ÍNDICE

0. INTRODUCCIÓN

La **Estadística** es la rama de las Matemáticas que **obtiene, organiza, presenta y describe un conjunto de datos** con el propósito de facilitar el uso, generalmente con el apoyo de tablas, medidas numéricas o gráficas. Estas técnicas son utilizadas en el proceso de investigación, en la etapa donde el investigador necesita procesar y analizar los datos recolectados en dicho estudio.

Estadística significa **ciencia del Estado**, y proviene del término alemán **Statistik**. ¿Por qué la ciencia del Estado? Porque en sus orígenes la estadística se utilizaba exclusivamente con fines estatales, en el sentido de que los gobiernos de las distintas naciones tenían (y tienen) la **necesidad**, por razones de organización, **de conocer las características de su población** para gestionar el pago de impuestos, el reclutamiento de soldados, el reparto de tierras o bienes, la prestación de servicios públicos etc. Esta necesidad llevó a los gobernantes a establecer sistemas para recoger y procesar de alguna manera la información obtenida, es decir, a hacer estadísticas sobre la población.

0.1. ¿Cómo surge la Estadística?

Normalmente **los primeros estudios estadísticos** que se hacían eran los **censos**, que son estudios descriptivos sobre todos los integrantes de una población. La elaboración de censos comenzó en la Edad Antigua, y sigue dándose en nuestros días.

Con el tiempo y el desarrollo científico surgieron alternativas a los censos: **las encuestas a sólo đuna parte de la población** y la posterior generalización a toda la población de los resultados obtenidos para la muestra, pero para ello fue necesario el desarrollo de la **Teoría de la Probabilidad** (rama de las Matemáticas), de la **Inferencia Estadística y del Muestreo** (ramas de la Estadística) que se dio en la Edad Moderna y Contemporánea.

Más...

0.2. Breve historia de la Estadística

En **Egipto** la actividad estadística comenzó con la Dinastía I, en el **año 3050 a.C**. El historiador griego Herodoto indica que los faraones ordenaban realizar censos de riqueza y población para planificar la construcción de las pirámides. El faraón de la Dinastía XIX Ramses II (1279 – 1213 a.C.) mandó elaborar un censo para establecer un nuevo reparto de tierras.

En **China**, en el año **2238 a.C.** el emperador Yao manda elaborar un censo general que recogió datos sobre la actividad agrícola, industrial y comercial.

En la **antigua Grecia** también se realizaron censos para cuantificar la distribución y posesión de la tierra y otras riquezas, organizar el servicio militar y determinar el derecho a voto de los ciudadanos.

Los censos y la actividad estadística tuvieron especial importancia en la **antigua Roma**. Durante el Imperio Romano se establecieron registros de nacimientos y defunciones, y se elaboraron estudios sobre los ciudadanos del Imperio, sus tierras y riquezas. El rey romano Servio Tulio (578 – 535 a.C.) elaboró un **catastro de todos los dominios de Roma**.

Más...

0.3. Ejemplos de aplicaciones de la estadística

Hoy en día, en la práctica totalidad de los países se crean oficinas de estadística y otros órganos similares que se encargan de elaborar las estadísticas oficiales del país, por ejemplo estadísticas sobre la **tasa de paro, índices de precios, actividad económica** (producto interior bruto, actividad industrial), estadísticas sobre **sanidad y educación, turismo, población** etc.

La oficina de estadística de España es el INE (**Instituto Nacional de Estadística**). En 1988 se crea el **IGE** (Instituto Galego de Estatística).

El **Sistema Estadístico Europeo** (SEE) está formado por **Eurostat** (la oficina de estadística de la UE), las oficinas de estadística de todos los estados miembros (los diferentes INE) y otros organismos que elaboran estadísticas europeas.

El SEE **garantiza que las estadísticas europeas** elaboradas en todos los Estados miembros de la Unión Europea sean **fiables**, siguiendo unos criterios y **definiciones comunes** y tratando los datos de la manera adecuada para que sean siempre **comparables** entre los distintos países de la UE.

Ejercicio 1: Entra en las páginas web del INE y del Eurostat para responder razonadamente:

a) ¿Cuál es el dato de paro de España del último mes? ¿Y la media de la UE? ¿Y en Galicia?

b) Busca en el apartado de paro de Eurostat los datos de paro de todos los países. ¿En qué formatos se puede presentar esta información?

c) ¿Qué país de la UE tiene la mayor tasa de desempleo? ¿Y el que menos? ¿En cuál de los formatos resulta más fácil encontrar estos datos?

Ejercicio 2: ***Todo era mejor en el pasado, ¿verdadero o falso?***

En la página web del INE encontrarás un enlace a los vídeos ganadores del concurso **European Statistics Competition 2024** en los que intentan responder esta pregunta.
Busca 4 datos similares a los que aparecen en los vídeos y da tu propia respuesta a esta pregunta.

Ejercicio 3: Entra en el **portal educativo del IGE** en el apartado "**Erros estatísticos**". Intenta encontrar los errores en las noticias en base a los datos estadísticos oficiales.

Envía un pantallazo de tu respuesta y de la página dónde has encontrado los datos al AV.

La estadística descriptiva es el primer paso para el **análisis de datos**. Enormes volúmenes de datos, de diferentes formatos y fuentes son recogidos por nuestros dispositivos móviles y ordenadores, sensores,… Es lo que se conoce como **Big Data**.

En este primer paso se estudia la calidad de los datos y si son completos, su estructura, el tipo de datos,… La rama del conocimiento que se encarga del tratamiento de los datos es el **Data Science** o Ciencia de los Datos.

Un nuevo avance en la interpretación y utilización de grandes conjuntos de datos es la **Inteligencia Artificial**.

Estas nuevas ramas del conocimiento se estudian actualmente como Grados universitarios independientes.

Otro concepto empresarial importante es **Data-Driven Company**. Una empresa Data-Driven es aquella que basa su estrategia basada en el análisis y la interpretación de datos. Amazon, Netflix, Starbucks, Uber, Tesla, Inditex, ...son compañías que han basado su éxito en los datos.

1. VARIABLES ESTADÍSTICAS UNIDIMENSIONALES

Un **estudio estadístico** consiste en recoger y analizar datos para extraer conclusiones utilizando la estadística. Para ello es necesario determinar:

La **población**, esto es, el conjunto formado por todos los individuos a los que va dirigido el estudio. Cuando por algún motivo no se puede recoger la información del total de la población (si es demasiado grande, recoger todos los datos resulta costoso,…) se elige una **muestra** que es una parte de esta.

La **variable estadística** es la característica de los individuos que se quiere analizar.

Las variables estadísticas pueden ser:

Cuantitativas: si representan un carácter medible y sus valores son numéricos.

Estas a su vez se clasifican en:

- **Discretas:** solo pueden tomar un número finito o infinito numerable de valores. Es el caso de la edad, del número de hijos o número de visitas diarias a una página web.
- **Continuas:** pueden tomar cualquier valor dentro de un intervalo dado. Por ejemplo, la temperatura, la estatura de los alumnos del IES Saturnino Montojo o el tiempo que estamos conectados a una red social.

Cualitativas: si representan una cualidad no medible numéricamente. Por ejemplo, el color de tu coche, lugar de nacimiento o marca del teléfono móvil.

1.1. Tabla de frecuencias

Los **datos recogidos** en un estudio estadístico se **organizan** habitualmente en una **tabla o distribución de frecuencias**. Si el número de datos recogidos es muy elevado se utilizan medios informáticos: hojas de cálculo o bases de datos.

Si la variable es **cualitativa** recogeremos los datos con el número de veces que aparece en la muestra (o el porcentaje).

Ejemplo 1: Alumnos matriculados en enseñanza no obligatoria en centros sostenidos con fondos públicos en Ferrol en 2024. (Fuente: Consellería de Educación)

Nivel educativo	N.º alumnos	%
Educación infantil	1110	16,3%
Educación primaria	3009	44,2%
Educación secundaria obligatoria	2692	39,5%
	N = 6811	100%

Supongamos una **variable cuantitativa discreta** de la que se han tomado N valores distintos x_1, x_2, …, x_N. En la tabla de frecuencias se incluye:

- **frecuencia absoluta, f_i,** número de veces que aparece en la muestra.
- **frecuencia absoluta acumulada, F_i,** suma de las frecuencias absolutas de los datos menores o iguales que uno determinado: $\mathbf{F_k = f_1 + \ldots + f_k = \sum_{i=1}^{k} f_i}$.
- **frecuencia relativa, hi,** cociente entre la frecuencia absoluta de un dato y el número total, N, de datos: $\mathbf{h_i = f_i / N}$.

  Frecuentemente nos interesa la frecuencia relativa en porcentaje: $\mathbf{h_i\,\% = f_i / N \times 100}$.
- **frecuencia relativa acumulada, Hi,** suma de las frecuencias relativas de los datos menores o iguales que uno determinado: $\mathbf{H_k = h_1 + \ldots + h_k = \sum_{i=1}^{k} h_i}$.

Ejemplo 2: Fuente: INE "Vehículos de los hogares"

N.º de coches por hogar en Ferrol	f_i	F_i	h_i	H_i
0	5170	5170	0,20	0,20
1	12095	17265	0,46	0,66
2	7584	24849	0,29	0,95
3 o más	1617	26466	0,06	1,01*
	N = 26466		1,01*	

* Deberían sumar 1, no es así debido a los redondeos.

a) ¿Cuántos hogares tienen menos de 2 coches? ¿Qué porcentaje representan sobre el total?

El número de hogares con menos de 2 coches es la suma de los hogares que no tienen coche y los que tienen 1: $F_2 = 17265$. El porcentaje es $H_2\% = 66\%$.

b) ¿Qué proporción de hogares tienen 3 o más coches? ¿En porcentaje? $h_4 = 0,06$ y $h_4\% = 6\%$

Para representar información de una **variable continua** procedente de una muestra se suelen agrupan las observaciones en intervalos que se denominan **intervalos de clase L_i**. Este procedimiento supone, de hecho, una pérdida de información y se utiliza para simplificar los cálculos.

La **marca de clase** es el punto medio del intervalo y se toma como el dato que representa a dicho intervalo: $x_i = \frac{L_{i-1} + L_i}{2}$.

Ejemplo 3: Datos de población de Ferrol en Julio 2023 por edades (Fuente: IGE "Cifras poboacionais. Grandes concellos"). Se han tabulado agrupándolos en intervalos de 10 años y calculando las marcas de clase como la media de los extremos del intervalo:

Intervalos de edades	x_i	f_i	F_i	h_i	H_i
[0,10)	$\frac{0+10}{2}=5$	3612			
[10,20)	15	5522			
[20,30)	25	5824			
[30,40)	35	6453			
[40,50)	45	9397			
[50,60)	55	9817			
[60,70)	65	9603			
[70,80)	75	8281			
[80,90)	85	4661			
[90,100]	95	1424			
		N = 64594			

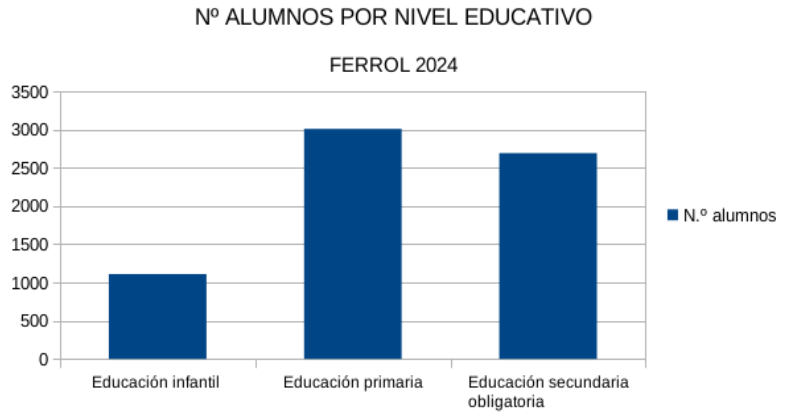
Comprueba los datos originales y cómo se han recogido en la tabla en intervalos.

1.2. Gráficos estadísticos

Los **gráficos estadísticos** constituyen una herramienta muy útil para **representar la información** de forma sencilla, clara y precisa, al igual que facilitan comparar datos, mostrar tendencias y marcar diferencias. Algunos de los más utilizados son los siguientes:

- El **diagrama de barras** es válido para variables **cualitativas** y **cuantitativas discretas**. Se puede utilizar la frecuencia absoluta, f_i, o relativa, h_i. Las barras pueden tener orientación horizontal o vertical.

Gráfico de barras para los datos del ejemplo 1 con datos cualitativos.



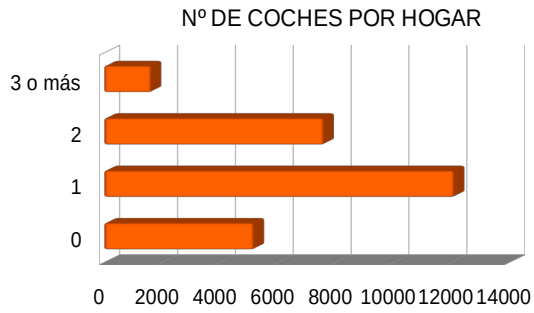
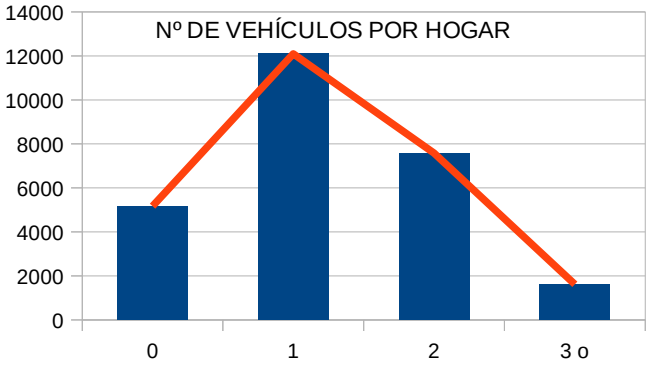
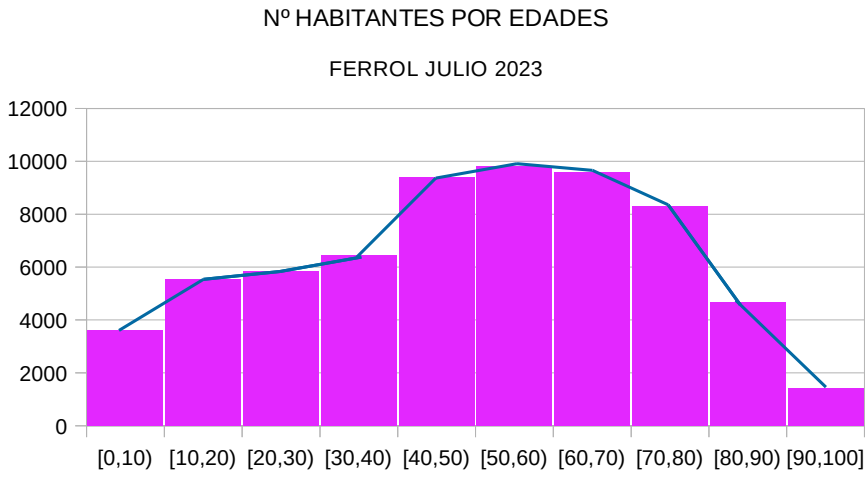


Gráfico de barras para los datos del ejemplo 2 con datos cuantitativos discretos.

- Si unimos los extremos superiores de las barras con una línea poligonal obtenemos el **polígono de frecuencias** correspondiente.

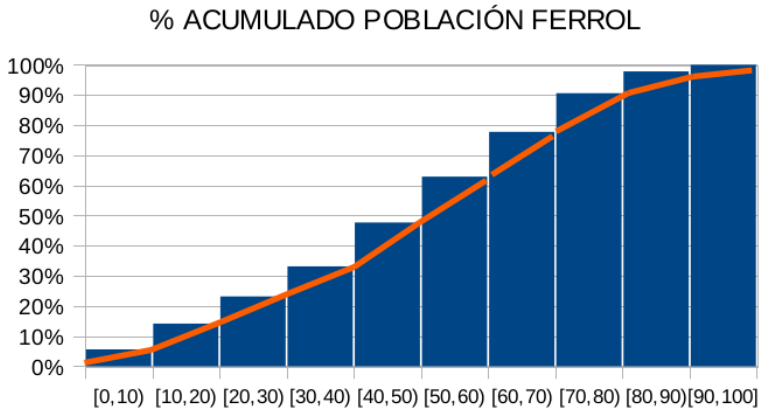


- Los **histogramas** se utilizan para representar variables **cuantitativas continuas**. En uno de los ejes se colocan los intervalos y en el otro eje se colocan de manera contigua, un rectángulo para cada intervalo con área proporcional a cada frecuencia representada.



Al ser todos los intervalos de la misma longitud basta igualar la altura de los rectángulos a la frecuencia absoluta para que las áreas sean proporcionales. También podemos dibujar el polígono de frecuencias uniendo los puntos medios de la parte superior de los rectángulos.

- El **histograma de frecuencias acumuladas** utiliza la columna Fi o Hi en el eje Y.



También se puede dibujar el **polígono de frecuencias acumuladas** uniendo los vértices de la esquina superior derecha de los rectángulos.

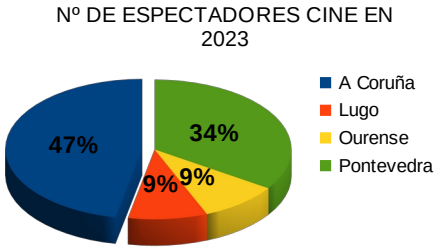
- El **diagrama de sectores** es una representación de las frecuencias relativas de las variables cualitativas o cuantitativas. Los ángulos de los sectores son proporcionales a las frecuencias relativas h_i, por tanto, se calculan $\alpha_i = h_i \times 360^o$.

Ejemplo 4: Millones de espectadores y millones de euros en recaudación en 2023 (Fuente: IGE - Ministerio de Cultura y Deporte. Estadística de Cinematografía)

	TOTAL		PELÍCULA NACIONAL		PELÍCULA EXTRANJERA	
	Nº espectadores	Recaudación	Nº espectadores	Recaudación	Nº espectadores	Recaudación
España	76,7	493,1	13,4	82,4	63,3	410,7
Galicia	3,2	20,1	0,6	3,7	2,6	16,4
A Coruña	1,5	9,5	0,3	1,7	1,2	7,8
Lugo	0,3	1,8	0,1	0,4	0,2	1,4
Ourense	0,3	1,5	0,1	0,3	0,2	1,2
Pontevedra	1,1	7,3	0,2	1,3	0,9	6

Mostraremos los datos del número de espectadores totales por provincias en un diagrama de sectores:

¶	fi¶	hi¶	αi = hi x 360º.¶
A Coruña¶	1,5¶	0,46875¶	168,75º¶
Lugo¶	0,3¶	0,09375¶	33,75º¶
Ourense¶	0,3¶	0,09375¶	33,75º¶
Pontevedra¶	1,1¶	0,34375¶	123,75º¶
¶	**N = 3,2¶**	1¶	360º¶

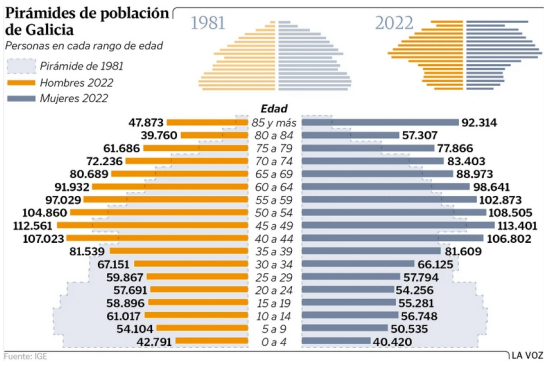


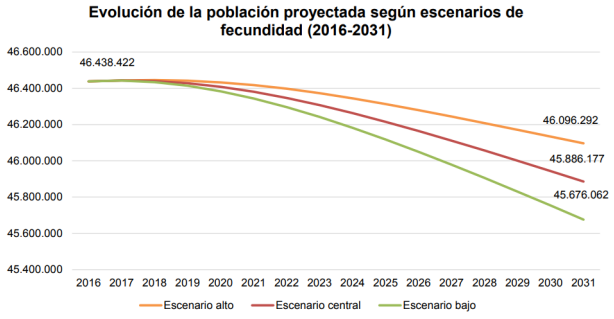
- El **pictograma** es un gráfico estadístico que se suele utilizar para caracteres cualitativos y que en lugar de barras para representar las frecuencias, utiliza dibujos o gráficos alusivos a cada atributo y cuya dimensión sea proporcional a la frecuencia absoluta.

El siguiente pictograma representa la cantidad de personas que asistieron a cine el semestre pasado.

= 5.000 personas

Otros gráficos estadísticos son las **pirámides de población**, las **series de tiempo**, los **cartogramas**, ...





Noticia de La Voz de Galicia

Fuente: INE

2. MEDIDAS DE CENTRALIZACIÓN

En muchos casos la tabla de la distribución de frecuencias y la información visual de los gráficos no son suficientes para tener una información completa de los datos. Se hacen necesarias entonces unas **medidas numéricas que permitan resumir dichos datos** de forma sencilla y eficaz, como las **medidas de centralización**, que son las que dan una idea de la ubicación del centro de la distribución.

2.1. Media aritmética

La **media aritmética** se representa por $\overline{x}$ y se calcula sumando todos los datos $x_1, x_2, \ldots, x_N$, y dividiendo entre el número de ellos:

$$\overline{x}=\frac{x_1+x_2+\ldots+x_N}{N}=\frac{\sum_{i=1}^{N} x_i}{N}$$

Ejemplo 5: Las notas de un alumno en métodos estadísticos son 8, 9, 7, 4 y 10. Si la nota media se calcula sin ponderar: $\overline{x}$ = (8 + 9 + 7 + 4+ 10) / 5 = 7,6

En el caso de tener datos tabulados, la frecuencia absoluta nos indica el número de veces que se repite cada dato. Si los datos están agrupados en intervalos utilizaremos las marcas de clase:

$$\overline{x}=\frac{x_1\cdot f_1+x_2\cdot f_2+\ldots+x_N\cdot f_N}{N}=\frac{\sum_{i=1}^{N} x_i\cdot f_i}{N}$$

Ejemplo 6: En un test de métodos estadísticos puntuado sobre 5 las notas han sido 3, 4, 5, 5, 2, 3, 3, 4, 4, 4, 5, 5, 5, 2, 3, 4. Como los resultados se repiten los ordenamos y contamos los repetidos 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5. Ahora podemos hacer la media:

$$\overline{x}=\frac{2\cdot 2+3\cdot 4+4\cdot 5+5\cdot 5}{16}=\frac{61}{16}=3{,}8125$$

Si tabulamos los datos del ejemplo 6 añadiríamos esta columna para calcular la media:

x_i	f_i	$x_i f_i$
2	2	4
3	4	12
4	5	20
5	5	25
	N = 16	$\sum_{i=1}^{N} x_i\cdot f_i$ =61

Si los datos se presentan en intervalos, se utilizan las marcas de clase.

2.2. Mediana

La **mediana**, M_e, es el valor que ocupa la posición central. Se calcula ordenando los datos de menor a mayor y escogiendo el dato central.

Si el número de datos es **impar**, habrá un dato que sea el central y, por tanto, será la mediana. Si el número de datos es **par**, existirán dos datos centrales y la mediana será su media aritmética:

$$M_e = x_{\left[\frac{N+1}{2}\right]}, \quad si\ N\ es\ impar, \qquad M_e = \frac{x_{\left[\frac{N}{2}\right]} + x_{\left[\frac{N}{2}+1\right]}}{2}, \quad si\ N\ es\ par$$

(donde $x_{[i]}$ indica el dato con su posición una vez ordenados los datos)

Vamos a ver un ejemplo de cada caso:

Ejemplo 7: Las notas de un alumno en matemáticas son 8, 9, 7, 4 y 10. La profesora de matemáticas decide poner la nota con la mediana. Ordena los datos: 4, 7, **8**, 9 y 10 y obtiene la nota: $M_e = 8$.

¿Es más beneficioso para el alumno la media aritmética o la mediana? ¿En todos los casos? Razona tu respuesta.

Ejemplo 8: Las notas de un alumno en métodos estadísticos son 3, 4, 5, 5, 10 y 4. La profesora de métodos decide poner la nota con la mediana. Ordena los datos: 3, 4, **4, 5**, 5 y 10 y obtiene la nota: $M_e=(4+5)/2 = 4{,}5$.

Observa que **la mediana** es una medida de centralización que **no se ve afectada por valores extremos**.

En caso de **datos** recogidos en **tabla de frecuencias** la mediana es el primer dato o marca de clase, x_i, cuya frecuencia relativa acumulada supera el 0,5, $H_i > 0{,}5$. En caso de que se iguale el 0,5 (50%) se hará la media entre el dato y el siguiente:

Ejemplo 7		
xi	**fi**	**Hi**
4	1	0,2
7	1	0,4
8	1	0,6
9	1	0,8
10	1	1
	N = 5	

–> $M_e = 8$

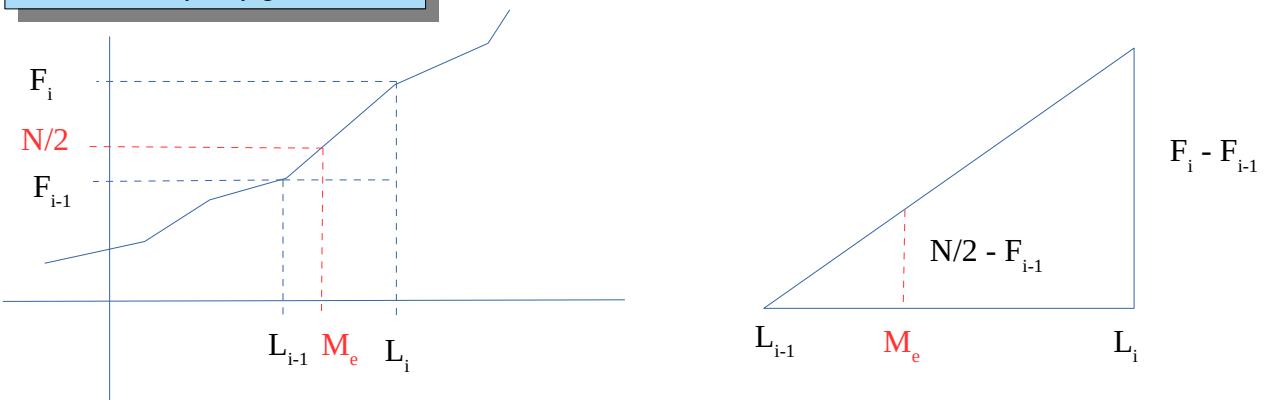
Ejemplo 8		
xi	**fi**	**Hi**
3	1	0,17
4	2	0,5
5	2	0,83
10	1	1
	N = 6	

–> $M_e = 4{,}5$

En caso de **variables continuas** presentadas en **intervalos** se puede calcular la mediana del mismo modo o con más precisión interpolando linealmente en el intervalo de la mediana en el polígono de frecuencias acumuladas:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - F_{i-1}}{F_i - F_{i-1}} \cdot (L_i - L_{i-1})$$

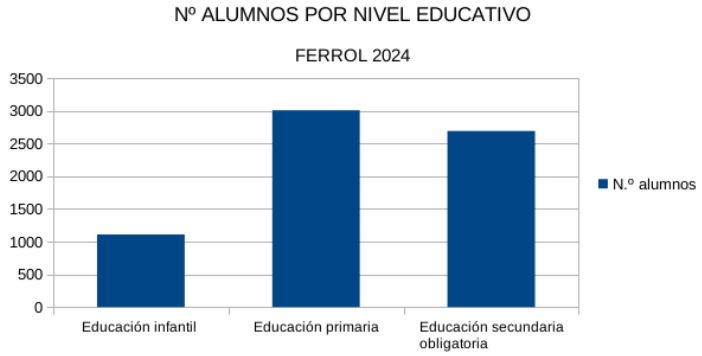
donde: L_{i-1} = Límite inferior del intervalo mediana
F_{i-1} = Frecuencia acumulada anterior al intervalo mediana
F_i = Frecuencia acumulada del intervalo mediana
N = Total de datos



Por triángulos semejantes: $\frac{M_e - L_{i-1}}{N/2 - F_{i-1}} = \frac{L_i - L_{i-1}}{F_i - F_{i-1}} \Rightarrow M_e = L_{i-1} + \frac{N/2 - F_{i-1}}{F_i - F_{i-1}} \cdot (L_i - L_{i-1})$

2.3. Moda

La moda, $\mathbf{M_o}$, es el valor o dato con mayor frecuencia. Su utilidad es importante en variables cualitativas, ya que es el único parámetro estadístico de tendencia central que se puede calcular.



En el ejemplo de alumnos matriculados en enseñanzas obligatorias en Ferrol en el curso 2024-25 vemos que la moda es la educación primaria:

Puede existir más de una moda. Así una distribución con dos modas se llama bimodal.

3. MEDIDAS DE POSICIÓN

Los **cuantiles** permiten determinar la posición de un dato en relación con los demás cuando los valores están ordenadosde menor a mayor:

Los **cuartiles** son los tres valores que dividen a la serie de datos en cuatro partes iguales. Se representan por Q_1, Q_2 y Q_3 y se llaman primer, segundo y tercer cuartil, respectivamente. El segundo cuartil, Q_2, coincide con la mediana, $Q_2 = M_e$.

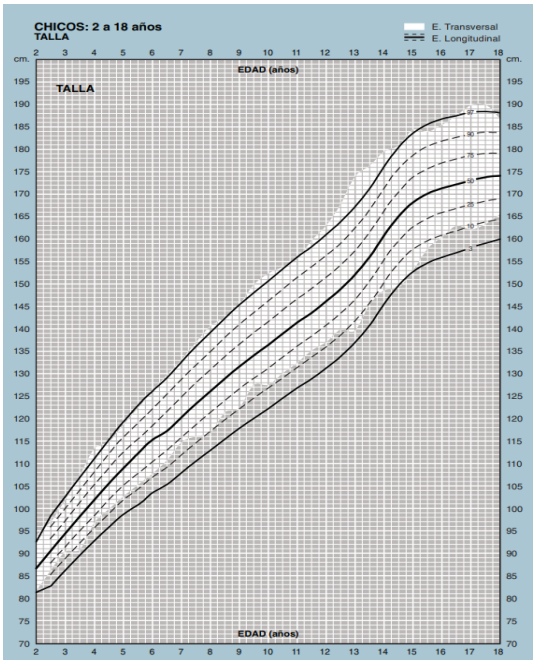
Ejemplo 9: Las edades de los pacientes de una consulta médica son: 25, 35, 54, 30, 25, 26, 45, 40, 32, 44, 65, 20, 52, 65 y 22. Ordenamos los datos y los dividimos en cuatro partes iguales:

20, 22, 25, **25**, 26, 30, 32, **35**, 40, 44, 45, **52**, 54, 65 y 65

Los datos coloreados dividen las edades de los pacientes en cuatro partes iguales: el 25% de los datos son menores que $Q_1 = 25$ años, el 50% de los datos son menores que $Q_2 = M_e = 35$ años y el 75% son menores que $Q_3 = 52$ años.

Los **percentiles** son los noventa y nueve valores que dividen la serie de datos en cien partes iguales. Se representan por P_1, P_2, …, P_{99}. Observa que $P_{25} = Q_1$, $P_{50} = Q_2 = M_e$ y $P_{75} = Q_3$.

Ejemplo 10: Esta medida de posición se utiliza para visualizar los datos de crecimiento de 0 a 18 años. Si compruebas tu cartilla de vacunación encontrarás gráficas de percentiles. Una de estas gráficas es la siguiente:



A partir de datos recogidos de chicos de 2 a 18 años se representan y calculan los percentiles.
Se marcan los percentiles más destacados: P_3, P_{10}, P_{25} (primer cuartil), P_{50} (mediana), P_{75} (tercer cuartil), P_{90} y P_{97}.

Por ejemplo, a los 17 años:
$P_3 = 158$ cm $\quad P_{10} = 163$ cm
$P_{25} = Q_1 = 168$ cm $\quad P_{50} = M_e = 173$ cm
$P_{75} = 179$ cm $\quad P_{90} = 183$ cm
$P_{97} = 188$ cm
Es decir, el 97% de los chicos de 17 años miden 188 cm o menos.

3.1. Diagrama de caja y bigotes

El **diagrama de caja y bigotes,** en inglés **boxplot o box and whisker plot,** es un gráfico basado en los cuartiles que da información de la dispersión y la simetría de los datos. Se construye dibujando:

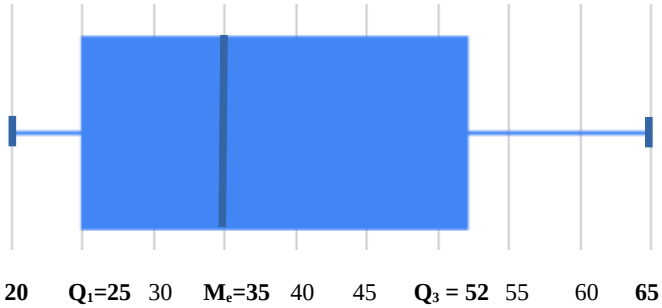
1º) Una **caja**, rectángulo, entre el primer y el tercer cuartil, con una línea vertical sobre la mediana.

2º) Los **bigotes** son unos segmentos cuyo extremo inferior es el máximo entre el primer dato, $x_{[1]}$, y $Q_1 - 1{,}5\,(Q_3 - Q_1)$ mientras que el extremo superior es el mínimo entre el dato mayor, $x_{[N]}$, y $Q_3 + 1{,}5\,(Q_3 - Q_1)$.
Si hay **datos que queden por encima o por debajo de los extremos de los bigotes**, se los representa con puntos. Estos puntos se conocen como **valores atípicos**.

Ejemplo: En el **ejemplo 9** de las edades de los pacientes en una consulta teníamos los siguientes datos ordenados:

20, 22, 25, 25, 26, 30, 32, 35, 40, 44, 45, 52, 54, 65 y 65 $Q_1 = 25$ años, $Q_2 = M_e = 35$ años y $Q_3 = 52$ años.

El diagrama de caja y bigotes que representa estos datos es:

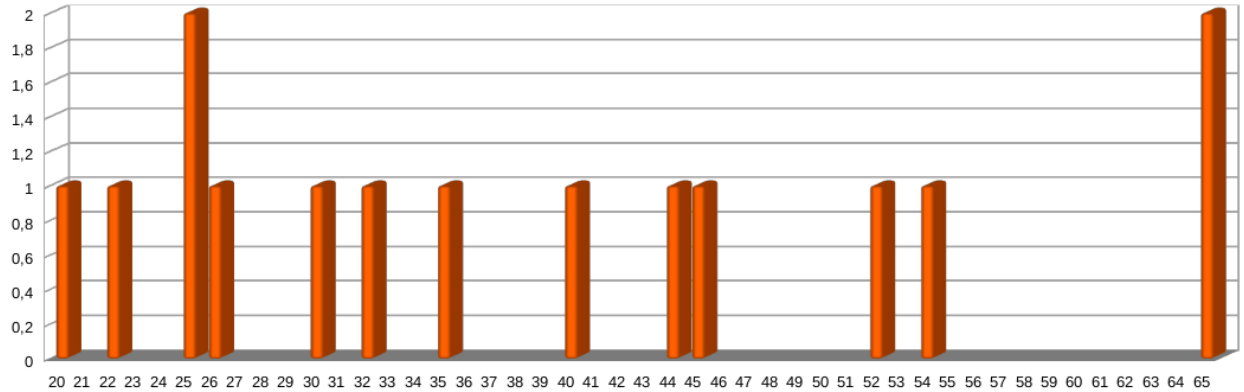


max (20, 25 – 1,5(52 – 25)) =20 min (65, 52 + 1,5(52 – 25)) = 65

Analizamos el diagrama de caja y bigotes de las edades de los pacientes:

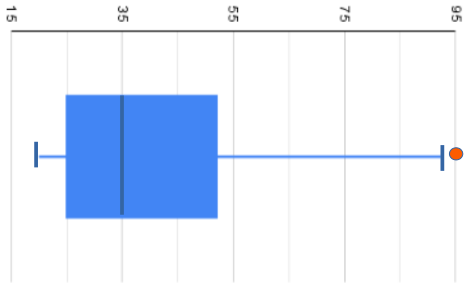
- La mediana no está centrada. Los valores menores que la mediana están más concentrados que los mayores.
- El rango intercuartílico es 52 – 25 = 27 años. El 50% de los datos centrales difieren en 27 años.
- No hay valores atípicos, todos los valores están entre los extremos de los bigotes.

Es un gráfico que ofrece un resumen más compacto de los datos que los diagramas de barras o histogramas. Comparamos con el diagrama de barras de la distribución de las edades:



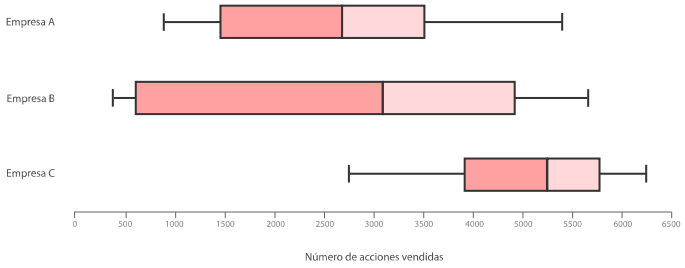
Ejemplo: Imagina que hubo un error al apuntar un dato y uno de los pacientes tenía 95 años en lugar de 65 años.

Actualizamos el diagrama de caja y bigotes:



El bigote derecho es ahora:
min (95, 52 + 1,5(52 – 25)) = 92,5 años

El paciente de 95 años es un valor atípico.



En muchas ocasiones los diagramas de caja y bigotes se utilizan para comparar datos de un solo vistazo:

4. MEDIDAS DE DISPERSIÓN

Las **medidas de dispersión** permiten conocer el grado de agrupamiento de los datos entorno a las medidas de centralización.

4.1. Recorrido o rango

Recorrido o rango, R: es la diferencia entre el mayor y el menor valor de la variable.

4.2. Rango intercuartílico

Rango intercuartílico, R_i: es la diferencia entre el primer y el tercer cuartil.

4.3. Varianza

La **varianza** es la media de los cuadrados de las desviaciones de los datos con respecto a la media, $x_i - \overline{x}$.

Varianza: $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\overline{x})^2 f_i}{N} = \frac{\sum_{i=1}^{N} x_i^2 f_i}{N} - \overline{x}^2$$

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

$$\frac{\sum_{i=1}^{N} x_i^2 f_i}{N} - 2\overline{x}\,\overline{x} + \overline{x}^2\frac{N}{N} = \frac{\sum_{i=1}^{N} x_i^2 f_i}{N} - \overline{x}^2$$

4.4. Desviación típica

La **desviación típica** es la raíz cuadrada de la varianza.

Desviación típica: $$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2 f_i}{N}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 f_i}{N} - \bar{x}^2}$$

Ejemplo 11: En un estudio sobre la distancia recorrida para ir al puesto de trabajo se han recogido los siguientes datos:

Distancia en km	[0, 10)	[10, 25)	[25, 50)	[50, 100)	[100, 200]
f_i	20	23	30	22	5

Calculamos la varianza y la desviación típica:

Intervalos	**Marca de clase x_i**	**f_i**	**$x_i f_i$**	**$x_i^2 f_i$**
[0, 10)	5	20	100	500
[10, 25)	17,5	23	402,5	7043,75
[25, 50)	37,5	30	1125	42187,5
[50, 100)	75	22	1650	123750
[100, 200)	150	5	750	112500
TOTAL		**N = 100**	**4027,5**	**285981,25**

La media es $\bar{x} = \frac{4027,5}{100} = 40,275$ km

La varianza es $\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 f_i}{N} - \bar{x}^2 = \frac{285981,25}{100} - 40,275^2 = 1237,7369$

La desviación típica es $\sigma = \sqrt{1237,74} = 35,18$ km

4.5. Coeficiente de variación

El **coeficiente de variación** sirve para comparar la dispersión (variación) de conjuntos de datos de medidas diferentes o con medias aritméticas diferentes.

Coeficiente de variación: $CV = \frac{\sigma}{\bar{x}}$

Ejemplo: En un segundo estudio sobre la distancia recorrida para ir al trabajo se ha obtenido una media de 4500 m y una desviación típica de 2000 m. Utilizamos el coeficiente de variación para comparar la dispersión de los datos en los dos estudios:

$CV_1 = 35,18 / 40,75 = 0,86$ $\qquad$ $CV_2 = 2000 / 4500 = 0,44$

El coeficiente de variación no tiene unidades y nos permite comparar los dos estudios que están en unidades distintas. Los datos del segundo estudio están más concentrados alrededor de la media, por tanto, será un dato más representativo que el del primer estudio.

4.6. Análisis de las medidas estadísticas

Los pasos para analizar las medidas estadísticas son:

1º) Comparamos las medidas de centralización: media, mediana y moda:

2º) Analizamos el diagrama de caja y bigotes.

3º) Comparamos la desviación típica con la media teniendo en cuenta:

En muchas ocasiones en un conjunto de datos se cumple que:
En el intervalo $(\bar{x}-\sigma, \bar{x}+\sigma)$ se encuentra aproximadamente el 68% de los datos.
En el intervalo $(\bar{x}-2\sigma, \bar{x}+2\sigma)$ se encuentra aproximadamente el 95% de los datos.
En el intervalo $(\bar{x}-3\sigma, \bar{x}+3\sigma)$ se encuentra aproximadamente el 99% de los datos.

Un **dato** es **atípico** si está fuera de estos intervalos.

Ejemplo: En el ejemplo 11 podemos calcular los intervalos a partir de la media, $\bar{x}$ = 40,275 km y la desviación típica, σ = 35,18 km:

$(\bar{x}-\sigma, \bar{x}+\sigma)$ = (40,275 – 35,18, 40,275 + 35,18) = (5,095, 75,455)

$(\bar{x}-2\sigma, \bar{x}+2\sigma)$ = (40,275 – 2·35,18, 40,275 + 2·35,18) = (-30,085, 110,635)

$(\bar{x}-3\sigma, \bar{x}+3\sigma)$ = (40,275 – 3·35,18, 40,275 + 3·35,18) = (-65,265, 145,815)

Los datos del intervalo [100, 200] pueden considerarse datos atípicos.

5. EJERCICIOS RESUELTOS

Estos dos ejercicios modelo muestran el cálculo de los parámetros estadísticos estudiados.

5.1. Modelo de ejercicio con variable cuantitativa discreta

Una empresa de telefonía móvil hizo un estudio entre sus clientes sobre el número de horas que hablan por teléfono al día de cara a elaborar nuevas promociones. Los datos obtenidos fueron los siguientes:

2, 3, 1, 4, 5, 2, 3, 1, 1, 5, 2, 3, 4, 1, 1, 1, 1, 1, 2, 3, 4, 1, 2, 3, 2, 2, 2, 3, 4, 5, 1, 2, 3, 2, 1, 10

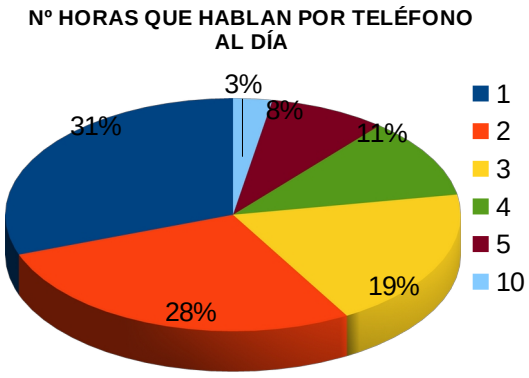
a) Elabora la tabla de frecuencias. ¿Qué porcentaje de clientes habla menos de 3 horas diarias?
b) Realiza una gráfica adecuada para estos datos.
c) Calcula las medidas de centralización: media, mediana y moda.
d) Calcula las medidas de posición: cuartiles, percentil 10 y percentil 95. Dibuja el diagrama de caja y bigotes.
e) Calcula las medidas de dispersión: recorrido, rango intercuartílico, varianza, desviación típica.
f) Calcula el CV y compara con los clientes de otra compañía que obtuvieron de media 3,5 horas y varianza 1 hora.
g) Analiza las medidas estadísticas obtenidas.

SOLUCIÓN:

a) Elabora la tabla de frecuencias.

x_i	f_i	h_i	F_i	H_i	$x_i f_i$	$x_i^2 f_i$	h_i * 360º
1	11	11/36 = 0,31	11	0,31	11	11	112º
2	10	10/36 = 0,28	21	0,59	20	40	101º
3	7	7/36 = 0,19	28	0,78	21	63	68º
4	4	4/36 = 0,11	32	0,89	16	64	40º
5	3	3/36 = 0,08	35	0,97	15	75	29º
10	1	1/36 = 0,03	36	1	10	100	11º
	N = 36	**1**			**93**	**353**	**361º**

b) Diagrama de sectores *Diagrama de barras*



c) Calcula las medidas de centralización: media, mediana y moda.

Media: $\bar{x}=\dfrac{\sum_{i=1}^{N} x_i f_i}{N}$ $\qquad \bar{x}=\dfrac{93}{36}=2{,}58\,\overline{3}\simeq 2{,}58$

Mediana: Es el valor que ocupa la posición central de los datos, después de ordenarlos, o la media de los datos centrales, si el número de datos es par.

Buscamos el primer valor x_i que cumple $F_i > N/2$ o $H_i > 0{,}5$ $\qquad$ **$M_e = 2$**

Si Fi = N/2 o Hi = 0,5, calculamos la media de este valor y el siguiente → $M_e=\dfrac{x_i+x_{i+1}}{2}$

Moda: es el dato o datos que aparecen con mayor frecuencia. $\qquad$ **$M_o = 1$.**

d) Calcula las medidas de posición: cuartiles, percentil 10 y percentil 95. Dibuja el diagrama de caja y bigotes.

Primer cuartil, Q_1: el primer dato x_i que cumple $F_i > N/4$ o $H_i > 0{,}25$ $\qquad$ **$Q_1 = 1$.**

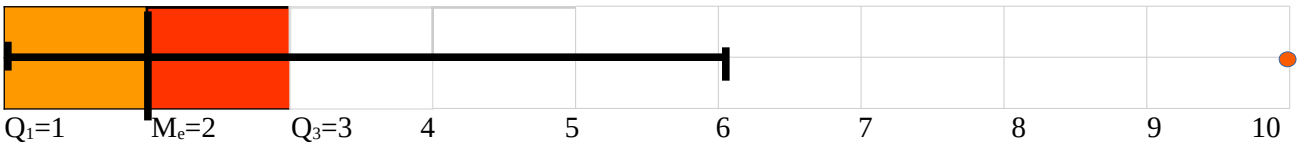
Tercer cuartil, Q_3: el primer dato x_i que cumple $F_i > 3N/4$ o $H_i > 0{,}75$ $\qquad$ **$Q_3 = 3$.**

El segundo cuartil Q_2 es la mediana.

Percentil 10, P_{10}: el primer dato x_i que cumple $F_i >$ 10% de N o $H_i > 0{,}1$ $\qquad$ **$P_{10} = 1$.**

Percentil 90, P_{90}: el primer dato x_i que cumple $F_i >$ 90% de N o $H_i > 0{,}9$ $\qquad$ **$P_{90} = 5$.**

Diagrama de caja y bigotes:



Bigote inferior: máximo (1, 1 - 1,5 * R.I. = -2) = 1
Bigote superior: mínimo (10, 3 + 1,5 * R.I. = 6) = 6

e) Calcula las medidas de dispersión: recorrido, rango intercuartílico, varianza, desviación típica.

Recorrido o rango, R: R = 10 – 1 = 9

Rango intercuartílico, RI: R.I. = $Q_3 - Q_1$ = 3 – 1 = 2

Varianza: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\bar{x})^2 f_i}{N}=\dfrac{\sum_{i=1}^{N} x_i^2 f_i}{N}-\bar{x}^2$ –> $\sigma^2=\dfrac{353}{36}-2{,}58^2=3{,}149\ \ horas^2$

Desviación típica: $\sigma = \sqrt{\dfrac{\sum_{i=1}^{N}(x_i-\bar{x})^2 f_i}{N}}=\sqrt{\dfrac{\sum_{i=1}^{N} x_i^2 f_i}{N}-\bar{x}^2}$ $\qquad \sigma=\sqrt{3{,}149}=1{,}77\ \ horas$

f) Calcula el CV y compara con los clientes de otra compañía que obtuvieron de media 3,5 horas y varianza 1 hora.

Coeficiente de variación: CV = σ / $\bar{x}$ $\qquad$ **CV = 1,77 / 2,58 = 0,686 = 68,6%**

Otra compañía: **CV = 1 / 3,5 = 0,286 = 28,6%**

Los datos en la segunda compañía está más concentrados con respecto a la media y por lo tanto la media es más representativa.

g) Analiza las medidas estadísticas obtenidas.

1º) Comparamos las medidas de centralización: media, mediana y moda:

La media es 2,58, la mediana 2 y la moda 1, por tanto, no coinciden las medidas de centralización lo que nos permite concluir que la distribución de frecuencias no es simétrica. Como podemos comprobar en el diagrama de barras.

2º) Analizamos el diagrama de caja y bigotes:

La caja del diagrama no está centrada con respecto a los datos, la distribución no es simétrica. El 50% de los datos se encuentra entre los valores 1 y 3. El valor 10 es atípico puesto que es mayor que el bigote superior.

3º) Comparamos la desviación típica con la media, así como el CV:

La desviación de los datos respecto a la media es de 1,77 horas, el CV del 68,6% indica que los datos están ligeramente dispersos y la media no es del todo representativa.

En el intervalo $(\bar{x}-3\sigma, \bar{x}+3\sigma)=(2{,}58-3\cdot 1{,}77, 2{,}58+3\cdot 1{,}77)=(-2{,}73, 7{,}89) \Rightarrow (0, 8)$ debería encontrarse el 99% de los datos, sin embargo en nuestro caso sería el 97% debido a que tenemos un valor atípico, 10.

5.2. Modelo de ejercicio con variable cuantitativa continua

Un estudio acerca del peso de un grupo de adolescentes ha obtenido los siguientes datos:
60, 55, 53, 65, 70, 68, 45, 67, 78, 67, 56, 65, 48, 67, 68, 79, 67, 56, 55, 49, 50, 56, 67, 77, 55

a) Reparte los datos en intervalos y calcula las marcas de clase. Elabora la tabla de frecuencias.
b) Realiza una gráfica adecuada para estos datos.
c) Calcula las medidas de centralización: media, mediana y moda.
d) Calcula las medidas de posición: cuartiles, P_{20} y P_{95}. Dibuja el diagrama de caja y bigotes.
e) Calcula las medidas de dispersión: recorrido, rango intercuartílico, varianza y desviación típica.
f) Calcula el CV y compara con los adolescentes de otro estudio que obtuvo de media 60 kg y desviación típica 5 kg.
g) Analiza las medidas estadísticas obtenidas.

SOLUCIÓN:

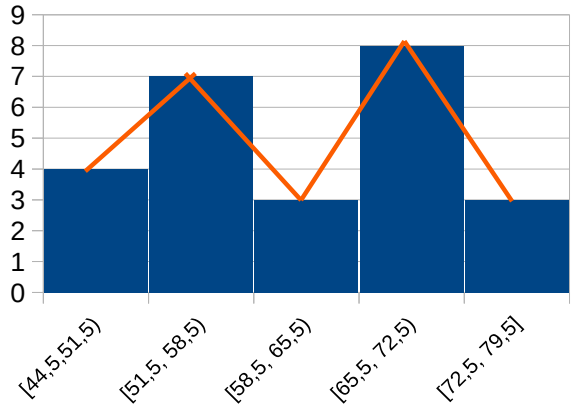
a) Reparte los datos en intervalos y calcula las marcas de clase. Elabora la tabla de frecuencias.

Se elige el primer extremo de los intervalos procurando que las marcas de clase, c_i, sean números enteros. Los intervalos tendrán el extremo izquierdo cerrado y el derecho abierto.

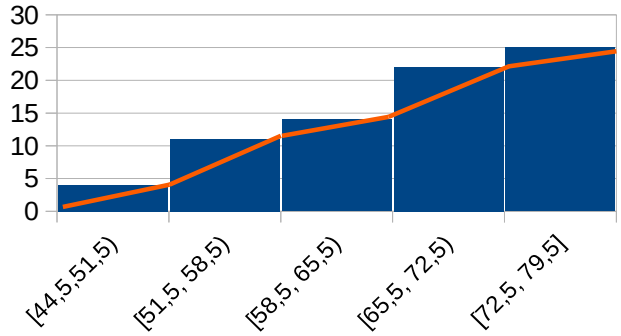
Intervalo $[L_{i-1}, L_i)$	Marca de clase $x_i = \frac{L_{i-1}+L_i}{2}$	f_i	h_i	F_i	H_i	$x_i f_i$	$x_i^2 f_i$
[44,5, 51,5)	48	4	0,16	4	0,16	192	9216
[51,5, 58,5)	55	7	0,28	11	0,44	385	21175
[58,5, 65,5)	62	3	0,12	14	0,56	186	11532
[65,5, 72,5)	69	8	0,32	22	0,88	552	38088
[72,5, 79,5]	76	3	0,12	25	1	228	17328
		N=25	1			1543	97339

b) Gráficos adecuados.

Histograma y polígono de frecuencias



Histograma y polígono de frecuencias acumuladas



c) Calcula las medidas de centralización: media, mediana y moda.

Media: $\bar{x}=\dfrac{\sum_{i=1}^{N} x_i f_i}{N}=\dfrac{1543}{25}=61{,}72$ kg

Mediana: Es la marca de clase del intervalo que ocupa la posición central de los datos, después de ordenarlos.

N/2 = 12,5 buscamos el primer intervalo $[L_{i-1}, L_i)$ que cumple $F_i > N/2$, su marca de clase x_i es la mediana: **M_e = 62 kg**

Un cálculo más preciso de la mediana se obtiene interpolando linealmente en el intervalo de la mediana en el polígono de frecuencias acumuladas:

$$M_e=L_{i-1}+\frac{\frac{N}{2}-F_{i-1}}{F_i-F_{i-1}}\cdot(L_i-L_{i-1})$$

donde: L_{i-1} = Límite inferior del intervalo mediana
F_{i-1} = Frecuencia acumulada anterior al intervalo mediana
F_i = Frecuencia acumulada del intervalo mediana
N = Total de datos

$M_e=58{,}5+\dfrac{\frac{25}{2}-11}{14-11}\cdot(65{,}5-58{,}5)=58{,}5+3{,}5=62\ \ kg$ en este ejemplo coincide con el método simple.

Moda o intervalo modal: es el dato o datos que aparecen con mayor frecuencia, si trabajamos con las marcas de clase o intervalo o intervalos modales en caso de trabajar con los intervalos.

Intervalo modal: [65,5, 72,5) **Moda: 69 kg**

d) Calcula las medidas de posición: cuartiles, percentil 20 y percentil 95. Dibuja el diagrama de caja y bigotes.

Primer cuartil, Q_1: marca de clase x_i del primer intervalo que cumple $F_i > N/4$: **Q_1 = 55 kg**

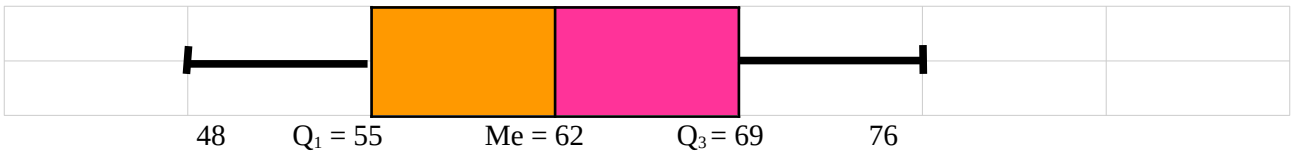
Tercer cuartil, Q_3: marca de clase x_i del primer intervalo que cumple $F_i > 3N/4$: **Q_3 = 69 kg**

El segundo cuartil Q_2 es la mediana.

Percentil 20, P_{20}: marca de clase xi del primer intervalo que cumple F_i > 20% de N: **P_{20} = 55 kg**

Percentil 95, P_{95}: marca de clase x_i del primer intervalo que cumple F_i > 95% de N: **P_{95} = 76 kg**

Diagrama de caja y bigotes:



Bigote inferior: máximo (48, 55 - 1,5 * R.I. = 34) = 48
Bigote superior: mínimo (76, 69 + 1,5 * R.I. = 90) = 76

e) Calcula las medidas de dispersión: recorrido, rango intercuartílico, varianza, desviación típica.

Recorrido o rango, R: R = 79 – 45 = 34

Rango intercuartílico, R I: R.I. = $Q_3 - Q_1 = 69 - 55 = 14$

Varianza: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\bar{x})^2 f_i}{N} = \dfrac{\sum_{i=1}^{N} x_i^2 f_i}{N} - \bar{x}^2 = \dfrac{97339}{25} - 61{,}72^2 = 84{,}2016$

Desviación típica: $\sigma = \sqrt{\dfrac{\sum_{i=1}^{N}(x_i-\bar{x})^2 f_i}{N}} = \sqrt{\dfrac{\sum_{i=1}^{N} x_i^2 f_i}{N} - \bar{x}^2} = \sqrt{84{,}2016} = 9{,}1761 \ \ kg$

f) *Calcula el CV y compara con los adolescentes de otro estudio que obtuvo de media 60 kg y desviación típica 5 kg.*

Coeficiente de variación: CV = σ / $\bar{x}$ = 9,1761 / 61,72 = 0,1487 =14,87%

Otro estudio: CV = 5 / 60 = 8,33%

Los datos del segundo estudio están más concentrados con respecto a la media y por lo tanto la media es más representativa.

g) Analiza las medidas estadísticas obtenidas.

1º) Comparamos las medidas de centralización: media, mediana y moda:

La media es 61,72 kg, la mediana 62 kg y la moda 69 kg, por tanto, coinciden la media y la mediana pero no la moda lo que nos permite concluir que la distribución de frecuencias es bastante simétrica. Como podemos comprobar en el diagrama de barras.

2º) Analizamos el diagrama de caja y bigotes:

La caja del diagrama está centrada con respecto a los datos, la distribución es bastante simétrica. El 50% de los datos se encuentra entre los valores 55 kg y 69 kg. No hay valores atípicos.

3º) Comparamos la desviación típica con la media, así como el CV:

La desviación de los datos respecto a la media es de unos 9 kg, el CV del 15% indica que los datos están bastante concentrados respecto a la media y esta es bastante representativa.

En el intervalo $(\bar{x}-3\sigma, \bar{x}+3\sigma) = (61{,}72-3\cdot 9{,}1761, 61{,}72+3\cdot 9{,}1761) = (34{,}19, 89{,}2483) \Rightarrow (45, 79)$ debería encontrarse el 99% de los datos, en nuestro caso sería el 100% lo que nos confirma que no hay datos atípicos y la media es representativa.