8. Estadística Descriptiva

Introducción

La Estadística Descriptiva es la rama de las Matemáticas que recolecta, presenta y caracteriza un conjunto de datos (por ejemplo la edad de una población, altura de los estudiantes de una escuela, la temperatura en los meses de verano ...) con el fin de describir apropiadamente las diversas características de ese conjunto.

El problema que surge a la hora de realizar dicho trabajo es cómo interpretar la “avalancha de cifras” de la realidad, para lo cual, es necesario utilizar los procedimientos de cálculo eficaces y los métodos de representación de los datos adecuados.

Actualmente, las técnicas estadísticas se utilizan en todos los campos del conocimiento.

Fases de un estudio estadístico

Un estudio estadístico pasa por las siguientes fases:

- Determinación del objeto de estudio: población, muestra, y variables a estudiar.
- Recogida y organización de los datos.
- Tratamiento y presentación de los datos: medidas de centralización y dispersión, representación gráfica.
- Interpretación y análisis.

Conceptos generales

Población

La población es el conjunto de individuos u objetos de una misma naturaleza, sobre los que se va a efectuar una investigación relativa a una característica variable, con el fin de averiguar sus tendencias principales y cómo se distribuye en la población.

Muestra

Puesto que en general una población es un conjunto muy extenso, se selecciona al azar una muestra, es decir, un subconjunto de la población sobre el que se efectuarán las mediciones estadísticas cuyos resultados serán extrapolados a toda la población.

Para que el estudio estadístico sea correcto, la muestra debe ser cuidadosamente seleccionada. Se dice que la muestra elegida debe ser representativa de toda la población.

Variable estadística

La característica cuya distribución desea estudiarse en una población se llama variable estadística. Existen distintos tipos.

- Variable cualitativa: es una variable que no es cuantificable o medible de forma numérica, o lo que es lo mismo, que no toma valores numéricos.

  Son ejemplos el color de ojos, el deporte favorito de los adolescentes, o el destino preferido de vacaciones de verano de los españoles.

- Variable cuantitativa: es una variable que puede cuantificarse o medirse numéricamente, es decir, toma valores numéricos. Según sean estos, puede clasificarse en discreta o continua.

  - Una variable cuantitativa es discreta cuando entre dos valores determinados no puede tomar cualquier otro valor. Los casos más típicos son aquellas variables que solo pueden tomar valores enteros.

    Son ejemplos de variables cuantitativas discretas el número de hijos, la edad, o el año de nacimiento.

- Una variable cuantitativa es continua cuando entre dos valores determinados, puede tomar cualquier otro valor intermedio. Son variables cuyos valores son números reales.

  Son ejemplos de variables cuantitativas continuas la estatura, la temperatura, o el tiempo de reacción a un estímulo.

Ejercicio 1 *Define los conceptos de población, muestra y variable estadística.*

Ejercicio 2 *Explica cómo se clasifican las variables estadísticas, y pon ejemplos de cada tipo.*

Ejercicio 3 *Clasifica los siguientes aspectos relativos a un camión como variables estadísticas: longitud, carga máxima, color de la cabina, número de ruedas, número de ejes, número de puertas, altura máxima, carburante empleado.*

Ejercicio 4 *Se quiere hacer un estudio sobre la estatura de los alumnos de 3º de ESO del CPI da Cañiza, y se mide a los alumnos de 3º A. Especifica cuáles son la población, la muestra, y la variable estadística del estudio, y clasifica la variable estadística.*

Ejercicio 5 *En una localidad en la que el* $55\,\%$ *de la población tiene menos de cuarenta años, se desea hacer una encuesta para determinar hasta qué punto está normalizada la práctica de comprar ropa por internet. Se entrevista a* 100 *personas, de las cuales:* 5 *tienen menos de veinticinco años,* 20 *entre veinticinco y cuarenta, y el resto, más de cuarenta. ¿Es representativa la muestra?. Justifica la respuesta.*

Recogida y organización de los datos. Tablas estadísticas

Cuando se procede a un estudio estadístico sobre una característica concreta de una población, una vez seleccionada una muestra representativa, deben organizarse los resultados obtenidos.

Ejemplo 1*:*

Supongamos que se desea estudiar el número de hijos que tienen las mujeres entre veinte y cuarenta años de una localidad. Seleccionamos una muestra de 60 *mujeres, y estos son los datos obtenidos:*

$0, 1, 0, 2, 4,$	$1, 1, 3, 2, 2,$	$0, 2, 1, 1, 3$
$1, 1, 1, 2, 1,$	$1, 0, 4, 2, 1,$	$3, 1, 0, 1, 2$
$1, 0, 0, 2, 1,$	$0, 1, 0, 3, 1,$	$0, 3, 1, 4, 3$
$1, 2, 1, 3, 0,$	$1, 0, 3, 0, 2,$	$2, 0, 0, 1, 0$

En este caso la variable de estudio es el número de hijos, de tipo discreto. Lo que se hace es presentar las anteriores respuestas (los valores que toma la variable) en una tabla llamada Tabla de Frecuencias.

Tabla de frecuencias

- En general, a una variable aleatoria se la representa con la letra X. Dicha variable toma distintos valores, cada uno de los cuales es un resultado de una medición o encuesta. En el caso de que pueda tomar n valores distintos, lo que se hace es organizarlos de menor a mayor en la primera columna de la tabla de frecuencias.

  Con X_i representamos el iésimo valor que toma la variable, y con x_i cuánto vale ese valor o respuesta. Es decir, con $X_i = x_i$ representamos que el iésimo valor de la variable X es x_i.

- La frecuencia absoluta de un valor de la variable es el número de veces que se la variable toma dicho valor. Lo denotaremos f_i.

- La frecuencia relativa de un valor de la variable es el cociente entre la frecuencia absoluta, y el número total de observaciones (o tamaño muestral). Lo denotaremos h_i.

- La frecuencia absoluta acumulada de un valor x_i de la variable es el número de respuestas con valor menor o igual a x_i. Lo denotaremos F_i.

- La frecuencia relativa acumulada de un valor x_i es el cociente entre la frecuencia absoluta acumulada y el tamaño muestral. Lo denotaremos H_i.
- En el ejemplo 1:
  - La variable $X=$ "Número de hijos" toma los valores $X_1 = 0$, $X_2 = 1$, $X_3 = 2$, $X_4 = 3$ y $X_5 = 4$.
  - Hay 22 mujeres que contestan que tienen 1 hijo, es decir, que la frecuencia absoluta de $X_2 = 1$ es $f_2 = 11$.
  - Hay 49 mujeres que tienen dos o menos hijos. Es decir, hay 49 valores de la variable X para los cuales la variable toma un valor menor o igual que $X_3 = 2$. Por tanto $F_3 = 49$.
  - Las frecuencias relativas, realmente no son más que tantos por uno.

X_i	f_i	F_i	h_i	H_i
0	16	16	$\frac{16}{60}$	$\frac{16}{60}$
1	22	38	$\frac{22}{60}$	$\frac{38}{60}$
2	11	49	$\frac{11}{60}$	$\frac{49}{60}$
3	8	57	$\frac{8}{60}$	$\frac{57}{60}$
4	3	60	$\frac{3}{60}$	$\frac{60}{60}$

Agrupación de datos

En algunos casos, cuando una variable cuantitativa discreta toma un amplio rango de valores, o cuando la variable es continua, es necesario agrupar los datos en intervalos, para facilitar tanto los cálculos como la representación gráfica.

Ejemplo 2:

Supongamos que tomamos una muestra de 100 *hombres adultos de una población para estudiar cómo se distribuye la estatura.*

En este caso la variable es continua. Es evidente, que lo más fácil sería dividir el rango de los posibles valores en intervalos, por ejemplo de 10 cm de longitud: $[1.50, 1.60)$, $[1.60, 1.70)$, $[1.70, 1.80)$, $[1.80, 1.90)$, $[1.90, 2.00)$,$[2.00, 2.10]$. Luego bastaría registrar el número de observaciones pertenecientes a cada intervalo.

Una vez hecho esto, se asumiría que todas las medidas pertenecientes a un intervalo dado, serían iguales al punto medio de dicho intervalo. Dicha cantidad se llama marca de clase del intervalo, y es el valor con el que se trabaja para calcular los parámetros estadísticos de media, desviación media, desviación típica y varianza.

Por ejemplo, la siguiente podría ser la tabla de frecuencias de la distribución de la estatura:

I_i	X_i	f_i	F_i	h_i	H_i
$[1.50, 1.60)$	1.55	2	2	0.02	0.02
$[1.60, 1.70)$	1.65	10	12	0.10	0.12
$[1.70, 1.80)$	1.75	38	50	0.38	0.50
$[1.80, 1.90)$	1.85	35	85	0.35	0.85
$[1.90, 2.00)$	1.95	10	95	0.10	0.95
$[2.00, 2.10]$	2.05	5	100	0.05	1

Parámetros estadísticos: Medidas de centralización

Las medidas de centralización nos proporcionan un valor de referencia en torno al cual se distribuyen los datos. Son la media, la moda y la mediana.

La media

La media aritmética, $\overline{x}$, es el "centro de gravedad" de la distribución de los datos. Se calcula sumándolos todos y diviendo entre el número de observaciones, n.

Utilizando las columnas de las tablas de frecuencias, el cálculo de la media se reduce mucho:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{n}$$

Basta añadir una columna más a la tabla de frecuencias. En el ejemplo 1 se haría así:

X_i	f_i	F_i	h_i	H_i	$x_i \cdot f_i$
0	16	16	$\frac{16}{60}$	$\frac{16}{60}$	0
1	22	38	$\frac{22}{60}$	$\frac{38}{60}$	22
2	11	49	$\frac{11}{60}$	$\frac{49}{60}$	22
3	8	57	$\frac{8}{60}$	$\frac{57}{60}$	24
4	3	60	$\frac{3}{60}$	$\frac{60}{60}$	12
Total	60	–	1	–	80

En este caso, el número medio de hijos es $\overline{x} = \frac{80}{60} = 1.33$.
Para trabajar con datos agrupados, como en el ejemplo 2, se haría igual, utilizando las marcas de clase.

La moda

La moda, M_o, es el valor x_i de la variable que más se repite. Es decir, el que tiene la frecuencia absoluta (o relativa) más alta.
Es el único parámetro estadístico que se puede calcular en el caso de las variables cualitativas.
En el ejemplo 1 de la distribución del número de hijos, $M_o = 1$.
En el ejemplo 2 de la distribución de la estatura, se hablaría de intervalo modal, y este sería $[1.70, 1.80)$. Determinaremos la moda gráficamente.

La mediana

La mediana, M_e, es el valor que ocupa la posición central en la distribución de los datos, una vez que estos se ordenan de menor a mayor. Es decir, el un valor tal que el 50 % de las observaciones son menores o iguales que él.
El cálculo de la mediana depende del tipo de variable cuantitativa:

- Si la variable es discreta, la mediana será el valor que divida a la muestra en dos partes iguales, o lo que es lo mismo, si la muestra está ordenada, la mediana será el valor que deje el mismo número de elementos a su derecha que a su izquierda. Pero claro, el cálculo dependerá de que el tamaño muestral sea par o impar.
  - Si el tamaño de la muestra n es impar, consultamos la tabla de frecuencias para ver qué valor x_i ocupa la posición $\frac{n+1}{2}$.
  - Si el tamaño de la muestra n es par, realmente no existirá un valor que divida a los datos en dos partes iguales, por tanto, buscamos en la tabla de frecuencias qué valores ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$. La mediana será el promedio de dichos valores.
    Así, el ejemplo 1, dado que hay 60 observaciones, consultamos la tabla de frecuencias para encontrar los valores de la variable que ocupan las posiciones 30 y 31 en la muestra ordenada. Si nos fijamos en la columna de las frecuencias absolutas acumuladas, vemos que para $X_1 = 0$, $F_1 = 16$, y que para $X_2 = 1$, $F_2 = 38$. Por tanto sabemos que los dos valores buscados son 1 y 1. A continuación los promediamos, y obtenemos que $M_e = 1$
- Si la variable es continua (o los datos están agrupados), localizamos el intervalo que contiene a la mediana. De la definición de mediana se deduce que será el primero cuya frecuencia relativa acumulada sea mayor o igual a 0.5.

  El valor de la mediana lo calcularemos gráficamente.

Propiedades de las medidas de centralización

- La media presenta el inconveniente de que se ve influenciada por todos los valores de la muestra, lo que la hace sensible a la presencia de valores extremos y anómalos, con lo cual, su valor puede perder representatividad. Se dice por esto que es una medida poco robusta. Este problema se agrava si la muestra es pequeña.
- La ventaja de la mediana y la moda frente a la media, es que son medidas de centralización robustas.
- La mediana cumple la propiedad de que siempre se encuentra comprendida entre la moda y la media. Dependiendo de la posición relativa de la media, mediana y moda, podemos hablar de distintos de asimetrías en la distribución de los datos.

Ejercicio 6 *Define los conceptos de media, moda y mediana. ¿Para qué sirven las medidas de centralización?*

Ejercicio 7 *Las calificaciones de 36 alumnos en Matemáticas han sido las siguientes:*

$5, 2, 4, 9, 7, 6,$	$4, 5, 6, 5, 7, 3,$	$7, 5, 5, 8, 2, 6,$
$10, 5, 6, 10, 4, 5$	$7, 6, 7, 3, 5, 6,$	$6, 9, 6, 1, 4, 10$

Calcular la moda, la mediana y la media aritmética.

Ejercicio 8 *Calcula la media, la moda y la mediana de los siguientes grupos de datos:*

a) $2, 2, 2, 3, 3, 4, 4, 5, 6, 7$

b) $45, 46, 49, 49, 50$

c) $1, 20, 20, 22, 23$

Parámetros estadísticos: Medidas de posicionamiento

Cuartiles

Los cuartiles, Q_1, Q_2, y Q_3, son los valores tales que el 25 %, 50 % y 75 % respectivamente de los valores de la variable son inferiores a él.

- Obviamente, el segundo cuartil Q_2, coincide con la mediana.
- Los cuartiles son los valores que dividen a la muestra ordenada en cuatro partes iguales.
- En nuestro ejemplo 1, consultando la tabla de frecuencias veríamos que $Q_1 = 0$, $Q_2 = 1$ y $Q_3 = 2$. Por tanto, concluiríamos que el 25 % de las mujeres de la población de estudio no tiene hijos, un 50 % de las mujeres tiene como mucho un hijo, y que un 75 % de la población no tiene más de dos hijos (o que sólo el 25 % tiene dos hijos o más).
- La diferencia $Q_3 - Q_1$ se llama rango o recorrido intercuartílico.

Percentiles

Los percentiles son la generalización de las anteriores medidas de posicionamiento, y se usan para dividir la muestra en 100 partes iguales.

Una vez ordenados los datos de menor a mayor, un percentil indica el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones.

- El percentil i-ésimo, se representa P_i, (donde la i toma valores del 1 al 99), y significa que el $i\,\%$ de las observaciones son menores o iguales que P_i, y el $(100 - i)\,\%$ restante son mayores o iguales que P_i. Suelen usarse en el caso de variables continuas.

  Por ejemplo, el percentil P_{20} es el valor bajo el cual se encuentran el 20 % las observaciones.

  En nuestro caso, dado que tenemos 60 observaciones, el percentil P_{20} es el valor de la variable que cumple que 12 (el 20 % de 60) de las observaciones son menores o iguales que él. Consultando la tabla de frecuencias vemos que $P_{20} = 0$.
- Los percentiles P_{25}, P_{50} y P_{75} corresponden respectivamente a los cuartiles Q_1, Q_2 y Q_3.
- Los percentiles P_{10}, P_{20}, P_{30},...,P_{90}, que dividen a la muestra en 10 partes iguales, reciben el nombre de deciles.

Ejercicio 9 *Completa las tablas de frecuencias de las tablas de la página 6 , y calcula los cuartiles.*

X_i	f_i
1	4
2	6
3	8
4	5
5	4

Figura 1: Ejercicio 9 a)

X_i	f_i
12	8
15	14
18	13
21	5
24	2

Figura 2: Ejercicio 9 b)

X_i	f_i
5	8
15	12
25	13
30	5
35	2

Figura 3: Ejercicio 9 c)

Parámetros estadísticos: Medidas de dispersión

Las medidas de dispersión se utilizan para cuantificar la representatividad de las medidas de centralización (especialmente la representatividad de la media).

Rango o recorrido

Es la diferencia entre los valores máximo y mínimo observados.

Desviación media

La desviación media es un concepto similar al de error relativo medio, es decir, dada una muestra, tomando como valor verdadero la media aritmética de las observaciones, $\overline{x}$, lo que hacemos es promediar el error absoluto de cada una de las observaciones:

$$DM = \frac{\sum_{i=1}^{n} |x_i - \overline{x}| \cdot f_i}{n}$$

Varianza

La fórmula para el cálculo de la desviación media presenta inconvenientes de manejo en el caso de diseño de estrategias matemáticas destinadas a la reducción de errores. Por eso se utiliza el concepto de varianza o error cuadrático medio, s^2, que consiste en sustituir el cálculo de los errores absolutos, por el cuadrado de las distancias de las observaciones al valor tomado como verdadero:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 f_i}{n}$$

No es difícil demostrar que la anterior fórmula conduce al mismo resultado que la expresión:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 f_i}{n} - \overline{x}^2$$

En general, es esta última la que se utiliza para el cálculo de la varianza, ya que añadiendo dos columnas más a la tabla de frecuencias, agiliza mucho los cálculos.

En nuestro ejemplo 1, podríamos presentar la tabla como sigue (prescindimos de las columnas de las frecuencias relativas, por no necesitarlas en este ejercicio):

X_i	f_i	F_i	$x_i \cdot f_i$	x_i^2	$x_i^2 f_i$
0	16	16	0	0	0
1	22	38	22	1	1
2	11	49	22	4	44
3	8	57	24	9	72
4	3	60	12	16	48
Total	60	–	80	–	165

Por tanto: $s^2 = \dfrac{165}{60} - 1.33^2 = 0.98$

La varianza tiene las siguientes propiedades:

- Dado que una varianza es una suma de cuadrados, la varianza siempre será un valor positivo.
- Al igual que la media, el valor de la varianza es muy sensible a la presencia de valores extremos.
- Cuanto más pequeño sea el valor de la varianza, mayor será la proximidad de los datos en torno a la media. Es decir, más representativa será la media.

Desviación típica

Es la raíz cuadrada de la varianza. Es decir:

$$s = \sqrt{\frac{\sum\limits_{i=1}^{n} x_i^2 f_i}{n} - \overline{x}^2}$$

Tiene las mismas propiedades que la varianza.

Coeficiente de variación

El coeficiente de variación, CV, se utiliza para cuantificar la importancia de la dispersión en una muestra. Es el cociente entre la desviación típica y la media.

$$CV = \frac{s}{|\overline{x}|}$$

Con respecto al coeficiente de variación, debemos hacer las siguientes observaciones:

- Su cálculo no tiene sentido si la media aritmética es cero o un valor muy cercano a cero.
- No tiene unidades, o lo que es lo mismo, es independiente de las unidades de medida en las que se expresaron los datos.
- También es frecuente calcularlo como $CV = 100 \cdot \dfrac{s}{|\overline{x}|}$, y expresarlo de modo porcentual.
- Se utiliza para comparar las dispersiones de dos distribuciones distintas.

Ejercicio 10 *¿Para qué sirven las medidas de dispersión?. Cita tres medidas de dispersión*

Ejercicio 11 *Define el concepto de coeficiente de variación de Pearson, y explica para qué se utiliza.*

Ejercicio 12 *Se tienen las siguientes distribuciones de notas de inglés:*

- *En 3ºA:* $2, 3, 3, 4, 5, 5, 5, 5, 5, 5, 6, 7$
- *En 3ºB:* $2, 2, 2, 3, 3, 4, 7, 7, 7, 7, 7, 8$

Calcula la media, la desviación típica y el coeficiente de variación, y utiliza los resultados para comparar las dos distribuciones.

Representación gráfica de los datos

Variables cualitativas

- **Diagrama de sectores circulares**

  Consiste en representar cada uno de los valores de la variable como un sector circular, cuyo ángulo central será directamente proporcional a la frecuencia del valor.

  Consideremos el siguiente ejemplo: *Supongamos que en una clase anotamos el color de ojos de cada alumno, obteniendo que* 10 *de ellos tienen los ojos marrones,* 6 *azules, y* 4 *verdes.*

  Podemos recopilar estos datos en una tabla como la de la figura 4 (página 8).

Color de ojos	azul	verde	negro
Nº de alumnos	6	4	10

Figura 4: Distribución del color de ojos

  Obtendríamos el diagrama de la figura 5 (página 8).

- **Diagrama de barras**

  En un eje horizontal se sitúan los distintos valores de la variable estadística. A cada uno de esos valores le corresponderá un rectángulo de altura proporcional a la frecuencia.

  En el caso de la tabla 4 del ejemplo de la distribución del color de ojos, obtendríamos un diagrama como el de la figura 6 (página 8).

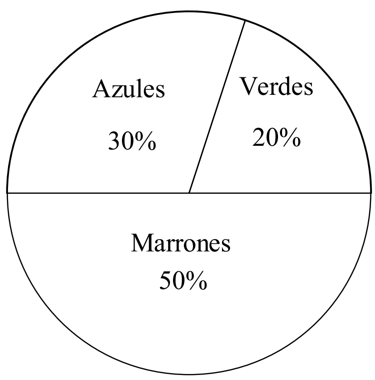

Figura 5: Diagrama de sectores

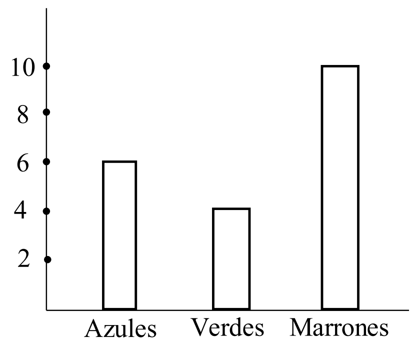

Figura 6: Diagrama de barras (de frecuencias absolutas)

Variables cuantitativas discretas

- **Diagrama de sectores circulares**

  Se efectúa de la misma forma que en el caso de las variables cualitativas, pero debe valorarse la conveniencia de usarlo o no en función de la cantidad de valores distintos que tome la variable. Si estos fuesen demasiados, habría que optar por agruparlos, o por otro tipo de representación gráfica.

- **Diagrama de barras de frecuencias**

  Se hacen igual que en el caso de las variables cualitativas, pero en este caso, las barras no son rectángulos, sino segmentos.

  Para el ejemplo 1 de la distribución del número de hijos, el gráfico de la figura 7 (página 9) sería el diagrama de barras de frecuencias absolutas.

- **Polígonos de frecuencias**

  Se construyen a partir de los diagramas de barras, uniendo los extremos superiores de los segmentos.

  En la figura 8 (página 9) se muestra el polígono de frecuencias correspondiente al anterior ejemplo 1 de la distribución del número de hermanos en un grupo de alumnos.

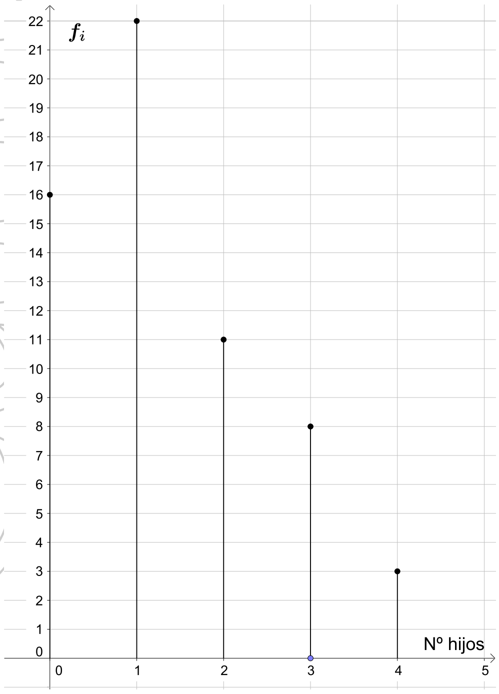

Figura 7: Diagrama de barras

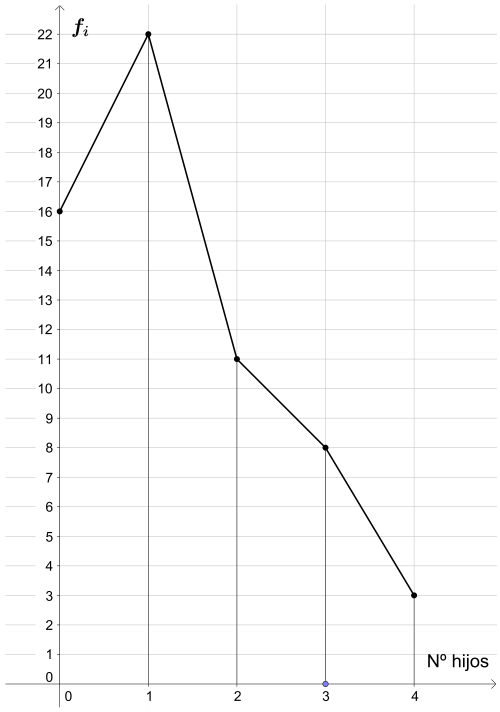

Figura 8: Polígono de frecuencias

- **Polígono de frecuencias acumuladas** Para la construcción del polígono de frecuencias acumuladas debe tenerse en cuenta el concepto de frecuencia acumulada (puede hacerse con las frecuencias absolutas, o con las relativas).

  En el caso de la distribución del ejemplo 1, el gráfico de la figura 9 (página 9) sería el polígono de frecuencias relativas acumuladas.

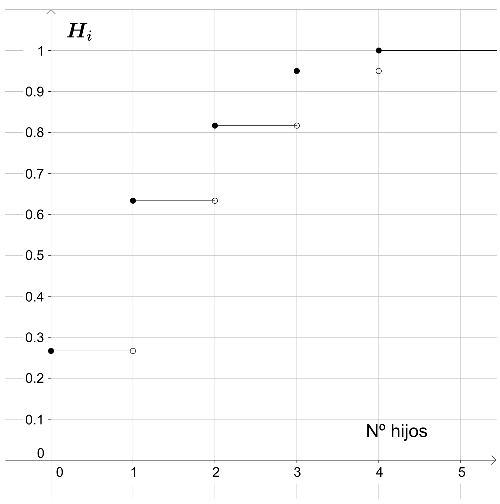

Figura 9: Polígono de frecuencias

CPI da Cañiza

- **Diagrama de cajas y bigotes**

Una gráfica de este tipo consiste en una caja rectangular, donde los lados más largos muestran el recorrido intercuartílico. Este rectángulo está dividido por un segmento vertical que indica dónde se posiciona la mediana y por lo tanto su relación con los cuartiles primero y tercero (recordemos que el segundo cuartil coincide con la mediana).

Esta caja se ubica a escala sobre un segmento que tiene como extremos los valores mínimo y máximo de la variable.

Las lineas que sobresalen de la caja se llaman bigotes, que tienen tienen un límite de prolongación de 1.5 veces el rango intercuartílico, de modo que cualquier dato o caso que no se encuentre dentro de este rango es marcado e identificado individualmente.

Ejemplo 3*:*

Supongamos que queremos estudiar cómo se distribuye la edad entre los 20 *trabajadores de una empresa. Realizamos una encuesta, y nos encontramos con que estas son las edades:*

$$20, 23, 24, 24, 24, 25, 29, 31, 31, 33, 34, 36, 36, 37, 39, 39, 40, 40, 41, 45$$

- El valor mínimo de la muestra es 20, y el máximo 45.
- Para calcular los cuartiles, debemos dividir la muestra en cuatro partes iguales, cada una de las cuales tendrá 5 observaciones:
  I) $Q_1 = \dfrac{x_5 + x_6}{2} = \dfrac{24+25}{2} = 24.5$
  II) $Q_2 = \dfrac{x_{10} + x_{11}}{2} = \dfrac{33+34}{2} = 33.5$
  III) $Q_3 = \dfrac{x_{15} + x_{16}}{2} = \dfrac{39+39}{2} = 39$
  IV) El rango intercuartílico vale $Q_3 - Q_1 = 14.5$, por tanto, la longitud máxima de los bigotes es $14.5 \cdot 1.5 = 21.75$. Es decir, en este caso, no tenemos valores anómalos que deban ser marcados individualmente.
- Dibujando a escala, obtenemos el diagrama de la figura 10 (página 10)

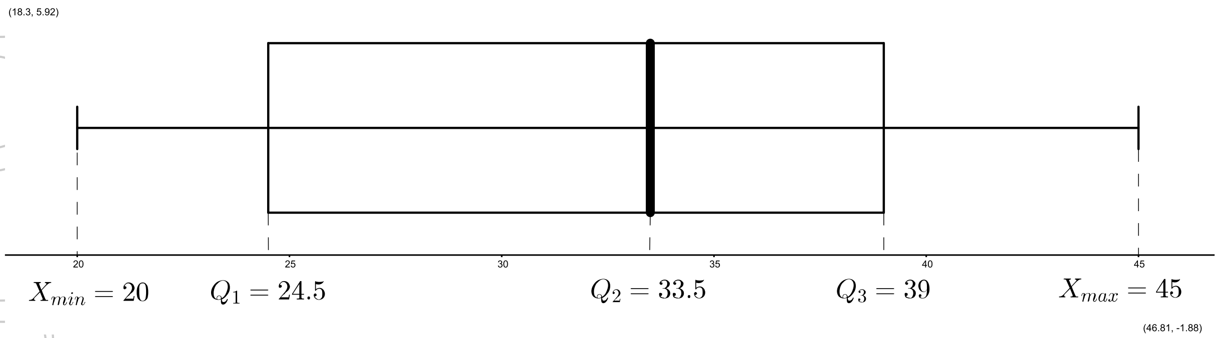

Figura 10: Diagrama de cajas y bigotes

- Analizando el diagrama vemos que:
  - La parte izquierda de la caja es mayor que la de la derecha; ello quiere decir que las edades comprendidas entre el 25 % y el 50 % de la población está más dispersa que entre el 50 % y el 75 %.
  - El bigote de la izquierda (X_{min}, Q_1) es más corto que el de la derecha (Q_3, X_{max}); por ello el 25 % de los más jóvenes están más concentrados que el 25 % de los más mayores.
  - El rango intercuartílico vale $Q_3 - Q_1 = 14.5$; es decir, el 50 % de la población está comprendido en un rango de edades de 14.5 años de amplitud.

Ejercicio 13 *Para las dos distribuciones del ejercicio 12, construye:*

a) Los diagrama de barras.

b) Los diagramas de cajas y bigotes.

c) Comenta los diagramas de bigotes.

Variables cuantitativas continuas

- **Histogramas y polígonos de frecuencias**

  El histograma es el equivalente a los diagramas de barras de las variables discretas, y también pueden ser de frecuencias absolutas o relativas y acumuladas.

  La diferencia con otras representaciones gráficas similares, es que es el área de los rectángulos lo que debe ser proporcional a la frecuencia representada. Con esto hay que tener especial cuidado si los intervalos en que se dividió la muestra no tienen la misma longitud.

  Vamos a presentar los histogramas de frecuencias relativas, y frecuencias relativas acumuladas para el ejemplo 2, de la distribución de las estaturas, así como los polígonos de frecuencias:

  - En la figura 11 (página 11) vemos el histograma de frecuencias relativas ordinarias, y cómo obtener gráficamente el valor de la moda de la distribución. Utilizando herramientas de semejanza de triángulos, podemos obtener el valor numérico de dicha moda.
  - En la figura 12 (página 12) vemos como obtener el polígono de frecuencias a partir del histograma de frecuencias.

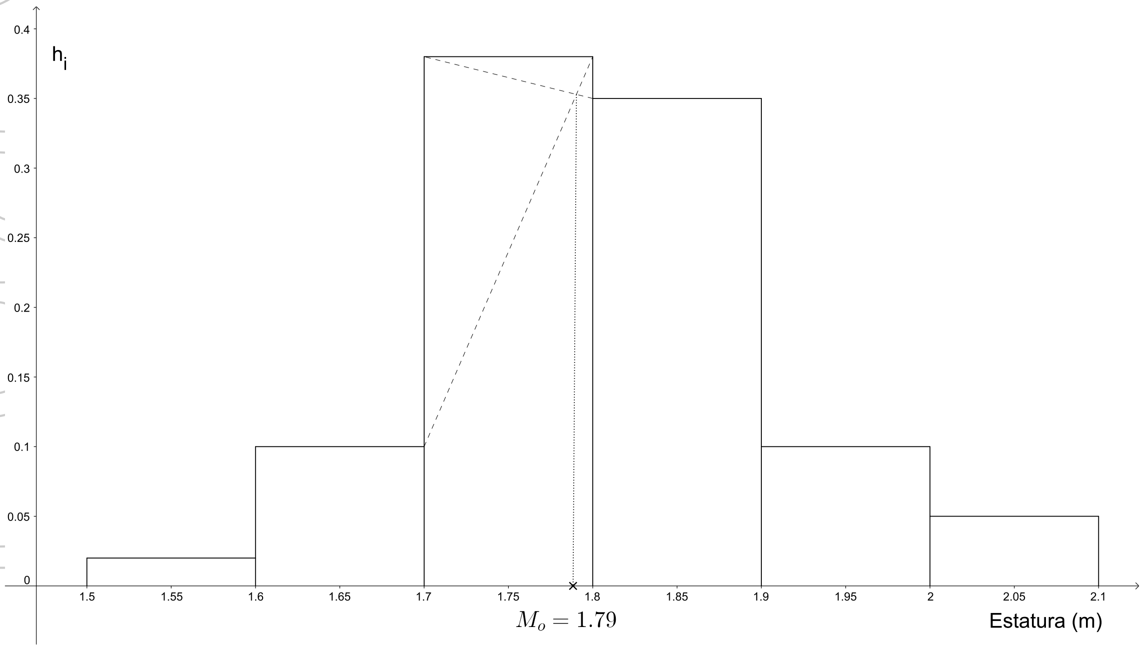

Figura 11: Cálculo de la moda con datos agrupados

  - En la figura 13 (página 12) vemos el histograma de frecuencias relativas acumuladas, y también el polígono de frecuencias relativas acumuladas. Además, vemos cómo obtener gráficamente los cuartiles, cuyos valores numéricos pueden determinarse utilizando herramientas de semejanza de triángulos.

- **Diagramas de cajas y bigotes**

  Se hacen igual que en el caso discreto.

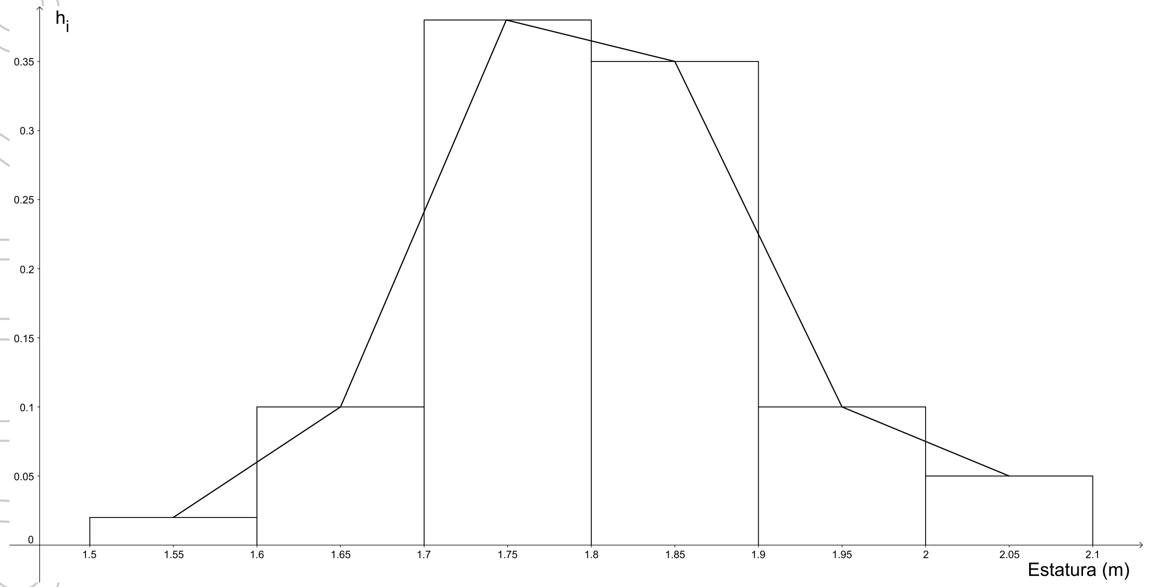

Figura 12: Histograma y polígono de frecuencias

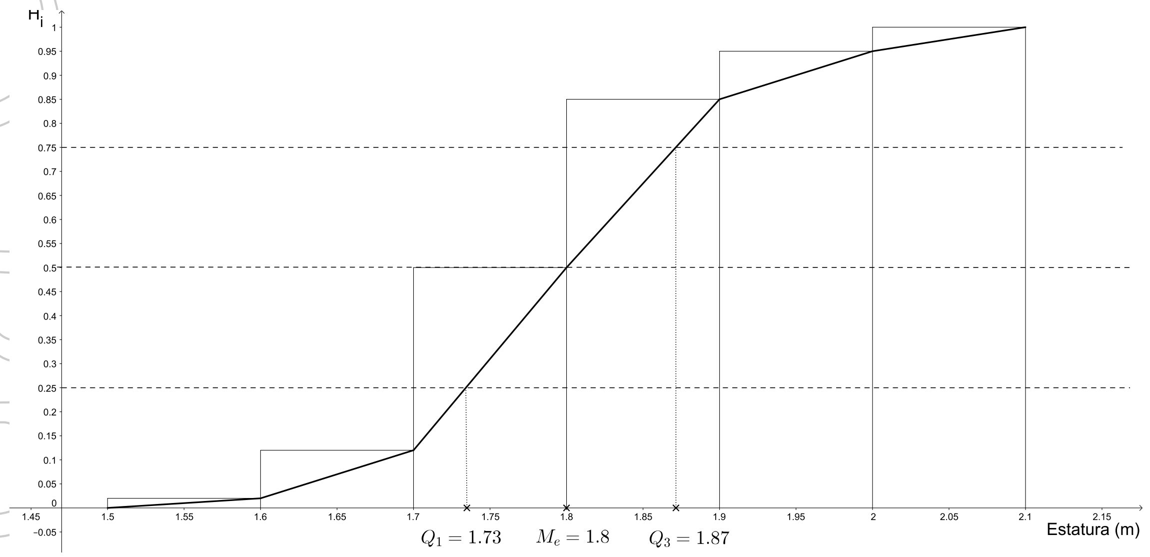

Figura 13: Cálculo de cuartiles con datos agrupados

Ejercicios de variables cualitativas

Ejercicio 14 *Un niño se dedica a anotar cuántos coches de cada color ve cuando vuelve del colegio a su casa, obteniendo los siguientes resultados:* 12 *blancos,* 25 *rojos,* 10 *negros,* 20 *grises metalizado,* 5 *azules.*

Representa adecuadamente los datos en una tabla, y en un diagrama de sectores circulares.

Ejercicio 15 *Se quiere estudiar cómo se distribuye el color de ojos en un instituto de secundaria. Se registra el color de ojos de los alumnos de 2ºA de ESO y se obtienen las siguientes respuestas:* 5 *alumnos con ojos azules,* 2 *alumnos con ojos negros,* 4 *alumnos con ojos verdes,* 9 *alumnos con ojos castaños*

a) Determina: la población del estudio, la muestra y el tamaño muestral.

b) ¿Cuál es la variable aleatoria que se estudia?

c) Construye la tabla de frecuencias.

CPI da Cañiza

d) *Dibuja el diagrama de barras y el diagrama de sectores circulares.*

e) *¿Cuál es la moda de la distribución?*

Ejercicio 16 *Se quiere estudiar cómo prefieren emplear los jóvenes españoles de entre* 14 *y* 20 *años su tiempo libre. Para ello se realiza una encuesta en el instituto de secundaria IES Estudio Mucho, obteniéndose los siguientes resultados:* 40 *alumnos prefieren ver la tele,* 10 *alumnos prefieren leer un libro,* 150 *alumnos prefieren estar con sus amigos,* 75 *prefieren contectarse a internet,* 25 *prefieren jugar a la consola.*

a) *Determina la población del estudio, la muestra y el tamaño muestral.*

b) *¿Cuál es la variable aleatoria que se estudia?*

c) *Construye la tabla de frecuencias.*

d) *Dibuja el diagrama de barras y el diagrama de sectores circulares.*

e) *¿Cuál es la moda de la distribuión?*

Ejercicio 17 *Se quiere estimar el resultado electoral de unas elecciones en una comunidad autónoma con una población de* 3 *millones de habitantes. Se realiza una encuesta sobre intención de voto a* 10000 *personas, de la que se obtiene la siguiente tabla:*

Partido	*A*	*B*	*C*
Nº votos	2500	6000	1500

a) *¿Cuál es el tamaño muestral?*

b) *¿Cuál es la variable aleatoria que se estudia?*

c) *Construye la tabla de frecuencias.*

d) *Dibuja el diagrama de barras y el diagrama de sectores circulares.*

e) *¿Cuál es la moda de la distribución?*

Ejercicios de variables cuantitativas

Ejercicio 18 *Estas son las notas de la primera evaluación de inglés de tres alumnos:*

Alumno A	4	6	5	5
Alumno B	1	5	9	5
Alumno C	4.5	5.5	4.5	4.5

a) *Calcula la nota media de cada uno.*

b) *Justifica con el parámetro estadístico adecuado, en qué caso la media es más representativa, y en cuál menos.*

Ejercicio 19 *Calcula la moda y la mediana en cada caso.*

a) $50, 60, 60$

b) $12, 12, 22, 32$

c) $10, 20, 30, 40, 20$

d) $35, 25, 35, 25, 25, 25$

Ejercicio 20 *¿Cuál o cuáles de los datos siguientes se pueden considerar una observación atípica en cada una de las dos series?. Observa qué le ocurre a la media de los datos según excluyas o incluyas dichos datos.*

a) $4, 5, 6, 5, 7, 8, 4, 5, 7, 5, 12, 6, 5$

b) $8, 1, 9, 9, 8$

Ejercicio 21 *Se dispone de los siguientes datos relativos a las estaturas (en cm) de los trabajadores de una empresa:*

174, 158, 150, 185, 186, 178, 166, 185, 199, 183, 175, 173, 175, 164, 176,
159, 190, 173, 189, 163, 169, 198, 182, 203, 195, 183, 174, 177, 202, 193

a) *Organízalos en intervalos de* 10 *cm, desde* 150 *a* 210*, y completa la tabla como necesites para calcular eficientemente la altura media y la desviación típica.*

b) *Construye el histograma de frecuencias, y obtén la moda.*

c) *Construye el polígono de frecuenias acumuladas, y obtén los cuartiles.*

d) *Construye el diagrama de cajas y bigotes.*

Ejercicio 22 *Se ha realizado una encuesta en* 30 *hogares, en la que se les pregunta el número de individuos que conviven en el domicilio habitualmente. Las respuestas obtenidas han sido las siguientes:*

4, 4, 1, 3, 5, 3, 2, 4, 1, 6, 2, 3, 4, 5, 5, 6, 2, 3, 3, 2, 2, 1, 8, 3, 5, 3, 4, 7, 2, 3

a) *¿De qué tipo es la variable de estudio?*

b) *Construye la tabla de frecuencias de la variable obteniendo las frecuencias absolutas, relativas, y sus correspondientes acumuladas.*

c) *Construye el diagrama de barras de frecuencias absolutas, y el de frecuencias acumuladas.*

d) *Calcula la media, la moda, y la mediana.*

e) *Calcula los cuartiles.*

f) *Calcula la desviación típica.*

g) *¿En qué porcentaje de hogares viven menos de cinco personas?*

Ejercicio 23 *Una entidad bancaria dispone de* 50 *sucursales en el territorio nacional, y ha observado el número de empleados que hay en cada una de ellas para un estudio posterior. Las observaciones obtenidas han sido:*

12, 8, 39, 41, 51, 16, 39, 15, 31, 42, 43, 24, 15, 31, 33, 52, 46, 17, 17, 16 46, 35, 34, 32, 21
31, 25, 26, 59, 42, 15, 13, 34, 46, 45, 38, 49, 18, 40, 31, 42, 32, 27, 23, 45, 37, 41, 23, 30, 19

a) *¿De qué tipo es la variable de estudio?*

b) *Agrupa los datos en intervalos de amplitud* 10*, y construye la tabla de frecuencias.*

c) *Representa el histograma y el polígono de frecuencias acumuladas.*

d) *Calcula la media, la moda y la mediana.*

e) *Utilizando el parámetro estadístico adecuado, comenta la representatividad de la media.*

f) *Calcula los cuartiles y representa el diagrama de cajas y bigotes. Coméntalo.*

Ejercicio 24 *Se realiza un estudio en una ciudad sobre la capacidad hotelera, y se obtienen los siguientes resultados:*

Plazas	*Nº hoteles*
Menos de 20	25
Entre 20 *y* 40	50
Entre 40 *y* 60	55
Entre 60 *y* 80	20

a) *Calcula la media de la variable.*

b) Representa el histograma de frecuencias y el de frecuencias relativas acumuladas.

c) Calcula la moda y la mediana.

d) ¿Qué porcentaje de hoteles tiene menos de 40 *plazas?*

e) Calcula la desviación típica de la variable.

Ejercicio 25 *Se ha realizado un estudio entre* 100 *mujeres mayores de* 25 *años, observándose el número de hijos de las mismas. El resultado ha sido recogido en la siguiente tabla:*

N^o hijos	0	1	2	3	4	5	6
N^o mujeres	13	20	25	20	11	7	4

a) ¿Cuál es la variable estudiada?. ¿De qué tipo es?

b) Construye la tabla de frecuencias.

c) Calcula las medidas de centralización.

d) Calcula la desviación típica y comentar con ayuda del parámetro estadístico adecuado, la representatividad de la media.

e) Representa el diagrama de barras.

f) Representa el diagrama de cajas y bigotes, y coméntalo.

Ejercicio 26 *La siguiente tabla expresa el número de coches vendidos durante una semana, por cada uno de los* 50 *concesionarios que una determinada firma tiene en España*

N^o coches vendidos	1	3	4	6	10
N^o concesionarios	5	12	20	8	5

a) ¿Cuál es la variable estudiada y de qué tipo es?

b) Construye la tabla de frecuencias.

c) Calcula las medidas de centralización. ¿Crees que la media es representativa?. Justifica la respuesta.

d) Calcula la desviación típica. ¿Presenta la variable una dispersión alta?

Ejercicio 27 *La distribución del importe de las facturas por reparación de carrocería, de una muestra de* 80 *vehículos en un taller, viene dada por la tabla siguiente:*

Importe (euros)	N^o facturas
$[0, 60)$	10
$[60, 120)$	20
$[120, 180)$	40
$[180, 240]$	10

a) ¿Cuál es la variable de estudio y de qué tipo es?

b) Construye la tabla de frecuencias absolutas y relativas, y acumuladas.

c) Calcula la media.

d) Representa los histogramas de frecuencias relativas y frecuencias relativas acumuladas.

e) Calcula la mediana y la moda.

f) Calcula los cuartiles.

g) Calcula la desviación típica.

Ejercicio 28 *Las estaturas en centímetros de 32 jóvenes son:*

155, 178, 170, 165,	173, 168, 160, 166,	176, 169, 158, 170,	179, 161, 164, 156
170, 171, 167, 151,	163, 158, 164, 174,	176, 164, 154, 157,	155, 162, 179, 182

a) *Agrupa los datos en tres intervalos de amplitud* 10*, empezando en* 150*, y construye la tabla de frecuencias absolutas, y frecuencias absolutas acumuladas.*

b) *Representa el histograma de frecuencias absolutas, y el de frecuencias absolutas acumuladas.*

c) *Calcula las medidas de centralización y la desviación típica.*

d) *Utilizando el parámetro estadístico adecuado, comenta la representatividad de la media.*

e) *Calcula los cuartiles y utilízalos para construir el diagrama de cajas y bigotes. Coméntalo.*

Ejercicio 29 *El volumen de exportaciones de una empresa tiene una media mensual de* 650000 *dólares, con desviación típica de* 92500 *dólares. La misma empresa vende mensualmente, en el mercado interior, un promedio de* 50 *millones de dólares, con desviación típica de* 4.3 *millones. ¿Qué mercado es más estable, el mercado interior o el exterior? Justifica la respuesta.*

Ejercicio 30 *Los gastos mensuales de una empresa A tienen una media de* 100000 *euros y una desviación típica de* 12500 *euros. En otra empresa más pequeña B, la media es* 15000 *euros y la desviación típica* 2500 *euros. Calcula mediante el coeficiente de variación, cuál de las dos tiene más variación relativa.*

Ejercicio 31 *Los salarios por hora de los obreros de dos empresas A y B, son los que se dan en la siguiente tabla:*

Salarios(euros)	*Empresa A*	*Empresa B*
$[550 - 750)$	10	7
$[750 - 1050)$	32	20
$[1050 - 1550)$	57	37
$[1550 - 2550]$	54	78

a) *Calcula el salario medio y el más frecuente en cada empresa.*

b) *¿Cuál es el salario que no es superado por el* 50 % *en cada empresa?*

c) *¿Cuál de las dos empresas tiene mayor homogeneidad salarial? Justifica la respuesta.*

d) *Construye el histograma de frecuencias acumuladas para cada empresa, y calcula los cuartiles para cada empresa.*

e) *Representa en un mismo dibujo los diagramas de cajas y bigotes para cada empresa, y comenta las diferencias que observas.*

Ejercicio 32 *En la siguiente tabla se tienen los puntos totales conseguidos por cada uno de los jugadores de dos equipos de baloncesto en la pasada liga:*

EQUIPO A	315	355	420	392	457	480	387	340
EQUIPO B	444	432	416	388	368	367	352	360

a) *Calcula la media y la desviación típica de cada equipo.*

b) *¿En qué equipo la media es más representativa? Justifica la respuesta.*