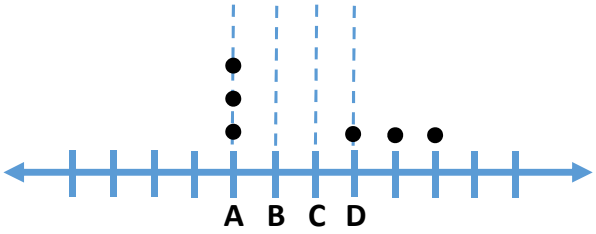
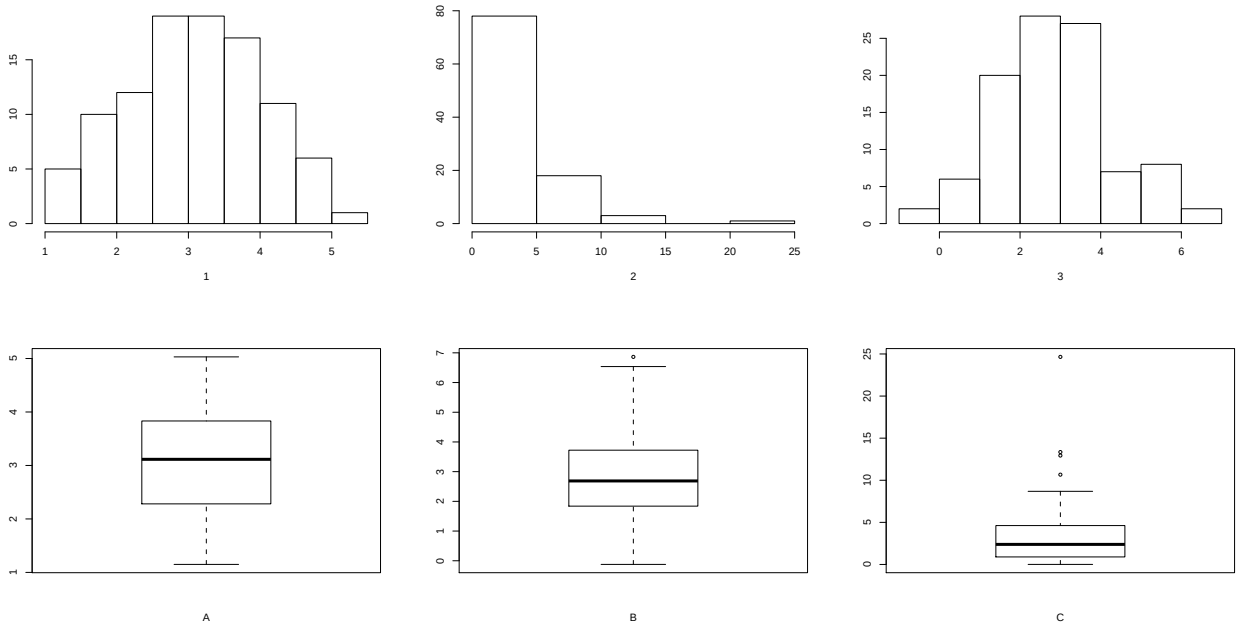
Problemas y cuestiones de estadística descriptiva

1. Si la muestra está formada por los puntos que se ven en la figura, ¿en cuál de las posiciones A, B, C, D está la media?



2. La siguiente figura muestra histogramas y diagramas de cajas para tres conjuntos de datos diferentes:



(a) Determina razonadamente el diagrama de cajas al que corresponde cada histograma.
(b) Para cada conjunto de datos, determina si la media y la mediana son aproximadamente iguales o no. En este último caso especifica cuál de las dos medidas es mayor.

3. Consideramos la siguiente lista de medidas utilizadas en estadística:

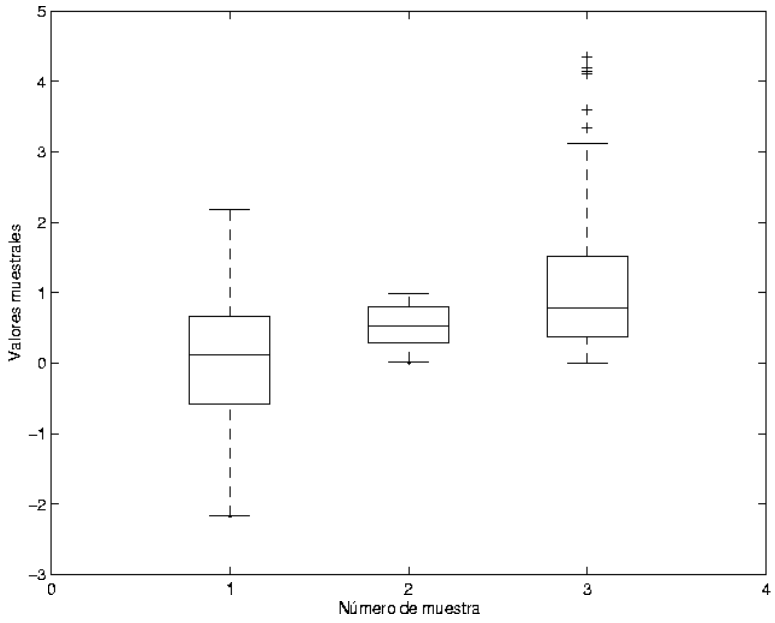
coeficiente de correlación; varianza; media; cuartil 1; coeficiente de variación; covarianza; mediana; rango intercuartílico; rango; desviación típica.

(a) Clasifica las cantidades de la lista anterior en alguno de los tres grupos siguientes:
   1. Medidas de posición de la distribución de un conjunto de datos.
   2. Medidas de dispersión de la distribución de un conjunto de datos.
   3. Cantidades no utilizadas para medir ni la posición ni la dispersión.

(b) De las medidas de la lista, enumera todas aquellas cuyo valor no dependa de las unidades en las que se expresen los datos (es decir, las medidas adimensionales).

4. Se ha registrado el número de clientes diarios de un restaurante de comida rápida durante 30 días, tanto en fin de semana como de lunes a viernes. Para los fines de semana (8 días) se obtuvo un número de clientes medio de 389,56, mientras que para los días entre semana (22 días) se obtuvo una media de 402,19. Calcula el número medio de clientes globales para los 30 días.

5. La siguiente figura muestra los diagramas de cajas correspondientes a tres muestras:



Determina razonadamente si las siguientes afirmaciones son verdaderas o falsas:

(a) Las tres muestras corresponden a distribuciones bastante simétricas.
(b) Una de las muestras parece proceder de una distribución normal de esperanza cero y varianza uno.
(c) El primer cuartil de la muestra 2 es menor que la mediana de las otras dos muestras.

6. Determina razonadamente si las siguientes afirmaciones son verdaderas o falsas:

(a) Como son dos medidas de posición, la media y la mediana del mismo conjunto de datos siempre toman valores parecidos.
(b) Si se transforman de gramos a kilos las unidades de medida de un conjunto de datos sobre pesos, la correspondiente varianza no cambia.
(c) Si se añade un punto de regalo a todas las notas de los alumnos de una clase, la desviación típica de la notas de la clase no cambia.

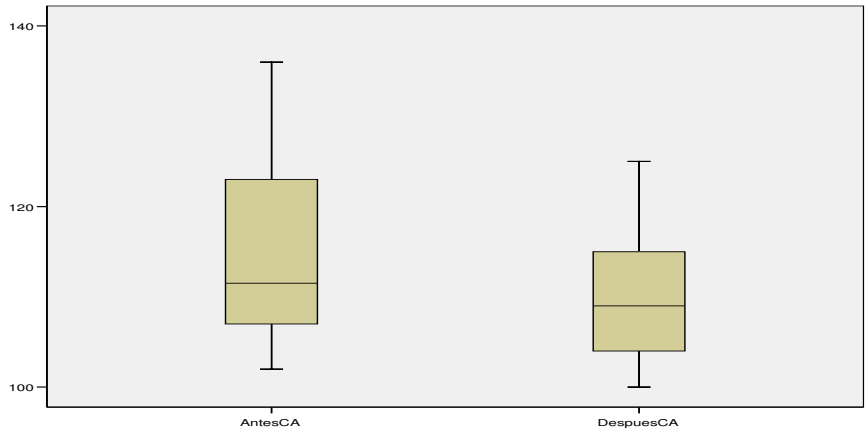
7. El agua de los ríos contiene pequeñas concentraciones de mercurio que se pueden ir acumulando en los tejidos de los peces. Se ha realizado un estudio en los ríos Wacamaw y Lumber en Carolina del Norte (EE.UU.), analizando la cantidad de mercurio que contenían 171 ejemplares capturados de una cierta especie de peces. En cada ejemplar capturado se han medido las siguientes variables: `LONG` ≡ Longitud (en cm) del pez; `PESO` ≡ Peso (en g) del pez; `CONC` ≡ Concentración (en ppm) de mercurio. Los resultados de un primer análisis descriptivo son los siguientes:

Estadísticos

		LONG	PESO	CONC
N	Válidos	171	171	171
	Perdidos	0	0	0
	Media	39,9708	1147,9123	1,1918
	Error típ. de la media	,65132	66,95359	,05825
	Mediana	39,0000	873,0000	,9300
	Desv. típ.	8,51715	875,53176	,76166
	Varianza	72,542	766555,869	,580
	Rango	39,80	4308,00	3,49
	Mínimo	25,20	203,00	,11
	Máximo	65,00	4511,00	3,60
Percentiles	25	33,3000	491,0000	,5900
	50	39,0000	873,0000	,9300
	75	46,2000	1455,0000	1,6000

(a) Calcula el coeficiente de variación de las tres variables. ¿Qué se deduce sobre la dispersión de los valores que toman?
(b) Comparando los valores de la media y la mediana, ¿cuál de las tres distribuciones parece ser más simétrica?
(c) Verdadero o falso: Al menos para 100 peces, la concentración de mercurio es superior a 0.93 ppm.
(d) Verdadero o falso: La longitud de aproximadamente 42 peces es mayor que 25.20 cm y menor que 33.3 cm.
(e) ¿Cuál es el rango intercuartílico de la variable que mide el peso de los peces?

8. Para determinar el efecto del consumo de calcio sobre la tensión arterial, se midió la tensión arterial sistólica de 10 personas antes y después de recibir un suplemento de calcio en su dieta durante 12 semanas. La figura siguiente recoge los diagramas de cajas correspondientes a los datos registrados antes y después del tratamiento:



Determina razonadamente si las siguientes afirmaciones son verdaderas o falsas:

(a) Al menos 6 de las tensiones registradas después del tratamiento son inferiores a 120.
(b) El rango intercuartílico de las tensiones registradas antes del tratamiento es inferior al rango intercuartílico de las tensiones registradas después.
(c) Al menos el 75 % de las tensiones registradas antes del tratamiento es inferior a la máxima tensión registrada después.

9. En un examen de matemáticas realizado por un grupo de alumnos se ha obtenido una nota media de 110 puntos con desviación típica de 10 y en otro examen de estadística realizado por el mismo grupo se ha obtenido una nota media de 25 puntos con desviación típica de 0,5. El coeficiente de correlación entre los resultados de ambos exámenes es $r = 0{,}85$. ¿Qué puntuación en el examen de estadística predecirías para un alumno que ha obtenido 125 puntos en el examen de matemáticas?

10. Considera la recta de mínimos cuadrados $y = \beta_0 + \beta_1 x$. Si alguien nos dice cuál es el valor de β_1, ¿cuáles de las siguientes operaciones conducen a calcular el valor correcto de β_0?

(a) Se sustituye $x = y = 0$ en la ecuación de la recta y se despeja β_0.
(b) Se sustituye x por $\bar{x}$ e y por $\bar{y}$ en la ecuación de la recta y se despeja β_0.
(c) Si (x_0, y_0) es uno de los puntos de la muestra que se ha usado para calcular la recta, se sustituye x por x_0 e y por y_0 en la ecuación de la recta y se despeja β_0.
(d) Ninguna de las tres opciones anteriores lleva a calcular correctamente β_0.

11. Una manera de controlar la correcta producción de insulina consiste en medir la concentración de péptido C en la sangre, puesto que éste se libera cuando se produce insulina. En un estudio sobre diabetes se ajustó un modelo de regresión lineal a una muestra de 43 individuos con el fin de estudiar el logaritmo de la concentración de péptido C (pmol/ml) en función de la edad. Los datos se analizaron con SPSS con los siguientes resultados:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,464[a]	,215	,196	,6461

a. Variables predictoras: (Constante), edad

Coeficientes[a]

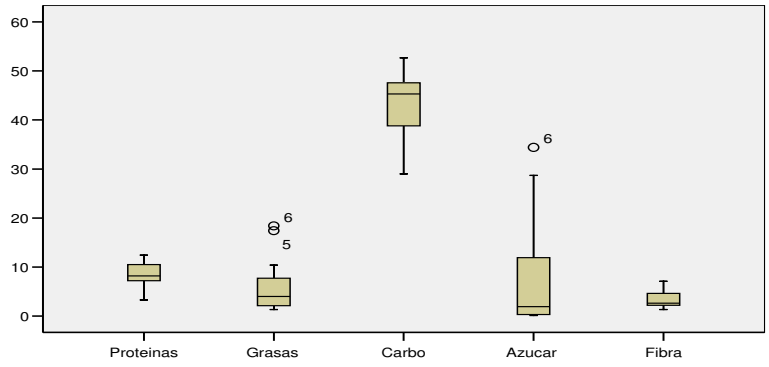
Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	3,996	,245		16,338	,000
	edad	,083	,025	,464	3,352	,002

a. Variable dependiente: logpeptidoC

Escribe la ecuación de la recta de mínimos cuadrados y predice el valor de la variable respuesta para un individuo de 2 años de edad. ¿Qué puedes decir sobre la fiabilidad de esta predicción?

12. En un estudio sobre las propiedades nutricionales de los alimentos de una panadería se ha analizado la información contenida en las etiquetas de 25 productos. En primer lugar se ha llevado a cabo un análisis descriptivo básico de las siguientes variables relativas a 100 g de cada producto: `Proteinas` (proteínas en g), `Grasas` (grasas en g), `Carbo` (carbohidratos en g), `Azucar` (azúcares en g) y `Fibra` (fibra en g). Los resultados fueron los siguientes:

		Proteinas	Grasas	Carbo	Azucar	Fibra
N	Válidos	25	25	25	25	25
	Perdidos	0	0	0	0	0
Media		8,3720	5,6480	43,4800	8,0360	3,3720
Mediana		8,2000	4,0000	45,3000	1,9000	2,6000
Desv. típ.		2,68631	4,65789	5,87927	10,93114	1,66871
Mínimo		3,30	1,30	29,00	,10	1,30
Máximo		12,40	18,40	52,70	34,40	7,10
Percentiles	25	6,4500	2,1000	38,5500	,2500	2,2000
	50	8,2000	4,0000	45,3000	1,9000	2,6000
	75	10,5000	8,0000	47,7500	16,1000	4,7000



Determina razonadamente si las siguientes afirmaciones son verdaderas o falsas:

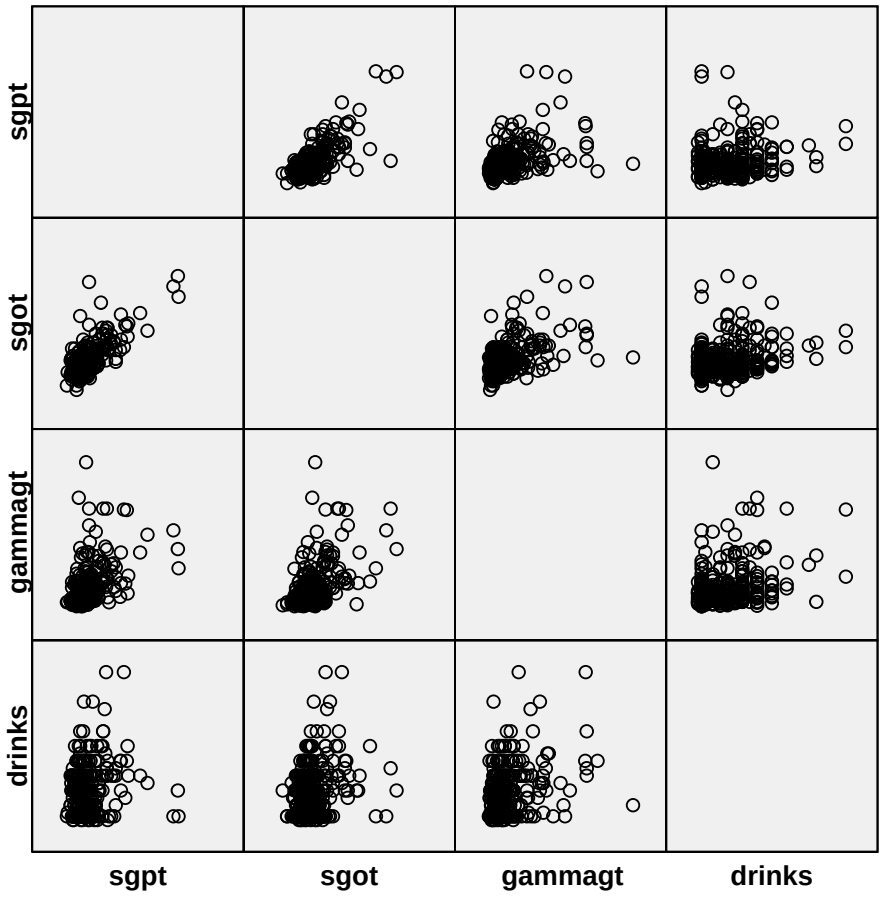
(a) En los diagramas de cajas se observa que valores altos en carbohidratos están asociados con valores bajos en fibra, por lo que la correlación entre ambas variables es negativa.
(b) Con la información disponible podemos afirmar que todas las variables tienen una distribución bastante simétrica.
(c) Con la información disponible podemos concluir que el producto etiquetado como 5 es un dato atípico para todas las variables.

13. Se realizó un estudio sobre afecciones hepáticas que pueden aparecer a causa de la ingesta excesiva de alcohol midiendo 3 variables de interés en la analítica de un total de 345 varones.

- x_1 = `sgpt` (alanina aminotransferasa).
- x_2 = `sgot` (aspartato aminotransferasa).
- x_3 = `gammagt` (gamma-glutamil transpeptidasa).

Además de estas 3 variables en la analítica de la sangre se consideró la variable

- x_4 = `drinks` (número de medias pintas de cerveza o equivalentes en bebidas alcohólicas consumidas al día).



Estadísticos

		sgpt	sgot	gammagt	drinks
N	Válidos	345	345	345	345
	Perdidos	0	0	0	0
	Media	30,4058	24,643	38,284	3,4551
	Mediana	26,0000	23,000	25,000	3,0000
	Desv. típ.	19,51231	10,0645	39,2546	3,33784
	Mínimo	4,00	5,0	5,0	,00
	Máximo	155,00	82,0	297,0	20,00
Percentiles	25	19,0000	19,000	15,000	,5000
	50	26,0000	23,000	25,000	3,0000
	75	34,0000	27,000	46,500	6,0000
	90	52,0000	35,000	82,800	8,0000
	95	67,7000	44,400	115,000	10,0000
	99	131,9000	71,780	203,000	16,0000

Contesta razonadamente a las siguientes preguntas:

(a) Verdadero o falso: Aproximadamente 259 personas de entre las encuestadas beben el equivalente a 6 medias pintas o más de cerveza al día.
(b) Verdadero o falso: La persona que más alcohol consume al día (lo correspondiente a 20 medias pintas de cerveza) tiene un nivel de gamma-glutamil transpeptidasa en sangre de 297,0.
(c) ¿Qué variables tienen la mayor correlación entre ellas?
(d) Verdadero o falso: Todas las variables consideradas son simétricas.
(e) Verdadero o falso: Mediante una regresión lineal podemos predecir el valor de la variable `drinks` si conocemos el valor de la variable `sgot`.

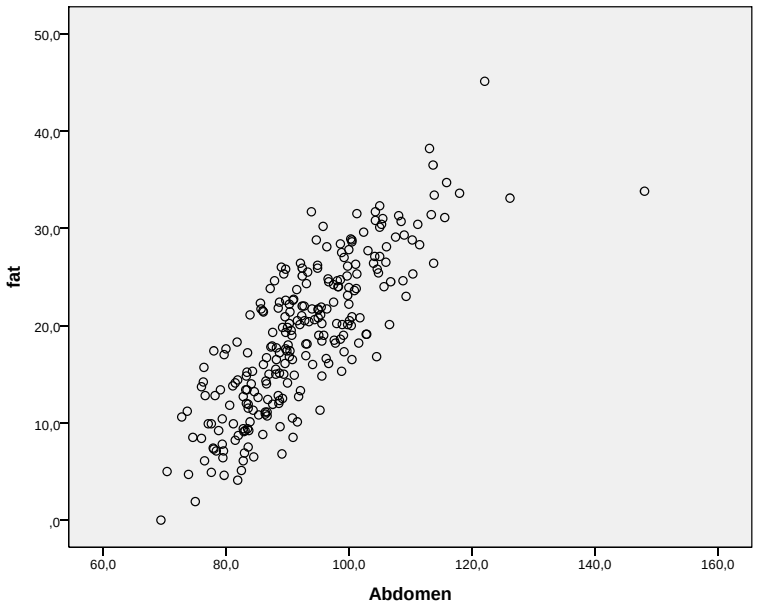
14. Un gran número de estudios afirman que el porcentaje de grasa corporal es un buen indicador de salud. En este ejercicio se pretende analizar la relación entre el porcentaje de grasa corporal y el perímetro abdominal. Se calculó de forma precisa el porcentaje de grasa corporal (variable `fat`) de 252 hombres mediante una técnica de pesado bajo el agua y utilizando la fórmula de Brozek. Además, se midió su perímetro abdominal (variable `Abdomen`) en centímetros desde el ombligo y al nivel de la cresta ilíaca.

Los resultados de los análisis estadísticos de estos datos se encuentran después de las preguntas.

(a) Calcula la varianza de la variable `Abdomen`. ¿En qué unidades se expresa?

(b) ¿Aproximadamente cuántos hombres de la muestra tienen un porcentaje de grasa corporal superior al $24,6\,\%$?

(c) Identifica cada una de las observaciones 39, 41 y 216 en el diagrama de dispersión de las variables `Abdomen` y `fat`.



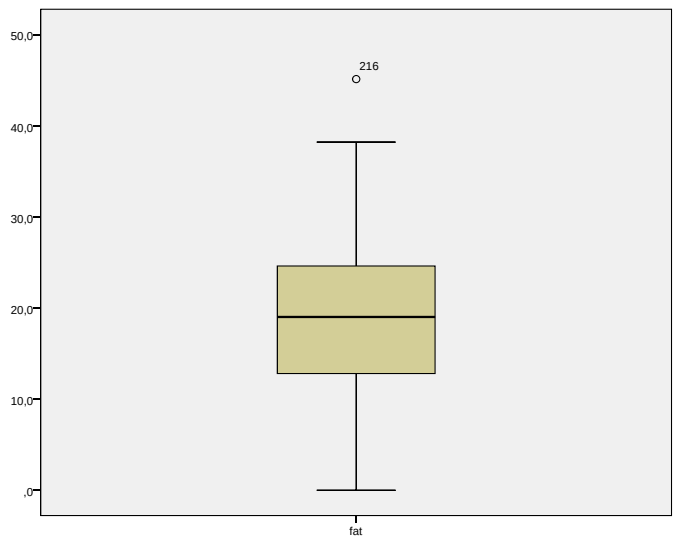
(d) Si conocemos que el valor de la variable `Abdomen` es 100 en un hombre, ¿qué valor podemos predecir para la variable `fat` en ese mismo hombre?

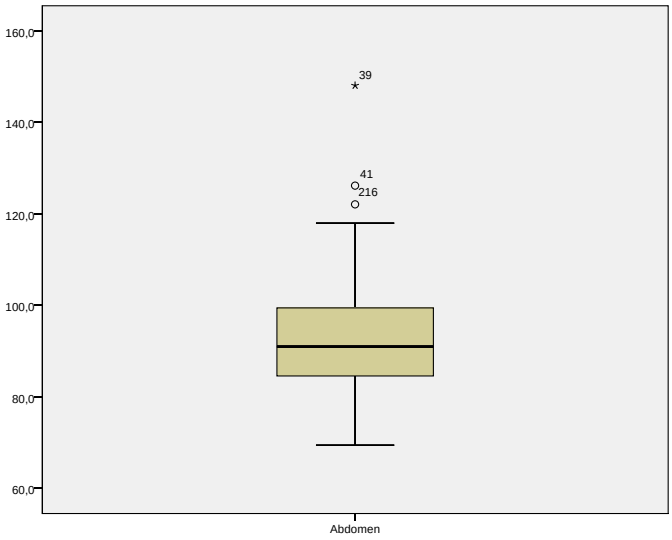
Estadísticos

		fat	Abdomen
N	Válidos	252	252
	Perdidos	0	0
Media		18,938	92,556
Mediana		19,000	90,950
Desv. típ.		7,7509	10,7831
Mínimo		,0	69,4
Máximo		45,1	148,1
Percentiles	25	12,800	84,525
	50	19,000	90,950
	75	24,600	99,575

Correlaciones

		fat	Abdomen
Correlación de Pearson	fat	1,000	,814
	Abdomen	,814	1,000
Sig. (unilateral)	fat	.	,000
	Abdomen	,000	.
N	fat	252	252
	Abdomen	252	252





Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-35,197	2,462		-14,294	,000
	Abdomen	,585	,026	,814	22,134	,000

a. Variable dependiente: fat

Problemas y cuestiones de modelos de probabilidad

15. El "tiempo de vida activa (en días)" de un plaguicida, X, viene representado por la función de densidad:

$$f(x) = \begin{cases} \frac{1}{500} e^{-\frac{x}{500}} & \text{si } x > 0 \\ 0 & \text{en el resto} \end{cases}$$

Calcular la mediana del tiempo de vida activa. ¿Cuál es su significado?

16. La variable aleatoria X="Tiempo transcurrido (en horas) hasta el fallo de una pieza" tiene función de densidad

$$f(x) = \begin{cases} \frac{1}{15000}\, e^{-\frac{x}{15000}} & \text{si } x > 0 \\ 0 & \text{en el resto.} \end{cases}$$

(a) Calcular el tiempo medio transcurrido hasta el fallo.
(b) Calcular el porcentaje de piezas que duran entre 10000 y 15000 horas.

17. Suponiendo que la probabilidad de que un niño que nace sea varón es 0,50, hallar la probabilidad de que una familia de 6 hijos tenga

(a) por lo menos una niña,
(b) por lo menos un niño,
(c) por lo menos dos niños y una niña.

18. Una compañía de seguros con 10000 asegurados halla que el 0,005% de la población fallece cada año de un cierto tipo de accidente.

(a) Hallar la probabilidad de que la compañía tenga que pagar a más de tres asegurados, por dicho accidente, en un año determinado.
(b) ¿Cuál es el número medio de accidentes por año?

19. La probabilidad de que un individuo tenga una reacción alérgica al inyectarle un suero es 0,001. Hallar la probabilidad de que en 2000 individuos tengan reacción alérgica

(a) exactamente tres,
(b) más de 2.

20. Se considera que la variable aleatoria "Kg. de algodón recogidos por parcela" sigue una distribución $N(\mu = 100; \sigma = 10)$.

Hallar el porcentaje de parcelas en las que el número de Kg. recogidos será inferior a 115.

21. En 1969 se descubrió que los faisanes de Montana (Estados Unidos) padecían una apreciable contaminación por mercurio debida a que habían comido semillas tratadas para su crecimiento con metilo de mercurio. Se sabe que el nivel de mercurio (medido en ppm) de un faisán seleccionado aleatoriamente en la población es una variable aleatoria con distribución $N(\mu = 0,25; \sigma = 0,10)$.

(a) Calcula la probabilidad de que, al seleccionar aleatoriamente un faisán de la población, su nivel de mercurio supere 0,30 ppm.
(b) Si se seleccionan aleatoria e independientemente 100 faisanes, clacula la probabilidad de que al menos 45 de ellos tengan un nivel de mercurio superior a 0,25 ppm.
(c) Si se seleccionan aleatoria e independientemente cuatro faisanes, calcula la probabilidad de que su nivel medio de mercurio sea superior a 0,30 ppm.

22. Un zoólogo estudia una cierta especie de ratones de campo. Para ello captura ejemplares de una población grande en la que la proporción de dicha especie es p.

(a) Si $p = 0,30$, hallar la probabilidad de que en 6 ejemplares capturados haya al menos 2 de los que le interesan.
(b) Si $p = 0,03$, calcular la probabilidad de que en 200 haya exactamente 3 de los que le interesan.
(c) Si $p = 0,40$, calcular la probabilidad de que en 200 haya entre 75 y 110 de los que le interesan.

23. Se sabe que el nivel de tensión sanguínea diastólica (en mmHg) en una población es una variable con distribución normal de media $\mu = 87$ y desviación típica $\sigma = 7,5$. Un individuo se clasifica como *hipertenso* si su presión es mayor de 90 mmHg.

(a) Calcula la probabilidad de que un individuo seleccionado al azar en esta población sea *hipertenso*.
(b) Si se seleccionan aleatoriamente 100 individuos de la población, calcula la probabilidad aproximada de que entre ellos haya más de 40 *hipertensos*.

(c) Calcula el valor aproximado del primer cuartil de la población, es decir, el valor Q_1 tal que la tensión sanguínea del 25 % de los individuos de la población es menor que Q_1.
(d) Si se seleccionan aleatoriamente 9 individuos de la población, calcula la probabilidad de que su tensión media sea superior a 90 mmHg.

24. La capacidad de enrollar la lengua está controlada por una pareja de genes: el gen E que determina su enrollamiento y el gen e que lo impide. El gen E es dominante, de modo que una persona Ee será capaz de enrollar la lengua.

En una ciudad grande se sabe que aproximadamente el 40 % no puede enrollar la lengua y el 60 % si puede.

Si elegimos 200 personas al azar, ¿cuál es la probabilidad de que más de 70 no puedan enrollar su lengua?

25. En una granja dedicada a la helicicultura se crían dos tipos de caracoles: el común y el romano. La velocidad (en metros por hora) del caracol común de jardín (*Helix aspersa*) sigue una distribución $N(\mu = 50;\ \sigma = 5)$. La velocidad (en metros por hora) del caracol romano (*Helix pomatia*) sigue una distribución $N(\mu = 42;\ \sigma = 5)$.

(a) Calcular el porcentaje de caracoles comunes de jardín que recorren menos de 60 metros en un hora.
(b) Calcular la probabilidad de que un caracol de jardín elegido al azar recorra menos espacio en una hora que un caracol romano.

26. Se sabe que los niveles de triglicéridos (en mg/dL) en una población, tanto para los hombres como para las mujeres, tienen distribución normal. Para los hombres la distribución es $N(\mu = 100; \sigma = 30)$, y para las mujeres la distribución es $N(\mu = 90; \sigma = 25)$.

(a) Seleccionando un hombre al azar, ¿cuál es la probabilidad de que su nivel de triglicéridos sea inferior a 130 mg/dL?
(b) Si se seleccionan aleatoria e independientemente un hombre y una mujer, ¿cuál es la probabilidad de que el nivel de triglicéridos de la mujer sea superior al del hombre?

27. Una línea eléctrica se avería cuando la tensión sobrepasa la capacidad de la línea. Si la tensión es $N(100;\ 20)$ y la capacidad es $N(140;\ 10)$, calcular la probabilidad de avería.

28. La concentración de ácido úrico en sangre (mg/dl) en la población de pacientes con síndrome de apnea-hipopnea durante el sueño (SAHS) sigue una distribución normal, $N(\mu = 6,30; \sigma = 1,50)$. Se recomienda que la concentración de ácido úrico se mantenga por debajo de 7,20 mg/dl ya que, por encima de ese valor, comienzan a surgir problemas (cálculos renales, gota, ...).

(a) Calcula el porcentaje de pacientes de la población anterior, cuya concentración de ácido úrico se mantiene por debajo de 7,20 mg/dl.
(b) Si se seleccionan 50 pacientes al azar en esa población, calcula la probabilidad aproximada de que más de 30 de ellos tengan un nivel de ácido úrico aceptable (por debajo de 7,20 mg/dl).
(c) Si se seleccionan al azar dos pacientes en esa población, ¿cuál es la probabilidad de que la diferencia de ácido úrico entre ellos sea inferior a 1 mg/dl?

Problemas y cuestiones de estimación puntual

29. Dada una muestra aleatoria de tamaño n de una variable X, calcular el estimador de máxima verosimilitud y el del método de los momentos, en los siguientes casos:

(a) $X \sim$ Bernoulli de parámetro p.
(b) $X \sim$ Poisson (λ).
(c) $X \sim$ Exponencial (λ); es decir, $f_\lambda(x) = \lambda e^{-\lambda x}$, para $x > 0$ $(\lambda > 0)$.
(d) $X \sim N(\mu, \sigma)$.

30. La proporción de genes dañados en un tejido celular tras una sesión de radiación es una variable aleatoria continua X con función de densidad $f(x) = \theta x^{\theta-1}$, para valores de $x \in (0,1)$, siendo $\theta > 0$ un parámetro desconocido que depende del tipo de tejido.

(a) Dada una muestra aleatoria $(X_1, \ldots, X_n)$ de X, calcúlese el estimador de θ por el método de los momentos y por el método de máxima verosimilitud.
(b) Tras analizar una muestra de 3 tejidos celulares, se obtuvieron los valores 0,10, 0,15 y 0,25. ¿Cuáles son la estimaciones concretas de θ con estos datos con los dos métodos?

31. En una gran piscifactoría hay una proporción desconocida, p, de cierto tipo de truchas. Para obtener información sobre esa proporción desconocida, vamos a ir sacando peces al azar hasta obtener una trucha de ese tipo, en tres ubicaciones diferentes:

En la primera ubicación se obtiene la primera trucha de ese tipo en la décima extracción.

En la segunda ubicación se obtiene la primera trucha de ese tipo en la decimoquinta extracción.

En la tercera ubicación se obtiene la primera trucha de ese tipo en la decimoctava extracción.

Escribir la función de verosimilitud y obtener la estimación de máxima verosimilitud de p.

32. Un modelo genético para las moscas de cierta variedad nos dice que pueden ser de tres tipos: homocigóticas AA (con probabilidad p^2), homocigóticas BB (con probabilidad q^2) y heterocigóticas AB (con probabilidad $2pq$), donde naturalmente $p + q = 1$.

En una muestra aleatoria de 100 moscas obtenemos 10 de tipo AA, 50 de tipo BB, y 40 de tipo AB.

Hallar la estimación de máxima verosimilitud de p con los datos obtenidos.

33. Una variable relacionada con el número de mutaciones en una secuencia de ADN puede tomar los valores 0, 1 ó 2 con probabilidades $(1+2\theta)/3$, $(1-\theta)/3$ y $(1-\theta)/3$ respectivamente, donde θ es un parámetro desconocido. Se han obtenido 60 observaciones independientes de esta variable resultando la siguiente tabla de frecuencias absolutas:

Valores	0	1	2
Frecuencias	25	20	15

(a) Estima el valor de θ a partir de las observaciones disponibles usando el método de los momentos.
(b) Estima el valor de θ a partir de las observaciones disponibles usando el método de máxima verosimilitud.

34. En el artículo *"A nanomaterial-based breath test for distinguishing gastric cancer from benign gastric conditions"* publicado en *British Journal of Cancer* en 2013 se describe un análisis de aliento sencillo que se puede usar para el diagnóstico precoz de cáncer de estómago, a partir de la medición de cinco sustancias: 2-propenonitrilo, 2-butoxietanol, furfural, 6-metil-5-hepten-2-ona e isopreno.

La sensibilidad de la prueba (probabilidad de dar positivo teniendo cáncer) es del 89 %, y su especificidad (probabilidad de dar negativo estando sano) es del 94 %.

Este análisis de aliento se aplica a un grupo aleatorio de 50 personas de una población de riesgo, obteniéndose 10 resultados positivos y 40 negativos.

(a) Estima q="Proporción de resultados positivos", razonando tu respuesta.
(b) Halla la relación entre p="Proporción de personas con cáncer de estómago" en esa población de riesgo y q="Proporción de resultados positivos", y utilízala para estimar p.

35. Para estudiar la proporción p de caballos afectados por la peste equina se les va a someter a una prueba. Se sabe que la prueba resulta positiva si el animal está enfermo. Además, si el animal está sano, hay una probabilidad 0.04 de que la prueba resulte positiva.

(a) Estudia la relación entre la probabilidad p de que un caballo esté enfermo y la probabilidad q de que la prueba resulte positiva.
(b) Si se realizó la prueba a 500 caballos y resultó positiva en 95 casos, ¿cuál es el estimador de máxima verosimilitud de q? A partir del resultado del apartado (a), calcula una estimación de p.

36. Unos laboratorios desarrollan una prueba sencilla para detectar la *gripe del pollo*. La prueba tiene una fiabilidad muy aceptable: proporciona un 4% de falsos positivos (prueba positiva cuando el pollo está sano) y un 0% de falsos negativos (prueba negativa cuando el pollo está enfermo).

En una granja avícola, se detecta un brote de *gripe del pollo*. Mediante la utilización de la prueba sencilla que se ha descrito anteriormente, se quiere estimar la incidencia de la enfermedad en esa granja. Para esto, se seleccionan al azar 100 pollos, se les efectúa la prueba y se obtienen 20 casos positivos. Estimar la proporción p de pollos enfermos en la granja, explicando todo el proceso seguido.

37. **La diferencia entre intención de voto directo y estimación de voto.**

En cierto país con dos partidos políticos, A y B, se lleva a cabo un sondeo sobre opinión electoral (lo podemos llamar barómetro) en el que se pregunta:

Suponiendo que mañana se celebrasen elecciones, ¿a qué partido votaría?

Los resultados son: 23% votaría A, 27% votaría B, y 50% de indecisos. Por tanto, en la **intención de voto directo**, saldría triunfador el partido B.

Pero hay un 50% de indecisos, de los cuales se puede intentar predecir el comportamiento, ya que en ese mismo sondeo, se hacen otras preguntas mediante las cuales se puede hacer un perfil (conservador o progresista) de las personas, y además, se les pregunta qué votaron en las últimas elecciones.

Supongamos que los resultados de ese sondeo son los siguientes:

(i) Del 50% de indecisos, un 30% tiene un perfil progresista. De estos, un 10% votaron A en las últimas elecciones, y un 60% votaron B (el resto no votaron o votaron en blanco).
(ii) Del 50% de indecisos, un 70% tiene un perfil conservador. De estos, un 60% votaron A en las últimas elecciones, y un 10% votaron B (el resto no votaron o votaron en blanco).

A partir de toda esta información, y suponiendo que los indecisos terminarán manteniendo un patrón de comportamiento similar al de las últimas elecciones, se puede calcular la **estimación de voto**.

Problemas y cuestiones de intervalos de confianza

38. En un informe leemos que un intervalo de confianza para la puntuación media de los estudiantes españoles en un test de inglés es (267,8, 276,2) con una confianza del 95 %.

(a) Verdadero o falso: El 95 % de los estudiantes han tenido puntuaciones entre 267,8 y 276,2
(b) ¿Cuál fue la puntuación media de los estudiantes de la muestra utilizada para calcular el intervalo?
(c) ¿Es correcto afirmar que la puntuación media de los estudiantes españoles está en el intervalo (267,8, 276,2)?
(d) ¿Es correcto decir que la puntuación media de los estudiantes españoles pertenece al intervalo (267,8, 276,2) con probabilidad 0,95?

39. Con unos datos muestrales hemos calculado el intervalo de confianza para la media de una población normal al 95 %. Si ahora calculamos el intervalo de confianza al 99 %, tenemos:

◯ La amplitud del nuevo intervalo disminuirá.
◯ La amplitud del nuevo intervalo aumentará.
◯ La amplitud del nuevo intervalo será la misma, pero el error aumentará.
◯ La amplitud del nuevo intervalo será la misma, pero el error disminuirá.

40. Estamos interesados en la concentración, X, de particulas contaminantes (expresada en ppm) en situaciones de altas presiones atmosféricas. Medimos dicha concentración en 6 estaciones de la Red de Control de la Calidad del Aire, tras cinco días de altas presiones, encontrando una concentración media de 150 ppm, con una cuasidesviación típica de 16 ppm.

(a) Calcula un intervalo de confianza para la concentración media de partículas contaminantes en esas condiciones, con un nivel de confianza del 95 %. ¿Qué hipótesis has necesitado para calcular este intervalo?
(b) ¿En cuántas estaciones deberíamos medir dicha concentración, para estimar la concentración media con un error inferior a 5 ppm (con el mismo nivel de confianza)?

41. En un estudio[1] se analizaron los factores que determinan la cantidad de algunos micronutrientes en la sangre mediante una muestra de 314 pacientes. Se considera la variable logBeta (logaritmo de la cantidad de β-caroteno en el plasma sanguíneo, expresada en ng/ml) en dos grupos de personas; fumadores y no fumadores. Los resultados obtenidos son los siguientes (algunos datos han sido suprimidos de la salida):

		N	Media	Desviación típ.	Error típ. de la media
logBeta	Fumadores	158	4,8520	,76648	AA
	No Fumadores	156	5,0685	,71422	BB

¿Permiten los datos afirmar a nivel $\alpha = 0{,}05$ que la media del logaritmo de la cantidad de β-caroteno es más alta en individuos que no fuman que en individuos fumadores? Enumera las condiciones bajo las cuales es válida tu respuesta. Calcula los valores de AA y BB en la tabla anterior.

42. Las toxinas producidas por cianobacterias (microorganismos procariotas capaces de realizar fotosíntesis oxigénica) en los embalses han sido reconocidas como un problema de salud (Chorus & Bartram, 2002); por ello, la nueva legislación europea sobre aguas de baño incluye el seguimiento y control de estas bacterias en los embalses.

(a) De 33 embalses muestreados en España, 17 presentan una gran abundancia de cianobacterias, frente a 16 que no presentan una cantidad preocupante. A partir de estos datos, dar un intervalo de confianza para estimar la proporción de embalses españoles con una gran abundancia de cianobacterias (con un nivel de confianza del 90 %).
(b) ¿Cuántos embalses habría que muestrear para que el error en la estimación de esa proporción quede por debajo de 0,04?

43. Un equipo de investigadores quiere estimar la proporción p de vacas que sufren el mal de las vacas locas en una gran explotación ganadera, mediante un intervalo con un error máximo de 0,015 y nivel de confianza 0,95. ¿A cuántas vacas deben analizar para alcanzar aproximadamente este objetivo, sabiendo que en un pequeño sondeo orientativo (muestra piloto) resultó que el 15 % de las vacas estaban afectadas por la enfermedad?

[1] Nierenberg *et al.* (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology* , 130, 511–521.

44. Se quiere estimar la proporción de manatíes en el Caribe que han sido heridos por hélices de barcos. ¿A cuántos manatíes tendremos que examinar para asegurar que la estimación tiene un error máximo del 10 % con un nivel de confianza del 95 %?

45. Se admite que el número de microorganismos en una muestra de 1 mm cúbico de agua de un río sigue una distribución de Poisson de parámetro λ. En 40 muestras se han detectado, en total, 833 microorganismos. Calcula un estimador puntual y un intervalo de confianza al 90 % para λ.

46. Nueve personas participan en el estudio de un producto que intenta reducir el apetito (clorofenilpiperacina). Cada uno de ello recibe este producto durante 2 semanas y placebo durante otras 2 semanas (naturalmente, el orden de los períodos de 2 semanas es aleatorio y ellos no lo conocen). Al final de cada período, se les pide que expresen su sensación de hambre (en una escala del 0 al 150). Los resultados son los siguientes:

Individuo	1	2	3	4	5	6	7	8	9
Después del producto	79	48	52	15	61	107	77	54	5
Después del placebo	78	54	142	25	101	99	94	107	64

(a) Hallar un intervalo de confianza al 95 % para la diferencia de las sensaciones medias de hambre con el producto y con placebo (asumir Normalidad).
(b) Lo mismo, pero trabajando (equivocadamente) con las muestras como si fueran independientes (asumir Normalidad e igualdad de varianzas).

Problemas y cuestiones de contrastes de hipótesis

47. Una farmacéutica desea sacar al mercado un antiinflamatorio con un nuevo tipo de ibuprofeno que reduzca el tiempo que tarda en hacer efecto (tiempo de efecto) respecto al de los medicamentos genéricos. Llamamos μ_N y μ_G al tiempo medio de efecto del nuevo medicamento y del genérico, respectivamente. Para poder sacar al mercado este nuevo medicamento tendremos que contrastar:

- ◯ $H_0 : \mu_N \leq \mu_G$.
- ◯ $H_0 : \mu_N = \mu_G$.
- ◯ $H_0 : \mu_G \leq \mu_N$.
- ◯ Ninguna de las restantes.

48. El ordenador nos proporciona un p-valor para un contraste con el que rechazamos la hipótesis nula, H_0, con significación $\alpha = 0{,}05$. Ahora queremos decidir si rechazar o aceptar H_0 con $\alpha = 0{,}01$.

- ◯ Tendremos que calcular nuevamente el p-valor para rechazar o aceptar H_0 con $\alpha = 0{,}01$.
- ◯ Siempre rechazamos H_0 con el nivel de significación $\alpha = 0{,}01$.
- ◯ Nunca rechazamos H_0 con el nivel de significación $\alpha = 0{,}01$.
- ◯ La decisión dependerá del valor del p-valor que nos dio el ordenador inicialmente.

49. La concentración media de dióxido de carbono en el aire a cierta altura es habitualmente de unas 355 p.p.m. (partes por millón). Se sospecha que esta concentración es mayor en la capa de aire más próxima a la superficie. Para contrastar esta hipótesis se analizó el aire en 20 puntos elegidos aleatoriamente a una misma altura cerca del suelo. Resultó una media muestral de 580 p.p.m. y una cuasi-desviación típica muestral de 180. Suponiendo normalidad para las mediciones, ¿proporcionan estos datos suficiente evidencia estadística, al nivel 0.01, a favor de la hipótesis de que la concentración es mayor cerca del suelo? Indicar razonadamente si el p-valor es mayor o menor que 0,01.

50. Un fabricante de materiales para insonorización produce dos tipos A y B. De los 1000 primeros lotes vendidos, 560 fueron del tipo A. ¿Proporcionan estos datos suficiente evidencia estadística (al nivel de significación 0.01) para concluir que los consumidores prefieren mayoritariamente el tipo A?

51. Se están estudiando dos colonias de ñúes azules, una que vive en un parque de Tanzania, y otra que vive en un parque de Kenia. Parece que la altura en Tanzania es mayor que la altura en Kenia. Se estudia una muestra de 10 ñúes en Tanzania, obteniéndose una altura media muestral de 130 cm con una cuasi-varianza muestral de 80, y otra muestra de 15 ñúes en Kenia, obteniéndose una altura media muestral de 124 cm con una cuasi-varianza muestral de 75. Asumiendo Normalidad para las alturas en las dos colonias, se pide:

(a) Con un nivel de significación de 0,10, ¿podemos aceptar igualdad de varianzas de las alturas en las dos colonias?
(b) ¿Disponemos de suficiente evidencia muestral para asegurar que la altura media en Tanzania es mayor que en Kenia (al nivel de significación 0,10)?

52. Se han analizado con SPSS los datos del fichero `mercurio.txt` con el fin de analizar si el nivel medio de contaminación por mercurio en los dos ríos es o no diferente. La salida obtenida ha sido la siguiente:

Estadísticos de grupo

	RIO	N	Media	Desviación típ.	Error típ. de la media
CONC	,00	73	1,0781	,64861	,07591
	1,00	98	1,2764	,82915	,08376

Prueba de muestras independientes

		Prueba T para la igualdad de medias						
							95% Intervalo de confianza para la diferencia	
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
CONC	Se han asumido varianzas iguales	-1,694	169	,092	-,19835	,11712	-,42954	,03285
	No se han asumido varianzas iguales	-1,755	168,570	,081	-,19835	,11304	-,42150	,02481

(a) ¿Existe evidencia estadística para afirmar al nivel $\alpha = 0{,}05$ que el nivel medio de concentración en el río Wacamaw (1) es superior al nivel medio en el río Lumber (0)?
(b) Indica las suposiciones previas necesarias para garantizar la validez del procedimiento empleado.

53. Se ha llevado a cabo un estudio para determinar si los niveles de colesterol medidos en personas que han sufrido un infarto son diferentes a los medidos en personas que no lo han sufrido. Para ello, se midió el colesterol de un grupo de enfermos (grupo 1) dos días después de sufrir un infarto. Como control, también se midió el colesterol de un grupo de personas sanas (grupo 2). Los datos fueron analizados con SPSS, con los siguientes resultados (algunos valores han sido sustituidos por letras):

Estadísticos de grupo

	grupo	N	Media	Desviación típ.	Error típ. de la media
colesterol	1	28	253,9286	47,71049	9,01644
	2	30	193,1333	22,30004	4,07141

Prueba de muestras independientes

		Prueba T para la igualdad de medias						
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
							Inferior	Superior
colesterol	Se han asumido varianzas iguales	BB	56	,000	60,79524	AA	41,41849	80,17199

(a) Calcula los valores AA y BB que han sido suprimidos en la salida anterior.
(b) A partir de la información disponible, ¿es posible afirmar a nivel $\alpha = 0{,}01$ que los niveles medios de colesterol son diferentes en la población de pacientes de infarto y en la población de personas sanas? Escribe claramente la hipótesis nula, la hipótesis alternativa y el p-valor del contraste que hay que realizar para contestar a la pregunta. Escribe también bajo qué supuestos las fórmulas utilizadas son válidas.
(c) Calcula un intervalo de confianza de nivel 0,95 para el nivel medio de colesterol en la población de personas sanas.

54. Para determinar el efecto del consumo de calcio sobre la tensión arterial, se midió la tensión arterial sistólica de 10 personas antes y después de recibir un suplemento de calcio en su dieta durante 12 semanas. Se analizaron los datos con SPSS, con los siguientes resultados:

Correlaciones de muestras relacionadas

	N	Correlación	Sig.
AntesCA y DespuesCA	10	,602	,065

Estadísticos de muestras relacionadas

	Media	N	Desviación típ.	Error típ. de la media
AntesCA	114,9000	10	10,83667	3,42685
DespuesCA	109,9000	10	7,79530	2,46509

Prueba de muestras relacionadas

	Diferencias relacionadas			t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media			
AntesCA - DespuesCA	5,00000	8,74325	2,76486	1,808	9	,104

(a) A partir de los resultados anteriores, ¿puede afirmarse a nivel $\alpha = 0{,}05$ que la tensión arterial es significativamente menor después del tratamiento que antes? Escribe claramente la hipótesis nula, la hipótesis alternativa y el p-valor del contraste aplicado para responder a la pregunta anterior.
(b) Calcula un intervalo de confianza de nivel 95% para la diferencia poblacional media entre la tensión anterior y la posterior al tratamiento.

55. El maíz común no tiene la cantidad de aminoácido lisina que necesitan los animales en su pienso. Sin embargo, unos científicos han desarrollado una variedad de maíz transgénico con alto contenido en lisina dedicado a pienso animal. En un experimento, un grupo de 20 pollos recibió una ración del maíz transgénico mientras que otros 20 pollos recibieron una ración de maíz común. Se registraron las ganancias en peso (en gramos) de los 40 pollos y se procesaron los datos obtenidos con SPSS obteniéndose los siguientes resultados:

	GRUPO	N	Media	Desviación típ.	Error típ. de la media
PESO	comun	20	366,3000	50,8052	11,3604
	trans	20	402,9500	42,7286	9,5544

		Prueba T para la igualdad de medias				
		t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia
PESO	Se han asumido varianzas iguales	-2,469	38	,018	-36,6500	

(a) La opción de SPSS utilizada para producir esta salida ha sido
- ◯ Prueba T para una muestra
- ◯ Prueba T para muestras independientes
- ◯ Prueba T para muestras relacionadas

(b) Se desea encontrar evidencia empírica de que la ganancia media de peso de los pollos con dieta transgénica es superior a la de los pollos con dieta común. El p-valor del contraste correspondiente es ◯ $0,018$; ◯ $0,036$; ◯ $0,009$.

(c) La hipótesis nula de que la dieta que siguen los pollos no influye en su ganancia media de peso se rechaza al nivel de significación α si
◯ $\alpha > 0,018$; ◯ $\alpha/2 > 0,018$; ◯ $\alpha < 0,018$.

(d) En la salida de SPSS se ha omitido el error típico de la diferencia de medias. Este error típico
- ◯ Es aproximadamente igual a $14,8440$.
- ◯ No se puede calcular con los datos disponibles.
- ◯ Es aproximadamente igual a $20,9148$.

Problemas y cuestiones de bondad de ajuste

56. Después de lanzar un dado 300 veces, se han obtenido las siguientes frecuencias:

	1	2	3	4	5	6
Frecuencias	43	49	56	45	66	41

Al nivel de significación 0,05, ¿se puede aceptar que el dado es regular?

57. En 1778, H. Cavendish realizó una serie de 29 experimentos con objeto de medir la densidad de la tierra. Sus resultados, tomando como unidad la densidad del agua, fueron:

5'50	5'61	4'88	5'07	5'26	5'55	5'36	5'29	5'58	5'65
5'57	5'53	5'62	5'29	5'44	5'34	5'79	5'10	5'27	5'39
5'42	5'47	5'63	5'34	5'46	5'30	5'75	5'68	5'85	

Al nivel de significación 0.05, ¿se puede aceptar que la densidad de la tierra se ajusta a una distribución Normal?

Observación:

Utilizar los intervalos A_1="Menor o igual que 5,30", A_2=(5,30; 5,45] , A_3=(5,45; 5,60] , A_4="Mayor que 5,60", y el hecho de que, a partir de los datos, se obtiene que $\hat{\mu} = 5,45$ y que $\hat{\sigma} = 0,22$.

58. Nos dicen que un programa de ordenador genera observaciones de una distribución $N(0;1)$. Como no estamos seguros de ello, obtenemos una muestra aleatoria de 450 observaciones, mediante dicho programa, obteniéndose los siguientes resultados:

30 observaciones menores que -2;

80 observaciones entre -2 y -1;

140 observaciones entre -1 y 0;

110 observaciones entre 0 y 1;

60 observaciones entre 1 y 2;

30 observaciones mayores que 2.

¿Se puede aceptar, al nivel $\alpha = 0,01$, que el programa funciona correctamente?

59. La tabla que aparece a continuación, muestra la frecuencia de la cifra final del *gordo* de 200 sorteos de la Lotería de Navidad:

Cifra final	0	1	2	3	4	5	6	7	8	9
Frecuencia	20	8	13	20	27	30	26	20	20	16

¿Se puede aceptar, al nivel de significación del 1 %, que todas las terminaciones son igualmente probables?

60. Se desea estudiar el número de erratas por página que se producen en la edición de una enciclopedia, antes de su lanzamiento. Para ello se analizan 200 páginas seleccionadas al azar, encontrando los siguientes resultados:

Número de erratas por página	0	1	2	3
Número de páginas	150	42	5	3

(a) Asumiendo que el "número de erratas por página" sigue un modelo de Poisson, calcula un intervalo (al 90 % de confianza) para el número medio de erratas por página en toda la enciclopedia.
(b) Con los datos obtenidos, ¿es realmente aceptable, al nivel de significación del 10 %, que el "número de erratas por página" sigue una distribución de Poisson?

61. Un modelo genético para las moscas de cierta variedad nos dice que pueden ser de tres tipos: homocigóticas AA (con probabilidad p^2), homocigóticas BB (con probabilidad q^2) y heterocigóticas AB (con probabilidad $2pq$), donde naturalmente $p + q = 1$.

En una muestra aleatoria de 100 moscas obtenemos 10 de tipo AA, 50 de tipo BB, y 40 de tipo AB.

¿Se ajustan los datos a dicho modelo genético, al nivel de significación 0,05?

62. Se clasificaron 1000 individuos de una población según el sexo y según fueran normales o daltónicos.

	Masculino	Femenino
Normal	442	514
Daltónicos	38	6

Según un modelo genético, las probabilidades deberían ser:

$$\frac{1}{2}p \quad \frac{1}{2}p^2 + pq$$

$$\frac{1}{2}q \qquad \frac{1}{2}q^2$$

donde $q = 1 - p$ = proporción de genes defectuosos en la población.

A partir de la muestra se ha estimado que $q = 0,087$. ¿Concuerdan los datos con el modelo, al nivel de significación 0,05?

63. Un Ayuntamiento decide poner 4 contenedores para reciclar papel en una zona de la ciudad, con la idea de que sean utilizados por la misma cantidad de personas (aproximadamente). Para ver si esto es cierto, hace una encuesta en la zona a 300 personas, preguntándoles que contenedor utilizan. Los resultados obtenidos son los siguientes:

El contenedor 1 es utilizado por 80 personas.

El contenedor 2 es utilizado por 70 personas.

El contenedor 3 es utilizado por 85 personas.

El contenedor 4 es utilizado por 65 personas.

(a) Como consecuencia de estos resultados, ¿resulta aceptable que los 4 contenedores tienen el mismo nivel de utilización? Dar una respuesta razonada, con un nivel de significación de 0.10.
(b) El p-valor del contraste anterior, ¿es inferior o superior a 0.10? Dar una respuesta razonada.