test.
3. Calcular la especificidad del test.

Solución

El espacio muestral del experimento correspondiente a la aplicación del test diagnóstico aparece en el siguiente árbol

Como la probabilidad el espacio muestral es 1, se puede deducir que la probabilidad que falta $P(E, +) = 1 - 0.01 - 0.02 - 0.9 = 0.07$.

1. La prevalencia de la enfermedad es

$$P(E) = P(E, +) + P(E, -) = 0.07 + 0.01 = 0.08,$$

es decir, un 8% de las personas de la población tienen la enfermedad.

2. La sensibilidad el test es

$$P(+/E) = \frac{P(E \cap +)}{P(E)} = \frac{0.07}{0.08} = 0.875.$$

3. La especificidad del test es

$$P(-/\bar{E}) = \frac{P(\bar{E} \cap -)}{P(\bar{E})} = \frac{0.90}{1 - 0.08} = 0.9783.$$

Ejercicio 3.45. ★ Según la clasificación de la New York Heart Association, el grado funcional de insuficiencia cardiaca se clasifica en 4 categorías dependiendo del esfuerzo físico para que se produzca disnea (dificultad respiratoria o falta de aire):

- A la categoría A pertenecen los pacientes en los que la disnea se produce sólo en niveles de esfuerzo altos.
- A la categoría B pertenecen los que la disnea se produce en niveles de esfuerzo medianos.
- A la categoría C pertenecen los que la disnea se produce en niveles de esfuerzo pequeños.
- A la categoría D pertenecen los que la disnea se produce incluso en reposo.

En un hospital se está investigando la evolución en el grado funcional de insuficiencia cardíaca como consecuencia de un tipo determinado de intervención en el corazón. Para los pacientes en los que se procedería a realizar la intervención, se observó que el 10% pertenecían a la categoría A , el 20% a la categoría B , el 30% a la categoría C y el 40% a la D . Después de la intervención todos los pacientes de la categoría A siguieron en A ; el 50% de los de B pasó a A y el resto siguió en B ; el 30% de los de C pasó a A , el 40% pasó a B y el resto se quedó en C ; el 10% de los de D pasó a A , el 30% pasó a B , el 40% pasó a C y el resto siguió en D .

Se pide:

1. Si se toma al azar un paciente de dicho hospital que cumple los criterios para la intervención, ¿cuál es la probabilidad de que después de la misma esté en la categoría C ?

- Si sabemos que un paciente después de intervenido pertenece a la categoría B , ¿cuál es la categoría de la que resulta más probable que proceda?
- Si el hospital trabaja con un total de 10000 pacientes intervenidos, ¿cuántos en ningún caso han pertenecido a la categoría C , ya sea antes o después de la intervención? ¿Y cuántos han pertenecido a la categoría A , ya sea antes o después de la intervención?

Solución

Llamando a , b , c y d a los sucesos consistentes en pertenecer a las categorías A , B , C , y D respectivamente antes de la intervención, y A , B , C , y D a los sucesos consistentes en pertenecer a esas categorías después, el árbol correspondiente al experimento es el siguiente

- De acuerdo al árbol de probabilidad se tiene

$$P(C) = P(c, C) + P(d, C) = 0.3 \cdot 0.3 + 0.4 \cdot 0.4 = 0.25.$$

- Si la persona está en la categoría B después de la intervención, está claro que antes de la intervención sólo podía estar en b , c o d . Las respectivas probabilidades condicionadas son:

$$\begin{aligned} P(b/B) &= \frac{P(b \cap B)}{P(B)} = \frac{P(b, B)}{P(b, B) + P(c, B) + P(d, B)} = \frac{0.2 \cdot 0.5}{0.2 \cdot 0.5 + 0.3 \cdot 0.4 + 0.4 \cdot 0.3} = \\ &= \frac{0.1}{0.34} = 0.2941, \\ P(c/B) &= \frac{P(c \cap B)}{P(B)} = \frac{P(c, B)}{P(b, B) + P(c, B) + P(d, B)} = \frac{0.12}{0.34} = 0.3529, \\ P(d/B) &= \frac{P(d \cap B)}{P(B)} = \frac{P(d, B)}{P(b, B) + P(c, B) + P(d, B)} = \frac{0.12}{0.34} = 0.3529. \end{aligned}$$

Luego es igualmente probable que proceda de las categorías c o d .

- La probabilidad de que una persona no haya pertenecido a la categoría C ni antes ni después de la intervención es

$$\begin{aligned} P(\bar{c} \cap \bar{C}) &= P(a, A) + p(b, A) + p(b, B) + p(d, A) + p(d, B) + P(d, D) = \\ &= 0.1 \cdot 1 + 0.2 \cdot 0.5 + 0.2 \cdot 0.5 + 0.4 \cdot 0.1 + 0.4 \cdot 0.3 + 0.4 \cdot 0.2 = \\ &= 0.54. \end{aligned}$$

Por tanto, si en la población hay 10000 pacientes intervenidos, el número de personas que no han pertenecido a la categoría C ni antes ni después de la intervención será $10000 \cdot 0.54 = 5400$.

Por otro lado, la probabilidad de que una persona haya estado en la categoría A antes o después de la intervención es

$$\begin{aligned} P(a \cup A) &= P(a, A) + p(b, A) + p(c, A) + p(d, A) = \\ &= 0.1 \cdot 1 + 0.2 \cdot 0.5 + 0.3 \cdot 0.3 + 0.4 \cdot 0.1 \\ &= 0.33. \end{aligned}$$

y el número de personas que han estado en la categoría A antes o después de la intervención es $10000 \cdot 0.33 = 3300$.

Ejercicio 3.46. ★ Dos cajas contienen bolas con diferentes colores. La primera caja contiene 3 bolas blancas y 2 negras, y la segunda caja contiene 2 bolas azules, 1 bola roja y 1 bola verde. Construir el espacio muestral de los siguientes experimentos aleatorios:

1. Tomar una bola al azar de cada caja.
2. Tomar dos bolas al azar de cada caja.

Ejercicio 3.47. ★ Las leyes de Morgan establecen que dados dos sucesos aleatorios A y B de un mismo espacio muestral, $\overline{A \cup B} = \bar{A} \cap \bar{B}$ y $\overline{A \cap B} = \bar{A} \cup \bar{B}$. Demostrar ambas igualdades usando diagramas de Venn.

Ejercicio 3.48. ★ Calcular la probabilidad de los siguientes sucesos del experimento aleatorio consistente en lanzar 3 monedas:

1. Obtener exactamente una cara.
2. Obtener exactamente dos cruces.
3. Obtener dos o más caras.
4. Obtener alguna cruz.

Solución

1. $P(1 \text{ cara}) = 0.375$.
2. $P(2 \text{ cruces}) = 0.375$.
3. $P(2 \text{ o más caras}) = 0.5$.
4. $P(\text{alguna cruz}) = 0.875$.

Ejercicio 3.49. ★ Para evaluar la efectividad de un test diagnóstico se aplicó el test a una muestra de personas con los siguientes resultados:

	Test +	Test -
Enfermos	2020	140

	Test +	Test -
Sanos	80	7760

Calcular para este test:

1. La sensibilidad y la especificidad.
2. Los valores predictivo positivo y negativo.
3. La probabilidad de un diagnóstico acertado.

Solución

Llamando E y $\bar{E}$ a los eventos consistentes en estar enfermo y sano respectivamente.

1. Sensibilidad $P(+|E) = 0.9352$ y especificidad $P(-|\bar{E}) = 0.9898$.
2. VPP $P(E|+) = 0.9619$ y VPN $P(\bar{E}|-) = 0.9823$.
3. $P((E \cap +) \cup (\bar{E} \cap -)) = 0.978$.

Ejercicio 3.50. ★ Se dispone de dos test A y B para diagnosticar una enfermedad. El test A tiene una sensibilidad del 98% y una especificidad del 80%, mientras que el test B tiene una sensibilidad del 75% y una especificidad del 99%.

1. ¿Qué test es mejor para confirmar la enfermedad?
2. ¿Qué test es mejor para descartar la enfermedad?
3. A menudo un test se utiliza para descartar una enfermedad rara un gran número de personas aparentemente sanas. Este tipo de test se conocen como *test de cribado*. ¿Cuál de los dos test funcionaría mejor como test de cribado? ¿Cuál es el valor predictivo positivo (VPP) de este test si la prevalencia de la enfermedad es 0.01? ¿Y si la prevalencia de la enfermedad es 0.2?
4. El valor predictivo positivo de un test de cribado no suele ser alto. ¿Cómo se pueden combinar ambos test para tener una confianza mayor en el diagnóstico de la enfermedad? Calcular la probabilidad a posteriori de tener la enfermedad con la combinación de ambos test si el resultado de ambos test es positivo y la prevalencia de la enfermedad es 0.01.

Solución

1. El test B ya que tiene una mayor especificidad.
2. El test A ya que tiene una mayor sensibilidad.
3. El test A funcionará mejor como un test de cribado.
Para una prevalencia de 0.01 el VPP es $P(E|+) = 0.0472$ y el VPN es $P(\bar{E}|-) = 0.9997$.
Para una prevalencia de 0.2 el VPP es $P(E|+) = 0.5506$ y el VPN es

$$P(\bar{E}|-) = 0.9938.$$

4. Aplicando primero el test A a todo el mundo y luego el test B a los resultados positivos de A .

$$P(E|+A \cap +B) = 0.7878.$$

Ejercicio 3.51. En una población con 10000 personas hay 15 con tuberculosis en un momento concreto. Después de 5 años, 5 de ellas murieron y otras 3 cogieron la tuberculosis. Suponiendo que el tamaño de la población no cambió en esos 5 años, calcular

1. La prevalencia inicial de la tuberculosis.
2. La incidencia acumulada de tuberculosis en el periodo de 5 años.
3. La tasa de incidencia (riesgo absoluto) de tuberculosis ese mismo periodo.
4. La prevalencia de la tuberculosis al final de ese periodo.

Solución

1. $P(T) = 0.0015$.
2. $R(T) = 0.0003$, es decir, 3 por cada 10000 personas en 5 años.
3. $R(T) = 0.00006$, es decir, 0.6 por cada 10000 personas-año.
4. $P(T) = 0.0013$.

Ejercicio 3.52. ★ En un brote de varicela, esta fue diagnosticada a 18 de 160 niños vacunados y en 30 de 70 niños no vacunados.

1. Calcular e interpretar el riesgo relativo de sufrir varicela en niños vacunados.
2. Calcular e interpretar el odds ratio (oportunidad relativa) de sufrir la varicela en niños vacunados.
3. ¿Qué medida de asociación es más apropiada para este tipo de estudio, el riesgo relativo o el odds ratio?

Solución

1. $RR(V) = 0.2625$.
2. $OR(V) = 0.169$.
3. El odds ratio, ya que se trata de un estudio retrospectivo.

Ejercicio 3.53. Un ensayo clínico trata de determinar el efecto de un programa de ejercicios de preparación para el parto sobre los traumas perineales. La siguiente tabla resume los resultados.

	Episiotomía	No episiotomía
--	-------------	----------------

Madres siguiendo el programa de ejercicios	62	358
Madres no siguiendo el programa de ejercicios	328	512

1. Calcular el riesgo absoluto y relativo de episiotomía en las madres que siguieron el programa de ejercicios. ¿Qué se puede decir sobre la efectividad del programa de ejercicios de preparación al parto?
2. Calcular e interpretar el odds ratio de episiotomía en las madres que siguieron el programa de ejercicios.
3. ¿Qué medida de asociación es más apropiada para este tipo de estudio, el riesgo relativo o el odds ratio?

Solución

1. $R(E) = 0.1476$ y $RR(E) = 0.378$.
2. $OR(E) = 0.2703$.
3. Ambas son válidas.

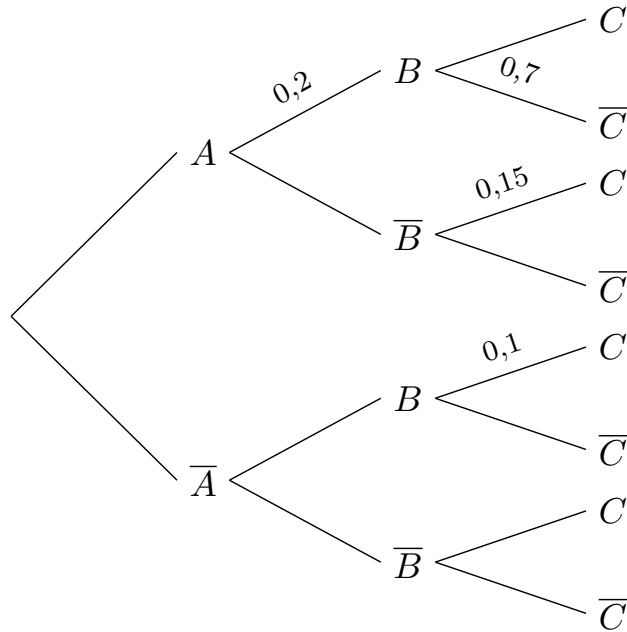
Ejercicio 3.54. En una ciudad se han detectado muchos casos de personas intoxicadas en la última semana y se sospecha que la causa pueda ser el consumo de una determinada marca de agua embotellada. Para confirmar la asociación entre el consumo de este agua y la intoxicación se tomó una muestra de 750 personas intoxicadas y otra de 1250 sin intoxicar. En las personas intoxicadas se observó que 420 habían consumido el agua embotellada de esa marca, mientras que en las personas sin intoxicar, 80 habían consumido agua embotellada de la marca. Se pide

1. Calcular la incidencia de la intoxicación en las personas que consumieron agua embotellada y en las que no.
2. Con estos datos, ¿es posible calcular la prevalencia de la intoxicación?
3. Calcular el riesgo relativo de intoxicación al consumir agua embotellada e interpretarlo.
4. Calcular el odds ratio de intoxicación al consumir agua embotellada.
5. ¿Qué medida de asociación es más adecuada para este estudio? Justificar la respuesta

Solución

1. $R_T(E) = 0.84$ and $R_C(E) = 0.22$.
2. No, porque se trata de un estudio retrospectivo.
3. $RR(E) = 3.8182$.
4. $OR(E) = 18.6136$.
5. El odds ratio, porque no se puede estimar la prevalencia de la intoxicación.

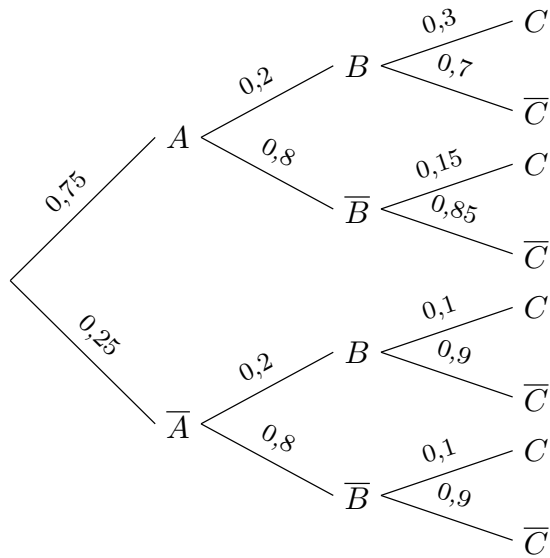
Ejercicio 3.55. El siguiente árbol corresponde al espacio probabilístico del experimento consistente en tomar una persona al azar de una población y comprobar si había tenido o no tres enfermedades A , B y C .



Sabiendo que el 1.5% de la población ha tenido las tres enfermedades, que el 54% no ha tenido ninguna, y que las enfermedades A y B son independientes, se pide:

1. Completar el árbol etiquetando las ramas con las probabilidades correspondientes.
2. Calcular la probabilidad de tener la enfermedad C .
3. Calcular la probabilidad de tener la enfermedad B si se tiene la enfermedad C .
4. Calcular la probabilidad de no tener la enfermedad A si no se tienen las otras dos.
5. ¿Son independientes los sucesos B y C ?

💡 Solución



- 1.
2. $P(C) = 0.12$.
3. $P(B|C) = 0.25$.
4. $P(\bar{A}|\bar{B} \cap \bar{C}) = 0.7606$.
5. No, porque $P(B|C) \neq P(B)$.

4 Variables aleatorias discretas

Ejercicio 4.1. Sea X una variable aleatoria discreta cuya ley de probabilidad es

X	4	5	6	7	8
$f(x)$	0.15	0.35	0.10	0.25	0.15

1. Calcular y representar gráficamente la función de distribución.
2. Obtener:
 1. $P(X < 7.5)$.
 2. $P(X > 8)$.
 3. $P(4 \leq X \leq 6.5)$.
 4. $P(5 < X < 6)$.

Solución

1.

$$F(x) = \begin{cases} 0 & \text{si } x < 4, \\ 0.15 & \text{si } 4 \leq x < 5, \\ 0.5 & \text{si } 5 \leq x < 6, \\ 0.6 & \text{si } 6 \leq x < 7, \\ 0.85 & \text{si } 7 \leq x < 8, \\ 1 & \text{si } 8 \leq x. \end{cases}$$

2. $P(X < 7.5) = 0.85$, $P(X > 8) = 0$, $P(4 \leq x \leq 6.5) = 0.6$ y $P(5 < X < 6) = 0$.

Ejercicio 4.2. Sea la variable aleatoria X con la siguiente función de distribución:

$$F(x) = \begin{cases} 0 & \text{si } x < 1, \\ 1/5 & \text{si } 1 \leq x < 4, \\ 3/4 & \text{si } 4 \leq x < 6, \\ 1 & \text{si } 6 \leq x. \end{cases}$$

Se pide:

1. Distribución de probabilidad.
2. Calcular la siguientes probabilidades:
 1. $P(X = 6)$.
 2. $P(X = 5)$.
 3. $P(2 < X < 5.5)$.
 4. $P(0 \leq X < 4)$.
3. Calcular la media.
4. Calcular la desviación típica.

Solución

X	1	4	6
$f(x)$	1/5	11/20	1/4

- 1.
2. $P(X = 6) = 1/4$, $P(X = 5) = 0$, $P(2 < X < 5.5) = 11/20$ y $P(0 \leq X < 4) = 1/5$.
3. $\mu = 3.9$.
4. $\sigma = 1.6703$.

Ejercicio 4.3. ★ Se realiza un experimento aleatorio consistente en inyectar un virus a tres tipos de ratas y observar si sobreviven o no. Se sabe que la probabilidad de que viva el primer tipo de rata es 0.5, la de que viva el segundo es 0.4 y la de que viva el tercero 0.3. Se pide:

1. Construir la variable aleatoria que mida el número de ratas vivas y su función de probabilidad.
2. Calcular la función de distribución.
3. Calcular $P(X \leq 1)$, $P(X \geq 2)$ y $P(X = 1.5)$.
4. Calcular la media y la desviación típica. ¿Es representativa la media?

Solución

X	0	1	2	3
$f(x)$	0.21	0.44	0.29	0.06

1.

2.

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ 0.21 & \text{si } 0 \leq x < 1, \\ 0.65 & \text{si } 1 \leq x < 2, \\ 0.94 & \text{si } 2 \leq x < 3, \\ 1 & \text{si } 3 \leq x. \end{cases}$$

3. $P(X \leq 1) = 0.65$, $P(X \geq 2) = 0.35$ y $P(X = 1.5) = 0$.

4. $\mu = 1.2$ ratas, $\sigma^2 = 0.7$ ratas² y $\sigma = 0.84$ ratas.

Ejercicio 4.4. Una tómbola asegura que en cada 1000 boletos hay 500 con “siga intentándolo”, 100 con un premio de 1€, 60 con un premio de 2€, 30 con un premio de 3€, y 10 con un premio de 10€. Un individuo decide comprar un boleto que cuesta 1€. Se pide:

1. Construir una variable aleatoria que mida la ganancia (o pérdida) con la compra de un boleto.
2. ¿Cual es la probabilidad de que pierda dinero?
3. ¿Qué ganancia espera obtener?

Ejercicio 4.5. La probabilidad de curación de un paciente al ser sometido a un determinado tratamiento es 0.85. Calcular la probabilidad de que en un grupo de 6 enfermos sometidos a tratamiento:

1. se curen la mitad.
2. se curen al menos 4.

Solución

Llamando X al número de pacientes curados de los 6 sometidos al tratamiento, se tiene que $X \sim B(6, 0.85)$.

1. $P(X = 3) = 0.041$.
2. $P(X \geq 4) = 0.9526$.

Ejercicio 4.6. ★ En una mesa de juego, en 1654, Meré propuso a Pascal la siguiente afirmación: “es más probable obtener al menos un as con cuatro dados, que al menos un doble as en veinticuatro tiradas de dos dados”.

Demostrar que Meré tenía razón.

Ejercicio 4.7. Se sabe que la probabilidad de que aparezca una bacteria en un mm^3 de cierta disolución es de 0.002. Si en cada mm^3 a los sumo puede aparecer una bacteria, determinar la probabilidad de que en un cm^3 haya como máximo 5 bacterias.

 Solución

Llamando X al número de bacterias en 1 cm^3 de disolución, se tiene $X \sim B(1000, 0.002) \approx P(2)$.
 $P(X \leq 5) = 0.9834$.

Ejercicio 4.8. Diez individuos entran en contacto con un portador de tuberculosis. La probabilidad de que la enfermedad se contagie del portador a un sujeto cualquiera es 0.10.

1. ¿Qué probabilidad hay de que ninguno se contagie?
2. ¿Qué probabilidad hay de que al menos dos se contagien?
3. ¿Cuántos se espera que contraigan la enfermedad?

 Solución

Llamando X al número de personas contagiadas.

1. $P(X = 0) = 0.3487$.
2. $P(X \geq 2) = 0.2639$.
3. $\mu = 1$.

Ejercicio 4.9. ★ Las matrículas de los coches constan de una parte numérica formada por cuatro cifras, y una parte literal. Se pide:

1. Hallar la probabilidad de que al pasar 30 coches haya menos de dos cuya parte numérica sea capicúa.
2. ¿Cuántos coches deben pasar para que la probabilidad de que alguno tenga la parte numérica capicúa sea mayor que 0.1?

Ejercicio 4.10. La probabilidad de que al administrar una vacuna dé una determinada reacción es 0.001. Si se vacunan 2000 personas, ¿Cuál es la probabilidad de que aparezca alguna reacción adversa?

 Solución

Llamando X al número de reacciones adversas, $P(X \geq 1) = 0.8648$.

Ejercicio 4.11. ★ El número medio de llamadas por minuto que llegan a una centralita telefónica es igual a 120. Hallar las probabilidades de los sucesos siguientes:

1. $A = \{\text{durante 2 segundos lleguen a la centralita menos de 4 llamadas}\}.$
2. $B = \{\text{durante 3 segundos lleguen a la centralita 3 llamadas como mínimo}\}.$

 Solución

1. Si X es el número de llamadas en 2 segundos, entonces $X \sim P(4)$ y $P(X < 4) = 0.4335$.
2. Si Y es el número de llamadas en 3 segundos, entonces $Y \sim P(6)$ y $P(Y \geq 3) = 0.938$.

Ejercicio 4.12. En un laboratorio de análisis de alimentos se sabe, de experiencias anteriores, que la probabilidad de que una muestra de café contenga plomo en cantidades superiores a las permitidas por la legislación vigente es 0.2. Si se reciben 50 muestras de café, ¿cuál es la probabilidad de rechazar al menos 5? ¿Cuál será el número medio de muestras que rechazaremos?

Ejercicio 4.13. Un proceso de fabricación de un fármaco produce por término medio 6 fármacos defectuosos por hora. ¿Cuál es probabilidad de que en un hora se produzcan menos de 3 fármacos defectuosos? ¿Y la de que en la próxima media hora se produzcan más de un fármaco defectuoso?

 Solución

Llamando X al número de fármacos defectuosos en 1 hora, se tiene $X \sim P(6)$ y $P(X < 3) = 0.062$.
Llamando Y al número de fármacos defectuosos en $1/2$ hora, se tiene $Y \sim P(3)$ y $P(Y > 1) = 0.8009$.

Ejercicio 4.14. Una mecanógrafa comete, en promedio, una errata cada 2000 caracteres que escribe. Suponiendo que escribe un folio con treinta líneas y setenta caracteres por línea, ¿cuál es la probabilidad de que cometa más de un error en dicho folio?

 Solución

Llamando X al número de errores en un folio, se tiene que $X \sim B(2100, 1/2000) \approx P(1.05)$, y $P(X > 1) = 0.2826$.

Ejercicio 4.15. Un examen de tipo test consta de 10 preguntas con tres respuestas posibles para cada una de ellas. Se obtiene un punto por cada respuesta acertada y se pierde medio punto por cada pregunta fallada. Un alumno sabe tres de las preguntas del test y las contesta correctamente, pero no sabe las otras siete y las contesta al azar. ¿Qué probabilidad tiene de aprobar el examen?

 Solución

Llamando X al número de preguntas acertadas de las 7 contestadas al azar, se tiene $X \sim B(7, 1/3)$ y $P(X \geq 4) = 0.1733$.

Ejercicio 4.16. ★ Un equipo de fútbol tiene 7 delanteros. Se sabe que por término medio, cada delantero se pierde 5 partidos por lesión en una temporada de 40 partidos. Suponiendo que todos los delanteros tienen la misma probabilidad de lesionarse, se pide:

1. ¿Cuál es la probabilidad de que en un partido determinado tenga menos de 5 delanteros en condiciones de jugar?
2. ¿Cuál es la probabilidad de que a lo largo de la temporada haya más de un partido en que tenga menos de 5 delanteros en condiciones de jugar?

Ejercicio 4.17. Para reforestar un bosque se compran árboles a un vivero en el que el 4% de los árboles suele morir debido a una enfermedad. Si la repoblación se efectúa por parcelas en las que se ponen 10 árboles, se pide:

1. Calcular la probabilidad de que no muera ningún árbol en una parcela.
2. Calcular la probabilidad de no mueran más de 2 árboles en una parcela.
3. Si en total se reforestan 3000 parcelas, ¿cuál es la probabilidad de que haya alguna en la que mueran más de dos árboles?

Ejercicio 4.18. ★ En un estudio sobre un determinado tipo de parásito que ataca el riñón de las ratas, se sabe que el número medio de parásitos en cada riñón es 3. Se pide:

1. Calcular la probabilidad de que una rata tenga más de 8 parásitos.
(Nota: se supone que una rata normal tiene dos riñones).
2. Si se tienen 10 ratas, ¿cuál es la probabilidad de que haya al menos 9 con parásitos?

 Solución

1. Si X es el número de parásitos en una rata, $X \sim P(6)$ y $P(X > 8) = 0.1528$.
2. Si Y es el número de ratas con parásitos en un grupo de 10 ratas, entonces $Y \sim B(10, 0.9975)$ y $P(Y \geq 9) = 0.9997$.

Ejercicio 4.19. ★ Se ha comprobado experimentalmente que una de cada 20 billones de células expuestas a un determinado tipo de radiación muta volviéndose cancerígena. Sabiendo que el cuerpo humano tiene aproximadamente 1 billón de células por kilogramo de tejido, calcular la probabilidad de que una persona de 60 kg expuesta a dicha radiación desarrolle cáncer. Si la radiación ha afectado a 3 personas de 60 kg, ¿cuál es la probabilidad de que desarrolle el cáncer más de una?

 Solución

Llamando X al número de mutaciones, $P(X > 0) = 0.9502$.

Llamando Y al número de personas que desarrollan el cáncer, $P(Y \geq 1) = 0.9999$.

Ejercicio 4.20. ★ En un servicio de urgencias de cierto hospital se sabe que, en media, llegan 2 pacientes a la hora. Calcular:

1. Si los turnos en urgencias son de 8 horas, ¿cuál será la probabilidad de que en un turno lleguen más de 5 pacientes?
2. Si el servicio de urgencias tiene capacidad para atender adecuadamente como mucho a 4 pacientes a la hora, ¿cuál es la probabilidad de que a lo largo de un turno de 8 horas el servicio de urgencias se vea desbordado en alguna de las horas del turno?

 Solución

1. Llamando X al número de pacientes en un turno, se tiene que $X \sim P(16)$ y $P(X > 5) = 0.9986$.
2. Llamando Y al número de horas en el que el servicio se vea desbordado porque lleguen más de 4 pacientes, se tiene que $Y \sim B(8, 0.0527)$ y $P(Y \geq 1) = 0.3515$.

Ejercicio 4.21. En un Parque Nacional se contabilizan 15 lince. Si sabemos, por estudios previos, que mueren en promedio 1 de cada 10 individuos a lo largo de un año (ya sea por accidentes, caza de furtivos o por causas naturales):

1. ¿Cuál es la probabilidad de que en el Parque Nacional se contabilicen más de 2 muertes de lince en un año?
2. Suponiendo un periodo de 12 años, ¿cuál es la probabilidad de que en el Parque Nacional haya algún año en el que mueran 2 lince?

Ejercicio 4.22. ★ El síndrome de Turner es una anomalía genética que se caracteriza porque las mujeres tienen sólo un cromosoma X . Afecta aproximadamente a 1 de cada 2000 mujeres. Además, aproximadamente 1 de cada 10 mujeres con síndrome de Turner, como consecuencia, también sufren un estrechamiento anormal de la aorta. Se pide:

1. En un grupo de 4000 mujeres, ¿cuál es la probabilidad de que haya más de 3 afectadas por el síndrome de Turner? ¿Y de que haya alguna con estrechamiento de aorta como consecuencia de padecer el síndrome de Turner?
2. En un grupo de 20 chicas afectadas por el síndrome de Turner, ¿cuál es la probabilidad de que menos de 3 sufran un estrechamiento anormal de la aorta?

 Solución

1. Si X es el número de mujeres afectadas por el síndrome de Turner en el grupo de 4000 mujeres, entonces $X \sim B(4000, 1/2000) \approx P(2)$ y $P(X > 3) = 0.1429$.
Si Y es el número de mujeres con estrechamiento de la aorta en el grupo de 4000 mujeres, entonces $Y \sim B(4000, 1/20000) \approx P(0.2)$ y $P(Y > 0) = 0.1813$.
2. Si Z es el número de mujeres con estrechamiento de la aorta en el grupo de 20 mujeres con el síndrome de Turner, entonces $Z \sim B(20, 1/10)$ y $P(Z < 3) = 0.6769$.

Ejercicio 4.23. Por estudios previos se sabe que, en una comarca, hay dos tipos de larvas que parasitan, de forma completamente independiente, los chopos, y que producen su muerte. Si la larva de tipo A está parasitando un 15% de los chopos, y la B un 30%, y en una zona concreta de la comarca hay 10 chopos:

1. ¿Qué probabilidad hay de que estén siendo parasitados por A más de dos?
2. ¿Qué probabilidad hay de que estén libres de B más de 8?
3. ¿Qué probabilidad hay de que más de 1 tenga los dos tipos de larva?
4. ¿Qué probabilidad hay de que más de 3 tengan algún tipo de larva?

 Solución

1. X_A es el número de chopos parasitados por larvas del tipo A , entonces $X_A \sim B(10, 0.15)$ y $P(X_A > 2) = 0.1798$.
2. Si $X_{\bar{B}}$ es el número de chopos no parasitados por larvas del tipo B , entonces $X_{\bar{B}} \sim B(10, 0.7)$ y $P(X_{\bar{B}} > 8) = 0.1493$.
3. Si llamamos $X_{A \cap B}$ al número de chopos parasitados por larvas de ambos tipos, entonces $X_{A \cap B} \sim B(10, 0.045)$ y $P(X_{A \cap B} > 1) = 0.0717$.
4. Si llamamos $X_{A \cup B}$ al número de chopos parasitados por algún tipo de larva, entonces $X_{A \cup B} \sim B(10, 0.405)$ y $P(X_{A \cup B} > 3) = 0.6302$.

Ejercicio 4.24. ★ En un Parque Regional se contabilizan 10 parejas de buitre leonado. Si sabemos que el 70% de las parejas de esta especie logran que alguna de sus crías sobreviva:

1. ¿Cuál es la probabilidad de que 8 parejas de buitre leonado del Parque logren que alguna de sus crías sobreviva?
2. ¿Cuál es la probabilidad de que alguna pareja logre que alguna de sus crías sobreviva?
3. Si sabemos que en dicho Parque Regional nidifican, en promedio, 8 parejas al año, ¿cuál es la probabilidad de que en un año concreto nidifiquen más de 6?

Ejercicio 4.25. Al lanzar 100 veces una moneda, ¿cuál es la probabilidad de obtener entre 40 y 60 caras?

Ejercicio 4.26. El trastorno de pánico aparece en 1 de cada 75 personas. ¿Cuál es la probabilidad de que en un grupo de 100 personas aparezca alguna con trastorno de pánico? ¿Cuál es el número esperado de personas con trastorno de pánico en ese grupo?

 Solución

Llamando X al número de personas que sufren trastorno del pánico en el grupo de 100 personas, se tiene que $X \sim B(100, 1/75)$ y $P(X \geq 1) = 0.7379$.

Ejercicio 4.27. Se sabe que 2 de cada 1000 pacientes son alérgicos a un fármaco A , y que 6 de cada 1000 lo son a un fármaco B . Además, el 30% de los alérgicos a B , también lo son a A . Si se aplican los dos fármacos a 500 personas,

1. ¿Cuál es la probabilidad de que no haya ninguna con alergia a A ?
2. ¿Cuál es la probabilidad de que haya al menos 2 con alergia a B ?
3. ¿Cuál es la probabilidad de que haya menos de 2 con las dos alergias?
4. ¿Cuál es la probabilidad de que haya alguna con alergia?

 Solución

1. Llamando X_A al número de personas alérgicas al fármaco A en el grupo de 500 personas, se tiene que $X_A \sim B(500, 0.002) \approx P(1)$ y $P(X_A = 0) = 0.3678$.
2. Llamando X_B al número de personas alérgicas al fármaco B en el grupo de 500 personas, se tiene que $X_B \sim B(500, 0.006) \approx P(3)$ y $P(X_B \geq 2) = 0.8009$.
3. Llamando $X_{A \cap B}$ al número de personas alérgicas a ambos fármacos $A \cap B$ en el grupo de 500 personas, se tiene que $X_A \sim B(500, 0.0018) \approx P(0.9)$ y $P(X_{A \cap B} < 2) = 0.7725$.
4. Llamando $X_{A \cup B}$ al número de personas alérgicas a alguno de los fármacos $A \cup B$ en el grupo de 500 personas, se tiene que $X_A \sim B(500, 0.0062) \approx P(3.1)$ y $P(X_{A \cup B} \geq 1) = 0.9550$.

Ejercicio 4.28. En una clase hay 40 alumnos de los cuales el 35% son fumadores. Si se toma una muestra aleatoria con reemplazamiento de 4 alumnos, ¿cuál es la probabilidad de que haya al menos 1 fumador? ¿Cuál sería dicha probabilidad si la muestra se hubiese tomado sin reemplazamiento?

 Solución

Si X es el número de fumadores en una muestra aleatoria con reemplazamiento de tamaño 4, entonces $X \sim B(4, 0.35)$ y $P(X \geq 1) = 0.8215$.

Si la muestra es sin reemplazamiento $P(X \geq 1) = 0.8364$.

Ejercicio 4.29. ★ Se sabe que por término medio 2 de cada 10000 niños que nacen son albinos.

1. Si en una región nacen cada año 22000 niños ¿cuál es la probabilidad de que un año nazcan al menos 4 albinos?
2. ¿Cuál es la probabilidad de que en esa región, en un periodo de 10 años no nazca ningún niño albino?

 Solución

1. Llamando X al número de niños albinos que nacen en un año, se tiene que $X \sim B(22000, 2/10000) \approx P(4.4)$ y $P(X \geq 4) = 0.6406$.

2. Llamando Y al número de niños albinos que nacen en 10 años, se tiene que $Y \sim B(220000, 2/10000) \approx P(44)$ y $P(Y = 0) = 0$.

Ejercicio 4.30. ★ Suponiendo una facultad en la que hay un 60% de chicas y un 40% de chicos:

1. Si un año van 6 alumnos a hacer prácticas en un hospital, ¿qué probabilidad hay de que vayan más chicos que chicas?
2. En un período de 5 años, ¿cuál es la probabilidad de que más de 1 año no haya ido ningún chico?

 Solución

1. Si consideramos un total de $n = 6$ alumnos, con un 60% de chicas y un 40% de chicos, la variable X que es el número de alumnos chicos que van a hacer las prácticas de un total de 6, sigue una distribución binomial de 6 intentos y probabilidad de éxito igual a 0.4: $X \sim B(6, 0.4)$. Como nos piden la probabilidad de que haya más chicos que chicas, eso se consigue si el número de chicos es 4 o más. Por lo tanto, nos piden la probabilidad de que

X sea mayor o igual que 4:

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6)$$

Calculando estas probabilidades con la función de probabilidad de la variable binomial, tenemos

$$P(X = 4) = \binom{6}{4} \cdot 0.4^4 \cdot (1 - 0.4)^{6-4} = 12 \cdot 0.4^4 \cdot 0.6^2 = 0.1382,$$

$$P(X = 5) = \binom{6}{5} \cdot 0.4^5 \cdot (1 - 0.4)^{6-5} = 6 \cdot 0.4^5 \cdot 0.6 = 0.0369,$$

$$P(X = 6) = \binom{6}{6} \cdot 0.4^6 \cdot (1 - 0.4)^{6-6} = 1 \cdot 0.4^6 \cdot 0.6^0 = 0.0041$$

Y sumando los tres resultados obtenidos:

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) = 0.1382 + 0.0369 + 0.0041 = 0.1792.$$

2. Si consideramos, en un total de 5 años, la probabilidad de que más de un año no haya ido ningún chico, la variable a tener en cuenta Y será el número de años en el total de 5 sin ningún chico, y de nuevo esta variable sigue una distribución binomial, esta vez con 5 intentos, cuyo éxito viene dado por la probabilidad de que en un año concreto no haya ningún chico: $Y \sim B(5, p)$, donde $p = P(X = 0)$.

$$p = P(X = 0) = \binom{6}{0} \cdot 0.4^0 \cdot (1 - 0.4)^{6-0} = 1 \cdot 0.4^0 \cdot 0.6^6 = 0.0467.$$

Por lo tanto, $Y \sim B(5, 0.0467)$.

Además, nos piden la probabilidad de más de una año en el total de 5; es decir:

$$P(Y > 1) = 1 - P(Y \leq 1) = 1 - P(Y = 0) - P(Y = 1).$$

Aplicando, de nuevo, la fórmula de la función de probabilidad de la binomial, obtenemos:

$$P(Y = 0) = \binom{5}{0} \cdot 0.0467^0 \cdot (1 - 0.0467)^{5-0} = 1 \cdot 0.0467^0 \cdot 0.9533^5 = 0.7873,$$

$$P(Y = 1) = \binom{5}{1} \cdot 0.0467^1 \cdot (1 - 0.0467)^{5-1} = 5 \cdot 0.0467^1 \cdot 0.9533^4 = 0.1928.$$

Teniendo lo anterior en cuenta, la probabilidad que nos piden vale:

$$P(Y > 1) = 1 - P(Y = 0) - P(Y = 1) = 1 - 0.7873 - 0.1928 = 0.0199.$$

Ejercicio 4.31. ★ La probabilidad de que en un grupo de 5 individuos mayores de 70 años todos padezcan arterioesclerosis cerebral es de 12,5 por mil.

1. ¿Cuál es la probabilidad de padecer la enfermedad entre los mayores de 70 años?
2. En un grupo de 1000 personas, ¿cuál es la probabilidad de que padezcan la enfermedad más de 450?

Ejercicio 4.32. ★ ¿Cuánto habría que restar a cada pregunta errada en un examen de tipo test de 5 preguntas con cuatro opciones y sólo una correcta, para que un individuo que responda al azar tenga una puntuación esperada de 0?

 Solución

Si llamamos X al número de preguntas acertadas, está claro que X sigue una distribución binomial $B(5, 1/4)$ ya que el examen tiene 4 preguntas, y la probabilidad de acertar cualquiera de ellas al azar es $1/4$ ya que hay cuatro opciones y sólo una es la correcta.

Ejercicio 4.33. Se sabe que el 6.8% las personas presentan a lo largo de su adolescencia un trastorno de hiperactividad, de los cuales tres cuartas partes son mujeres. Si en la población hay el mismo número de hombres y mujeres, se pide:

1. Calcular la probabilidad de que en una muestra de tres hombres, haya alguno que haya tenido hiperactividad en su adolescencia.
2. Calcular la probabilidad de que en una muestra de 2 hombres y 2 mujeres, haya alguno que haya tenido hiperactividad en su adolescencia.

 Solución

1. Si llamamos X al número de hombres que han tenido hiperactividad en su adolescencia en una muestra de 3 hombres, se tiene que $X \sim B(3, 0.034)$ y $P(X \geq 1) = 0.0986$.
2. Si llamamos X_H al número de hombres que han tenido hiperactividad en su adolescencia en una muestra de 2 hombres y X_M al número de mujeres que han tenido hiperactividad en su adolescencia en una muestra de 2 mujeres, entonces $X_H \sim B(2, 0.034)$ y $X_M \sim B(2, 0.102)$. Entonces $P(X_H \geq 1 \cup X_M \geq 1) = 0.2475$.

Ejercicio 4.34. ★ A un hospital llegan pacientes por la mañana a efectuarse extracciones de sangre. Se ha medido la frecuencia de llegada de los mismos en intervalos de 15 minutos. La distribución de probabilidad (medida de forma frecuentista) se muestra en la siguiente tabla:

X	0	1	2	3	4	5	6
$P(x)$	0.1	0.2	0.25	0.15	0.15	0.1	0.05

Se pide:

1. Calcular la probabilidad de que en un intervalo de 15 minutos lleguen 2 o más personas, y probabilidad de que lleguen menos de 8 personas.
2. ¿Cuál es el número medio esperado de personas que llegarán a sacarse sangre cada 15 minutos?
3. Suponiendo que el número de personas que llegan a sacarse sangre en 15 minutos sigue una distribución de Poisson de media la calculada en el apartado anterior, ¿cuál es la probabilidad de que llegue alguna en 15 minutos? ¿Y de que llegue alguna en 5 minutos?

Solución

1. Teniendo en cuenta que nos dan la distribución de probabilidad de la variable aleatoria discreta X que expresa el número de pacientes que llegan en 15 minutos, nos están pidiendo $P(X \geq 2)$ y $P(X < 8)$. Estas probabilidades son:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X = 0) - P(X = 1) = 1 - 0.1 - 0.2 = 0.7,$$

$$P(X < 8) = 1 - P(X \geq 8) = 1.$$

2. El número esperado es la media de la variable aleatoria:

$$\mu = \sum x f(x) = 0 \cdot 0.1 + 1 \cdot 0.2 + 2 \cdot 0.25 + 3 \cdot 0.15 + 4 \cdot 0.15 + 5 \cdot 0.1 + 6 \cdot 0.05 = 2.55 \text{ personas.}$$

3. Suponiendo que X sigue una distribución de Poisson con $\lambda = 2.55$, $X \sim P(2.55)$, la probabilidad de que llegue alguna persona en 15 minutos vendrá dada por:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-2.55} \frac{2.55^0}{0!} = 0.9219.$$

Para la segunda pregunta, teniendo en cuenta que el número medio de personas que llegan cada 15 minutos es 2.55, en 5 minutos llegarán en media

$2.55/3 = 0.85$ personas, y, por tanto, tenemos una nueva variable aleatoria Y , que seguirá una distribución de Poisson con $\lambda' = 0.85$.

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - e^{-0.85} \frac{0.85^0}{0!} = 0.5726.$$

Ejercicio 4.35. ★ En las siguientes tablas, indicar razonadamente, en los caso que sea posible, los valores de h que deben ponerse en cada tabla para que se tenga una distribución de probabilidad:

x	$f(x)$
-2	0.3
5	h
8	0.1

x	$f(x)$
1	-0.2
3	0.7
4	h

x	$f(x)$
2	h
3	0.5
4	0.6

En las tablas que constituyan una distribución de probabilidad:

1. Representar gráficamente la función de distribución.
2. Calcular media y desviación típica.
3. Calcular la mediana.
4. Si a los valores de x se multiplican por una constante $k < 0$, ¿cómo se ve afectada la media? ¿Y la desviación típica?

Solución

La segunda tabla no puede ser una distribución de probabilidad pues $f(1) = -0.2$ y la función de probabilidad no puede tomar valores negativos. Por otro lado, la tercera tabla tampoco puede ser una distribución de probabilidad ya que la suma

de todas las probabilidades debe ser 1, y para ello debería ser $h = -0.1$, lo cual no es posible al no poder tomar valores negativos. Así pues la única tabla que puede ser una distribución de probabilidad es la primera, y como la suma de todas las probabilidades tiene que ser 1, $0.3+h+0.1 = 1$, se deduce que $h = 0.6$. Trabajaremos pues, con la distribución

x	$f(x)$
-2	0.3
5	0.6
8	0.1

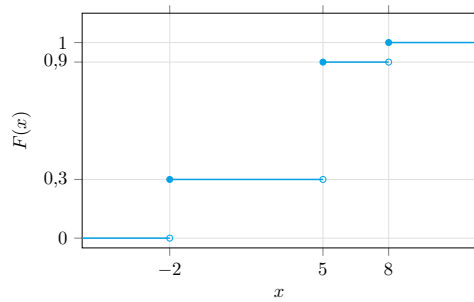
1. La función de distribución se define como $F(x_0) = P(X \leq x_0)$, y por tanto, mide probabilidades acumuladas. Acumulando las probabilidades de la tabla anterior tenemos

x	$f(x)$	$F(x)$
-2	0.3	0.3
5	0.6	0.9
8	0.1	1

O lo que es lo mismo, expresado como una función a trozos

$$F(x) = \begin{cases} 0, & \text{si } x < -2; \\ 0.3, & \text{si } -2 \leq x < 5; \\ 0.9, & \text{si } 5 \leq x < 8; \\ 1, & \text{si } x \geq 8. \end{cases}$$

La gráfica de esta función es la siguiente



2. Calculamos los estadísticos que nos piden

$$\begin{aligned} \mu &= \sum x_i f(x_i) = -2 \cdot 0.3 + 5 \cdot 0.6 + 8 \cdot 0.1 = 3.2, \\ \sigma^2 &= \sum x_i^2 f(x_i) - \mu^2 = (-2)^2 \cdot 0.3 + 5^2 \cdot 0.6 + 8^2 \cdot 0.1 - 3.2^2 = 22.6 - 10.24 = 12.36, \\ \sigma &= \sqrt{12.36} = 3.516. \end{aligned}$$

3. La mediana es el valor que deja acumulada una probabilidad 0.5, es decir, $F(\text{med}) = 0.5$, y mirando en la función de distribución, el valor donde se consigue acumular esta probabilidad es el 5.
4. Sea $Y = kX$ donde $k < 0$. Por las propiedades de las transformaciones lineales de variables aleatorias, tenemos que $\mu_y = k\mu_x$, y por tanto la media quedará también multiplicada por la constante k . Para la desviación típica tenemos que $\sigma_y = |k|\sigma_x$ y la desviación típica quedará multiplicada por el valor absoluto de k .

Ejercicio 4.36. ★ En una empresa el número de días al año que los empleados están de baja es, por término medio, 5. Suponiendo que un año tiene 240 días laborables y que cada mes tiene 20, se pide:

1. Calcular el porcentaje de empleados que no faltarían más de 5 días al año.
2. Calcular la probabilidad de que un empleado falte algún día en un mes.
3. ¿Cual es la probabilidad de que en un año haya más de 2 meses en los que haya faltado alguna vez?

 Solución

1. Sea X la variable que mide el número de días de baja al año de cada empleado. Entonces $X \sim B(240, 5/240)$, pero como $n = 240 > 30$ y $p = 5/240 < 0.1$, podemos aproximarla como una distribución Poisson $P(5)$. La probabilidad de que un empleado no falte más de 5 días al año es

$$\begin{aligned} P(X \leq 5) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) = \\ &= e^{-5} \frac{5^0}{0!} + e^{-5} \frac{5^1}{1!} + e^{-5} \frac{5^2}{2!} + e^{-5} \frac{5^3}{3!} + e^{-5} \frac{5^4}{4!} + e^{-5} \frac{5^5}{5!} = \\ &= 0.0067 + 0.0337 + 0.0842 + 0.1404 + 0.1755 + 0.1755 = 0.616, \end{aligned}$$

es decir, un 61.6%.

2. Sea Y la variable que mide el número de días de baja al mes de cada empleado. Entonces $Y \sim B(20, 5/240)$, y la probabilidad de que algún empleado falte algún día en un mes es

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y < 1) = 1 - P(Y = 0) = 1 - \binom{20}{0} \left(\frac{5}{240}\right)^0 \left(1 - \frac{5}{240}\right)^{20} = \\ &= 1 - \left(\frac{235}{240}\right)^{20} = 0.3437. \end{aligned}$$

3. Sea ahora Z la variable que mide el número de meses al año en que un empleado falta alguna vez. Entonces, como la probabilidad de que un empleado falte alguna vez en un mes, según el apartado anterior es 0.3437, tenemos que $Z \sim B(12, 0.3437)$. Así pues, la probabilidad que nos piden es

$$\begin{aligned} P(Z > 2) &= 1 - P(Z \leq 2) = 1 - P(Z = 0) - P(Z = 1) - P(Z = 2) = \\ &= 1 - \binom{12}{0} 0.3437^0 0.6563^{12} - \binom{12}{1} 0.3437^1 0.6563^{11} - \binom{12}{2} 0.3437^2 0.6563^{10} = \\ &= 1 - 0,0064 - 0,0401 - 0,1156 = 0.8379. \end{aligned}$$

Ejercicio 4.37. ★ Sabiendo que la prevalencia de la isquemia cardíaca es del 1%, y que la aplicación de un test diagnóstico para detectar la isquemia cardíaca tiene una sensibilidad del 90%, y una especificidad del 95%. Calcular:

1. Los valores predictivos, tanto el positivo como el negativo.
2. La probabilidad de diagnóstico acertado.

3. Si tenemos un grupo de 10 enfermos de isquemia cardíaca, ¿cuál es la probabilidad de que diagnostiquemos la enfermedad a menos de 8?

Ejercicio 4.38. ★ Un test diagnóstico para una enfermedad devuelve un 1% de resultados positivos, y sus valores predictivos positivo y negativo valen, respectivamente, 0.95 y 0.98. Se pide:

1. ¿Cuál es la prevalencia de la enfermedad?
2. ¿Cuánto valen la sensibilidad y la especificidad del test?
3. Si aplicamos el test a 12 individuos enfermos, ¿qué probabilidad hay de que se equivoque en alguno de ellos?
4. Si aplicamos el test a 12 individuos, ¿qué probabilidad hay de que acierte en todos?

 Solución

1. $P(E) = 0.0293$.
2. Sensibilidad $P(+|E) = 0.3242$ y especificidad $P(-|\bar{E}) = 0.9995$.
3. Llamando X al número de diagnósticos erróneos en 12 individuos enfermos, $P(X \geq 1) = 1$.
4. Llamando Y al número de diagnósticos acertados en 12 individuos, $P(X = 12) = 0.7818$.

Ejercicio 4.39. ★ La probabilidad de que en un grupo de 5 individuos mayores de 70 años todos padezcan arterioesclerosis cerebral es de 12.5 por mil.

1. ¿Cuál es la probabilidad de padecer la enfermedad entre los mayores de 70 años?
2. En un grupo de 1000 personas, ¿cuál es la probabilidad de que padezcan la enfermedad más de 450?

Ejercicio 4.40. ★ Si sabemos, por estudios previos, que las cepas que provocarán la gripe del siguiente otoño-invierno afectarán a un 20% de la población:

1. ¿Cuál es la probabilidad de que en una población de 10000 habitantes queden infectados menos de 1900?
2. Suponiendo que se vacunan los 10000 habitantes y sabiendo, por estudios previos, que la vacuna inmuniza al 98% de los vacunados, ¿Cuál es la probabilidad de que queden sin inmunizar menos de 180?
3. De nuevo, suponiendo que se han vacunado los 10000 habitantes y teniendo en cuenta que, por estudios previos, la vacuna produce reacciones alérgicas en uno de cada 5000 casos, ¿cuál es la probabilidad de que se produzca alguna reacción alérgica en dicha población?

5 Variables aleatorias continuas

Ejercicio 5.1. Una variable aleatoria continua X tiene una función de densidad dada por:

$$f(x) = \begin{cases} k(6 - 3x) & \text{si } 0 \leq x \leq 2, \\ 0 & \text{si } x < 0 \text{ o } x > 2. \end{cases}$$

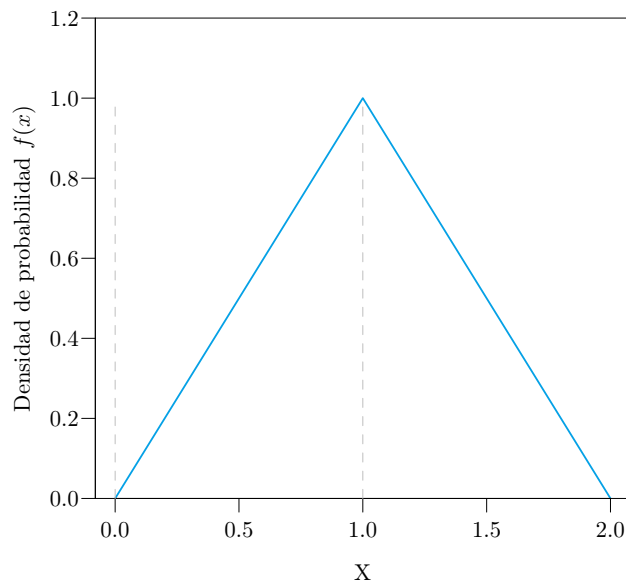
1. Determinar el valor de k .
2. Hallar $P(X \leq 1)$; $P(X > 2)$; $P(X = 1/4)$; $P(1/3 \leq X \leq 2/3)$.
3. Calcular μ y σ .
4. Hallar la función de distribución $F(x)$.

Solución

1. $k = 1/6$.
2. $P(X \leq 1) = 0.75$, $P(X > 2) = 0$, $P(X = 1/4) = 0$ y $P(1/3 \leq X \leq 2/3) = 1/4$.
3. $\mu = 2/3$, $\sigma^2 = 2/9$ y $\sigma = \sqrt{2}/3$.
- 4.

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ x - \frac{x^2}{4} & \text{si } 0 \leq x \leq 2, \\ 1 & \text{si } 2 < x. \end{cases}$$

Ejercicio 5.2. ★ La función de densidad de la variable aleatoria continua X viene dada por la gráfica siguiente:



Calcular:

1. $P(\frac{1}{2} \leq X \leq \frac{5}{4})$.
2. Función de distribución.
3. Media y desviación típica.

💡 Solución

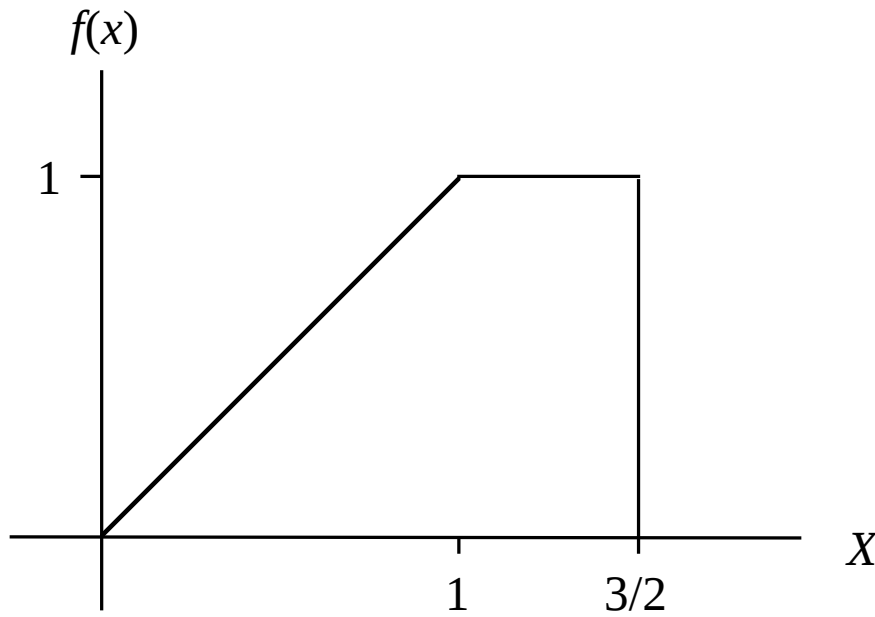
1. $P(\frac{1}{2} \leq X \leq \frac{5}{4}) = 19/32$.

2.

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ \frac{x^2}{2} & \text{si } 0 \leq x \leq 1, \\ -\frac{x^2}{2} + 2x - 1 & \text{si } 1 < x \leq 2, \\ 1 & \text{si } 2 < x. \end{cases}$$

3. $\mu = 1$, $\sigma^2 = 1/6$ y $\sigma = 1/\sqrt{6}$.

Ejercicio 5.3. ★ Dada la función de densidad dada por la siguiente gráfica,



calcular:

1. $P(X < 1)$, $P(X > 0)$, $P(X = 1/4)$, $P(1/2 \leq X \leq 3/2)$.
2. Media y desviación típica.

Ejercicio 5.4. ★ Se sabe que la duración de las baterías de un tipo de marcapasos (expresada en años) es una variable aleatoria cuya función de densidad es

$$f(x) = \begin{cases} 0 & \text{si } x < 0, \\ \frac{e^{-x/10}}{10} & \text{si } x \geq 0. \end{cases}$$

Se pide:

1. Comprobar que $f(x)$ es función de densidad.
2. Calcular la función de distribución.
3. Calcular la probabilidad de que la batería dure menos de 5 años, y de que dure entre 5 y 10 años.
4. Calcular la vida media de la batería.

💡 Solución

$$1. \int_{-\infty}^{\infty} f(x) dx = [-e^{-x/10}]_{-\infty}^{\infty} = 1.$$

2.

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ 1 - e^{-x/10} & \text{si } x \geq 0. \end{cases}$$

3. $P(X < 5) = 0.3935$ y $P(5 < X < 10) = 0.2387$.

4. $\mu = 10$ años.

Ejercicio 5.5. ★ La proporción de cierto aditivo en la gasolina determina su precio. Si en la producción de gasolina la proporción de aditivo es una variable aleatoria X con función de densidad $f(x) = 6x(1 - x)$ para $0 \leq x \leq 1$, de manera que si $x < 0,5$ la gasolina es del tipo I y se vende a 0.6 €/l, si $0.5 \leq x \leq 0.8$ la gasolina es del tipo II y se vende a 0.7 €/l, y si $x > 0.8$ la gasolina es del tipo III que se vende a 0.8 €/l, se pide:

1. Calcular la función de distribución de X .
2. Calcular los porcentajes de producción de cada tipo de gasolina.
3. Calcular el precio medio por litro.

Ejercicio 5.6. Un empleado suele acudir al trabajo en cualquier instante entre las 6 y las 7 con igual probabilidad. Se pide:

1. Calcular la función de densidad de la variable que mide el instante en que acude a trabajar y dibujarla.
2. Calcular la función de distribución y dibujarla.
3. Calcular la probabilidad de que llegue entre las 6 y cuarto y las 6 y media.
4. Calcular la hora media a la que se espera que llegue.

💡 Solución

1.

$$f(x) = \begin{cases} 0 & \text{si } x < 6, \\ 1 & \text{si } 6 \leq x \leq 7, \\ 0 & \text{si } 7 < x. \end{cases}$$

2.

$$F(x) = \begin{cases} 0 & \text{si } x < 6, \\ x - 6 & \text{si } 6 \leq x \leq 7, \\ 1 & \text{si } 7 < x. \end{cases}$$

3. $P(6.25 < X < 6.5) = 0.25$.

4. $\mu = 6.5$, es decir, a las 6 horas y media.

Ejercicio 5.7. Sea Z una variable aleatoria que sigue una distribución $N(0, 1)$. Determinar el valor de t en cada uno de los siguientes casos:

1. El área entre 0 y t es 0.4783.
2. El área a la izquierda de t es 0.6406.
3. El área entre -1.5 y t es 0.2313.

 Solución

1. $t = 2.02$.
2. $t = 0.36$.
3. $t = -0.53$.

Ejercicio 5.8. Hallar las siguientes probabilidades:

1. $P(-2.4 \leq Z \leq -1.2)$ si Z es $N(0, 1)$.
2. $P(|Z| > 1.2)$ si Z es $N(0, 1)$.
3. $P(1.3 \leq X \leq 3.3)$ si X es $N(2, 1)$.
4. $P(|X - 3| > 2)$ si X es $N(3, 4)$.

Ejercicio 5.9. ★ Supongamos una variable aleatoria X que sigue una distribución normal de desviación típica 1 ($\sigma = 1$), y media (μ) desconocida. Si nos dan como dato que la función de densidad en $x=1$ vale $\frac{1}{\sqrt{2\pi}}$, y teniendo en cuenta que la función de densidad viene dada por:

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

calcular $P(|X - 2| < 3)$.

Ejercicio 5.10. Se sabe que el nivel de colesterol en varones de más de 30 años sigue una distribución normal, de media 220 mg/dl y desviación típica 30 mg/dl. Realizando un estudio sobre 20000 varones mayores de 30 años,

1. ¿Cuántos se espera que tengan su nivel de colesterol entre 210 y 240 mg/dl?
2. Si un nivel de colesterol por encima de 250 mg/dl puede provocar trombosis, ¿cuántos tienen riesgo de trombosis?
3. ¿Cuál será el nivel de colesterol por encima del cual está el 20% de la población?

 Solución

1. $P(210 \leq X \leq 240) = 0.3781 \rightarrow 7561.3$ varones.
2. $P(X > 250) = 0.1587 \rightarrow 3173.1$ varones.
3. $P_{80} = 245.2486$ mg/dl.

Ejercicio 5.11. Entre los diabéticos, el nivel de glucosa en la sangre en ayunas, puede suponerse de distribución aproximadamente normal, con media 106 mg/100 ml y desviación típica 8 mg/100 ml.

1. Calcular la probabilidad de que una persona diabética elegida al azar tenga un nivel de glucosa inferior a 120mg/100ml.
2. ¿Qué porcentaje de diabéticos tendrá niveles entre 90 y 120 mg/100 ml?
3. Calcular e interpretar el primer cuartil del nivel de glucosa.

 Solución

1. $P(X \leq 120) = 0.9599$.
2. $P(90 < X < 120) = 0.9371$, es decir, un 93.71%.
3. 100.64 mg/100 ml.

Ejercicio 5.12. Se sabe que la distribución de probabilidad de una variable aleatoria X sigue la ley normal. Si conocemos $P(X > 3.6) = 0.3821$ y $P(X < 5.8) = 0.9192$, calcular μ y σ de la distribución normal.

Ejercicio 5.13. En una población con 40000 personas, se sabe que 2276 tienen entre 0.80 y 0.84 miligramos de bilirrubina por decilitro de sangre, y que 11508 tienen más de 0.84. Suponiendo que la concentración de bilirrubina en sangre sigue una distribución normal, se pide:

1. Calcular su media y su desviación típica.
2. Calcular el número de personas con más de 1 miligramo de bilirrubina por decilitro de sangre.

 Solución

1. $\mu = 0.7$ mg y $\sigma = 0.25$ mg.
2. $P(X > 1) = 0.1151$ y el número de personas con más de 1 mg de bilirrubina es $0.1151 \cdot 40000 = 4604$.

Ejercicio 5.14. ★ En un examen, el 63% de los alumnos ha obtenido una nota superior a 5, y el 44% entre 5 y 7. Suponiendo que las notas siguen una distribución normal:

1. Calcular la media y la desviación típica de las notas.
2. Calcular el porcentaje de alumnos con nota superior a 8.
3. ¿Cuál es la nota por encima de la cual está el 5% de los alumnos?

 Solución

Llamando X a la nota obtenida en el examen, se tiene que $X \sim N(\mu, \sigma)$.

1. $\mu = 5.55$ puntos y $\sigma = 1.65$ puntos.
2. $P(X > 8) = 0.0648$, es decir, el 6.48% de los alumnos han tenido una nota superior a 8.
3. 8.21 puntos.

Ejercicio 5.15. ★ Se realiza un estudio sobre la tensión arterial máxima en varones de edad comprendida entre 30 y 40 años. Como consecuencia del estudio se sabe que el 40% la tiene por encima de 140 mmHg y el 20% por debajo de 110 mmHg. Suponiendo que la tensión arterial máxima sigue una distribución normal, se pide:

1. Calcular la media y la desviación típica.
2. Calcular el porcentaje de hipertensos, si se considera hipertenso a una persona cuya tensión arterial máxima sea mayor que 150 mmHg.
3. ¿Cuál es el valor de la tensión arterial máxima por encima del cual se espera que esté el 10% de la población?

Ejercicio 5.16. ★ Un grupo de 100 alumnos realiza un examen. Una vez corregido, el profesor observa que la nota media es 4.2 y que 32 alumnos han tenido una calificación superior a 5. Suponiendo que las notas siguen una distribución normal, ¿cuántos alumnos habrán obtenido una calificación superior a 7?

 Solución

$$P(X > 7) = 0.0508 \rightarrow 5.1 \text{ alumnos.}$$

Ejercicio 5.17. ★ Se supone que la tensión arterial de los habitantes de una población de 20000 habitantes sigue una distribución normal, cuya media es 13 y su rango intercuartílico 4. Se pide:

1. ¿Cuántas personas tienen una tensión por encima de 16?
2. ¿Cuánto tendrá que disminuir la tensión de una persona que tiene 16 para situarse en el 40% de la población con tensión más baja?

 Solución

1. $P(X > 16) = 0.1587$ y el número de personas con tensión por encima de 16 mmHg es $0.1587 \cdot 20000 = 3174$.
2. Debe disminuir 3.75 mmHg.

Ejercicio 5.18. ★ En una población de 30000 individuos se está interesado en medir el volumen sanguíneo de sus individuos. Se sabe que la desviación típica de la población es 0.4 litros y que el 50% de los individuos tienen un volumen superior a 4.8 litros. ¿Cuántos individuos presentarán un volumen menor de 4.3 litros?

 Solución

El número de individuos con menos de 4.3 litros de sangre es 3168.

Ejercicio 5.19. Se sabe que el nivel de potasio en personas sanas sigue una distribución normal de media 4.4 mg/l y desviación típica 0.4 mg/l. En un estudio realizado sobre 20000 personas sanas:

1. ¿Cuántos se espera que tengan un nivel de potasio entre 3.7 y 5 mg/l?
2. ¿Cuántos se espera que tengan un nivel de potasio por encima de 5.3 mg/l?
3. ¿Cuál será el nivel de potasio por encima del cual está la cuarta parte de la población?

Ejercicio 5.20. Se consideran las variables aleatorias X_1 y X_2 . La variable X_1 sigue una distribución normal de media μ y desviación típica σ , mientras que la variable X_2 sigue también una distribución normal de media $\mu + 1$ y desviación típica σ . Si la probabilidad de que X_1 tome valores superiores a 14.2 es 0.5636, y la de que X_2 tome valores inferiores a 17.4 es 0.6103, se pide:

1. Hallar los valores de μ y σ .
2. Si se rechazan los individuos que están fuera del intervalo (12, 18), hallar los porcentajes de rechazo correspondientes a X_1 y X_2 .
3. Si se desea seleccionar el 20% de individuos que tengan los valores más altos de X_1 , ¿cuál será el valor de X_1 a partir del cual se seleccionarán?

 Solución

1. $\mu = 15$ y $\sigma = 5$.
2. $P(12 < X_1 < 18) = 0.4515$, luego el porcentaje de rechazos es $100\% - 45.15\% = 54.85\%$.
 $P(12 < X_2 < 18) = 0.4436$, luego el porcentaje de rechazos es $100\% - 44.36\% = 55.64\%$.
3. 19.21.

Ejercicio 5.21. ★ El peso de los recién nacidos no prematuros en una ciudad sigue una distribución normal de media y desviación típica desconocidas. Teniendo en cuenta que, de un total de 200 recién nacidos no prematuros, 15 han pesado más de 4 kg y 25 menos de 2,5 kg:

1. ¿Cuáles son la media y la desviación típica del peso?
2. ¿Cuántos niños no prematuros habrán nacido con un peso entre 3 y 3.5 kg?
3. Si los médicos consideran peligrosos los pesos por debajo del percentil 10, ¿cuál será dicho peso?, ¿cuántos niños habrán nacido con un peso por debajo de dicho percentil?

 Solución

Llamando X al peso de los recién nacidos no prematuros, se tiene que $X \sim N(\mu, \sigma)$.

1. $\mu = 3.17$ kg y $\sigma = 0.58$ kg.
2. 66 niños.
3. $P_{10} = 2.43$ kg y habrán nacido 20 niños por debajo de este peso.

Ejercicio 5.22. La cantidad de fuel en aguas costeras próximas al hundimiento de un petrolero sigue una distribución normal de media $20 \mu\text{g/l}$ y desviación típica $8 \mu\text{g/l}$. Si se considera que existe contaminación si el nivel de fuel excede los $5 \mu\text{g/l}$, ¿Cual es la probabilidad de que al realizar una medición al azar no se detecte la contaminación? ¿Y si en lugar de realizar una medición, realizamos dos y tomamos la media?

Ejercicio 5.23. ★ Una solución contiene virus bacteriófagos T_4 en una concentración de $4 \cdot 10^6$ por mm^3 . En la misma solución hay $2 \cdot 10^6$ bacterias por mm^3 . Suponiendo que todos los virus infectan bacterias y que se distribuyen al azar entre las mismas, se pide:

1. ¿Cuál es el porcentaje de bacterias que no están infectadas por el virus?
2. ¿Qué porcentaje de bacterias tendrá al menos 2 virus fijados sobre ellas?
3. Si tomamos un volumen pequeño de dicha solución en el que hay 4 bacterias, ¿cuál es la probabilidad de que alguna esté infectada?
4. Si tomamos un volumen en el que hay 10000 bacterias, ¿cuál es la probabilidad de que estén infectadas al menos 8600?

Ejercicio 5.24. El coeficiente intelectual es una puntuación derivada de los test de inteligencia que tiene media 100 puntos y desviación típica 15. Si se considera que una persona por encima de 145 es una superdotada, ¿qué porcentaje de superdotados habrá en la población? Si se considera que el 1% de las personas con menor coeficiente intelectual son deficientes, ¿por debajo de qué coeficiente estarán dichas personas?

 Solución

Llamando X al coeficiente intelectual, se tiene que $X \sim N(100, 15)$.
 $P(X > 145) = 0.0013$, luego el 0.13% de las personas son superdotadas.
 $P_1 = 65.10$, por debajo de este coeficiente las personas son deficientes.

Ejercicio 5.25. ★ Si se supone que la concentración de plomo en suelo sigue una distribución normal de media 320 mg/Kg y desviación típica desconocida, y sabiendo que el 25% de los suelos presenta una concentración de plomo por abajo de 240 mg/Kg:

1. ¿Cuánto vale la desviación típica de la distribución?
2. Suponiendo que niveles de Plomo por arriba de 250 mg/Kg son peligrosos para la salud, y se catalogan como contaminados todos los suelos que sobrepasen ese nivel, ¿cuál es la probabilidad de que un suelo analizado esté contaminado?
3. Si analizamos 10000 suelos diferentes, ¿cuántos esperamos que no estén contaminados?
4. ¿Cuánto vale el rango intercuartílico de la distribución? ¿Qué interpretación tienen?

Ejercicio 5.26. ★ Para la recuperación de una lesión en encinas parasitadas por un tipo de hongo se dispone de dos tratamientos alternativos A y B. El tiempo de recuperación con el procedimiento A sigue una distribución normal de media 16 meses y desviación típica 3 meses, mientras que con el procedimiento B la media es de 20 meses y la desviación típica de 1.

1. ¿Qué tanto por ciento de recuperaciones se producen con cada tratamiento entre 18 y 21 meses?
2. ¿Qué tratamiento consigue antes la recuperación del 90% de las encinas tratadas?. Justificar adecuadamente la respuesta.
3. Si a una encina le aplicamos los dos tratamientos, que suponemos que actúan de forma independiente, ¿cuál es la probabilidad de que cure antes de 19 meses?

Ejercicio 5.27. Calcular:

1. $P(T \leq 1.476)$ si $T \sim T(5)$.
2. $P(T \geq 0.69)$ si $T \sim T(16)$.
3. El valor t_0 tal que $P(T < t_0) = 0.995$, con $T \sim T(12)$.
4. El valor t_0 tal que $P(T > t_0) = 0.01$, con $T \sim T(8)$.

Ejercicio 5.28. Calcular:

1. $P(X \leq 5.23)$ si $X \sim \chi^2(12)$.
2. $P(X \geq 1.65)$ si $X \sim \chi^2(8)$.
3. El valor x_0 tal que $P(X < x_0) = 0.995$, con $X \sim \chi^2(18)$.
4. El valor x_0 tal que $P(X > x_0) = 0.25$, con $X \sim \chi^2(7)$.

Ejercicio 5.29. Calcular:

1. El valor f_0 tal que $P(F < f_0) = 0.9$, con $F \sim F(12, 8)$.
2. El valor f_0 tal que $P(F > f_0) = 0.025$, con $F \sim F(5, 7)$.

Ejercicio 5.30. ★ Suponiendo que la duración del embarazo sigue una distribución normal de media y desviación típica desconocidas, y teniendo en cuenta que el 80% de las mujeres dan a luz antes de 40 semanas, y que el 70% lo hacen después de 38 semanas y 2 días, se pide:

1. Calcular la media y la desviación típica de la distribución dadas en número de semanas.
2. Suponiendo un hospital en el que se han producido 2000 nacimientos, ¿cuántos lo habrán hecho con más de 282 días de gestación?
3. Suponiendo una mujer que ha tenido 2 embarazos y que la duración del segundo ha sido independiente del primero, ¿cuál es la probabilidad de que alguno de los partos se haya producido antes de los 275 días de embarazo?

Ejercicio 5.31. ★ El gasto mensual en medicamentos de las familias españolas antes de la crisis seguía una distribución normal $X \sim N(160, \sigma)$, mientras que ahora sigue una distribución normal $Y \sim N(\mu, 2\sigma)$. Sabiendo que antes de la crisis el 10% de las familias gastaba más de 200 euros y que después el 40% gastaba menos de 100 euros, se pide:

1. ¿Qué porcentaje de familias gastarán ahora entre 110 y 120 euros?
2. ¿Qué percentil de la distribución actual se corresponde con el tercer decil de la distribución de antes de la crisis?

Solución

1. Teniendo en cuenta que el gasto antes de la crisis seguía una distribución normal de media 160 euros y desviación típica desconocida, pero nos dan el dato de que antes de la crisis el 10% de las familias gastaba más de 200, fácilmente podemos aprovechar este dato para calcular la desviación típica, y utilizamos esta desviación típica y el dato de que el 40% de las familias gastan ahora menos de 100 euros para calcular la media del gasto después. Con ello, podemos calcular la probabilidad que nos piden: $P(110 \leq Y \leq 120)$. Utilizamos el dato de que $P(X > 200) = 0.1000$:

$$\begin{aligned} P(X > 200) &= 1 - P(X \leq 200) = 1 - P\left(Z \leq \frac{200 - 160}{\sigma}\right) = 1 - P\left(Z \leq \frac{40}{\sigma}\right) = \\ &= 1 - F\left(\frac{40}{\sigma}\right) = 0.1000 \Leftrightarrow F\left(\frac{40}{\sigma}\right) = 0.9000. \end{aligned}$$

Mediante la tabla de la distribución normal tipificada vemos que $F(1.28) = 0.8997$, que es la probabilidad acumulada más cercana al 0.9000 que buscamos. Por lo tanto:

$$\frac{40}{\sigma} = 1.28 \Leftrightarrow \sigma = \frac{40}{1.28} = 31.25 \text{ euros.}$$

Para calcular la media μ del gasto después, utilizamos el dato de que $P(Y < 100) = 0.4000$:

$$\begin{aligned} P(Y < 100) &= P\left(Z \leq \frac{100 - \mu}{2\sigma}\right) = P\left(Z \leq \frac{100 - \mu}{2 \cdot 31.25}\right) = P\left(Z \leq \frac{100 - \mu}{62.50}\right) = \\ &= F\left(\frac{100 - \mu}{62.50}\right) = 0.4000. \end{aligned}$$

De nuevo, buscamos en la tabla de la normal tipificada y observamos que $F(-0.25) = 0.4013$, que es la probabilidad acumulada más cercana al 0.4000 que buscamos. Por lo tanto:

$$\frac{100 - \mu}{62.50} = -0.25 \Leftrightarrow \mu = 100 + 15.625 = 115.625 \text{ euros.}$$

Por lo tanto, concluimos que $X \sim N(160, 31.25)$ y $Y \sim N(115.625, 62.50)$.

Por último, calculamos la probabilidad (el porcentaje) de que el gasto ahora esté entre 110 y 120 euros:

$$\begin{aligned} P(110 \leq Y \leq 120) &= P\left(\frac{110 - 115.625}{62.50} \leq Z \leq \frac{120 - 115.625}{62.50}\right) = P(-0.09 \leq Z \leq 0.07) = \\ &= F(0.07) - F(-0.09) \stackrel{1}{=} 0.5279 - 0.4641 = 0.0638 = 6.38\%. \end{aligned}$$

(1) Mirando en la tabla de la función de distribución de la normal estándar.

2. Para calcular el tercer decil de antes de la crisis, sabemos que $P(X \leq D_3) = 0.3000$. Por lo tanto:

$$P\left(Z \leq \frac{D_3 - 160}{31.25}\right) = F\left(\frac{D_3 - 160}{31.25}\right) = 0.3000$$

Mediante la tabla de la normal tipificada, la probabilidad acumulada más cercana es: $F(-0.52) = 0.3015$. Por lo tanto:

$$\frac{D_3 - 160}{31.25} = -0.52 \Leftrightarrow D_3 = 143.75 \text{ euros,}$$

y para ver qué porcentaje de familias consumen ahora, durante la crisis, menos de 143.75 euros:

$$P(Y \leq 143.75) = P\left(Z \leq \frac{143.75 - 115.625}{62.50}\right) = P(Z \leq 0.45) = F(0.45) = 0.6736.$$

Por lo tanto, los 143.75 euros son el percentil 67 de la distribución de gasto durante la crisis.

Ejercicio 5.32. Un test diagnóstico para niños de 10 años con problemas de lectura da puntuaciones que se distribuyen normalmente con una media de 80 y una varianza de 100. Se pide:

1. Dar la probabilidad de que un niño seleccionado al azar tenga una puntuación de:
 1. menos de 68;
 2. entre 75 y 90;
2. El test indica que el niño tiene un problema del lenguaje si su puntuación está por debajo del 10% de la población que realizó el test. ¿Por debajo de qué puntuación el test diagnosticará problemas de lenguaje?
3. Si se selecciona una muestra de 16 niños:
 1. ¿Cuántos se espera que tengan una puntuación por encima de 68?
 2. ¿Cuál es la probabilidad de que su puntuación media supere los 84 puntos?

Solución

Llamando X a la puntuación del test diagnóstico, se tiene que $X \sim N(80, 10)$.

1. $P(X < 68) = 0.1151$ y $P(75 < X < 90) = 0.5328$.
2. $P_{10} = 67.18$ puntos.
3. $P(X > 68) = 0.8849$ y el número esperado niños con puntuaciones por encima de 68 es $0.8849 \cdot 16 = 14.16$.
 $\bar{x} \sim N(80, 2.5)$ y $P(\bar{x} > 84) = 0.0548$.

Ejercicio 5.33. ★ El tiempo de recuperación de una determinada lesión de rodilla en futbolistas profesionales sigue una distribución normal de media 50 días y desviación típica 10 días. Si dentro de 65 días hay una final, ¿cuál es la probabilidad de que un jugador profesional que acaba de lesionarse la rodilla se pierda la final? Si las semifinales son dentro de 40 días y se acaban de lesionar la rodilla 4 jugadores, ¿cuál es la probabilidad de que alguno de ellos pueda jugar la semifinal?

Solución

Llamemos X a la variable aleatoria que mide el número de días que tarda en recuperarse un futbolista con la lesión de rodilla. Entonces, según el enunciado, $X \sim N(50, 10)$. La probabilidad de que un futbolista tarde más de 65 días en curarse es $P(X > 65)$. Para calcularla, tipificamos

$$P(X > 65) = P\left(\frac{X - 50}{10} > \frac{65 - 50}{10}\right) = P(Z > 1.5) = 1 - P(Z \leq 1.5) = 1 - F(1.5),$$

y mirando en la tabla de la función de distribución de la normal estándar, la probabilidad acumulada hasta 1.5 es 0.9332. Por tanto, $P(X > 65) = 1 - 0.9332 = 0.0668$.

Para responder a la segunda pregunta, necesitamos otra variable aleatoria que mida el número de jugadores que tardan menos de 40 días en curarse de entre 4 que tenemos lesionados. Llamemos Y a dicha variable. Puesto que Y responde a un esquema de repetición de pruebas (una por jugador), donde cada prueba consiste en ver si el jugador tarda menos de 40 días en curarse (éxito) o no (fracaso), entonces sigue una distribución binomial de parámetros $n = 4$ jugadores y $p = P(X < 40)$. Necesitamos, en primer lugar, calcular esta probabilidad. Al igual que antes, tipificando y mirando en la tabla de la función de distribución de la normal estándar obtenemos

$$P(X < 40) = P\left(\frac{X - 50}{10} < \frac{40 - 50}{10}\right) = P(Z < -1) = F(-1) = 0.1587.$$

Así pues, $Y \sim B(4, 0.1587)$, y la probabilidad que nos piden es

$$P(Y \geq 1) = 1 - P(Y < 1) = 1 - P(Y = 0) = 1 - \binom{4}{0} 0.1587^0 (1 - 0.1587)^4 = 1 - 0.8413^4 = 0.499.$$

Ejercicio 5.34. ★ Según el teorema central del límite, se sabe que para muestras grandes ($n \geq 30$) la media muestral $\bar{x}$ sigue una distribución normal $N(\mu, \sigma/\sqrt{n})$, donde μ es la media de la población y σ su desviación típica.

Se sabe que en una población la elongación del triceps sural tiene una media de 60 cm y una desviación típica de 15 cm. Si se toma una muestra de tamaño 30, ¿cuál es la probabilidad de obtener una media muestral mayor de 62? Si se considera que una muestra es atípica si su media muestral está por debajo del percentil 5 de su distribución, ¿se puede considerar atípica una muestra de 60 individuos con $\bar{x} = 57$?

Solución

Sea $\bar{X}$ la variable que mide la elongación media del triceps sural en muestras de tamaño 30. Por el teorema central del límite, puesto que tomamos una muestra grande, tenemos que $\bar{X} \sim N(\mu, \sigma/\sqrt{n}) = N(60, 15/\sqrt{30}) = N(60, 2.74)$. Así pues, la probabilidad que nos piden es

$$\begin{aligned} P(\bar{X} > 62) &\stackrel{(1)}{=} P\left(\frac{\bar{X} - 60}{2.74} > \frac{62 - 60}{2.74}\right) = P(Z > 0.73) = \\ &= 1 - P(Z \leq 0.73) = 1 - F(0.73) \stackrel{(2)}{=} 1 - 0.7673 = 0.2327. \end{aligned}$$

(1) Tipificando.

(2) Mirando en la tabla de la función de distribución de la normal estándar.

Para responder a la segunda pregunta, llamemos $\bar{Y}$ a la variable que mide la elongación media del triceps sural en muestras de tamaño 60. Al igual que antes, por el teorema central del límite, $N(\mu, \sigma/\sqrt{n}) = N(60, 15/\sqrt{60}) = N(60, 1.94)$. Para saber si una muestra es atípica, tenemos que calcular el percentil 5 de $\bar{Y}$. Dicho valor cumplirá $P(\bar{Y} \leq y_0) = 0.05$. Tipificando tenemos

$$P(\bar{Y} \leq y_0) = P\left(\frac{\bar{Y} - 60}{1.94} \leq \frac{y_0 - 60}{1.94}\right) = P\left(Z \leq \frac{y_0 - 60}{1.94}\right) = F\left(\frac{y_0 - 60}{1.94}\right) = 0.05.$$

Y buscando en la tabla de la función de distribución de la normal estándar, tenemos que

$$\frac{y_0 - 60}{1.94} = -1.65 \Leftrightarrow y_0 = -1.65 \cdot 1.94 + 60 = 56.8 \text{ cm.}$$

Como la media muestral obtenida es 57 y no está por debajo de 56.8 que es el percentil 5, concluimos que la muestra obtenida no se puede considerar atípica.

Ejercicio 5.35. ★ Para el tratamiento de una enfermedad se utilizan 3 fármacos diferentes: A en un 20% de los casos, B en un 30%, y C en un 50%. Los tratados con A curan en media al cabo de 6.2 días con una desviación típica de 1.1 días; los tratados con B en 7.1 días con una desviación típica de 1.3 y los tratados con C en 7.4 días con una desviación típica de 0.9. Con ello:

1. En qué fármaco es mayor el percentil 90, en B o en C ?
2. Qué probabilidad total hay de que tomado un enfermo al azar haya curado antes de 8.1 días?

3. Sabiendo que un enfermo ha tardado en curar más de 7.2 días, con qué fármaco es más probable que haya sido tratado? Justificar adecuadamente la respuesta.

 Solución

Llamando X al tiempo de curación y A , B y C a los sucesos consistentes en haber sido tratado respectivamente con cada uno de los fármacos:

1. El percentil 90 en el fármaco B es 8.7660 días y en C es 8.5534 días, así que es mayor en el fármaco B .
2. $P(X < 8.1) = 0.8161$.
3. $P(A/X > 7.2) = 0.0771$, $P(B/X > 7.2) = 0.2989$ y $P(C/X > 7.2) = 0.624$, así que es más probable que haya sido tratado con el fármaco C .

Ejercicio 5.36. ★ Se sabe que la concentración de urea láctica, en mg/cm^3 , en vacas sanas sigue una distribución normal de media 28 y desviación típica 1.5, mientras que en vacas con una determinada enfermedad sigue una distribución normal de media 34.5 y desviación típica 2. Para detectar la enfermedad se realiza un test diagnóstico que da positivo cuando el nivel de urea láctica está por encima de 32. Se pide:

1. ¿Qué sensibilidad $P(+/E)$ y qué especificidad $P(-/\bar{E})$ tiene el test diagnóstico?
2. Si el test da positivo en un 5% de los casos analizados, ¿cuál será el porcentaje de vacas enfermas en la población?
3. Teniendo en cuenta el porcentaje de vacas enfermas anterior, ¿cuál será la probabilidad de diagnóstico acertado con este test?

 Solución

1. $P(+/E) = 0.8944$ y $P(-/\bar{E}) = 0.9962$.
2. $P(E) = 0.0465$.
3. $P(E \cap +) + P(\bar{E} \cap -) = 0.0416 + 0.9498 = 0.99143$.

Ejercicio 5.37. ★ Sea Z una variable aleatoria continua que sigue un modelo de distribución normal estándar $N(0, 1)$. Calcular las siguientes probabilidades utilizando la tabla de la función de distribución.

1. $P(Z < 1.24)$
2. $P(Z > -0.68)$
3. $P(-1.35 \leq Z \leq 0.44)$

 Solución

1. $P(Z < 1.24) = 0.8925$.
2. $P(Z > -0.68) = 0.7517$.
3. $P(-1.35 \leq Z \leq 0.44) = 0.5815$.

Ejercicio 5.38. ★ Sea Z una variable aleatoria continua con distribución normal estándar. Determinar el valor de x en los siguientes casos utilizando la tabla de la función de distribución.

1. $P(Z < x) = 0.6406$.
2. $P(Z > x) = 0.0606$.
3. $P(0 \leq Z \leq x) = 0.4783$.
4. $P(-1.5 \leq Z \leq x) = 0.2313$.
5. $P(-x \leq Z \leq x) = 0.5467$.

 Solución

1. $x = 0.3601$.
2. $x = 1.5498$.
3. $x = 2.0198$.
4. $x = -0.5299$.
5. $x = 0.7499$.

Ejercicio 5.39. ★ Sea X una variable aleatoria continua con distribución normal $N(10, 2)$.

1. Calcular $P(X \leq 10)$.
2. Calcular $P(8 \leq X \leq 14)$.
3. Calcular el rango intercuartílico.
4. Calcular el tercer decil.

 Solución

1. $P(X \leq 10) = 0.5$.
2. $P(8 \leq X \leq 14) = 0.8186$.
3. $RI = 2.698$.
4. $D_3 = 8.9512$.

Ejercicio 5.40. ★ Un estudio trata de determinar el efecto de una dieta baja en grasas sobre el tiempo de vida de un tipo de ratas. Las ratas se dividieron en dos grupos, uno con una dieta normal y el otro con una dieta baja en grasas. Se supone que el tiempo

de vida sigue un modelo de distribución normal con igual varianza pero con distintas medias. Si el 20% de las ratas con dieta normal vivieron más de 12 meses, el 5% vivieron menos de 8 meses y el 85% de las ratas con dieta baja en grasas vivieron más de 11 meses,

1. ¿Cuál es la media y la desviación típica del tiempo de vida de las ratas con una dieta baja en grasas?
2. Si en el experimento había un 40% de ratas con dieta normal y un 60% de ratas con dieta baja en grasas, ¿cuál es la probabilidad de que una rata elegida al azar muera antes de 9 meses?

Solución

Llamando X al tiempo de vida de las ratas.

1. $\mu = 12.6673$ meses y $s = 1.6087$ meses.
2. $P(X < 9) = 0.068$.

Ejercicio 5.41. ★ Un test diagnóstico para detectar el dopaje en atletas da un resultado positivo cuando la concentración de una determinada sustancia en sangre es mayor de $4 \mu\text{g/ml}$. Si la distribución de la concentración de a sustancia en sangre en atletas dopados sigue un modelo de distribución normal con media $4.5 \mu\text{g/ml}$ y desviación típica $0.2 \mu\text{g/ml}$, y en atletas no dopados sigue un modelo de distribución normal con media $3 \mu\text{g/ml}$ y desviación típica $0.3 \mu\text{g/ml}$,

1. ¿cuál es la sensibilidad y la especificidad del test?
2. Si hay un 10% de atelas dopados en una competición, ¿cuál es el valor predictivo positivo del test?

Solución

Llamando D al suceso aleatorio consistente en estar dopado, X a la concentración de la sustancia en atletas dopados e Y a la concentración de la sustancia en atletas no dopados.

1. Sensibilidad $P(+|D) = P(X > 4) = 0.9938$ y especificidad $P(-|\bar{D}) = P(Y < 4) = 0.9996$.
2. VPP $P(D|+) = 0.9961$.

6 Intervalos de confianza

Ejercicio 6.1. Una muestra aleatoria de tamaño 81 extraída de una población normal con $\sigma^2 = 64$, tiene una $\bar{x} = 78$. Calcular el intervalo de confianza del 95% para μ .

 Solución

$$\mu \in 79 \pm 1.742 = (76.258, 79.742).$$

Ejercicio 6.2. Para determinar si un pescado es o no apto para el consumo por su contenido en Hg (mercurio), se realizan 15 valoraciones obteniendo una media de 0.44 ppm (partes por millón) de Hg, y una desviación típica de 0.08 ppm. Calcular los límites de confianza para la media, a un nivel de significación $\alpha = 0.1$.

 Solución

$$\mu \in 0.44 \pm 0.0376 \text{ ppm} = (0.4024\text{ppm}, 0.4776\text{ppm}).$$

Ejercicio 6.3. Se obtuvieron cinco determinaciones del pH de una solución con los siguientes resultados:

$$7.90 \quad 7.85 \quad 7.89 \quad 7.86 \quad 7.87.$$

Hallar unos límites de confianza de la media de todas las determinaciones del pH de la misma solución, al nivel de significación $\alpha = 0.01$.

 Solución

$$\mu \in 7.874 \pm 0.0426 = (7.8314, 7.9166).$$

Ejercicio 6.4. Se desea saber cuál debe ser el tamaño muestral mínimo de una muestra para poder realizar la estimación de la tasa media de glucosa plasmática de una determinada población, con un nivel de confianza 0.95 y pretendiendo una amplitud de 2.5 mg.

NOTA: En una muestra previa de tamaño 10 se obtuvo una desviación típica de 10 mg.

 Solución

249 individuos.

Ejercicio 6.5. Para que un fármaco sea efectivo, la concentración de un determinado principio activo debe ser 20 mg/mm^3 . Se recibe un lote de dicho fármaco y se analizan 10 para medir la concentración del principio activo, obteniendo los resultados siguientes:

17.6 19.2 21.3 15.1 17.6 18.9 16.2 18.3 19 16.4.

En vista de los resultados, ¿podremos rechazar el lote con una confianza 0.95 de no equivocarnos?

 Solución

$$17.96 \pm 1.278 \text{ mg/mm}^3 = (16.682 \text{ mg/mm}^3, 19.238 \text{ mg/mm}^3).$$

Ejercicio 6.6. ★ En un estudio sobre el consumo anual de litros de cerveza entre la población de menores de 18 años de una ciudad se obtuvo la siguiente muestra:

42 16 60 29 7 20 30 25 38 5.

Se pide:

1. Calcular el intervalo de confianza del 95% para la media. Si se considera que un consumo medio por encima de 40 litros es peligroso, ¿existen pruebas significativas para afirmar que la población de partida no está en peligro?
2. ¿Qué tamaño muestral mínimo hubiese sido necesario para conseguir un intervalo de confianza de amplitud 5?

Ejercicio 6.7. En una explotación minera se mide el contenido en mercurio de las rocas extraídas. Tras analizar 20 rocas, se obtiene un contenido medio del 10.8% y una desviación típica de 2.7%. Se pide:

1. Si, para que la explotación sea rentable, el porcentaje medio de contenido de mercurio debe ser superior al 10%, ¿existen pruebas para afirmar que la explotación será rentable?
2. ¿Y si para que la explotación sea rentable, el contenido de mineral debe tener cierta uniformidad ($\sigma < 3$)?

Ejercicio 6.8. Para ver si una población de aves se ha visto afectada por los vertidos tóxicos de una fábrica, se pretende estimar la proporción de aves contaminadas con metales pesados. Para ello se realiza un sondeo preliminar con 30 aves, de las que 5 resultan estar contaminadas. Se pide:

1. Construir un intervalo de confianza del 95% para la proporción de aves contaminadas.
2. Si se desea estimar la proporción con un error máximo de $\pm 2\%$, ¿qué tamaño muestral habría que tomar?

Ejercicio 6.9. Se realizó un estudio sobre el contenido de principio activo de un determinado fármaco a partir de una muestra, determinándose los siguientes resultados en mg/cm^3 :

$$46.4 - 46.1 - 45.8 - 47.0 - 46.1 - 45.9 - 45.8 - 46.9 - 45.2 - 46.0.$$

Obtener un intervalo de confianza del 95% para la varianza del contenido de principio activo de dicho fármaco, suponiendo que sigue una distribución normal.

 Solución

$$\sigma^2 \in (0.1356(\text{mg}/\text{cm}^3)^2, 0.9541(\text{mg}/\text{cm}^3)^2).$$

Ejercicio 6.10. Se desea hacer un estudio estadístico sobre el número de hematíes en las mujeres. Se selecciona una muestra de 25 mujeres y se obtiene un número medio de hematíes de 4300000 con una desviación típica de 200000 (en cada milímetro cúbico de sangre). Calcular el intervalo de confianza para la media y la varianza del número de hematíes de las mujeres en la población, con un nivel de significación 0.1.

 Solución

Intervalo de confianza del 95% para la media: $\mu \in 4.3 \cdot 10^6 \pm 69850$ hematíes = (4230150, 4369850).

Intervalo de confianza del 95% para la varianza: $\sigma^2 \in (2747 \cdot 10^{10}, 7246 \cdot 10^{10})$ hematíes².

Ejercicio 6.11. ★ Para determinar el nivel medio de colesterol en la sangre de una población, se realizaron análisis sobre una muestra de 8 personas, obteniéndose los siguientes resultados:

$$196 - 212 - 188 - 206 - 203 - 210 - 201 - 198$$

Hallar intervalos de confianza para la media y la varianza de nivel de colesterol con un nivel de significación 0.1, suponiendo que el nivel de colesterol en la población sigue una distribución normal.

Ejercicio 6.12. Para determinar la concentración media de albúmina en la sangre se realizaron mediciones sobre un grupo experimental obteniéndose los siguientes resultados, expresados en g/l:

$$38 - 42 - 46 - 37 - 49 - 42 - 40 - 36.$$

Obtener un intervalo de confianza para la varianza de la población con un nivel de significación 0.05.

 Solución

$$\sigma^2 \in (8.844\text{g}^2/\text{l}^2, 83.728\text{g}^2/\text{l}^2).$$

Ejercicio 6.13. El tiempo que tarda en hacer efecto un analgésico sigue una distribución aproximadamente normal. En una muestra de 20 pacientes se obtuvo una media de 25.4 minutos y una desviación típica de 5.8 minutos. Calcular el intervalo de confianza del 90% e interpretarlo. ¿Cuántos pacientes habría que tomar para poder estimar la media con una precisión de ± 1 minuto?

 Solución

Intervalo de confianza del 90% para μ : (23.0992, 27.7008).

Tamaño muestral para una precisión de ± 1 minuto: $n = 96$.

Ejercicio 6.14. En un estudio para el estado de la salud oral de una ciudad, se tomó una muestra elegida al azar de 280 varones entre 35 y 44 años y se contó el número de piezas dentarias en la boca. Tras la revisión pertinente, los dentistas informaron que había 70 individuos con 28 o más dientes. Se desea realizar una estimación por intervalo de confianza de la proporción de individuos de esta ciudad con 28 dientes o más, con un nivel de confianza 0.95.

 Solución

$$p \in 0.25 \pm 0.0507 = (0.1993, 0.3007).$$

Ejercicio 6.15. Leemos en una revista médica que la cuarta parte de los cancerosos de cierto tumor de estómago presentan vómitos, con una precisión o tolerancia del 10% y con una confianza del 99%. ¿Con cuántos pacientes se ha realizado el estudio?

Ejercicio 6.16. ★ Un país está siendo afectado por una epidemia de un virus. Para valorar la gravedad de la situación se tomaron 40 personas al azar y se comprobó que 12 de ellas tenían el virus. Determinar el intervalo de confianza para el porcentaje de infectados con un nivel de significación 0.05.

 Solución

$p \in (0.1580, 0.4420)$ con un 95% de confianza.

Ejercicio 6.17. Se desea obtener un intervalo de confianza del 95% para la diferencia de marcas obtenidas por chicos y chicas en una prueba física. Se toma una muestra de 50 chicas y 75 chicos, obteniendo las chicas una marca media de 76 y los chicos de 82. Además, se conocen las desviaciones típicas de las marcas obtenidas en las poblaciones de chicas y chicos, que son 6 y 8 respectivamente.

 Solución

$\mu_1 - \mu_2 \in 6 \pm 2.458 \text{ puntos} = (3.542 \text{ puntos}, 8.458).$

Ejercicio 6.18. Las temperaturas medias mensuales (en °C) durante el año 2001 en Madrid y Sevilla fueron:

Ciudad	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Madrid	7.2	8.4	12.2	13.7	16.7	23.3	24.2	25.5	20.4	16.2	8.1	4.2
Sevilla	12.1	13.6	16.8	19.0	20.9	27.0	26.6	28.2	24.7	21.3	13.8	11.5

¿Existen diferencias significativas en las temperaturas medias de ambas ciudades?

Ejercicio 6.19. Se está ensayando un nuevo procedimiento de rehabilitación para una cierta lesión. Para ello se trataron nueve pacientes con el procedimiento tradicional y otros nueve con el nuevo, y se midieron los días que tardaron en recuperarse, obteniéndose los siguientes resultados:

Método tradicional: 32 37 35 28 41 44 35 31 34

Método nuevo: 35 31 29 25 34 40 27 32 31

Se desea obtener un intervalo de confianza del 95% para la diferencia de las medias del tiempo de recuperación obtenido con ambos procedimientos. Se supone que los tiempos de recuperación siguen una distribución normal, y que las varianzas son aproximadamente iguales para los dos procedimientos.

 Solución

$$\mu_1 - \mu_2 \in 3.667 \pm 4.712 \text{ días} = (-1.045 \text{ días}, 8.379 \text{ días}).$$

Ejercicio 6.20. En un hospital pediátrico se comprobó que de 200 niños con un determinado síndrome, 48 murieron antes de cumplir un año de edad, mientras que sólo 25 de 125 niñas con el mismo síndrome murieron. ¿Se puede afirmar con cierta seguridad que el síndrome es más letal en los niños que en las niñas?

 Solución

$p_1 - p_2 \in 0.04 \pm 0.077 = (-0.037, 0.117)$ luego no se puede afirmar que el síndrome sea más letal en los niños que en las niñas con un 95% de confianza.

Ejercicio 6.21. Se ha realizado un estudio para investigar el efecto del ejercicio físico en el nivel de colesterol en la sangre. En el estudio participaron once personas, a las que se les midió el nivel de colesterol antes y después de desarrollar un programa de ejercicios. Los resultados obtenidos fueron los siguientes

Persona	Nivel previo	Nivel posterior
1	182	198
2	232	210
3	191	194
4	200	220
5	148	138
6	249	220
7	276	219
8	213	161
9	241	210
10	280	213
11	262	226

Hallar un intervalo de confianza del 90% para la diferencia del nivel medio de colesterol antes y después del ejercicio.

 Solución

$$\mu_{x_1 - x_2} \in 24.0909 \pm 19.4766 \text{ mg/dl} = (4.6143 \text{ mg/dl}, 43.5675 \text{ mg/dl}).$$

Ejercicio 6.22. Dos químicos *A* y *B* realizan 14 y 16 determinaciones, respectivamente, de plutonio. Los resultados obtenidos se muestran en la siguiente tabla

<i>A</i>	<i>B</i>
263.36	286.53
254.68	254.54
248.64	284.55
276.32	286.30
243.64	272.52
256.42	282.90
272.68	283.85
261.10	253.75
287.33	252.01
268.41	245.26
287.26	275.08
282.65	266.08
250.97	267.53
284.27	252.05
	253.82
	269.81

Se pide:

1. Calcular intervalos de confianza del 95% de confianza para cada caso.
2. ¿Se puede decir que existen diferencias significativas en la media?

Solución

1. Intervalo de confianza del 95% para la media de *A*: $\mu_A \in 266.98 \pm 8.681 = (258.299, 275.661)$.
Intervalo de confianza del 95% para la media de *B*: $\mu_B \in 267.91 \pm 7.677 = (260.233, 275.587)$.
2. $\mu_A - \mu_B \in -0.93 \pm 11.020 = (-11.95, 10.92)$, luego no hay diferencias significativas en las medias con un 95% de confianza.

Ejercicio 6.23. ★ Un equipo de investigación está interesado en ver si una droga reduce el colesterol en la sangre. Con tal fin se toma una muestra de 10 pacientes y determina el contenido de colesterol antes y después del tratamiento. Los resultados expresados en miligramos por cada 100 mililitros son los siguientes:

Paciente	1	2	3	4	5	6	7	8	9	10
Antes	217	252	229	200	209	213	215	260	232	216
Después	209	241	230	208	206	211	209	228	224	203

Se pide:

1. Construir la variable Diferencia que recoja la diferencia entre los niveles de colesterol antes y después del tratamiento, y calcular el intervalo de confianza con $1 - \alpha = 0.95$ para la media de dicha variable.
2. A la vista del intervalo anterior, ¿hay pruebas significativas de que la droga disminuye el nivel de colesterol en sangre?

 Solución

1. $\mu_{x_1-x_2} \in 7.4 \pm 7.572 \text{ mg/100ml} = (-0.172\text{mg/100ml}, 14.972\text{mg/100ml})$.
2. No se puede afirmar que la droga disminuya el colesterol con un 95% de confianza.

Ejercicio 6.24. ★ Se está ensayando un nuevo procedimiento de rehabilitación para una cierta lesión. Se sabe que de 80 deportistas tratados con el procedimiento tradicional, se recuperaron perfectamente 26, mientras que de los 20 tratados con el nuevo procedimiento se han recuperado 11. ¿Se puede afirmar con una confianza del 95% que el nuevo procedimiento es mejor que el tradicional?

Ejercicio 6.25. ★ En una muestra aleatoria de 200 personas, 114 están a favor de la fluoración de las aguas. Se pide:

1. Hallar el intervalo de confianza del 96% para la fracción de la población que está a favor de la fluoración de las aguas.
2. ¿Qué tamaño mínimo de muestras habría que tomar para tener una confianza del 96% de que la proporción muestral difiere menos de 0.02 de la proporción real de la población?

 Solución

1. $p \in 0.375 \pm 0.197 = (0.178, 0.572)$.
2. 2634 individuos.

Ejercicio 6.26. ★ Para ver si una campaña de publicidad sobre un fármaco ha influido en sus ventas, se tomó una muestra de 8 farmacias y se midió el número de fármacos vendidos durante un mes, antes y después de la campaña, obteniéndose los siguientes resultados:

Antes	147	163	121	205	132	190	176	147
Después	150	171	132	208	141	184	182	145

Obtener la variable diferencia y construir un intervalo de confianza para la media de la diferencia con un nivel de significación 0.05. ¿Existen pruebas suficientes para afirmar con un 95 % de confianza que la campaña de publicidad ha aumentado las ventas?

Ejercicio 6.27. ★ Para comparar la eficacia de dos tratamientos A y B en la prevención de repeticiones de infarto de miocardio, se aplicó el tratamiento A a 80 pacientes y el B a 60. Al cabo de dos años se observó que habían sufrido un nuevo infarto 14 pacientes de los sometidos al tratamiento A y 15 de los del B . Se pide:

1. Construir un intervalo de confianza del 95% para la diferencia entre las proporciones de personas sometidas a los tratamientos A y B que no vuelven a sufrir un infarto.
2. A la vista del resultado obtenido, razonar si con ese nivel de confianza puede afirmarse que uno de los tratamientos es más eficaz que el otro.

 Solución

- $p_A - p_B \in (-0.2126, 0.0626)$ con un nivel de confianza del 95%.
- No puede afirmarse que un tratamiento sea más eficaz que otro pues la diferencia de medias podría ser positiva, negativa o cero.

Ejercicio 6.28. La siguiente tabla muestra el porcentaje de industrias españolas y europeas según el consumo de energía primaria de las mismas durante el año 2002 (se estudiaron 1000 industrias).

Fuente energética	España	Resto de Europa
Petróleo	52.2%	40.4%
Carbón	15.2%	14.8%
Nuclear	13.0%	15.2%
Gas	12.8%	23.5%
Energías Renovables	6.5%	6.1%
Saldo Energético	0.2%	0%

Estudiar para qué energías el porcentaje de industrias de España es significativamente diferente del resto de Europa.

Ejercicio 6.29. ★ En un análisis de obesidad dependiendo del hábitat en niños menores de 5 años, se obtienen los siguientes resultados:

	Casos analizados	Casos con sobrepeso
Hábitat rural	1150	480
Hábitat urbano	1460	660

Se pide:

1. Construir un intervalo de confianza, con un nivel de significación 0.01, para la proporción de niños menores de 5 años con sobrepeso en el hábitat rural. Igualmente para el hábitat urbano.
2. Construir un intervalo de confianza, con un nivel de confianza del 95%, para la diferencia de proporciones de niños menores de 5 años con sobrepeso entre el hábitat rural y el urbano. A la vista del resultado obtenido, ¿se puede concluir, con un 95% de confianza, que la proporción de niños menores de 5 años con sobrepeso depende del hábitat?

Solución

1. $p_R \in (0.3799, 0.4548)$ y $p_U \in (0.4185, 0.4856)$ con un 99% de confianza.
2. $p_R - p_U \in (-0.0729, 0.0036)$ con un nivel de confianza del 95%, luego no se puede afirmar que haya diferencias en las proporciones de niños menores de 5 años con sobrepeso.

Ejercicio 6.30. Se ha realizado un estudio con 1000 mujeres que han dado a luz recientemente, elegidas al azar entre los registros de los diferentes hospitales de la comunidad de Madrid, para saber si un nuevo protocolo (visitas al médico y consumo de ciertos fármacos) resulta más efectivo para prevenir las infecciones (ya sean pre, intra o postparto). Del total, 750 han seguido el protocolo habitual, entre las cuales 35 han sufrido algún tipo de infección; mientras que 250 han seguido el protocolo nuevo y 9 de ellas han padecido alguna infección. ¿Se puede afirmar, con un 95% de confianza, que la proporción de mujeres que ha tenido algún tipo de infección ha sido diferente según el protocolo utilizado?

Ejercicio 6.31. ★ La siguiente tabla muestra los datos de emisiones de CO_2 y CH_4 (en Kg/hab) y el producto interior bruto per cápita (en miles US\$) de varios países en el último año:

País	CO_2	CH_4	PIB
Austria	7.60	0.97	38.40
España	6.73	0.81	30.12
Francia	5.71	0.94	33.19
EEUU	19.40	1.72	45.84
Alemania	9.80	0.83	34.18
Canadá	15.60	3.08	38.43
Italia	7.29	0.58	30.44
Japón	9.44	0.16	33.58
Australia	17.48	6.36	36.26
Reino Unido	8.99	0.76	35.13

País	CO ₂	CH ₄	PIB
Bolivia	1.05	3.44	40.13
Niger	0.1	0.12	0.67
Senegal	0.35	0.76	1.69
Pakistán	0.65	0.59	2.59
Filipinas	0.83	0.46	3.38
Perú	0.94	0.75	7.80
Túnez	2.17	0.48	7.47
Nepal	0.13	0.90	1.21
Nicaragua	0.7	0.32	2.62
Mauritania	0.97	0.85	2.01

¿Existen diferencias significativas en la emisión de CO₂ entre los países con un PIB superior a 10 mil US\$ y los países con un PIB inferior? ¿Y en las emisiones de CH₄? Justificar la respuesta.

Ejercicio 6.32. En una muestra de 250 estudiantes de una universidad, 146 hablaban inglés. ¿Entre qué valores estará el porcentaje de individuos de la universidad que hablan inglés, con un nivel de confianza del 90%?

 Solución

$p \in (0.5327, 0.6353)$ con un 90% de confianza.

Ejercicio 6.33. Si el porcentaje de individuos daltónicos de una muestra aleatoria es 18%, ¿cuál será el mínimo tamaño muestral necesario para conseguir una estimación del porcentaje de daltónicos con una confianza del 95% y un error menor del 3%?

 Solución

$n = 2520$.

Ejercicio 6.34. Un psicólogo está estudiando la concentración de una enzima en la saliva como un posible indicador de la ansiedad crónica. En un experimento se tomó una muestra de 12 neuróticos por ansiedad y otra de 10 personas con bajos niveles de ansiedad. En ambas muestras se midió la concentración de la enzima, obteniendo los siguientes resultados:

Con ansiedad:	2.60	2.90	2.60	2.70	3.91	3.15	3.94	2.46	2.91	3.88	3.55	3.96
Sin ansiedad:	2.37	1.10	2.55	2.64	2.20	2.12	2.47	2.90	1.66	2.72		

¿Se puede concluir a partir de estos datos que la población de neuróticos con ansiedad y la población de personas sin ansiedad son diferentes en el nivel medio de concentración de encimas? Justificar la respuesta.

 Solución

$\frac{\sigma_1^2}{\sigma_2^2} \in (0.3383, 4.7486)$ con un nivel de confianza del 95%, luego se puede suponer que las varianzas son iguales.
 $\mu_1 - \mu_2 \in (0.4296, 1.4510)$ con un 95% de confianza, luego se puede concluir que hay diferencias entre las medias.

Ejercicio 6.35. Las notas en Estadística de una muestra de 10 alumnos han sido:

6.3, 5.4, 4.1, 5.0, 8.2, 7.6, 6.4, 5.6, 4.3, 5.2

Dar una estimación puntual de la nota media, de la varianza y del porcentaje de aprobados en la clase.

 Solución

$\bar{x} = 5.81$ puntos, $\hat{s}^2 = 1.7721$ puntos² y $\hat{p} = 0.8$.

Ejercicio 6.36. Para estudiar si la estación del año influye en el estado de ánimo de la gente, se ha tomado una muestra 12 personas y se ha medido su nivel de depresión en verano e invierno mediante un cuestionario con puntuaciones de 0 a 100 (a mayor puntuación mayor depresión). Los resultados obtenidos fueron:

Invierno	65	72	84	31	80	61	75	52	73	79	85	71
Verano	60	51	81	45	62	53	70	52	64	51	67	62

¿Se puede afirmar que la estación del año influye en el estado de ánimo de la gente con un 99% de confianza? ¿Cómo influye?

 Solución

No se puede afirmar que la estación influya en el estado de ánimo ya que la media de la diferencia está en $(-0.7498, 19.0831)$ con un 99% de confianza.

Ejercicio 6.37. Para realizar una determinada tarea se necesitan personas con un cierto nivel de pericia. Una empresa está interesada en contratar a un grupo de personas que tengan la pericia necesaria para realizar la tarea, pero que además sean bastante

homogéneas en cuanto a la pericia, es decir, que su rendimiento sea parecido. Para ver si un grupo de alumnos que se han formado en dicha tarea cumplen los requisitos, se ha tomado una muestra y se les ha sometido una prueba para ver cuántas tareas son capaces de realizar con éxito en una hora. Los resultados obtenidos fueron:

16 - 12 - 14 - 21 - 11 - 15 - 17 - 15 - 18 - 24 - 10 - 14 - 17 - 14

Se pide:

1. Si la empresa busca un grupo de empleados capaces de realizar una media de al menos 12 tareas por hora, ¿se puede afirmar con una confianza del 95% que el grupo lo cumple?
2. Si la empresa busca que entre los trabajadores haya una dispersión media de $\sigma < 3$ tareas, ¿se puede afirmar con una confianza del 95% que el grupo lo cumple?

Solución

- Lo cumple ya que $\mu \in (13.4026, 17.7403)$ con un 95% de confianza.
- No lo cumple ya que $\sigma \in (2.7232, 6.0516)$ con un 95% de confianza.

Ejercicio 6.38. Se cree que el consumo de tabaco va ligado al consumo de alcohol y para corroborar esta hipótesis se ha realizado un estudio en el que se han obtenido los siguientes datos

	Bebedores	No bebedores
Fumadores	487	137
No fumadores	312	365

¿Se puede afirmar que existe relación entre el consumo de tabaco y el de alcohol? Justificar la respuesta.

Ejercicio 6.39. ★ Para analizar la eficacia de un determinado fármaco para aumentar las horas de sueño se realiza un estudio con 8 pacientes a los que se les controla durante todo un mes, en el que no toman el medicamento y se cuantifica la media de las horas de sueño para cada uno de ellos. El mismo proceso se lleva a cabo durante todo un mes en el que sí que toman el fármaco. Los resultados (en horas de sueño) aparecen en la siguiente tabla:

Paciente	Sin el fármaco	Con el fármaco
1	8.3	9.0
2	6.9	8.1
3	7.2	9.2
4	9.1	9.5

Paciente	Sin el fármaco	Con el fármaco
5	8.2	8.5
6	7.5	9.0
7	6.1	8.0
8	7.4	7.9

¿Se puede concluir con un nivel de confianza del 99% que el fármaco ha sido eficaz para aumentar las horas de sueño?

Solución

Para ver si el fármaco ha sido eficaz necesitamos calcular el intervalo de confianza para la media de la diferencia entre horas de sueño sin y con el fármaco. Al tratarse de datos pareados, construimos una nueva variable X que recoja esta diferencia.

Paciente	Sin el fármaco	Con el fármaco	$X = \text{Diferencia}$
1	8.3	9.0	-0.7
2	6.9	8.1	-1.2
3	7.2	9.2	-2.0
4	9.1	9.5	-0.4
5	8.2	8.5	-0.3
6	7.5	9.0	-1.5
7	6.1	8.0	-1.9
8	7.4	7.9	-0.5

Puesto que tenemos una muestra pequeña (8 pacientes), debemos utilizar el intervalo de confianza de la t de Student, cuya fórmula es

$$\bar{x} \pm t(n-1)_{\alpha/2} \frac{\hat{s}}{\sqrt{n}}.$$

Calculamos primero los estadísticos necesarios para construir el intervalo

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{-0.7 - \dots - 0.5}{8} = \frac{-8.5}{8} = -1.0625, \\ s^2 &= \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{(-0.7)^2 + \dots + (-0.5)^2}{8} - (-1.0625)^2 = \frac{12.29}{8} - 1.1289 = 0.4073, \\ \hat{s}^2 &= \frac{n}{n-1} \cdot s^2 = \frac{8}{7} \cdot 0.4073 = 0.4655, \\ \hat{s} &= \sqrt{0.4655} = 0.6823.\end{aligned}$$

Como nos piden una confianza del 99%, $\alpha = 0.01$ y el valor de $t(n-1)_{\alpha/2} = t(7)_{0.005}$ lo buscamos en la tabla de la función de distribución de la t de Student con 7 grados de libertad. Dicho valor es $t(7)_{0.005} = 3.499$.

Sustituyendo estos valores en la fórmula del intervalo de confianza, tenemos

$$-1.0625 \pm 3.499 \frac{0.6823}{\sqrt{8}} = -1.0625 \pm 0.8442,$$

con lo que el intervalo de confianza resultante es $(-1.9067, -0.2183)$.

Según este intervalo, al 99% de confianza existen pruebas significativas de que la media de la diferencia de horas es negativa y por tanto el número de horas de sueño después del fármaco aumenta por término medio. Podemos concluir, con un 99% de confianza, que el fármaco es eficaz.

Ejercicio 6.40. ★ Se quiere probar si la cirrosis hepática hace variar el índice de colinesterasa en suero. Se eligen 2 muestras aleatorias e independientes, una primera de 60 individuos normales, con media 1.6 y desviación típica 0.3, y la segunda de 50 individuos cirróticos, con media 1.1 y desviación típica 0.4. ¿Podemos concluir que existen diferencias significativas, con un 99% de confianza, entre las medias de la colinesterasa en individuos normales e individuos cirróticos?

Ejercicio 6.41. ★ Un grupo de investigadores obtuvo datos acerca de las concentraciones de amilasa en el suero de muestras de individuos sanos y de individuos hospitalizados, con el objetivo de determinar si la concentración media es, o no, diferente en ambas poblaciones. Las concentraciones, en unidades/ml, en 10 individuos sanos fueron:

100	103	96	93	91	104	93	99	88	91
-----	-----	----	----	----	-----	----	----	----	----

Y en 12 individuos enfermos fueron:

118	115	101	104	116	114	112	113	117	123	119	121
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Suponiendo que la concentración de amilasa en suero sigue una distribución normal, tanto en individuos sanos como hospitalizados, y que las varianzas son desconocidas pero iguales, se pide:

1. Calcular el intervalo de confianza para la diferencia de medias con un nivel de confianza del 95%.
2. ¿A qué conclusión deben llegar los investigadores sobre la igualdad o no de la concentración de amilasa? Justificar la respuesta.

Ejercicio 6.42. ★ En un estudio se pretende contrastar el efecto de una vacuna contra la alergia. Para ello, se dispone de un grupo experimental formado por 200 individuos y otro control formado por 300 individuos. Teniendo en cuenta que en el grupo experimental sufrieron alergia 20 personas, con una duración media de 20 días y una desviación típica de 6, y en el control la sufrieron 42, con una duración media de 25 días y una desviación típica de 7. Se pide:

1. Calcular el intervalo de confianza al 90% para la diferencia de proporciones de los afectados por la alergia entre los que han sido vacunados y los que no. ¿Se puede concluir que la vacuna ha disminuido significativamente la proporción de afectados por la alergia? Justificar adecuadamente la respuesta.
2. Suponiendo varianzas poblacionales iguales y que la duración de la alergia sigue una distribución normal, tanto en el grupo experimental como en el grupo control, calcular el intervalo de confianza del 95% para la diferencia entre la duración media de la alergia en el grupo experimental y en el grupo control. ¿Se puede concluir que la vacuna ha disminuido significativamente la duración de la alergia? Justificar adecuadamente la respuesta.

7 Contrastes de hipótesis

Ejercicio 7.1. Se sabe que una vacuna que se está utilizando al cabo de dos años sólo protege al 60% de las personas a las que se administró. Se desarrolla una nueva vacuna, y se quiere saber si al cabo de dos años protege a más personas que la primera. Para ello se seleccionan 10 personas al azar y se les inyecta la nueva vacuna. Establecemos que si más de 8 de los vacunados conservan la protección al cabo de dos años, entonces consideraremos la nueva vacuna mejor que la antigua. Se pide:

1. Calcular la probabilidad de cometer un error de tipo I.
2. Si la nueva vacuna protegiera a un 80% de las personas vacunadas al cabo de 2 años, ¿Cuál será la probabilidad de cometer un error de tipo II?

Repetir los cálculos si se toma una muestra de 100 personas y se establece que la vacuna es mejor si más de 75 de los vacunados conservan la protección al cabo de 2 años.

Nota: Aproximar la distribución binomial mediante una distribución normal.

Solución

Contraste: $H_0 : p = 0.6$, $H_1 : p > 0.6$. Muestra de tamaño 10:

- $P(\text{Rechazar } H_0/H_0) = 0.0464$.
- $P(\text{Aceptar } H_0/H_1) = 0.6242$.

Muestra de tamaño 100:

- $P(\text{Rechazar } H_0/H_0) = 0.0011$.
- $P(\text{Aceptar } H_0/H_1) = 0.1056$.

Ejercicio 7.2. Un fisioterapeuta afirma que con un nuevo procedimiento de rehabilitación que él aplica, determinada lesión tiene un tiempo de recuperación medio no mayor de 15 días. Se seleccionan al azar 36 personas que sufren dicho tipo de lesión para verificar su afirmación, y se obtiene un tiempo medio de recuperación de 13 días y una cuasivarianza de 9. ¿Contradice lo observado en la muestra la afirmación del fisioterapeuta para un $\alpha = 0.05$?

 Solución

Contraste para la media: $H_0 : \mu = 15$, $H_1 : \mu < 15$.

Región de aceptación para $\alpha = 0.05$: $-1.6444 < z$.

Estadístico del contraste: $z = -4$. Como cae fuera de la región de aceptación se rechaza la hipótesis nula y se confirma la afirmación del fisioterapeuta.

Ejercicio 7.3. Se decide retirar una cierta vacuna si produce más de un 10% de reacciones alérgicas. Se consideran 100 pacientes sometidos a la vacuna y se observan 15 reacciones alérgicas. ¿Debe retirarse la vacuna? (Utilizar un $\alpha = 0.01$).

 Solución

Contraste para la proporción: $H_0 : p = 0.1$, $H_1 : p > 0.1$.

Región de aceptación para $\alpha = 0.01$: $z > 2.3263$.

Estadístico del contraste: $z = 1.6667$. Como cae dentro de la región de aceptación, se acepta la hipótesis nula y se concluye que no hay pruebas suficientes para retirar la vacuna.

Ejercicio 7.4. Se utiliza un grupo de 150 pacientes para comprobar la teoría de que la vitamina C tiene alguna influencia en el tratamiento del cáncer. Los 150 pacientes fueron divididos en dos grupos de 75. Un grupo recibió 10 gramos de vitamina C y el otro un placebo cada día, además de la medicación habitual. De los que recibieron la vitamina C, 47 presentaban alguna mejoría al cabo de cuatro semanas, mientras que de los que recibieron el placebo, 43 experimentaron mejoría. Contrastar esta hipótesis.

 Solución

Contraste para la comparación de proporciones: $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$.

Región de aceptación para $\alpha = 0.05$: $-1.96 < z < 1.96$.

Estadístico del contraste $z = 0.6677$. Como cae dentro de la región de aceptación, se acepta la hipótesis nula y no se pudo concluir que la vitamina C tenga influencia en el tratamiento del cáncer.

Ejercicio 7.5. Se realizó en dos hospitales una encuesta entre los pacientes sobre la satisfacción con la atención recibida, calificándola de 0 a 100. En el hospital A rellenaron la encuesta 12 pacientes, obteniéndose una media de 85 y una cuasivarianza de 16, mientras que en el hospital B rellenaron la encuesta 10 pacientes, obteniéndose una media de 81 y una cuasivarianza de 25. ¿Puede concluirse que el nivel de satisfacción en el hospital A es mayor que en el B?

Nota: Hacer previamente un contraste de igualdad de varianzas.

Solución

Contraste de comparación de varianzas: $H_0 : \sigma_A = \sigma_B$, $H_1 : \sigma_A \neq \sigma_B$.

Región de aceptación para $\alpha = 0.05$: $0.2787 < f < 3.9121$.

Estadístico del contraste: $f = 0.64$. Como cae dentro de la región de aceptación, se acepta la hipótesis de que las varianzas son iguales.

Contraste de comparación de medias: $H_0 : \mu_A = \mu_B$, $H_1 : \mu_A > \mu_B$.

Región de aceptación para $\alpha = 0.05$: $t > 1.7247$ Estadístico del contraste: $t = 2.0863$. Como cae fuera de la región de aceptación, se rechaza la hipótesis nula y se concluye que hay pruebas significativas de que el nivel de satisfacción en el hospital A es mayor que en el B.

Ejercicio 7.6. ★ Se compararon los niveles de ácido ascórbico en plasma de mujeres embarazadas fumadoras y no fumadoras, obteniéndose los siguientes resultados expresados en gramos de ácido ascórbico por mililitro de plasma:

- Mujeres fumadoras: 0.97–0.72–1.00–0.81–0.62–1.32–1.24.
- Mujeres no fumadoras: 0.48–0.71–0.98–0.68–0.5.

Realizar un contraste de hipótesis para comprobar si el nivel de ácido ascórbico en la sangre de mujeres fumadoras es mayor que el de mujeres no fumadoras.

Solución

Contraste de comparación de varianzas: $H_0 : \sigma_1 = \sigma_2$, $H_1 : \sigma_1 \neq \sigma_2$.

Región de aceptación para $\alpha = 0.05$: $0.1606 < f < 9.1973$.

Estadístico del contraste: $f = 1.6592$. Como cae dentro de la región de aceptación, se acepta la hipótesis de que las varianzas son iguales.

Contraste de comparación de medias: $H_0 : \mu_1 = \mu_2$, $H_2 : \mu_1 > \mu_2$.

Región de aceptación para $\alpha = 0.05$: $t < 1.8125$.

Estadístico del contraste: $t = 2.0372$. Como cae fuera de la región de aceptación, se rechaza la hipótesis nula y se concluye que las mujeres fumadoras tienen mayor nivel de ácido ascórbico.

Ejercicio 7.7. Verificar la hipótesis de que el contenido medio de unos recipientes de ácido sulfúrico es de 10 litros, si los contenidos de una muestra aleatoria de 10 recipientes son 10.2, 9.7, 10.1, 10.3, 10.1, 10.1, 9.8, 9.9, 10.4, 10.3 y 9.8 litros. Utilizar un nivel de significación de 0.01 y suponer que la distribución de los contenidos es normal.

Ejercicio 7.8. Un fabricante de equipos de medida afirma que sus equipos pueden realizar al menos 12 mediciones más que los de la competencia sin necesidad de un nuevo ajuste. Para probar esta afirmación se realizan mediciones con 50 equipos de este fabricante y 50 de la competencia. En los suyos el número de mediciones hasta necesitar

un nuevo ajuste tuvo de media 86.7 y cuasidesviación típica 6.28, mientras que en los de la competencia estos valores fueron 77.8 y 5.61 respectivamente. Verificar la afirmación del fabricante con $\alpha = 0.05$.

 Solución

Contraste de comparación de medias: $H_0 : \mu_1 < \mu_2 + 12$, $H_1 : \mu_1 \geq \mu_2 + 12$.
Intervalo de confianza para la diferencia de medias: $\mu_1 - \mu_2 \in (6.5659, 11.2341)$ con un 95% de confianza, luego hay diferencias significativas entre el número medio de mediciones, pero no se puede afirmar que sean mayores de 12 mediciones.

Ejercicio 7.9. Para determinar si un nuevo suero detiene la leucemia, se seleccionan 9 ratones con leucemia en una fase avanzada. Cinco reciben el tratamiento y cuatro no. Los tiempos de supervivencia, en años, desde el momento que comenzó el experimento son los siguientes:

- Con tratamiento: 2.1 – 5.3 – 1.4 – 4.6 – 0.9.
- Sin tratamiento: 1.9 – 0.5 – 2.8 – 3.1.

¿Puede afirmarse con un $\alpha = 0.05$ que el suero es eficaz? Suponer que ambas distribuciones son normales con varianzas iguales.

Ejercicio 7.10. Un estudio afirma que el 70% de los habitantes de la capital lee diariamente algún periódico. ¿Estaríamos de acuerdo con las conclusiones de dicho estudio si al preguntar a 15 personas elegidas aleatoriamente, 8 leen diariamente algún periódico?

 Solución

Contraste para la proporción: $H_0 : p = 0.7$, $H_1 : p \neq 0.7$.
Región de aceptación para $\alpha = 0.05$: $-1.96 < z < 1.96$.
Estadístico del contraste: $z = -1.408590$. Como cae dentro de la región de aceptación, se acepta la hipótesis nula y se estaría de acuerdo con las afirmación del estudio.

Ejercicio 7.11. Un distribuidor de tabaco asegura que el 20% de los fumadores de su ciudad prefiere los cigarrillos de marca A. Se selecciona al azar una muestra de 20 fumadores, y 6 de ellos prefieren la marca A. ¿Qué conclusión se obtiene con $\alpha = 0.05$?

Ejercicio 7.12. En un estudio sobre el consumo de alcohol entre los jóvenes durante los fines de semana, se preguntó a 100 chicos y a 125 chicas, de los que 63 chicos y 59 chicas contestaron que consumían. En vista de estos datos, ¿existe alguna diferencia significativa entre las respuestas de chicos y chicas? Utilizar $\alpha = 0.10$.

 Solución

Contraste de comparación de proporciones: $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$.

Región de aceptación para $\alpha = 0.01$: $-1.6449 < z < 1.6449$.

Estadístico del contraste: $z = 2.4026$. Como cae fuera de la región de aceptación, se rechaza la hipótesis nula y se concluye que hay diferencias significativas entre el consumo de alcohol de chicos y chicas.

Ejercicio 7.13. Un fabricante de baterías para automóvil asegura que la duración de sus baterías tiene una distribución aproximadamente normal con desviación típica no superior a 0.9 años. Si una muestra aleatoria de 10 de estas baterías tiene una cuasidesviación típica de 0.7 años, ¿qué se puede concluir sobre la afirmación del fabricante?

 Solución

Contraste para la desviación típica: $H_0 : \sigma = 0.9$, $H_1 : \sigma < 0.9$.

Región de aceptación para $\alpha = 0.05$: $3.3251 < j$.

Estadístico del contraste: $j = 4$. Como cae dentro de la región de aceptación, no se puede rechazar la hipótesis nula y se concluye que no hay pruebas significativas de que sea cierta la afirmación del fabricante.

Ejercicio 7.14. En un estudio sobre el contenido de ortofósforo de las aguas de un río, se realizaron medidas en dos estaciones distintas. Se sacaron 15 muestras de la estación 1 y 12 de la estación 2. Las muestras de la estación 1 presentaron un contenido medio de ortofósforo de 3.84 mg/l y una cuasidesviación típica de 3.07 mg/l, mientras que las de la estación 2 tuvieron media 1.49 mg/l y una cuasidesviación típica 0.8 mg/l.

Se pide:

1. Calcular el intervalo de confianza para el cociente de varianzas.
2. Realizar el contraste de hipótesis de igualdad de varianzas.

Utilizar un $\alpha = 2$.

Ejercicio 7.15. Se ha desarrollado un aditivo para gasolina que reduce la emisión de CO_2 en la combustión. Para comprobar la efectividad del aditivo, se realiza un estudio en el que se mide en una muestra de 10 coches la cantidad de CO_2 emitida (en Kg/l), tanto con gasolina con aditivo, como con gasolina sin aditivo, obteniendo los siguientes resultados:

Sin aditivo:	0.38	0.42	0.41	0.39	0.45	0.47	0.44	0.38	0.40	0.50
Con aditivo:	0.38	0.36	0.38	0.32	0.39	0.45	0.39	0.39	0.35	0.48

Se pide:

1. Realizar un contraste de hipótesis para comprobar la efectividad del aditivo.
2. ¿Qué potencia tiene el contraste para detectar una reducción en la emisión de CO_2 de 0.2 Kg/l?

Ejercicio 7.16. Un psicólogo está estudiando la concentración de una encima en la saliva como un posible indicador de la ansiedad crónica. En un experimento se tomó una muestra de 12 neuróticos por ansiedad y otra de 10 personas con bajos niveles de ansiedad. En ambas muestras se midió la concentración de la encima, obteniendo los siguientes resultados:

Con ansiedad:	2.60	2.90	2.60	2.70	3.91	3.15	3.94	2.46	2.91	3.88	3.55	3.96
Sin ansiedad:	2.37	1.10	2.55	2.64	2.20	2.12	2.47	2.90	1.66	2.72		

¿Se puede concluir a partir de estos datos que la población de neuróticos con ansiedad y la población de personas sin ansiedad son diferentes en el nivel medio de concentración de encimas? Justificar la respuesta.

Solución

Contraste de comparación de varianzas: $H_0 : \sigma_1 = \sigma_2$, $H_1 : \sigma_1 \neq \sigma_2$.

Región de aceptación para $\alpha = 0.05$: $0.2787 < f < 3.9121$.

Estadístico del contraste: $f = 1.2138$. Como cae dentro de la región de aceptación, se acepta la hipótesis nula y se concluye que las varianzas son iguales.

Contraste de comparación de medias: $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$.

Región de aceptación para $\alpha = 0.05$: $-2.0860 < t < 2.0860$.

Estadístico del contraste: $t = 3.8407$. Como cae fuera de la región de aceptación, se rechaza la hipótesis nula y se puede concluir que hay diferencia entre la concentración media de encimas para neuróticos con ansiedad y sin ansiedad.

Ejercicio 7.17. En una investigación para ver la efectividad de una nueva droga antidepresiva, se ha tomado una muestra de 15 pacientes depresivos que han completado un cuestionario para detectar el nivel depresivo, antes y después de recibir la droga. En la puntuación del cuestionario los valores menores indican una mayor depresión. Los resultados obtenidos han sido:

Antes:	18	21	16	19	14	23	16	14	21	18	17	14	16	14	20
Después:	23	20	17	20	16	22	18	18	21	16	19	20	15	15	21

Realizar un contraste para averiguar si la droga tiene un efecto positivo sobre la depresión. ¿Qué tamaño muestral sería necesario para detectar una diferencia en la puntuación como la que hay entre las medias de las muestras?

Solución

Contraste para la media de la diferencia entre antes y después: $H_0 : \mu = 0$, $H_1 : \mu < 0$.

Región de aceptación para $\alpha = 0.05$: $-1.7613 < t$.

Estadístico del contraste: $t = -2.2563$. Como cae fuera de la región de aceptación, se rechaza la hipótesis nula y se puede afirmar que la droga reduce la depresión.

El tamaño muestral para detectar una diferencia de $\delta = \bar{x}_1 - \bar{x}_2 = 17.4 - 18.7333 = -1.3333$ con una potencia del 90% es $n = 26$ individuos.

Ejercicio 7.18. Un experimento pretende contrastar la teoría de que la memoria a corto plazo se ve afectada por la similitud entre los estímulos. El experimento consiste en leer en voz alta una secuencia de letras a un sujeto, quien después de una breve pausa debe repetir la secuencia. Si la teoría es correcta, habrá más errores en la lista que contenga letras que suenan de forma similar que si contiene letras que se parecen. A cada sujeto se le presentan dos tipos de secuencias, una con letras que suenan de forma similar y otra con letras que se escriben de forma parecida. Los errores producidos en cada secuencia son:

Errores de letras que suenan parecidas:	7	5	6	11	3	8	4	10	9
Errores de letras con similar escritura:	8	2	5	9	5	4	4	7	4

¿Se puede validar la teoría?

Solución

Contraste para la media de la diferencia entre los errores de letras que suenan parecidas y los errores en letras con similar escritura: $H_0 : \mu = 0$, $H_1 : \mu > 0$.

Región de aceptación para $\alpha = 0.05$: $t > 1.8595$.

Estadístico del contraste: $t = 2.1320$. Como cae fuera de la región de aceptación, se rechaza la hipótesis nula y se concluye que la teoría es cierta.

Ejercicio 7.19. Se sabe que el tiempo de reacción ante un estímulo sigue una distribución normal de media 30 ms y desviación típica 10 ms. Se cree que la alcoholemia aumenta el tiempo de reacción de los sujetos, y para comprobar esta hipótesis se ha tomado una muestra aleatoria de 40 individuos a los que se les ha inducido una alcoholemia de 0.8 g/l y en los que se ha apreciado un tiempo medio de respuesta de 35 ms y una desviación típica de 12 ms. ¿Se puede afirmar que una alcoholemia de 0.8 gm/l influye en el tiempo medio de respuesta con un riesgo $\alpha = 0.05$? ¿Y con un riesgo $\alpha = 0.01$?

¿Cuál será la potencia del contraste para detectar una diferencia en la media del tiempo de reacción de 4 ms? ¿Cuál debería ser el tamaño muestral para aumentar la potencia hasta un 90%?

 Solución

Contraste para la media: $H_0 : \mu = 30, H_1 : \mu > 30$.

Región de aceptación para $\alpha = 0.01$: $z < 2.3263$.

Estadístico del contraste: $z = 2.6020$. Como cae fuera de la región de aceptación, se rechaza la hipótesis nula tanto para $\alpha = 0.01$ y con mayor motivo para $\alpha = 0.05$, de manera que se concluye que la alcoholemia influye en el tiempo de respuesta.

Potencia del contraste para $\delta = 4$: $1 - \beta = 1 - 0.5966 = 0.4034$.

El tamaño muestral para, $\alpha = 0.05$, $\delta = 4$ y una potencia del 90% es $n = 80$.

Ejercicio 7.20. Se cree que el nivel medio de protrombina en plasma de una población normal tiene una media de 19mg/100ml y una desviación típica de 4mg/100ml. Para contrastar estas hipótesis se tomó una muestra de 8 individuos en los que se obtuvieron los siguientes niveles de protrombina en plasma:

$$16.3 - 18.4 - 20.0 - 17.6 - 15.4 - 23.7 - 17.8 - 19.5$$

¿Se pueden aceptar ambas hipótesis con un riesgo $\alpha = 0.1$?

 Solución

Contraste para la media: $H_0 : \mu = 19, H_1 : \mu \neq 19$.

Región de aceptación para $\alpha = 0.1$: $-1.8946 < t < 1.8946$.

Estadístico del contraste: $t = -0.4552$. Como cae dentro de la región de aceptación, se mantiene la hipótesis de que la media es 19mg/100ml.

Contraste para la varianza: $H_0 : \sigma = 4, H_1 : \sigma \neq 4$.

Región de aceptación para $\alpha = 0.1$: $2.1673 < j < 14.0671$.

Estadístico del contraste: $j = 2.8743$. Como cae dentro de la región de aceptación, también se mantiene la hipótesis de que la desviación típica es de 4mg/100ml.

Ejercicio 7.21. ★ Para ver si la ley antitabaco está influyendo en el número de cigarros que se fuman mientras se está en los bares se seleccionó una muestra en la que se midió el número de cigarros fumados por hora mientras se estaba en un bar antes de la entrada en vigor de la ley y otra muestra distinta en la que también se midió el número de cigarros fumados por hora después de la entrada en vigor de la ley (se entiende que con la ley en vigor los cigarros se fuman en el exterior de los bares). Los resultados aparecen en la siguientes tablas:

Tabla 7.4: Antes

Cigarros	Personas
0-1	12
1-2	21
2-3	20
3-4	8

Tabla 7.5: Después

Cigarros	Personas
0-1	22
1-2	18
2-3	8
3-4	4

Se pide:

1. Calcular el intervalo de confianza del 99% para el número medio de cigarros fumados por hora en los bares antes de la entrada en vigor de la ley. ¿Cuántos individuos serían necesarios para poder estimar dicha media con un margen de error no mayor de ± 0.1 cigarros por hora?
2. Contrastar si la nueva ley ha reducido significativamente el consumo medio de tabaco en los bares. ¿Cuánto vale el p -valor del contraste?

Solución

Sean X e Y las variables que miden el número medio de cigarros fumados por hora antes y después de la entrada en vigor de la ley respectivamente.

1. Puesto que se trata de una muestra grande de $n_x = 61$, la fórmula del intervalo de confianza para la media es

$$\bar{x} \pm z_{\alpha/2} \frac{\hat{s}}{\sqrt{n}}.$$

A partir de la tabla de frecuencias se calculan los estadísticos necesarios:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i n_i}{n_x} = \frac{0.5 \cdot 12 + \dots + 3.5 \cdot 8}{61} = \frac{115.5}{61} = 1.8934, \\ s_x^2 &= \frac{\sum x_i^2 n_i}{n_x} - \bar{x}^2 = \frac{0.5^2 \cdot 12 + \dots + 3.5^2 \cdot 8}{61} - 1.8934^2 = \frac{273.25}{61} - 3.585 = 0.8944, \\ \hat{s}_x^2 &= \frac{n_x}{n_x - 1} s_x^2 = \frac{61}{60} 0.8944 = 0.9093, \\ \hat{s}_x &= \sqrt{0.9093} = 0.9536.\end{aligned}$$

Como se pide un nivel de confianza del 99% se tiene que $\alpha = 0.01$ y $\alpha/2 = 0.005$, de modo que buscando en la tabla de la función de distribución de la normal estándar se tiene que $z_{\alpha/2} = 2.5758$. Así pues, sustituyendo en la fórmula del intervalo se obtiene

$$\bar{x} \pm z_{\alpha/2} \frac{\hat{s}}{\sqrt{n}} = 1.8934 \pm 2.5758 \frac{0.9536}{\sqrt{61}} = 1.8934 \pm 0.3145 = (1.5789, 2.2079).$$

Por otro lado, el número de individuos necesario para estimar la media con un margen de error no mayor de ± 0.1 cigarros por hora y una confianza del 99% es

$$n = \frac{4z_{\alpha/2}^2 \hat{s}^2}{A^2} = \frac{4 \cdot 2.5758^2 \cdot 0.9093}{(2 \cdot 0.1)^2} = 603.2974,$$

es decir, se necesitarían como mínimo 603 individuos.

2. Para contrastar si la nueva ley ha reducido significativamente el consumo medio de tabaco en los bares hay que realizar un contraste unilateral de comparación de medias

$$H_0 : \mu_x = \mu_y$$

$$H_1 : \mu_x > \mu_y$$

Como los tamaños muestrales son grandes, $n_x = 61$ y $n_y = 52$, y no se conocen las varianzas poblacionales, el estadístico de contraste es

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{s}_x^2}{n_x} + \frac{\hat{s}_y^2}{n_y}}},$$

que sigue una distribución normal estándar. Para calcularlo se necesitan, además de los estadísticos de X calculados en el apartado anterior, los estadísticos de Y , que son:

$$\begin{aligned}\bar{y} &= \frac{\sum y_j n_j}{n_y} = \frac{0.5 \cdot 22 + \dots + 3.5 \cdot 4}{52} = \frac{72}{52} = 1.3846, \\ s_y^2 &= \frac{\sum y_j^2 n_j}{n_y} - \bar{y}^2 = \frac{0.5^2 \cdot 22 + \dots + 3.5^2 \cdot 4}{52} - 1.3846^2 = \frac{145}{52} - 1.9171 = 0.8713, \\ \hat{s}_y^2 &= \frac{n_y}{n_y - 1} s_y^2 = \frac{52}{51} 0.8713 = 0.8884,\end{aligned}$$

de manera que sustituyendo en la fórmula del estadístico de contraste se obtiene

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{s}_x^2}{n_x} + \frac{\hat{s}_y^2}{n_y}}} = \frac{1.8934 - 1.3846}{\sqrt{\frac{0.9093}{61} + \frac{0.8884}{52}}} = 2.8447.$$

Como el estadístico sigue una distribución normal estándar, la región de aceptación para un nivel de significación $\alpha = 0.05$ es $Z \leq z_\alpha = 1.6449$, y como el estadístico cae fuera de esta región se rechaza la hipótesis nula y se concluye que hay pruebas significativas de que la nueva ley ha reducido el consumo medio de tabaco en los bares.

Finalmente, el p -valor del contraste es $P(Z > 2.8447) = 1 - P(Z \leq 2.8447) = 1 - F(2.8447) = 0.0022$.

Ejercicio 7.22. ★ En un estudio sobre el reparto de género de las tareas domésticas se ha medido el número medio de horas diarias que se destinan a las tareas domésticas en un grupo de personas, obteniendo los siguientes resultados:

Mujeres: 3.2 3.1 2.7 4.4 3.7 3.9 2.4 3.6
Hombres: 3.3 2.1 1.7 2.4 1.6 1.8 2.7

Contrastar si las mujeres dedican más tiempo que los hombres a las tareas domésticas. ¿Entre qué valores estará la diferencia del tiempo medio destinado a tareas domésticas entre mujeres y hombres para un 95% de confianza?

Solución

Estadísticos muestrales:

$$\bar{x}_M = 3.375, s_M^2 = 0.3744, \hat{s}_M^2 = 0.4279, \hat{s}_M = 0.654,$$

$$\bar{x}_H = 2.2286, s_H^2 = 0.3249, \hat{s}_H^2 = 0.379, \hat{s}_H = 0.616.$$

Contraste de comparación de varianzas: $H_0 : \sigma_M^2 = \sigma_H^2, H_1 : \sigma_M^2 \neq \sigma_H^2$,

Estadístico del contraste: $F = 1.129$,

Región de aceptación para $\alpha = 0.05$: $(0.1954, 5.6955)$.
 Por tanto, se acepta la hipótesis nula y se supone que las varianzas son iguales.
 Contraste de comparación de medias: $H_0 : \mu_M = \mu_H$, $H_1 : \mu_M > \mu_H$,
 Estadístico del contraste: $T = 3.479$.
 Región de aceptación para $\alpha = 0.05$: $(1.7709, \infty)$.
 Por tanto, se rechaza la hipótesis nula y se concluye que las mujeres dedican más tiempo que los hombres a tareas domésticas.
 Intervalo de confianza del 0.95% para $\mu_M - \mu_H$: $(0.4345, 1.8583)$.

Ejercicio 7.23. ★ Un estudio sobre la adicción a los videojuegos ha medido el número medio de horas diarias que un grupo de jóvenes pasa jugando a los videojuegos, obteniendo los siguientes resultados:

2.8 – 4.1 – 1.8 – 2.2 – 0.5 – 3.2 – 1.6 – 1.1 – 2.5 – 0.7 – 4.5 – 3.3 – 1.4 – 1.4 – 1.2 – 1.7

Se pide:

1. Calcular el intervalo de confianza para el número medio de horas diarias de juego.
2. Si un número de horas de juego superior a 2 horas puede suponer un alto riesgo de adicción, estimar mediante un intervalo de confianza del 90% el porcentaje de jóvenes que estarán en riesgo de adicción. ¿Qué tamaño muestral sería necesario para poder estimar dicho porcentaje con un margen de error no mayor de $\pm 5\%$?

Solución

1. Estadísticos muestrales: $\bar{x} = 2.125$, $\hat{s}^2 = 1.3913$, $\hat{s} = 1.18$. Intervalo de confianza del 0.95% para μ : $(1.4962, 2.7538)$.
2. Proporción muestral: $\hat{p} = 0.4375$. Intervalo de confianza del 0.9% para p : $(0.2335, 0.6415)$.
 Tamaño muestral necesario para estimar p con una confianza del 0.9 y un error $A = \pm 0.05 = 0.1$: $n = 266.342$.

8 Análisis de la varianza de un factor

Ejercicio 8.1. Se realiza un estudio para comparar la eficacia de tres programas terapéuticos para el tratamiento del acné. Se emplean tres métodos:

- A. Lavado, dos veces al día, con cepillo de polietileno y un jabón abrasivo, junto con el uso diario de 250 mg de tetraciclina.
- B. Aplicación de crema de tretinoína, evitar el sol, lavado dos veces al día con un jabón emulsionante y agua, y utilización dos veces al día de 250 mg de tetraciclina.
- C. Evitar el agua, lavado dos veces al día con un limpiador sin lípidos y uso de crema de tretinoína y de peróxido benzoílico.

En el estudio participan 35 pacientes. Se separó aleatoriamente a estos pacientes en tres subgrupos de tamaños 10, 12 y 13, a los que se asignó respectivamente los tratamientos I, II, y III. Después de 16 semanas se anotó para cada paciente el porcentaje de mejoría en el número de lesiones.

A	B	C
48.6	68.0	67.5
50.8	71.9	61.4
49.4	67.0	62.5
47.1	71.5	67.4
50.1	70.1	64.2
52.5	69.9	65.4
49.8	64.5	62.5
49.0	68.9	63.2
50.6	68.0	63.9
46.7	67.8	61.2
	68.3	64.8
	68.9	60.5
		62.3

Se pide:

1. Dibujar el diagrama de dispersión. ¿Qué conclusiones sacas de la nube de puntos?
2. Obtener la tabla de ANOVA correspondiente al problema. ¿Se puede concluir que los tres tratamientos tienen el mismo efecto medio con un nivel de significación de 0.05?

Ejercicio 8.2. Se sabe que se ha arrojado material tóxico a un río que desemboca en una gran área de pesca comercial en agua salada. Se pretende estudiar la forma en que el agua transporta el material tóxico midiendo la cantidad de material (en partes por millón) hallada en las ostras recogidas en tres lugares diferentes, desde la salida del estuario hasta la bahía donde se realiza la mayor parte de la pesca comercial. Los resultados obtenidos son los siguientes:

Estuario	Lejos de la bahía	Cerca de la bahía
15	19	22
26	15	26
20	10	24
20	26	26
29	11	15
28	20	17
21	13	24
26	15	
	18	

Contrastar si hay o no diferencias en la contaminación media de las ostras dependiendo del lugar en el que han sido recogidas, con un nivel de significación de 0.05.

Ejercicio 8.3. Se midió la frecuencia cardíaca (latidos por minuto) en cuatro grupos de adultos; controles normales (A), pacientes con angina (B), individuos con arritmias cardíacas (C) y pacientes recuperados del infarto de miocardio (D). Los resultados son los siguientes:

A	B	C	D
83	81	75	61
61	65	68	75
80	77	80	78
63	87	80	80
67	95	74	68
89	89	78	65
71	103	69	68
73	89	72	69
70	78	76	70
66	83	75	79
57	91	69	61

¿Proporcionan estos datos la suficiente evidencia para indicar una diferencia en la frecuencia cardíaca media entre esos cuatro tipos de pacientes?. Considerar $\alpha = 0.05$.

Ejercicio 8.4. Se sospecha que hay diferencias en la preparación del examen de selectividad entre los diferentes centros de bachillerato de una ciudad. Con el fin de comprobarlo,

de cada uno de los 5 centros, se eligieron 8 alumnos al azar, con la condición de que hubieran cursado las mismas asignaturas, y se anotaron las calificaciones que obtuvieron examen de selectividad. Los resultados fueron:

A	B	C	D	E
5.5	6.1	4.9	3.2	6.7
5.2	7.2	5.5	3.3	5.8
5.9	5.5	6.1	5.5	5.4
7.1	6.7	6.1	5.7	5.5
6.2	7.6	6.2	6.0	4.9
5.9	5.9	6.4	6.1	6.2
5.3	8.1	6.9	4.7	6.1
6.2	8.3	4.5	5.1	7.0

Estudiar si se confirma o no la sospecha de partida.

Ejercicio 8.5. Se midió la frecuencia respiratoria (inspiraciones por minuto) en ocho animales de laboratorio y con tres niveles diferentes de exposición al monóxido de carbono. Los resultados son los siguientes:

Bajo	Moderado	Alto
36	43	45
33	38	39
35	41	33
39	34	39
41	28	33
41	44	26
44	30	39
45	31	29

Con base en estos datos, ¿es posible concluir que los tres niveles de exposición, en promedio, tienen un efecto diferente sobre la frecuencia respiratoria? Tomar $\alpha = 0,05$.

Ejercicio 8.6. Se ensayan dos nuevas técnicas de aprendizaje frente a la tradicional para realizar una tarea. Para ver su efectividad se crean tres grupos de niños y se instruye a cada grupo con una de las técnicas. La puntuación obtenida en la realización de la tarea por parte de los niños de los tres grupos es:

Método 1	Método 2	Tradicional
85	76	65
80	62	67
83	71	69
77	68	55
		63
		58
		49
		44
		50
		43
		47
		57
		46

Método 1	Método 2	Tradicional
79	81	66

¿Existen diferencias significativas entre los tres métodos de aprendizaje? ¿Y entre los dos métodos nuevos? Justificar la respuesta.

Ejercicio 8.7. Para ver cómo afecta la estación del año sobre el grado de depresión, se tomó un grupo de 20 pacientes depresivos y se les distribuyó aleatoriamente cuatro grupos. A cada grupo se le sometió a distintos estímulos en una estación del año, midiendo el grado de depresión. Las puntuaciones obtenidas fueron (a mayor puntuación mayor depresión):

Primavera	7	6	5	5	6
Verano	4	3	4	4	5
Otoño	8	7	6	7	8
Invierno	8	6	7	5	6

De acuerdo a estos datos, ¿se puede afirmar que la estación del año influye en la depresión? Justificar la respuesta.

9 Contrastes de la chi-cuadrado

Ejercicio 9.1. En estudio se ha medido la objetividad de 300 hombres y 400 mujeres ejutivos en la evaluación de personal recomendado. La siguiente tabla refleja los resultados observados:

	Hombres	Mujeres
Objetivos	176	245
No objetivos	124	155

¿Existen diferencias en la objetividad entre hombres y mujeres? Justificar la respuesta.