Matemáticas Orientadas a las Enseñanzas Académicas. 3º ESO

Estadística Descriptiva

CPI da Cañiza. Departamento de Matemáticas

curso 2017-2018

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.

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 en tablas de frecuencias.
- Interpretación y análisis por medio del cálculo de medidas de centralización y dispersión.
- Representación gráfica de los datos.

Conceptos generales

- Población y Muestra.
- Variable Estadística.
  - Variable Estadística Cualitativa.
  - Variable Estadística Cuantitativa: Variables Discretas y variables Continuas.
- Tablas de frecuencias: Frecuencias absolutas y relativas, ordinarias y acumuladas.
- Parámetros estadísticos.
  - Medidas de centralización: Media, Moda y Mediana.
  - Medidas de dispersión: Rango, Varianza, Desviación Típica.
  - Medidas de posicionamiento: Cuartiles, Rango intercuartílico.
  - Coeficiente de variación.
- Representaciones gráficas.

Conceptos generales: Población

- 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 dicho conjunto.

Conceptos generales: Muestra

- Muestra: es 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.

Conceptos generales: Variable Estadística. Tipos

- Variable Estadística: Es la característica cuya distribución desea estudiarse en una población.
  Pueden ser cualitativas o cuantitativas.
  - 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.
    Ejemplos: color de ojos, deporte favorito de los adolescentes, 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.

Tipos de Variable Estadística Cuantitativa

- Variable Cuantitativa Discreta: es una variable que entre dos valores determinados no puede tomar cualquier otro valor. Los casos más típicos son aquellas variables que solo pueden tomar valores enteros.
  Ejemplos: número de hijos, la edad, o el año de nacimiento.
- Variable Cuantitativa Continua: es una variable que entre dos valores determinados puede tomar cualquier otro valor intermedio. Son variables cuyos valores son números reales.
  Ejemplos: la estatura, la temperatura, o el tiempo de reacción a un estímulo.

Tratamiento de Variables Cualitativas

Ejemplo 1

En una clase se desea estudiar cómo se distribuye el color de ojos, obteniendo que 10 de ellos tienen los ojos marrones, 6 azules, y 4 verdes.

- La variable X es "color de ojos".
  Cada respuesta de la variable se representa de modo genérico por X_i. En este caso, $X_1 =$ "marrón", $X_2 =$ "azul", $X_3 =$ verde.
- El número de veces que se presenta cada valor X_i se llama Frecuencia Absoluta, y lo representaremos por f_i.
- La Frecuencia Relativa del valor X_i es $h_i = \frac{f_i}{n}$ (n es el número de datos).

Tabla de Frecuencias. Cálculo de la Moda

- Construimos la tabla de frecuencias correspondiente.

X_i	f_i	h_i
azul	6	$\frac{6}{20}$
verde	4	$\frac{4}{20}$
marrón	10	$\frac{10}{20}$
TOTAL	20	1

- La Moda, M_o, es el valor de la variable con la frecuencia más alta. En este caso $M_o = \text{marrón}$

Representaciones gráficas de variables cualitativas

Cuando la variable es cualitativa, las formas más habituales para representar los datos gráficamente son los diagramas de barras y los de sectores circulares.

- 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.
- Diagrama de sectores circulares: En un círculo, cada valor de la variable se corresponde con un sector circular, cuyo ángulo central será proporcional a la frecuencia de la variable.

Representaciones gráficas de variables cualitativas

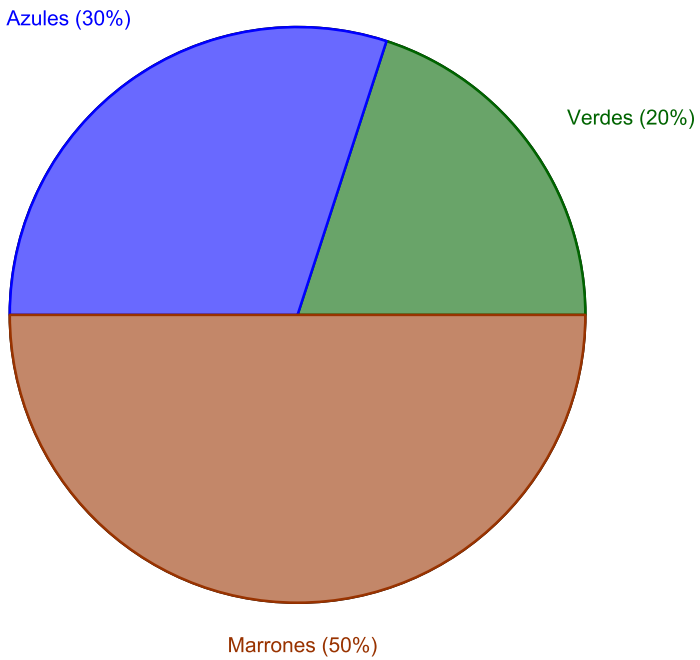


Figura: Diagrama de sectores

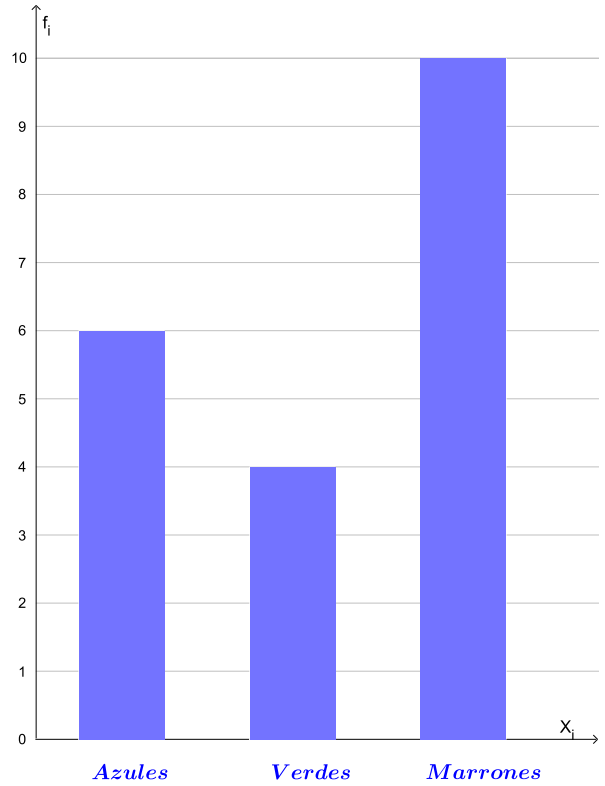


Figura: Diagrama de barras

Tratamiento de Variables Cuantitativas Discretas

Ejemplo 2

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

Tabla de Frecuencias

Es similar a la de las variables cualitativas, pero en el caso de las variables cuantitativas, existe también el concepto de frecuencia acumulada.

- Frecuencia Absoluta Acumulada F_i: Es el número de respuestas menores o iguales al valor de X_i.
- Frecuencia Relativa Acumulada H_i: Es el cociente $H_i = \frac{F_i}{n}$

Se utilizan para calcular cómodamente los valores de los cuartiles.

Tabla de Frecuencias

En este ejemplo, X es el número de hijos.
La tabla de frecuencias absolutas y relativas es la siguiente:

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	$\frac{3}{60}$	$\frac{60}{60}$
TOTAL	60	–	1	–

Parámetros Estadísticos: Medidas de Centralización (I)

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 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.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i \cdot f_i}{n}$$

Parámetros Estadísticos: Medidas de Centralización (II)

- 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.
- 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, es un valor tal que el 50 % de las observaciones son menores o iguales que él.
  Su cálculo depende de si el número de datos es par o impar.

Cálculo de las medidas de centralización (I)

- Media: se añade una columna a la tabla de frecuencias:

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

$$\text{Con lo cual } \overline{X} = \frac{\sum_{i=1}^{n} X_i \cdot f_i}{n} = \frac{80}{60} = 1'33$$

Cálculo de las medidas de centralización (II)

- Mediana:
  - En este caso hay 60 datos.
    Supongamos que tenemos las 60 respuestas ordenadas de menor a mayor $x_1, x_2, x_3, \ldots, x_{60}$. La respuesta que dividiese a la muestra en dos partes iguales sería la que ocupase la posición $30'5$, por eso se toma como mediana el promedio de las respuestas que están en posición 30 y 31.
    Consultando la columna de las F_i, vemos que $x_{30} = 1$, y $x_{31} = 1$, así que $M_e = 1$.
  - Si hubiesen sido 61 datos, la mediana sería el valor de x_{31}.
- Moda: $M_o = 1$.

Parámetros Estadísticos: Medidas de Dispersión

Las medidas de dispersión se utilizan para cuantificar la representatividad de las medidas de centralización (especialmente la de la media). Son el rango, la varianza, y la desviación típica.

- Rango: Es la diferencia entre los valores máximo y mínimo observados.
- Varianza: Es el error cuadrático medio cometido al tomar la media como valor de la variable.

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

- Desviación típica: Es la raíz cuadrada de la varianza. Se representa por s.

Cálculo de las medidas de dispersión (I)

En la práctica, s^2 se obtiene con la expresión $s^2 = \frac{\sum_{i=1}^{n} X_i^2 f_i}{n} - \overline{X}^2$.
Para facilitar el cálculo se añaden dos columnas más a la tabla de frecuencias.

X_i	f_i	F_i	h_i	H_i	$X_i \cdot f_i$	X_i^2	$X_i^2 \cdot f_i$
0	16	16	$\frac{16}{60}$	$\frac{16}{60}$	0	0	0
1	22	38	$\frac{22}{60}$	$\frac{38}{60}$	22	1	22
2	11	49	$\frac{11}{60}$	$\frac{49}{60}$	22	4	44
3	8	57	$\frac{8}{60}$	$\frac{57}{60}$	24	9	72
4	3	60	$\frac{3}{60}$	$\frac{60}{60}$	12	16	48
TOTAL	60	–	1	–	80	–	186

Cálculo de las medidas de dispersión (II)

En nuestro ejemplo tenemos:

- Rango: 4
- Varianza $s^2 = \frac{\sum_{i=1}^{n} X_i^2 f_i}{n} - \overline{X}^2 = \frac{186}{60} - 1'33^2 = 1'3311$
- Desviación típica $s = \sqrt{1'3311} = 1'1537$

Parámetros Estadísticos: Medidas de Posicionamiento (I)

- 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.
  - La diferencia $Q_3 - Q_1$ se llama rango o recorrido intercuartílico.
- El cálculo de los cuartiles es similar al de la mediana. Es decir, hay que tener en cuenta si al ir dividiendo la muestra el número de datos es par o impar. En nuestro caso:
  - $Q_1 = \frac{x_{14}+x_{15}}{2} = 0$
  - $Q_2 = M_e = \frac{x_{30}+x_{31}}{2} = 1$
  - $Q_3 = \frac{x_{44}+x_{45}}{2} = 2$
  - El rango intercuartílico vale 2.

Parámetros Estadísticos: Medidas de Posicionamiento (II)

Los percentiles son la generalización de las anteriores medidas de posicionamiento, y se usan para dividir la muestra en 100 partes iguales.
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. Son más usados en el caso de variables continuas.
- 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.

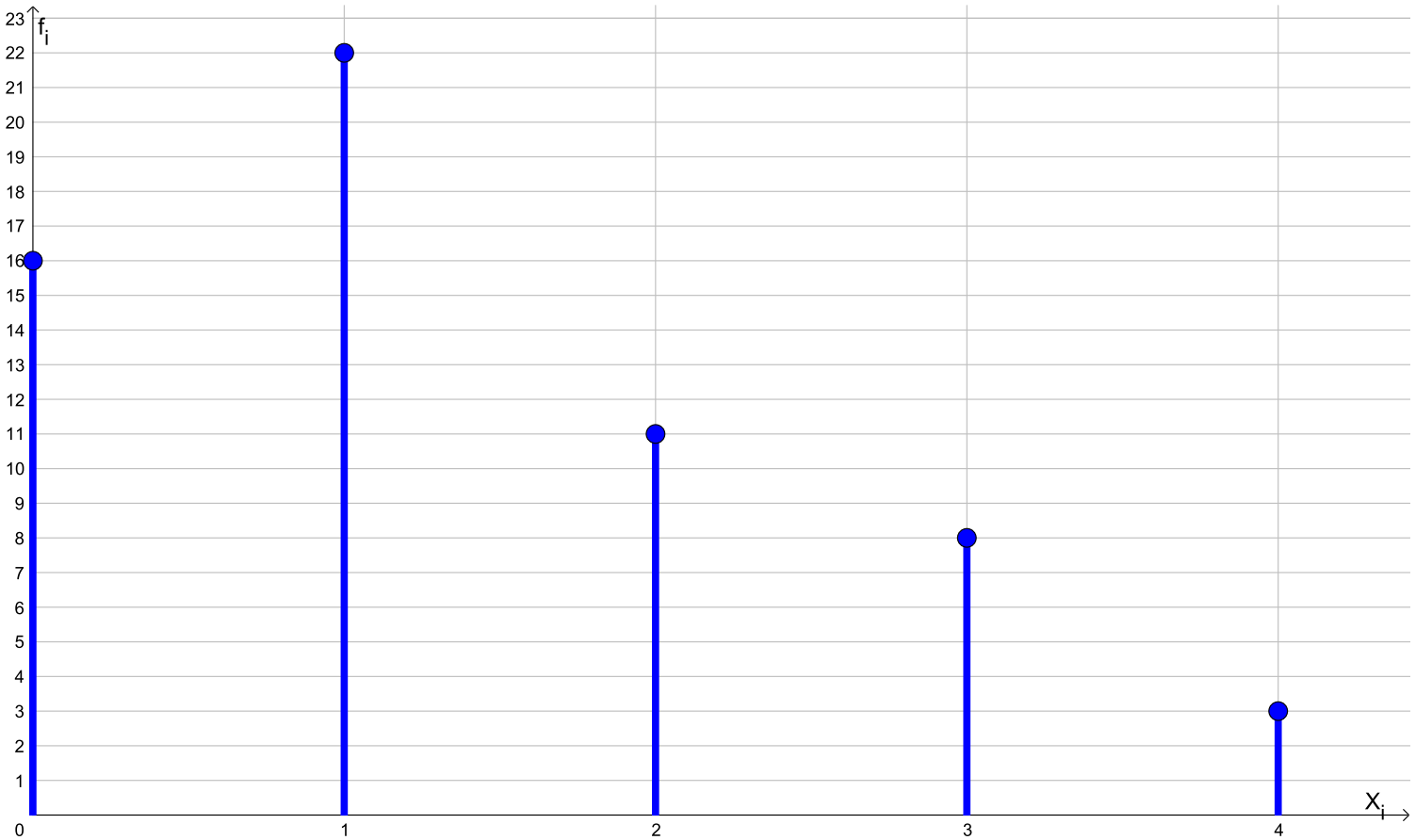
Parámetros Estadísticos: Medidas de Posicionamiento (III)

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 $P_{20} = 0$.

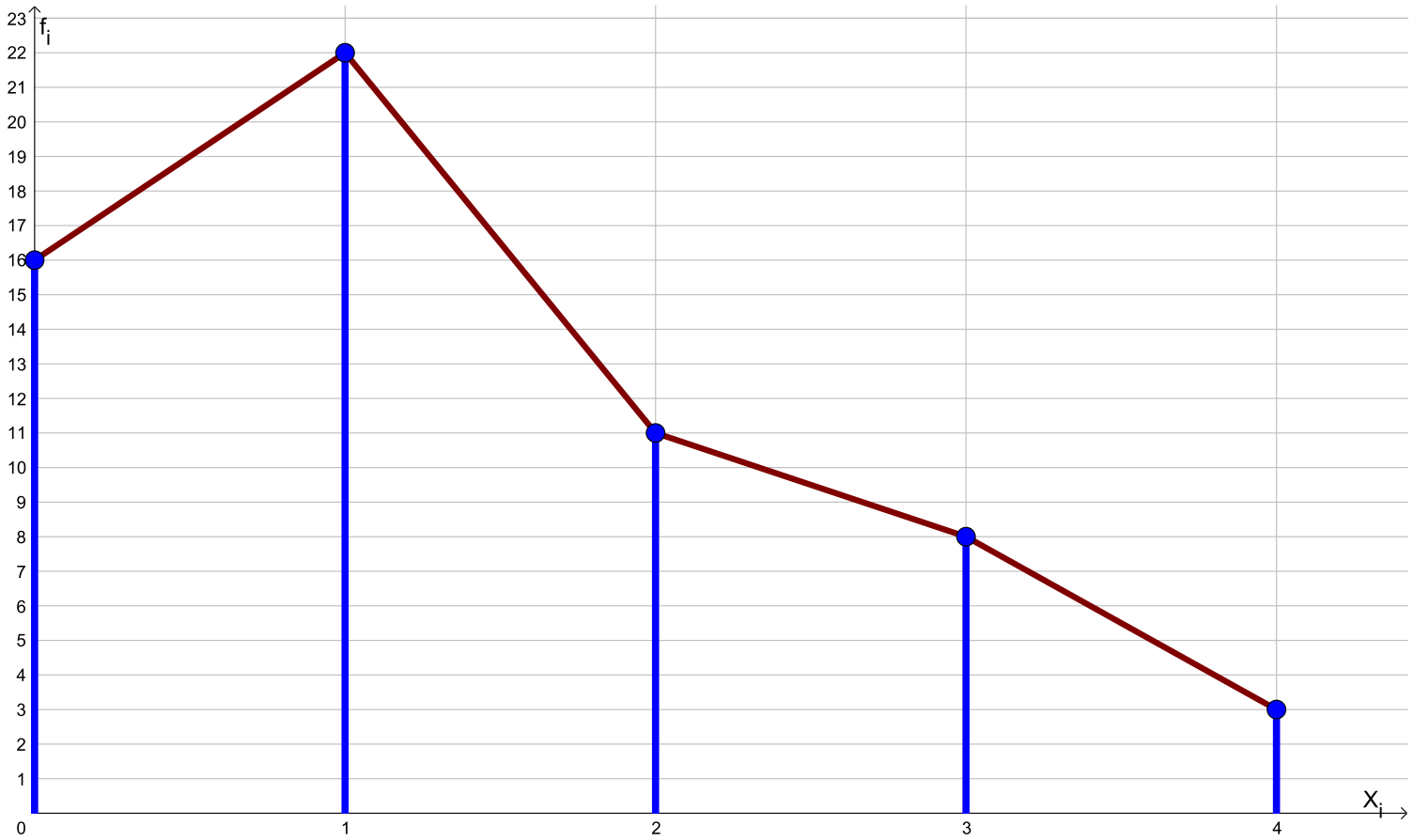
Representaciones gráficas de variables cuantitativas discretas

- 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.
- Polígonos de frecuencias ordinarias (absolutas o relativas): Se construyen a partir de los diagramas de barras, uniendo los extremos superiores de los segmentos.
- Polígonos de frecuencias acumuladas (absolutas o relativas).
- 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.
- Diagrama de cajas y bigotes.

Diagrama de barras



Polígono de frecuencias ordinarias



Polígono de frecuencias acumuladas

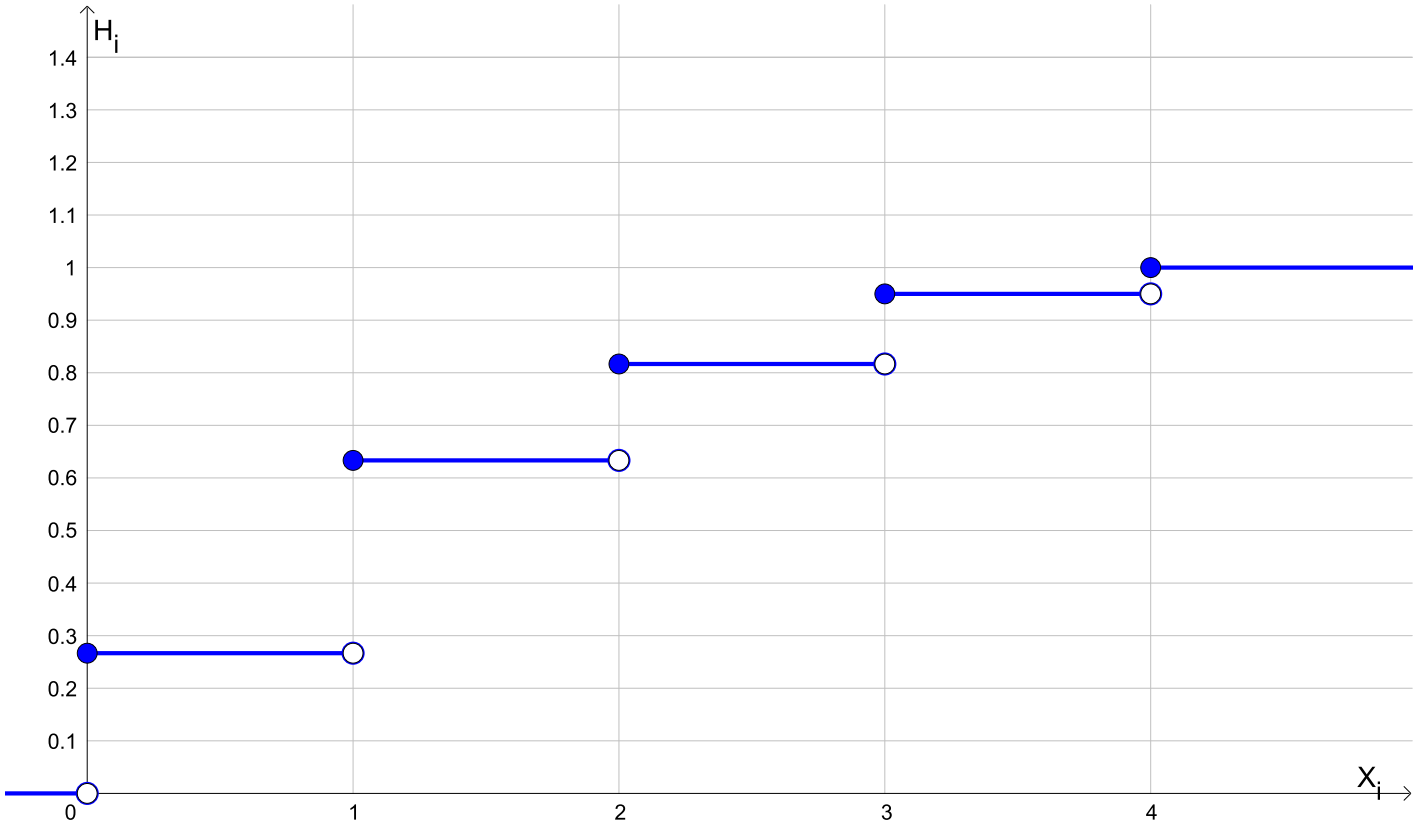


Diagrama de cajas y bigotes (I)

- 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.
- La 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. Tienen tienen un límite de prolongación de $1'5$ veces el rango intercuartílico, de modo que cualquier dato que no se encuentre dentro de este rango es marcado e identificado individualmente.

Diagrama de cajas y bigotes (II)

Ejemplo 3

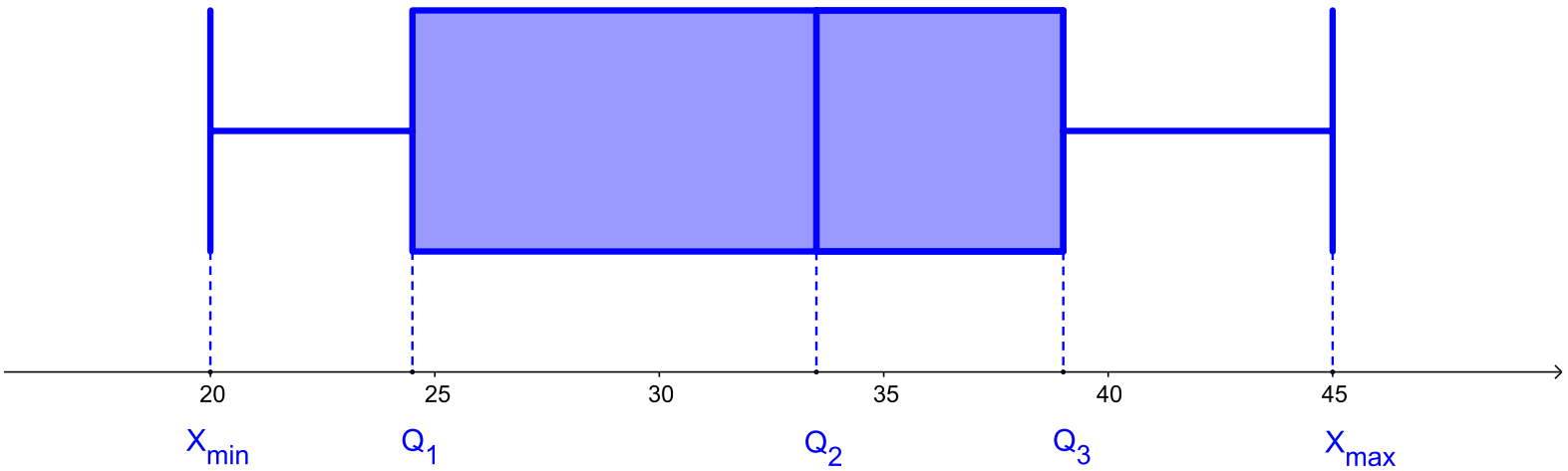
Supongamos que queremos estudiar cómo se distribuye la edad entre los 20 trabajadores de una empresa. Realizamos una encuesta, y nos encontramos con que estas son las edades:

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

Diagrama de cajas y bigotes (III)

- El valor mínimo de la muestra es 20, y el máximo 45.
- $Q_1 = \frac{x_5 + x_6}{2} = \frac{24 + 25}{2} = 24'5$
- $Q_2 = \frac{x_{10} + x_{11}}{2} = \frac{33 + 34}{2} = 33'5$
- $Q_3 = \frac{x_{15} + x_{16}}{2} = \frac{39 + 39}{2} = 39$
- 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 (porque no hay datos por debajo de $Q_1 - 21'75 = 2'75$, ni por encima de $Q_3 + 21'75 = 60'75$)

Diagrama de cajas y bigotes (IV)



$Q_1 = 24.5$
$Q_2 = 33.5$
$Q_3 = 39$
Rango intercuartílico : $Q_3 - Q_1 = 14.5$

Diagrama de cajas y bigotes (V)

Analizando el diagrama vemos que:

- La parte izquierda de la caja es mayor que la de la derecha: las edades comprendidas entre el 25 % y el 50 % de la población están más dispersas 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.

Tratamiento de Variables Cuantitativas Continuas

- Cuando una variable cuantitativa es continua (o si es discreta pero toma un amplio rango de valores), es necesario agrupar los datos en intervalos, para facilitar tanto los cálculos como la representación gráfica.
- En las tablas de frecuencias se incluirá una columna con los intervalos o clases.
- Se asumirá que todas las medidas pertenecientes a un intervalo dado, son 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.

Tratamiento de Variables Cuantitativas Continuas

Ejemplo 4

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.

Tabla de Frecuencias (I)

Supongamos que tenemos la siguiente 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
TOTAL	–	100	–	1	––

Tabla de Frecuencias (II)

Para calcular cómodamente los parámetros estadísticos de la media, la varianza y la desviación típica, debemos añadir columnas a la tabla.

l_i	X_i	f_i	F_i	h_i	H_i	$X_i \cdot f_i$	X_i^2	$X_i^2 \cdot f_i$
$[1'50, 1'60)$	$1'55$	2	2	$0'02$	$0'02$	$3,1$	$2'40$	$4'81$
$[1'60, 1'70)$	$1'65$	10	12	$0'10$	$0'12$	$16,5$	$2'72$	$27'23$
$[1'70, 1'80)$	$1'75$	38	50	$0'38$	$0'50$	$66'5$	$3'06$	$116'38$
$[1'80, 1'90)$	$1'85$	35	85	$0'35$	$0'85$	$64'75$	$3'42$	$119'79$
$[1'90, 2'00)$	$1'95$	10	95	$0'10$	$0'95$	$19'5$	$3'80$	$38'03$
$[2'00, 2'10]$	$2'05$	5	100	$0'05$	1	$10'25$	$4'20$	$21'01$
TOTAL	–	100	–	1	––	$180'60$	–	$327'23$

Medidas de centralización

- Media:
  $\overline{X} = \frac{\sum_{i=1}^{n} X_i \cdot f_i}{100} = \frac{180'60}{100} = 1'81$
- Moda: Se calcula gráficamente, utilizando el histograma de frecuencias ordinarias.
  Se llama clase o intervalo modal a aquel con mayor frecuencia.
- Mediana: Se calcula gráficamente, utilizando el polígono de frecuencias acumuladas.

Medidas de dispersión

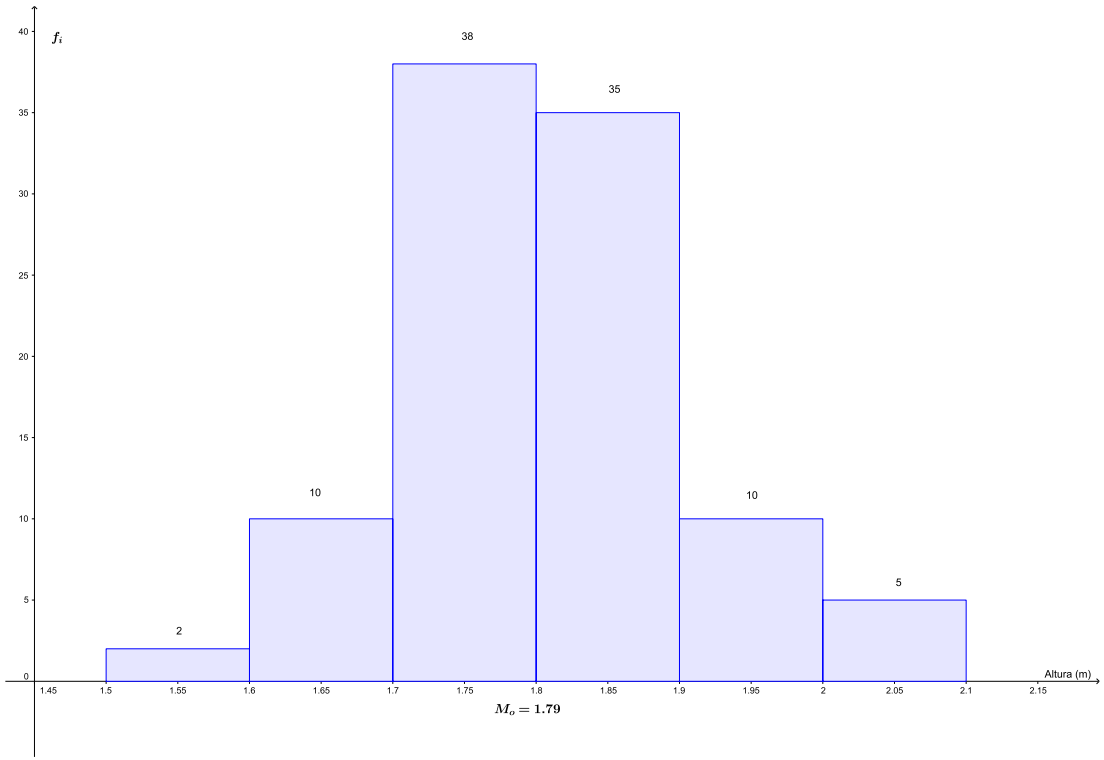
- Rango: $2'10 - 1'50 = 0'60$ m
- Varianza:
  $s^2 = \frac{\sum_{i=1}^{n} X_i^2 \cdot f_i}{100} - \overline{X}^2 = \frac{327'23}{100} - 1'81^2 = 0'01$
- Desviación típica:
  $s = \sqrt{s^2} = \sqrt{0'01} = 0'1$

Medidas de posicionamiento

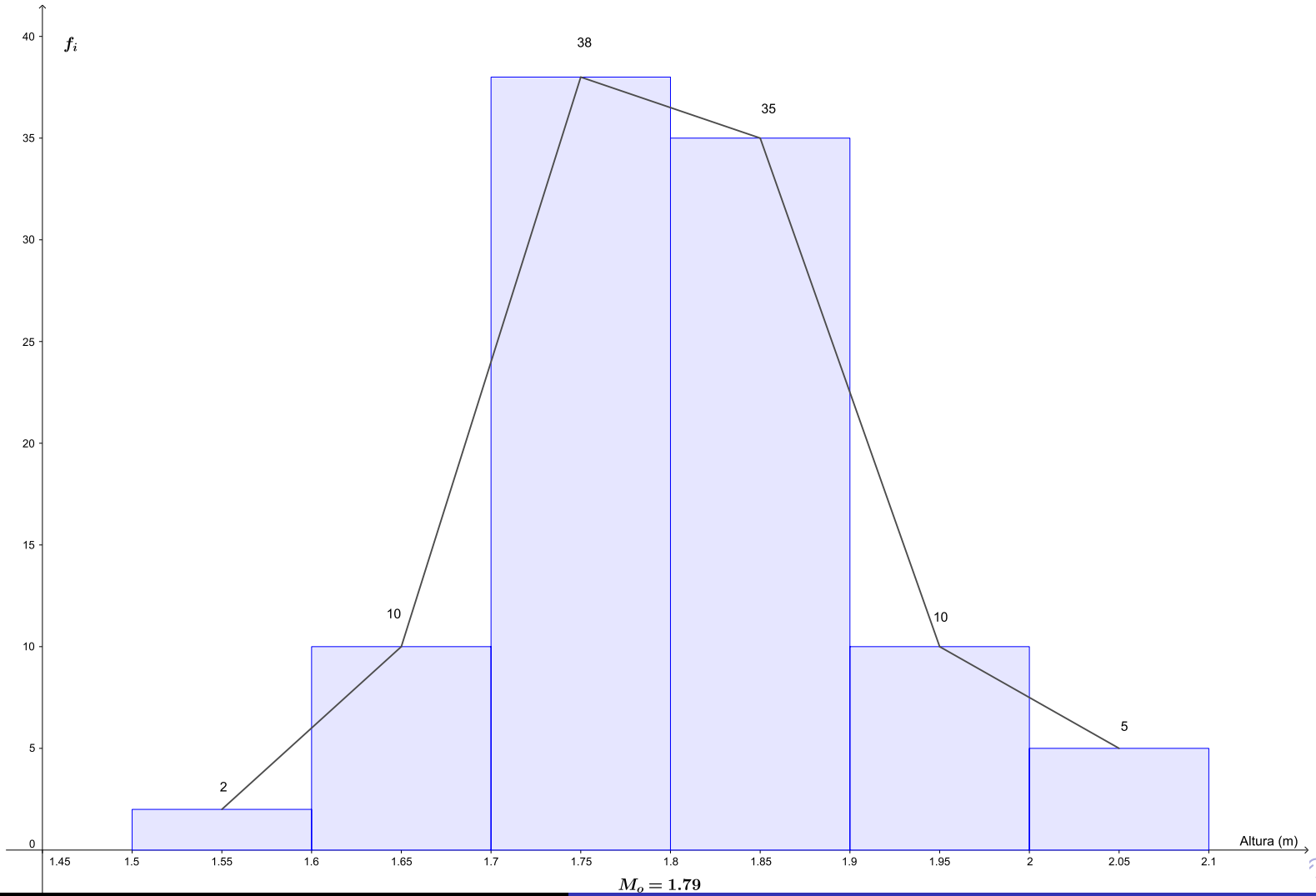
Los cuartiles se calculan gráficamente, utilizando el polígono de frecuencias acumuladas.

Representación gráfica (I): Histograma

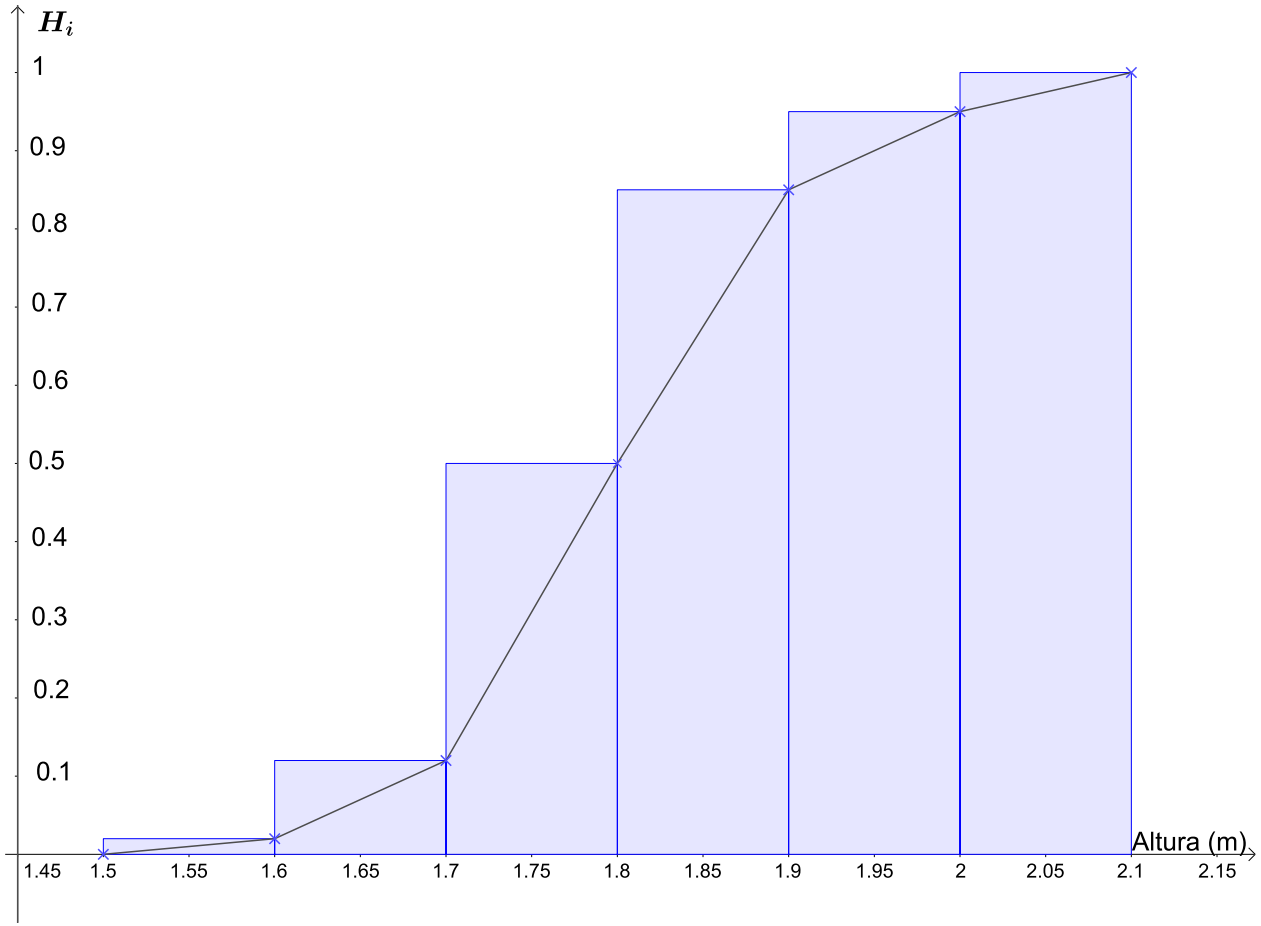
El área de cada rectángulo debe ser directamente proporcional a la frecuencia del intervalo.



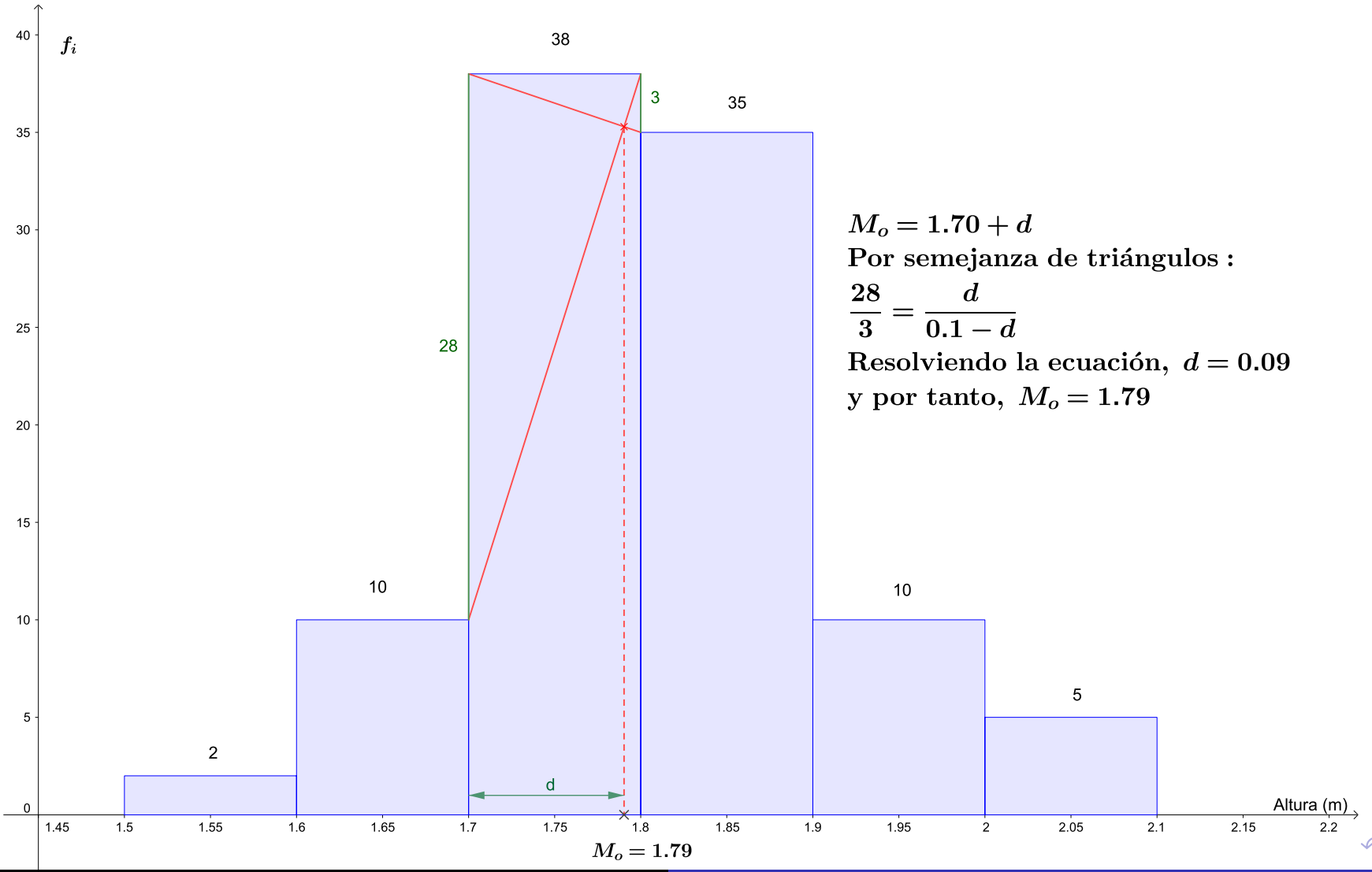
Representación gráfica (II): Polígono de Frecuencias



Representación gráfica (III): Polígono de Frecuencias Acumuladas



Cálculo gráfico de la moda



$M_o = 1.70 + d$

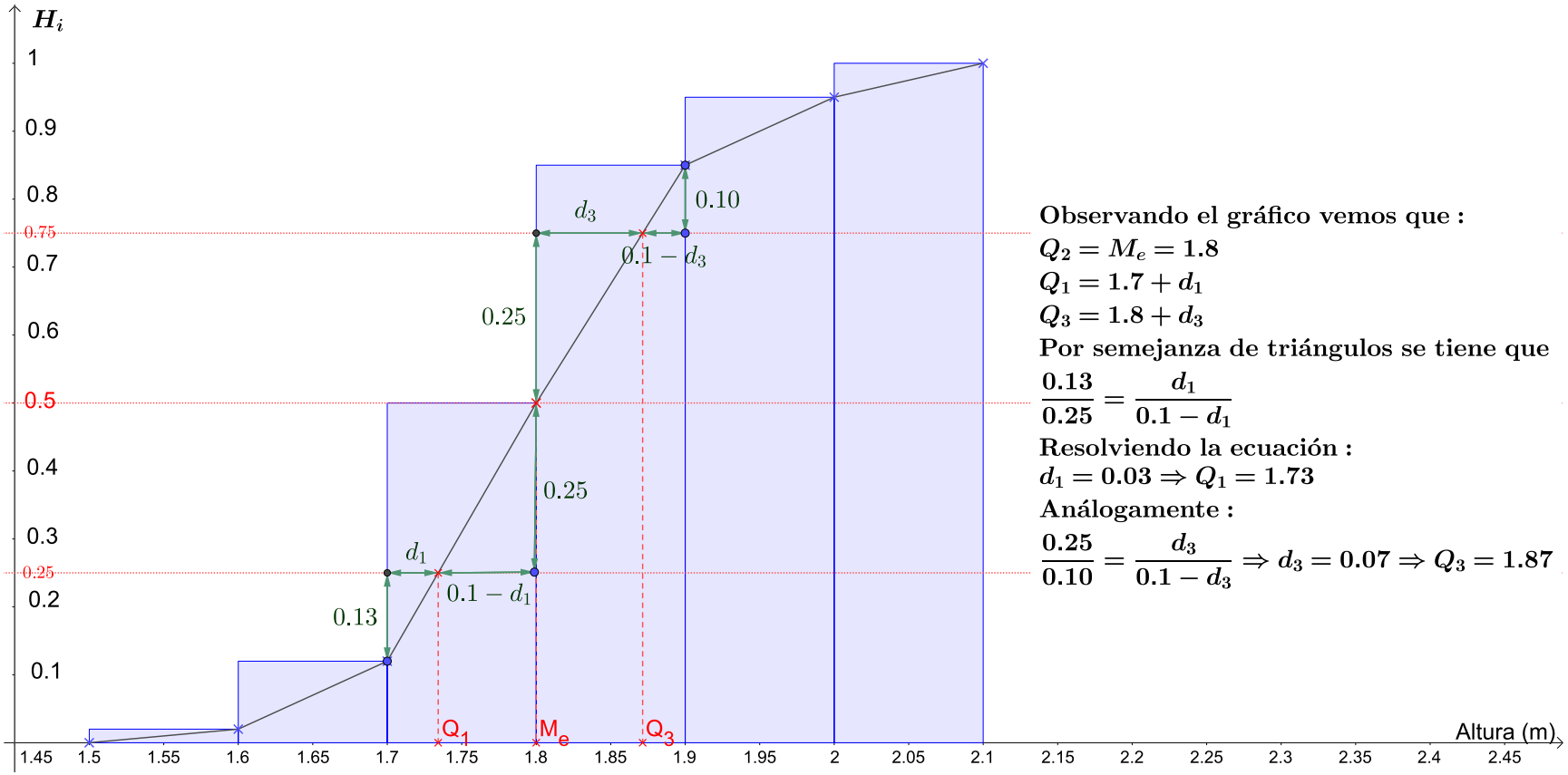
Por semejanza de triángulos :

$$\frac{28}{3} = \frac{d}{0.1 - d}$$

Resolviendo la ecuación, $d = 0.09$

y por tanto, $M_o = 1.79$

Cálculo gráfico de mediana y cuartiles



Observando el gráfico vemos que :

$Q_2 = M_e = 1.8$

$Q_1 = 1.7 + d_1$

$Q_3 = 1.8 + d_3$

Por semejanza de triángulos se tiene que

$$\frac{0.13}{0.25} = \frac{d_1}{0.1 - d_1}$$

Resolviendo la ecuación :

$d_1 = 0.03 \Rightarrow Q_1 = 1.73$

Análogamente :

$$\frac{0.25}{0.10} = \frac{d_3}{0.1 - d_3} \Rightarrow d_3 = 0.07 \Rightarrow Q_3 = 1.87$$

Propiedades de los parámetros estadísticos (I)

- 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 tipos de asimetrías en la distribución de los datos.

Propiedades de los parámetros estadísticos(II)

- 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.
- La desviación típica tiene las mismas propiedades que la varianza. Sus unidades coinciden con las unidades de la media.

El Coeficiente de Variación de Pearson

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

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

- 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 $CV = 100 \cdot \frac{s}{|\overline{x}|}$, y expresarlo de modo porcentual.