Tema 1: Estadística Descriptiva

1. La fiabilidad de un ordenador se mide en términos de la vida de un componente de hardware específico (por ejemplo, la unidad de disco). Con objeto de estimar la fiabilidad de un sistema en particular, se prueban 100 componentes de un ordenador hasta que fallan, y se registra su vida.

   (a) Determinar la población de interés, los individuos y la muestra.

   (b) Determinar el carácter, su tipo y las posibles modalidades.

   (c) ¿Cómo podría utilizarse la información de la muestra para estimar la fiabilidad del sistema?

2. En cada caso, determinar el tipo de distribución, organizar los datos en una tabla de frecuencias y representar gráficamente la distribución. También se pide, calcular algunas medidas de tendencia central, medidas de dispersión, de simetría y de apuntamiento.

   (a) Consumo de combustible (litros/100km a 90km/h) de seis automóviles de la misma marca.

   $$6'7 \quad 6'3 \quad 6'5 \quad 6'5 \quad 6'4 \quad 6'6$$

   (b) Resultados obtenidos en las pruebas de durabilidad de 80 lámparas eléctricas con filamento de tungsteno. La vida de cada lámpara se da en horas, aproximando las cifras a la hora más cercana.

854	1284	1001	911	1168	963	1279	1494	798	1599	1357	1090	1082
1494	1684	1281	590	960	1310	1571	1355	1502	1251	1666	778	1200
849	1454	919	1484	1550	628	1325	1073	1273	1710	1734	1928	1416
1465	1608	1367	1152	1393	1339	1026	1299	1242	1508	705	1199	1155
822	1448	1623	1084	1220	1650	1091	210	1058	1930	1365	1291	683
1399	1198	518	1199	2074	811	1137	1185	892	937	945	1215	905
1810	1265											

3. Calcular los valores que se piden en función de los datos:

   (a) Si $N = 2$, $\bar{x} = 2'6$ y $\sigma = 1'1$, ¿cuáles son los datos de la muestra?

   (b) Si $CV = 0'5$, $\bar{x} = 2$ y $m_3 = 14$, ¿cuánto vale μ_3?

4. Se considera la siguiente tabla de frecuencias donde las distintas modalidades están ordenadas de menor a mayor

x_i	n_i	N_i	f_i	F_i
	10			
0		15		0'3
3				
5			0'08	
20				0'8
25		46		
50				1

   (a) Completar la tabla estadística, utilizando los datos que ya contiene, y los valores de las siguientes medidas: N=50, $\bar{x}$=10, Me=4, Mo=10, Rango=51 y σ^2=201.

   (b) Determinar qué datos y medidas resultan irrelevantes para completar la tabla.

5. Consideremos la siguiente tabla de frecuencias

Clases	[20,30)	[30,40)	[40,50)	[50,60)	[60,80)	[80,100)
n_i	10	18	20	27	30	45

   (a) ¿Qué porcentaje de observaciones son superiores a la observación 53?

   (b) ¿Qué valor de la variable es superado por el 65% de la población?

6. Dada la tabla de frecuencias absolutas:

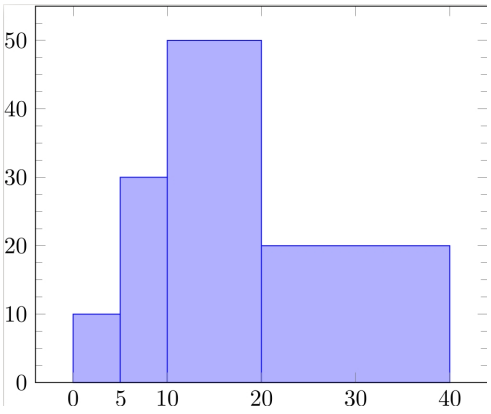
Clases	(-∞,2]	(2,4]	(4,5]	(5,6]	(6,7]	(7,10]	(10,∞)
n_i	5	7	5	6	4	7	6

   (a) Dibuja el correspondiente histograma.

   (b) A la vista del histograma. ¿Qué porcentaje de la población se encuentra entre 3 y 6?

   (c) ¿Cuál es el intervalo modal?

7. Sabiendo que los datos del siguiente histograma provienen de una muestra de 850 individuos.



   (a) Calcula la tabla de frecuencias.

   (b) Calcula los momentos centrales de orden 1, 2, 3 y 4.

   (c) Calcula los coeficientes de asimetría y curtosis de Fisher.

8. El tamaño de la muestra A es 10, la media y la mediana son respectivamente 16'5 y 13 y la varianza es 5'1. El tamaño de la muestra B es 20, la media y la mediana son respectivamente 11'4 y 10 y la varianza es 3'9. Consideremos la unión de las dos muestras, que denotaremos por C, cuyo tamaño es 30. Si es posible, calcule la media, la mediana y la varianza de la muestra C, y en otro caso, determine la posición aproximada de la medida desconocida.

9. El sueldo medio de los obreros de una fábrica es 1.500 euros. En las negociaciones del nuevo convenio colectivo se presentan dos alternativas: un aumento de 150 euros euros a cada obrero o un aumento del 10% del sueldo de cada uno. Estudiar qué modalidad es más social en el sentido de que iguala más los salarios.

10. Sea k un número entero positivo. Determine la media, la varianza y el sesgo en cada una de las siguientes muestras:

    (a) $M_1 = \{1, 2, 3, ..., k\}$

    (b) $M_2 = \{p, p + c, p + 2c, p + 3c, ..., p + kc\}$, con $p \in \mathbb{R}$.

11. En un examen final de Estadística, la puntuación media de 150 estudiantes fue de $7'8$, y la desviación típica de $0'8$. En Cálculo, la media fue $7'3$ y la desviación típica $0'76$. ¿En qué materia fue mayor la dispersión en términos absolutos? ¿y en términos relativos? Explicar la respuesta. Si un alumno obtuvo $7'5$ en Estadística y $7'1$ en Cálculo, ¿en qué examen sobresalió más?

12. En una muestra se obtienen los valores 2, 4, 6 y 8 de la variable X. Se pide:

    (a) Calcular la media y la varianza de los valores de la muestra.

    (b) Hallar los valores tipificados de la variable X y comprobar que la media de estos nuevos valores es 0 y la varianza es 1.

    (c) Demostrar que el resultado del apartado anterior constituye una propiedad de cualquier variable tipificada.

13. Las distribuciones de frecuencias de las variables X e Y son campaniformes y simétricas.Además, se sabe conocen los siguientes datos:

Variable X	Me=10	σ_x^2=4	N=2	$\sum x_i^4 f_i$=12416
Variable Y	Mo=8	σ_y^2=4	N=82	$\sum y_i^4 f_i$= 5648

    Determinar los dos valores de X, y comparar la dispersión y la curtosis de ambas variables.

14. **Sentido crítico.** Antes de extraer conclusiones de unos resultados estadísticos, conviene examinar detenidamente los valores numéricos obtenidos. El gran número de operaciones realizadas y el volumen de datos manejados son fuentes de error que inciden en los resultados. Un poco de sentido crítico puede ayudar a determinar si unos resultados son consistentes con los datos del problema. En este ejercicio se propone una serie de casos donde el resultado numérico no es correcto. Se trata de explicar razonadamente la inconsistencia del resultado en función de los datos.

    (a) El número medio de accesos a una página web es -3.

    (b) La mediana del número de hijos de las familias españolas es $2'1$.

    (c) La moda del número de hijos es $1'5$.

(d) El cuartil C_3 es 28 y el cuartil C_1 es 32.

(e) El centil P_1 es 32 y el decil D_1 es 28.

(f) La varianza es -100.

(g) La media es 10, la mediana 12 y la desviación típica es 0.

(h) La expresión $g_2 + 3$ toma un valor negativo.

15. **Modificar los datos de una muestra** En este ejercicio se va a estudiar el comportamiento de la media y la varianza cuando se pierde, se gana o se modifica algún dato de la variable. Se consideran los valores $\{2, 4, 6, 8\}$ obtenidos en una muestra. Se pide:

    (a) Calcular la media y la varianza.

    (b) En cada caso, obtener el nuevo valor de la media y la varianza sin tener que aplicar nuevamente las fórmulas a todos los datos:

    Caso1: Se descubre que el valor 8 observado es erróneo y se elimina.

    Caso2: Se cuenta con un nuevo valor, el 5, para la muestra.

    Caso3: Se descubre que el valor 8 observado es erróneo y se cambia por el verdadero valor que es el 9.

16. Estudiamos el tiempo de duración de un proceso donde, en algunos casos, el proceso ni siquiera comienza y, por tanto, el tiempo de duración es cero. Realizamos 200 pruebas y obtenemos un tiempo medio de 3'5 segundos con una varianza de 7.

    (a) Si el 23% de las pruebas fueron consideradas de tiempo 0. ¿Cuál es la media y la varianza de las restantes.

    (b) Si en las 200 pruebas se obtuvieron tiempos positivos y consideramos 50 nuevas pruebas de tiempo 0, ¿cuál es la nueva media y varianza para las 250 observaciones?

    (c) Obtener una fórmula que permita obtener la nueva media y varianza de una muestra cuando añadimos o eliminamos un número arbitrario de observaciones de valor 0.

17. **Media ponderada.**. Una generalización del concepto de media aritmética es la *media ponderada*. Se utiliza cuando se asocian ciertos valores $(w_1, w_2, \ldots, w_k)$, denominados pesos, a los valores $(x_1, x_2, \ldots, x_k)$ de la variable con el fin de dar más relevancia a unos que a otros.

$$\mathrm{MP} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

    El conjunto de los pesos $\{w_1, w_2, \ldots, w_k\}$ se denomina *ponderación*, y diremos que una ponderación es *propia* si todos los pesos son distintos de cero, es decir, $w_i \neq 0$ para todo $i = 1, \ldots, n$.

    Ahora, veamos un ejemplo: Si la nota final de una asignatura se obtiene mediante la realización de tres pruebas parciales con pesos 1, 2 y 2, indica que la prueba segunda y tercera tiene el doble de importancia que la primera. En este caso, un alumno cuyas notas hubiesen sido 7'5, 3'0 y 5'5, su nota final sería:

$$\frac{1 \cdot 7'5 + 2 \cdot 3'0 + 2 \cdot 5'5}{1 + 2 + 2} = \frac{24'5}{5} = 4'9$$

    Se pide:

    (a) ¿Qué nota tendría que haber sacado en la tercera prueba para aprobar la asignatura?

    (b) ¿Cuál habría sido su nota final si los pesos hubiesen sido 2, 1, y 1?

18. **Otras medias.** Aunque la media aritmética es la más utilizada, existen otras medidas de tendencia central que pueden resultar interesantes para determinados casos. Otro tipo de medias lo constituye un grupo denominado φ-medias que se obtienen aplicando la fórmula

$$\varphi^{-1}\left(\sum_{i=1}^{k} \varphi(x_i) f_i\right)$$

    para alguna función φ que sea continua y monótona en el intervalo de valores posibles de la variable. Las más usuales son la media cuadrática, armónica y geométrica que utilizan la función que se indica:

Media cuadrática	$MQ = \sqrt{\dfrac{x_1^2 n_1 + x_2^2 n_2 + \ldots + x_k^2 n_k}{N}}$	$\varphi(x) = x^2$
Media armónica	$H = \dfrac{N}{\dfrac{n_1}{x_1} + \dfrac{n_2}{x_2} + \ldots + \dfrac{n_k}{x_k}}$	$\varphi(x) = \dfrac{1}{x}$
Media geométrica	$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \ldots x_k^{n_k}}$	$\varphi(x) = \ln(x)$

Entre ellas se establece la siguiente relación:

$$H \leq G \leq \bar{x} \leq MQ$$

Se pide

(a) Calcular, si es posible, el valor de las cuatro medias anteriores para los valores 2, 6 y 10, y analiza los distintos resultados pensando que esos valores corresponden a las notas de los tres exámenes de una asignatura.

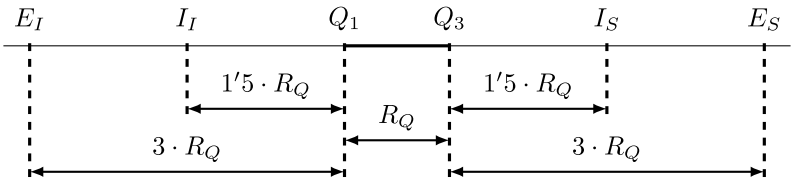
(b) Repetir el apartado anterior con los valores 0, 5 y 10.

(c) Buscar, en la bibliografía, las características de cada una de estas medias y sus aplicaciones.

(d) Definir una nueva φ-media utilizando la función exponencial y alguna función trigonométrica. Observación: Las funciones utilizadas han de ser monótonas en el rango de valores de la variable.

19. **Datos anómalos.** En ocasiones, hay muestras que contienen "observaciones anómalas", es decir, observaciones que están muy alejadas del cuerpo central de los datos. Este tipo de observaciones se pueden atribuir a varias causas: el dato se observa, se registra o se introduce incorrectamente; el dato proviene de una población distinta; el dato es correcto pero representa un suceso poco común, etc. Veamos un método para detectar posibles datos anómalos en una muestra utilizando el rango intercuartílico.

Primero se calculan Q_1 y Q_3 que determinan el rango intercuartílico R_Q. A partir de ellos se obtienen los valores $I_I = Q_1 - 1'5 \cdot R_Q$ e $I_S = Q_3 + 1'5 \cdot R_Q$ denominados cotas interiores inferior y superior. Estas cotas se localizan a una distancia de $1'5 \cdot R_Q$ por debajo de Q_1 en el caso de L_I y por encima de Q_3 en el caso de I_S. Por último, se calculan los valores $E_I = Q_1 - 3 \cdot R_Q$ y $E_S = Q_3 + 3 \cdot R_Q$ denominados cotas exteriores inferior y superior. Estas cotas se localizan a una distancia de $3 \cdot R_Q$ por debajo de Q_1 en el caso de E_I y por encima de Q_3 en el caso de E_S. Todo esto queda representado en la figura **??**.



Ahora, si los datos caen entre las cotas interiores y exteriores se denominan "posibles valores fuera de intervalo". Si los datos caen fuera de las cotas exteriores se denominan "valores fuera del intervalo muy probables".

Detectar los posibles datos anómalos de la siguiente muestra del tiempo (en segundos) de ejecución de 25 trabajos, en un ordenador.

$1'17$	$1'61$	$1'16$	$1'38$	$3'53$	$1'23$	$3'76$	$1'94$	$0'96$	$4'75$	$0'15$	$2'41$	$0'71$	$0'02$	$1'59$
$0'19$	$0'82$	$0'47$	$2'16$	$2'01$	$0'92$	$0'75$	$2'59$	$3'07$	$1'40$					

20. Con el fin de determinar la fecha de construcción de cierta iglesia bizantina se tomaron muestras de las maderas que se encontraban en el interior y exterior de dicha iglesia. Las fechas de las maderas fueron determinadas dando lugar a los siguientes datos:

$$\{1294,\ 1251,\ 1279,\ 1248,\ 1274,\ 1240,\ 1264,\ 1284,\ 1274,\ 1272,\ 1264,\ 1256,\ 1256,\ 1254,$$

$$1250,\ 1242,\ 1232,\ 1263,\ 1220,\ 1254,\ 1218,\ 1251,\ 1210\}$$

Construir el intervalo de valores que distan de la mediana en menos de 1.5 de la diferencia entre el percentil P_{82} y el decil D_3 de los datos. ¿Qué porcentaje de los datos están en ese intervalo? ¿Existe alguna observación anómala?