EDIP: Relación 1

Integrantes del grupo ordenados alfabéticamente por apellidos:
Shao Jie Hu Chen
Adrián Jaén Fuentes
Aarón Jerónimo Fernández
Noura Lachhab Bouhmadi
Laura Lázaro Soraluce

Problema 1

El número de hijos de las familias de una determinada barriada de una ciudad es una variable estadística de la que se conocen los siguientes datos:

x_i	n_i	N_i	f_i
0	80		0.16
1	110		
2		320	
3			0.18
4	40		
5			
6	20		

n_i : frecuencias absolutas
N_i : frecuencias absolutas acumuladas
f_i : frecuencias relativas

1. Completar la tabla de frecuencias.
2. Representar la distribución mediante un diagrama de barras y la curva de distribución.
3. Promediar los valores de la variable mediante diferentes medidas. Interpretarlas.

Apartado 1

Completar la tabla de de frecuencias.

En una poblacion de tamaño $n = 500$ familias de una determinada barriada, se ha observado una variable estadística $X = numero$ de hijos de cada familia que ha presentado $k = 7$ $(0, 1, 2, 3, 4, 5, 6)$ modalidades distintas con la siguiente distribución de frecuencias $(x_i\ ,\ n_i)$

x_i	n_i	N_i	f_i
0	80	80	0.16
1	110	190	0.22
2	130	320	0.26
3	90	410	0.18
4	40	450	0.08
5	30	480	0.06
6	20	500	0.04
		=500	=1

Los cálculos realizados para rellenar esta tabla se detallan a continuación:

$N_0 = n_0 = 80$

$f_0 = \frac{n_0}{n} \rightarrow n = \frac{n_0}{f_0} = \frac{80}{16} = 500$

$N_2 = n_0 + n_1 + n_2 = 320 \rightarrow n_2 = 320 - 190 = 130$

$f_2 = \frac{n_2}{n} = \frac{130}{500} = 0.26$

$f_3 = \frac{n_3}{n} \rightarrow n_3 = 500 * 0.18 = 90$

$N_3 = N_2 + n_3 = 320 + 90 = 410$

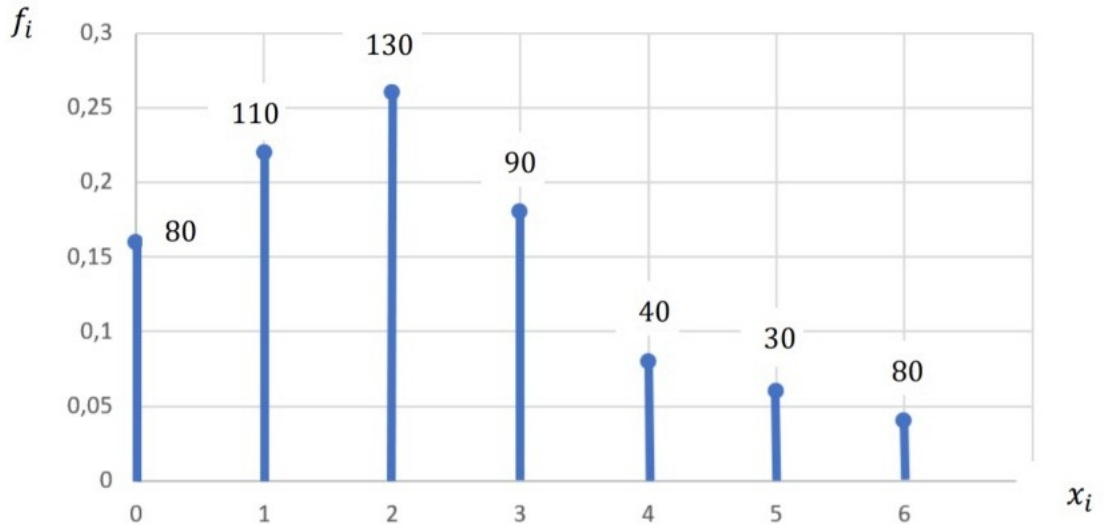
$N_5 = N_4 + n_5 \rightarrow n_5 = 500 - n_0 - n_1 + n_2 - n_3 + n_4 + n_6 = 30$

$N_5 = N_4 + n_5 = 450 + 30 = 480$

Apartado 2

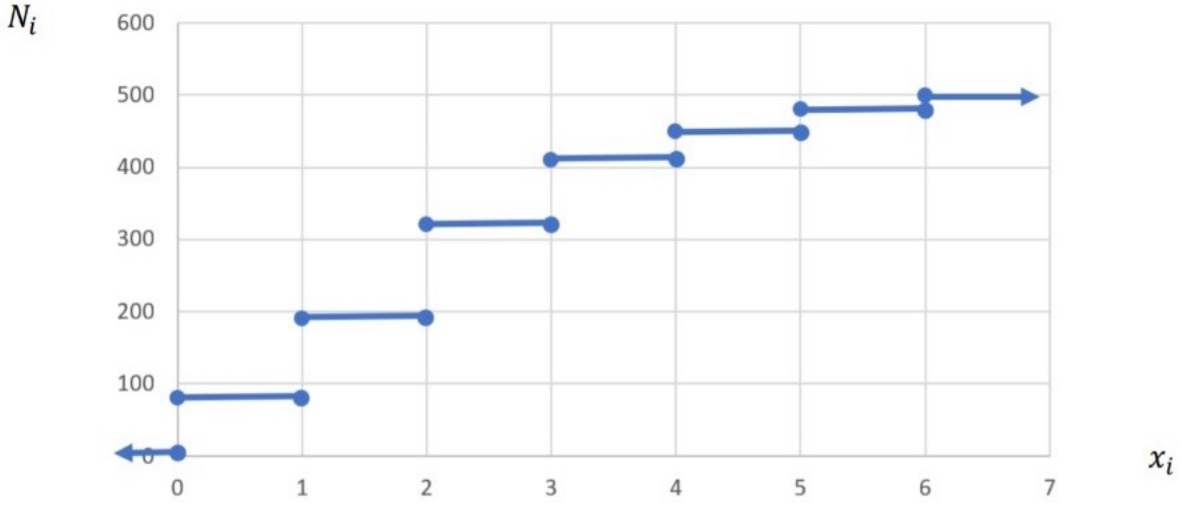
Representar la distribución mediante un diagrama de barras y la curva de distribución.

DIAGRAMA DE BARRAS:



Para el diagrama de barras en el eje X tenemos las modalidades (x_i) y en el eje Y tenemos las frecuencias relativas acumuladas de cada una de las modalidades (x_i).

CURVA DE DISTRIBUCIÓN:



Para la curva de distribucion en el eje X tenemos las modalidades (x_i) y en el eje Y tenemos las frecuencias absolutas acumuladas de cada una de las modalidades (x_i).

Apartado 3

Promediar los valores de la variable mediante diferentes medidas. Interpretarlas.

-MEDIA: en este calcularemos la única que tenga sentido. En este caso es la media aritmética:

$\bar{x} = \frac{1}{n} \sum n_i * x_i$

$\bar{x} = \frac{1}{500} * (110 + 260 + 270 + 160 + 150 + 120) = 2.14$ hijos

-MEDIANA: que es el valor que divide a los individuos de la población en dos efectivos iguales, y para ello buscamos el primer valor cuya frecuencia acumulada sea mayor o igual a $\frac{n}{2}$.

$Me = x_i : N_i > \frac{n}{2}$

$\frac{500}{2} = 250$

$N_3 > 250 \Rightarrow Me = 2hijos$

-Moda: es el valor de mayo frecuencia.

$Mo = x_i : x_i >= n_j \forall j = 1, .., 7 \Rightarrow Mo = 2hijos$

→La mediana y la media coinciden en este caso.

Problema 2

La puntuación obtenida por 50 personas que se presentaron a una prueba de selección, sumadas las puntuaciones de los distintos tests, fueron:

$$174, 185, 166, 176, 145, 166, 191, 175, 158, 156, 156, 187, 162, 172, 197, 181, 151,$$
$$161, 183, 172, 162, 147, 178, 176, 141, 170, 171, 158, 184, 173, 169, 162, 172, 181,$$
$$187, 177, 164, 171, 193, 183, 173, 179, 188, 179, 167, 178, 180, 168, 148, 173.$$

1. Agrupar los datos en intervalos de amplitud 5 desde 140 a 200 y dar la tabla de frecuencias.
2. Representar la distribución mediante un histograma, poligonal de frecuencias y curva de distribución.

POBLACIÓN: Las personas que se presentaron a la prueba de selección
TAMAÑO: 50
MODALIDADES: Los intervalos que contienen las notas que han obtenido las personas en la prueba

Apartado 1

Agrupar los datos en intervalos de amplitud 5 desde 140 a 200 y dar la tabla de frecuencias.

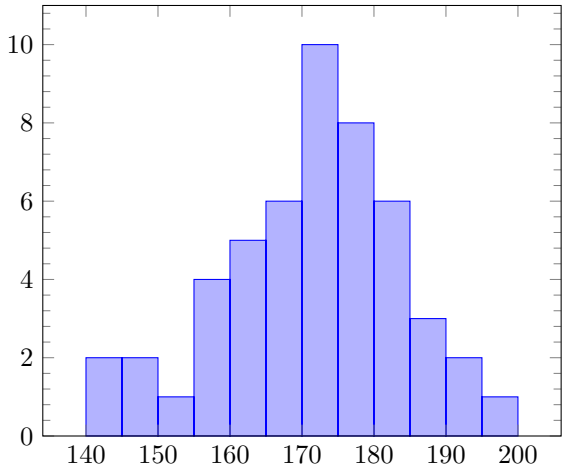
x_i	n_i	N_i	f_i	F_i
(140-145]	2	2	0,04	0,04
(145-150]	2	4	0,04	0,08
(150-155]	1	5	0,02	0,1
(155-160]	4	9	0,08	0,18
(160-165]	5	14	0,1	0,28
(165-170]	6	20	0,12	0,4
(170-175]	10	30	0,2	0,6
(175-180]	8	38	0,16	0,76
(180-185]	6	44	0,12	0,88
(185-190]	3	47	0,06	0,94
(190-195]	2	49	0,04	0,98
(195-200]	1	50	0,02	1

Establecemos los intervalos de amplitud 5, empezando con 140 y terminando con 200. Para cada intervalo, contamos cuantas personas han obtenido una puntuación que esté dentro de dicho intervalo. Estas son las frecuencias absolutas (n_i). Para la frecuencia absoluta acumulada (N_i), sumamos la frecuencia absoluta de la modalidad que estamos analizando a la frecuencia absoluta acumulada de la modalidad anterior. O lo que es lo mismo, sumamos las frecuencias absolutas de todas las modalidades hasta la que estamos analizando: $N_i = n_1 + n_2 + ... + n_i$. Para frecuencia relativa, dividimos cada n_i entre n, es decir, entre el número total de individuos de la población, la frecuencia absoluta acumulada de la última modalidad. Para la frecuencia relativa acumulada hacemos lo mismo que para la absoluta acumulada pero con la frecuencia relativa: sumamos las frecuencias relativas de todas las modalidades hasta la que estamos analizando: $F_i = f_1 + f_2 + ... + f_i$.

Apartado 2

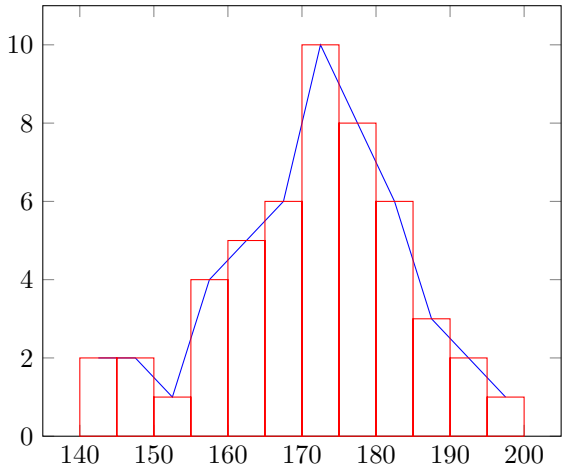
Representar la distribución mediante un histograma, poligonal de frecuencias y curva de distribución.

Histograma



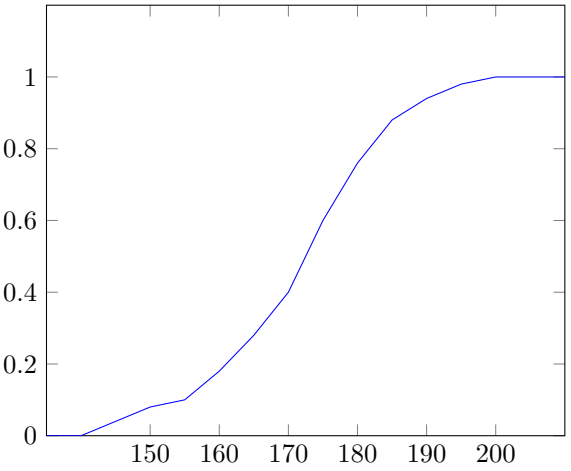
Para el histograma, en el eje X tenemos las modalidades (x_i) y en el eje Y un múltiplo de h_i ($h_i = \frac{fi}{a}$). En este caso, las amplitudes de todos los intervalos son iguales: 5.

Poligonal de frecuencias



Para el poligonal de frecuencias, los ejes son los mismos que para el histograma. En este caso, unimos los puntos que corresponden a las marcas de clase de los intervalos en el histograma. Para hallar dichas marcas de los intervalos, aplicamos la siguiente expresión: $c_i = \frac{e_{i-1}+e_i}{2}$.

Curva de distribución



En la curva de distribución, en el eje X tenemos los extremos superiores de los intervalos (e_i) y en el eje Y, $F(e_i)$. Pintamos la curva que une los puntos tal que $F(e_i) = \sum_{j=1}^{i} f_j = F_i$. En este caso tenemos que la curva es continua. $F(e_i)$ es 0 para valores menores que e_1 y 1 para valores mayores que el último intervalo e_k.

Problema 3

La distribución de la renta familiar en el año 2003 por comunidades autónomas se recoge en la siguiente tabla:

I_i	n_i	N_i	f_i	F_i	c_i	a_i	h_i
$(8300, 9300]$	2						
$, 10200]$		5					
				10/18		1100	
			2/18		12000		
	4						0.005/18
		18					0.002/18

n_i : frecuencias absolutas
N_i : frec. absolutas acumuladas
f_i : frecuencias relativas
F_i : frec. relativas acumuladas
c_i : marcas de clase
a_i : amplitudes
h_i : densidades de frecuencia

1. Completar la tabla.
2. Representar la distribución mediante un histograma, poligonal de frecuencias y curva de distribución.
3. ¿Cuántas comunidades presentan una renta menor o igual que 12700 euros? ¿Y cuántas superior a 11300 euros?

Apartado 1

Completar la tabla

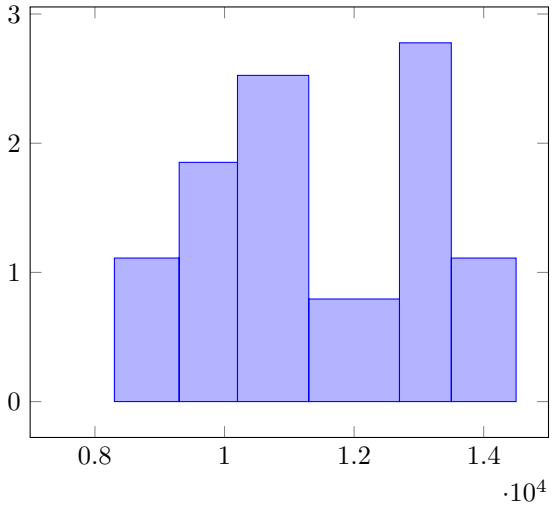
I	n_i	N_i	f_i	F_i	c_i	ai	h_i
(8300, 9300]	2	2	2/18	2/18	8800	1000	0.002/18
(9300, 10200]	3	5	3/18	5/18	9750	900	1/5400
(10200, 11300]	5	10	5/18	10/18	10750	1100	1/3960
(11300, 12700]	2	12	2/18	12/18	12000	1400	1/12600
(12700, 13500]	4	16	4/18	16/18	13100	800	0.005/18
(13500, 14500]	2	18	2/18	1	14000	1000	0.002/18

Apartado 2

Representar la distribución mediante un histograma, poligonal de frecuencias y curva de distribución.

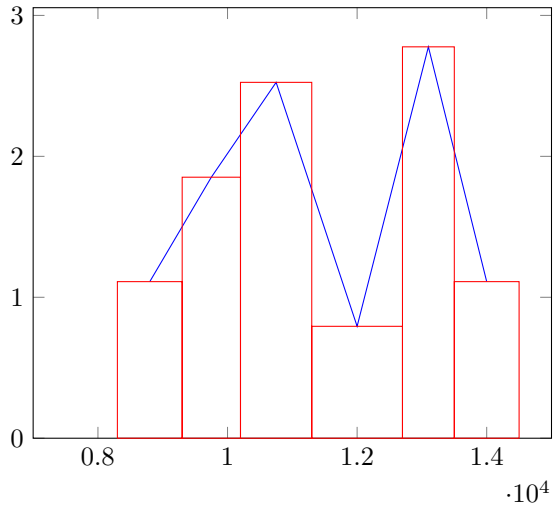
En el histograma se representan los I_i en el eje X y reescalando h_i multipliplicándolo por 10.000, para poder tener una mejor representación en el eje Y.

Histograma



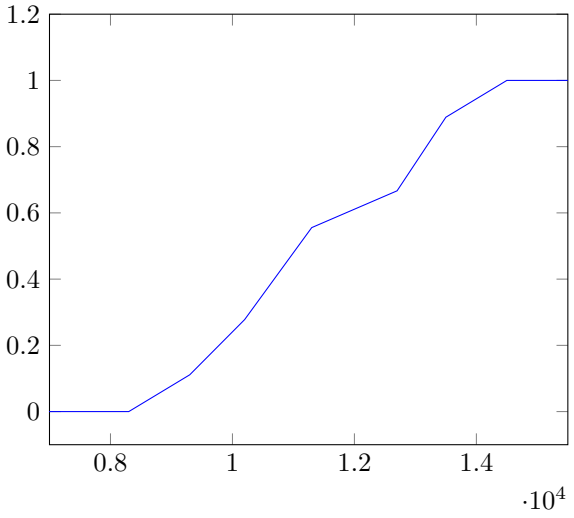
En la poligonal de frecuencias unimos los "techos", es decir las marcas de clase a_i, de cada intervalo I_i.

Poligonal de frecuencias



En la curva de distribución representamos los I_i en el eje X y las frecuencias absolutas acumuladas N_i en el eje Y.

Curva de distribución



Apartado 3

¿Cuántas comunidades presentan una renta menor o igual que 12700 euros? ¿Y cuántas superior a 11300 euros?

Como podemos ver en la tabla, las comunidades que presentan una renta menor o igual a 12700 corresponden con N_4, es decir, 12 comunidades. Las comunidades con una renta superior a 11300 corresponderían con la suma de las comunidades de los intervalos superiores, es decir, a la suma de n_4, n_5 y n_6. Finalmente, el resultado a esta última pregunta es 8 comunidades.

Problema 4

En una determinada empresa se realiza un estudio sobre la calidad de su producción. La distribución siguiente informa sobre el número de piezas defectuosas encontradas en 100 cajas examinadas con 50 unidades cada una de ellas:

Nº piezas defectuosas	0	1	2	3	4	5	6	7	8	9	10
Nº de cajas	6	9	10	11	14	16	16	9	4	3	2

1. Calcular el número medio de piezas defectuosas por caja.
2. ¿Cuantas piezas defectuosas se encuentran más frecuentemente en las cajas examinadas?
3. ¿Cuál es el número mediano de piezas defectuosas por caja?
4. Calcular los cuartiles de la distribución. Interpretarlos.
5. Calcular los deciles de orden 3 y 7. Interpretarlos.
6. Cuantificar la dispersión de la distribución utilizando diferentes medidas, interpretando los resultados y señalando las ventajas e inconvenientes de cada una.

En los lotes de 50 cajas de los productos producidos por una determinada empresa producidos durante un determinado período, con un total de $n = 100$ cajas, se ha observado el número de piezas defectuosas, presentando ésta un total de $k = 11$ modalidades, cuya distribución de frecuencias viene representado en la siguiente tabla:

x_i	n_i	N_i	$x_i n_i$	$n_i\|x_i - \bar{x}\|$	$n_i\|x_i - Me\|$	$n_i(x_i - \bar{x})^2$
0	6	6	0	26,16	27,00	114,058
1	9	15	9	30,24	31,50	101,606
2	10	25	20	23,60	25,00	55,696
3	11	36	33	14,96	16,50	20,346
4	14	50	56	5,04	7,00	1,814
5	16	66	80	10,24	8,00	6,554
6	16	82	96	26,24	24,00	43,034
7	9	91	63	23,76	22,50	62,726
8	4	95	32	14,56	14,00	52,998
9	3	98	27	13,92	13,50	64,589
10	2	100	20	11,28	11,00	63,619

Las últimas tres columnas se han obtenido a partir de los cálculos de los apartados 1 y 3.

Apartado 1

Calcular el número medio de piezas defectuosas por caja.

La variable estadística observada es discreta, presentando un total de 11 modalidades: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 y 10. En este caso, de las medias de las que disponemos, la única con significado estadístico es la aritmética, cuya expresión matemática viene determinada por:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

De la tabla podemos obtener, teniendo también en consideración que $n = 100$ cajas, que $\bar{x} = 4,36 \approx 4$ piezas. Esto es, en cada lote examinado, cabe esperar un total de 4 piezas defectuosas de media, aunque para cálculos empleamos 4,36 piezas.

Apartado 2

¿Cuantas piezas defectuosas se encuentran más frecuentemente en las cajas examinadas?

La medida estadística que proporciona información sobre el número de piezas defectuosas que más habitualmente se puede esperar de una caja es la moda. Se define como $Mo = max(x_i)$, con $i = 1, 2, 3, ..., k$. En este caso, se presentan dos modalidades que presentan la misma frecuencia absoluta. Por tanto, podemos afirmar que hay dos valores de moda en esta distribución, siendo

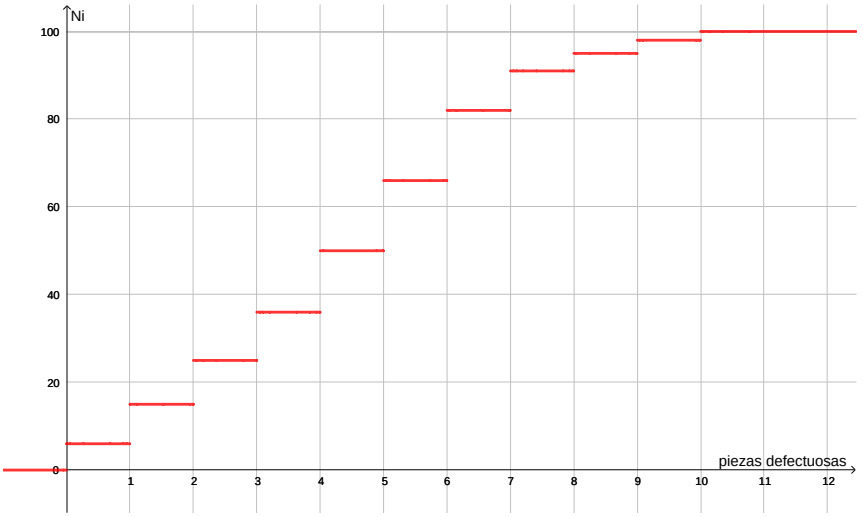
más frecuente encontrar un total de 5 y 6 piezas defectuosas por lote examinado.

Apartado 3

¿Cuál es el número mediano de piezas defectuosas por caja?

La mediana se define como un valor que divide a una determinada población en dos subgrupos con el mismo número de individuos: una de ellas cuyos individuos pertenecen a una modalidad cuya frecuencia absoluta es inferior a la mediana y otra con frecuencias absolutas superiores.

Para determinar la mediana en esta distribución de frecuencias, representamos la curva de distribución de la variable estadística:



Observamos que $\frac{n}{2} = 50$, coincidiendo justamente con el valor $F(x_5) = n/2$. Conviene pues tomar la media aritmética entre x_5 y x_6, de donde $Me = 4,5$ piezas defectuosas.

Apartado 4

Calcular los cuartiles de la distribución. Interpretarlos.

El cuartil es una medida que permite caracterizar el porcentaje de individuos de la población cuya variable estadística observada es inferior al $25\%, 50\%, 75\%$ de toda la población, respectivamente. Su cálculo es análogo al de la media (únicamente hay que sustituir $\frac{n}{2}$ por las fracciones correspondientes). De esta forma, deducimos las siguientes medidas:

- $Q_1 = 2,5$ piezas. Representa que un 25% de los lotes analizados poseen una cantidad de productos defectuosos inferiores o iguales a 2,5 piezas.
- $Q_2 = 4,5$ piezas. Representa que un 50% de los lotes analizados poseen una cantidad de productos defectuosos inferiores o iguales a 4,5 piezas.
- $Q_3 = 6$ piezas. Representa que un 75% de los lotes analizados poseen una cantidad de productos defectuosos inferiores o iguales a 6 piezas.

Apartado 5

Calcular los deciles de orden 3 y 7. Interpretarlos.

El decil tiene un significado análogo a la de los cuartiles. Por tanto, el decil es una medida que permite caracterizar el porcentaje de individuos de la población cuya variable estadística observada es inferior al $10\%, 20\%, ..., 90\%$ de toda la población. Su cálculo es, nuevamente, análogo a los casos anteriores:

- $D_3 = 3$ piezas. Representa que un 30% de los lotes analizados poseen una cantidad de productos defectuosos inferiores o iguales a 3 piezas.
- $D_7 = 6$ piezas. Representa que un 70% de los lotes analizados poseen una cantidad de productos defectuosos inferiores o iguales a 6 piezas.

Apartado 6

Cuantificar la dispersión de la distribución utilizando diferentes medidas, interpretando los resultados y señalando las ventajas e inconvenientes de cada una.

Algunas medidas de dispersión que podemos tomar sobre la variable estadística observada son las siguientes:

- Recorrido. Es la magnitud que determina la anchura total de la muestra tomada (esto es, toma la diferencia entre el valor más alto tomado de la variable y el menor valor observado). Su ventaja es que tiene un significado concreto, pero su inconveniente es que varía mucho con fluctuaciones muestrales. En este caso, su valor es $R = 10$ piezas. Esto quiere decir que hay una variación de 10 productos defectuosos entre los datos de la muestra.
- Recorrido intercuartílico. Es la magnitud que indica la longitud del intervalo que contiene al 50% central de los datos observados. Su virtud es que no varía mucho con fluctuaciones muestrales, pero su inconveniente es que no toma en consideración todos los datos de la población observada. En este caso, su valor es $R_I = 3,5$ piezas. Ello quiere decir que, del 50% central de la muestra, hay, a lo sumo, una diferencia de 3,5 piezas defectuosas entre ellos.
- Desviación absoluta media respecto a la media aritmética. Indica cómo están los datos distribuidos en función del valor promedio (la media aritmética). Su ventaja es que tiene un significado preciso, pero presenta el inconveniente de que varía mucho en función de fluctuaciones estadísticas (puesto que en su cálculo interviene la media aritmética). Su valor es, en este caso, $D_{\bar{x}} = 2$ piezas. Representa que los individuos de la población muestran una diferencia de media de 2 piezas defectuosas respecto a la media aritmética.
- Desviación absoluta media respecto a la mediana. Es la medida que representa la media de las distancias de los valores presentados durante la muestra respecto al 50%. Presenta la ventaja de que tiene un significado preciso, pero no es fácil de calcular. Su valor es $D_{Me} = 2$ piezas, en este caso. Esto significa que los individuos presentan, respecto a la mediana, una diferencia de 2 piezas defectuosas.
- Varianza. Es una medida que representa la variabilidad de una serie de datos respecto a su media. Presenta la ventaja de que en su cálculo intervienen todos las modalidades observadas de la variable estadística, pero su inconveniente es que su valor es sensible a fluctuaciones muestrales. Su valor numérico, en este caso, es $Var(X) = 5,85$ piezas al cuadrado. Dado que sabemos que la varianza está acotada por $\min((x_i - \bar{x})^2) < Var(X) < \max((x_i - \bar{x})^2)$, y sabemos que $\min((x_i - \bar{x})^2) = 0.130$ piezas al cuadrado y que $\max((x_i - \bar{x})^2) = 31,810$ piezas al cuadrado, deducimos que la distribución presenta una dispersión moderada.
- Desviación típica. Es una medida que proporciona información sobre el margen óptimo de diferencia entre los valores medidos de la variable respecto de la media aritmética. Su principal ventaja es que fluctúa poco con variaciones extremas de la muestra, pero tiene el inconveniente de que requiere gran capacidad de cómputo para poder calcularse. Su valor es, en este caso, $\sigma_x = 2,42$ piezas. Su significado es que, respecto a la media, la diferencia óptima que se debería encontrar en la distribución es de 2,42 piezas.
- Recorrido relativo. Informa sobre el recorrido de la variable respecto de su media aritmética. Su ventaja es que es fácil de calcular, pero tiene la desventaja de que varía mucho con las fluctuaciones estadísticas. En este caso, su valor es $R_R = 0,8$. Informa que la amplitud de las modalidades tomadas oscila en torno a 0,8 veces la media arimética.

- Coeficiente de variación. Representa la desviación típica de la muestra en función de la media aritmética. Su principal virtud es que comparar el grado de dispersión entre diferentes variables estadísticas, pero tiene el inconveniente de que no es válida para todas las variables estadísticas (no sería válida para aquellas con $\bar{x} = 0$ u). Su valor es, en este caso, $C.V.(X) = 0,55$. Se trata de una medida adimensional. Razonando análogamente al caso de la varianza, deducimos que la distribución es moderadamente homogénea.

Problema 5

Dadas las siguientes distribuciones:

$I_i^{(1)}$	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
$n_i^{(1)}$	12	13	11	8	6

$I_i^{(2)}$	$(0,1]$	$(1,3]$	$(3,6]$	$(6,10]$	$(10,12]$
$n_i^{(2)}$	1	6	7	12	2

Calcular para cada una de ellas:

1. Medias aritmética, armónica y geométrica.
2. El valor más frecuente.
3. El valor superado por el 50 % de las observaciones.
4. Recorrido, recorrido intercuartílico y desviación típica. Interpretarlos. ¿Qué distribución es más homogénea?

Escribamos las tablas con los datos que nos pueden interesar para la resolución del ejercicio:

$I_i^{(1)}$	$c_i^{(1)}$	$n_i^{(1)}$	$N_i^{(1)}$	$h_i = \frac{n_i}{a_i}$
$(0,1]$	0.5	12	12	12
$(1,2]$	1.5	13	25	13
$(2,3]$	2.5	11	36	11
$(3,4]$	3.5	8	44	8
$(4,5]$	4.5	6	50	6

$I_i^{(2)}$	$c_i^{(2)}$	$n_i^{(2)}$	$N_i^{(2)}$	$h_i = \frac{n_i}{a_i}$
$(0,1]$	0.5	1	1	1
$(1,3]$	2	6	7	3
$(3,6]$	4.5	7	14	2.333
$(6,10]$	8	12	26	3
$(10,12]$	11	2	28	1

Apartado 1

Medias aritmética, armónica y geométrica.

La media aritmética de una variable es la suma de sus valores entre en número total de observaciones. Como los datos están organizados en intervalos de clase, para calcular la media aritmética vamos a suponer que todos los datos de un intervalo son idénticos a la marca de clase de cada intervalo. Por tanto, la media la calculamos de la siguiente manera:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i$$

Para la primera distribución, $n = 50$, luego la media aritmética es 2.16 u.

Para la segunda, $n = 28$, luego la media aritmética será 5.786 u.

La media armónica se usa para promediar datos de magnitudes relativas. La definimos como la inversa de la media aritmética de los valores inversos de la variable (en nuestro caso, usamos las marcas de clase):

$$H = \frac{n}{\frac{n_1}{c_1} + \frac{n_2}{c_2} + ... + \frac{n_k}{c_k}} = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{c_i}}$$

Para la primera distribución: H = 1.229 u.

Para la segunda distribución: H = 3.399 u.

Finalmente la media geométrica se usa cuando se desea promediar datos de una variable que tiene efectos multiplicativos acumulativos en la evolución de una determinada característica con un valor inicial fijo.

Es la raíz n-ésima del producto de los n valores (o marcas de clase) de la distribución:

$$G = \sqrt[n]{\prod_{i=1}^{k} c_i^{n_i}}$$

Para no perder precisión en el resultado, la calcularemos sabiendo que el logaritmo de la media geométrica es la media aritmética de los logaritmos de los valores de la variable:

$$\log G = \log \sqrt[n]{\prod_{i=1}^{k} c_i^{n_i}} = \frac{1}{n}\sum_{i=1}^{k} n_i \log c_i$$

Para la primera distribución: G = 1.685 u.

Para la segunda distribución: G = 4,769 u.

Apartado 2

Se nos pide calcular la moda de las distribuciones, es decir, el valor de mayor frecuencia:

En el caso primero, observamos que se encuentra en el intervalo (1,2], luego haremos un promedio teniendo en cuenta los intervalos contiguos. Para ello apliquemos la interpolación lineal (pues estamos suponiendo que los datos están distribuidos uniformemente), además de suponer que es una distribución continua.

$$M_O = e_{i-1} + \frac{h_i - h_{i-1}}{2h_i - h_{i-1} - h_{i+1}}(e_i - e_{i-1})$$

Por tanto, la moda será: $M_O = 1.\overline{3}$.

En el segundo caso, tenemos dos modas, pues el valor máximo de h_i lo toman dos intervalos. De esta forma, deducimos que existen dos modas: la primera en el intervalo (1,3] y otra en el intervalo (6,10]. En el primer caso, la moda se calcula con expresión anterior empleando $e_{i-1} = 1, e_i = 3, h_{i-1} = 1, h_i = 3, h_{i+1} = 2.333$, deducimos que $M_O = 2.5u$. En el segundo, que proviene del intervalo (6, 10], aplicamos el mismo método teniendo en cuenta que $e_{i-1} = 6, e_i = 10, h_{i-1} = 2.333, h_i = 3, h_{i+1} = 2$, $M_O = 7u$.

Apartado 3

Para calcular la mediana hemos de observar las frecuencias absolutas acumuladas; tendremos que encontrar el punto que divida a la población en dos partes iguales, es decir, $n/2$:

Para la primera distribución: $n/2 = 25$ luego el 50% de la población supera el valor 2.

Para la segunda distribución: $n/2 = 14$ luego el 50% de la población supera el valor 6.

Apartado 4

1. Recorrido: Es la amplitud del intervalo en la que se mueven los valores de la variable, por tanto, se calcula restando el valor mas grande posible menos el más pequeño. Como en nuestro caso tenemos intervalos, cojamos el extremo superior del último intervalo de clase y le restamos el extremo inferior del primer intervalo de clase:

   Para la primera distribución será 5 y para la segunda será 12.

2. Recorrido intercuartílico: Es la magnitud que indica la longitud del intervalo que contiene al 50% central de los datos observados. Para calcularlo necesitamos los cuartiles 1 y 3:

   $$Qk = e_{i-1} + \frac{\frac{n}{k} - N_{i-1}}{n_i}(e_i - e_{i-1})$$

   Para la distribución 1 será $Q_1 = 1.038u, Q_3 = 3.187u; R_I = 2.1u$

   Para la distribución 2 será $Q_1 = 3u, Q_3 = 8.333u; R_I = 5.333u$

3. Desviación típica: es una medida de dispersión, una medida de cómo los valores individuales de el conjunto pueden diferir de la media. Es la raíz cuadrada de la media de las distancias al cuadrado de cada uno de los valores de la variable a la media de la distribución. También se puede calcular como la raíz cuadrada positiva de la varianza.

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

   Para la distribución 1 será $\sigma^2 = 1.744; \sigma = 1.321$

   Para la distribución 1 será $\sigma^2 = 8.522; \sigma = 2.919$

   Para ver cuál es mas homogénea, calculemos el Coeficiente de variación de Pearson:

   $$C.V._1 = \frac{1.321}{2.16} = 0.612$$

   $$C.V._2 = \frac{2.919}{5.786} = 0.504$$

   La segunda distribución es más homogénea.

Problema 6

Un móvil efectúa un recorrido de 100 km en dos sentidos. En uno va a una velocidad constante de V_1=60 km/h y en el otro va a una velocidad constante de V_2=70 km/h. Calcular la velocidad media del recorrido.

Consideremos un cuerpo que se mueve entre dos puntos A y B, alejados una distancia $d > 0$ (en particular, $d \neq 0$), con una velocidad v_1 para ir desde A hasta B y una velocidad v_2 para ir desde B hasta A. La velocidad de un cuerpo es una magnitud física que se define como la rapidez con la

que varía la posición. Su expresión matemática viene dado por: $v = \frac{\Delta s}{\Delta t}$, donde v es la velocidad del cuerpo, s es la distancia recorrida y t es el tiempo transcurrido.

La velocidad media del cuerpo viene determinado por la expresión $v_{media} = \frac{\Delta s}{\Delta t}$. Para determinar dicha velocidad, tengamos en cuenta que el cuerpo recorre dos veces la distancia d, una para ir desde A hasta B y otra para ir desde B hasta A. Por tanto, sabemos que $s = 2d$. Por otra parte, sabemos que el tiempo invertido sería $t = t_1 + t_2$, donde t_1 es el tiempo invertido para ir desde A a B y t_2 es el tiempo invertido para ir desde B hasta A. Por tanto, deducimos que $v_{media} = \frac{2d}{t_1+t_2}$. Obedeciendo a las expresiones $t_1 = \frac{d}{v_1}$ y $t_2 = \frac{d}{v_2}$, además de dividir numerador y denominador por la distancia $d > 0$, obtenemos:

$$v_{media} = \frac{2d}{\frac{d}{v_1} + \frac{d}{v_2}} \Rightarrow v_{media} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}}$$

Por tanto, tenemos que la velocidad media del móvil es la **media armónica** de las dos velocidades. En efecto, la media armónica es el promedio que se ha de tomar en este tipo de situaciones, pues se trata de una variable formada por cocientes de dos magnitudes (la posición y el tiempo). Como $v_1 = 60$ km/h y $v_2 = 70$ km/h, deducimos que la velocidad media del recorrido total sería:

$$v_{media} = 64,61 \text{ km/h}$$

Problema 7

Las acciones de una empresa han producido los siguientes rendimientos netos anuales:

Año	Rentabilidad
1994	12%
1995	10%
1996	7%
1997	6%
1998	5%

Obtener el rendimiento neto medio en esos cinco años.

En una población de tamaño $n = 5$ se ha observado una variable estadística $X = Rentabilidad$ de la empresa durante los años 1994-1998 que ha presentado $k = 5$ modalidades distintas con la siguiente distribución de frecuencias

x_i	n_i	N_i	c_i
5	1	1	1.05
6	1	2	1.06
7	1	3	1.07
10	1	4	1.1
12	1	5	1.2

Como se trata de una variable con rendimientos acumulativos, calcularemos la media geométrica:

$G = \sqrt[5]{1.05 \cdot 1.06 \cdot 1.07 \cdot 1.1 \cdot 1.12} - 1 = 1.079687 - 1 = 0.079687$

De lo que significa que el rendimiento neto medio entre 1994 y 1998 es del 7.969%.

Problema 8

Un profesor califica a sus alumnos según el criterio siguiente: 40% de suspensos, 30% de aprobados, 15% notables, 10% sobresalientes y 5% de matrículas. Las notas obtenidas son las siguientes:

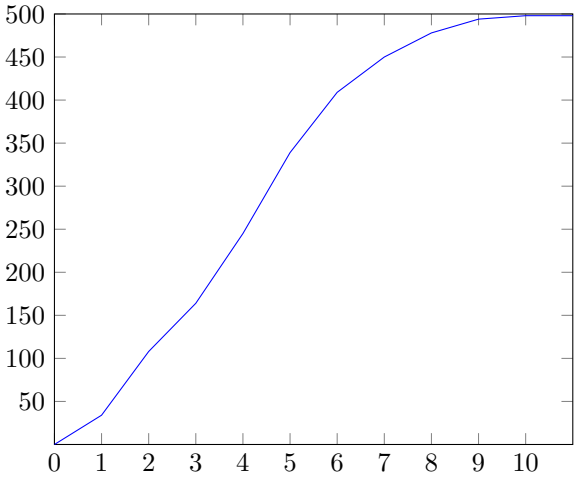
$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$	$(5,6]$	$(6,7]$	$(7,8]$	$(8,9]$	$(9,10]$
34	74	56	81	94	70	41	28	16	4

Calcular las notas máximas para obtener cada una de las calificaciones.

POBLACIÓN: Los alumnos.
TAMAÑO: $34+74+56+81+94+70+41+28+16+4=498$
MODALIDADES: Los intervalos que contienen las notas obtenidas en los exámenes.

x_i	n_i	N_i	f_i	F_i
(0,1]	34	34	$\frac{34}{498}$	$\frac{34}{498}$
(1,2]	74	108	$\frac{74}{498}$	$\frac{108}{498}$
(2,3]	56	164	$\frac{56}{498}$	$\frac{164}{498}$
(3,4]	81	245	$\frac{81}{498}$	$\frac{245}{498}$
(4,5]	94	339	$\frac{94}{498}$	$\frac{339}{498}$
(5,6]	70	409	$\frac{70}{498}$	$\frac{409}{498}$
(6,7]	41	450	$\frac{41}{498}$	$\frac{450}{498}$
(7,8]	28	478	$\frac{28}{498}$	$\frac{478}{498}$
(8,9]	16	494	$\frac{16}{498}$	$\frac{494}{498}$
(9,10]	4	498	$\frac{4}{498}$	1

Curva de distribución

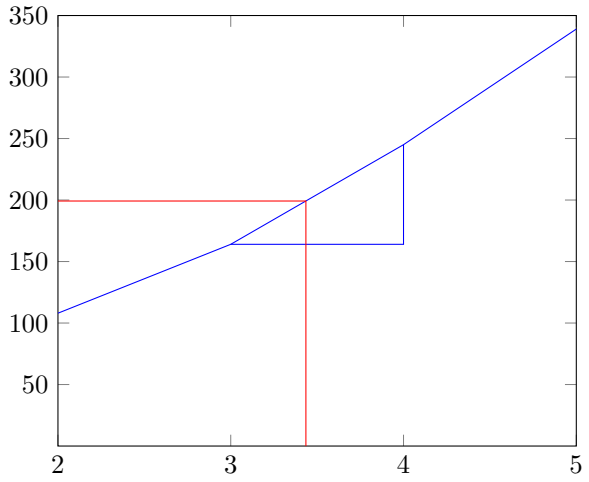


Para la curva de distribución, en el eje X tenemos los extremos de los intervalos y en el eje Y ponemos N_i, la frecuencia absoluta acumulada.

Para hallar las notas máximas para obtener cada una de las calificaciones calculamos los percentiles. La nota que fuese el percentil 30, por ejemplo, significa que el 30% de la población, tiene una nota igual o menor a ella, es decir, un suspenso en este caso.

Empezamos calculando la nota máxima para sacar un suspenso. Como el 40% son suspendidos, calculamos el percentil 40, o lo que es lo mismo, el decil 4. $P_{40}=D_4=\frac{nr}{100}=\frac{498*40}{100}=199.2$. Vemos en la tabla que en la columna de la frecuencia absoluta acumulada, N_4 es la inmediatamente mayor a 199.2. Así sabemos que el percentil 40 va a estar en el intervalo $I_4=(3,4]$. Vamos a la curva de distribución y hacemos semejanza de triángulos.

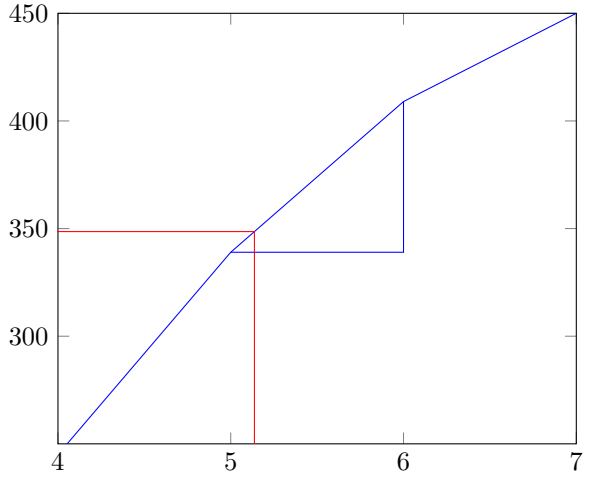
Curva de distribución - Percentil 40



Hacemos $\frac{P_{40}-3}{4-3} = \frac{199.2-164}{245-164}$. Así obtenemos que $P_{40} = 3.435$. Esto es que el 40% de la población tiene una nota inferior a esta, un 3.435 es la máxima nota dentro de los suspensos.

Para los aprobados, tenemos que buscar una nota tal que por debajo de ella estén todos los aprobados y los suspendidos, es decir, $30 + 40 = P_{70}$. De nuevo hacemos $P_{70} = D_7 = \frac{498*70}{100} = 348.6$. Vemos en la tabla que N_6 es la inmediatamente mayor a 348.6. El percentil 70 está en el intervalo $I_6 = (5, 6]$. Vamos a la curva de distribución y hacemos semejanza de triángulos.

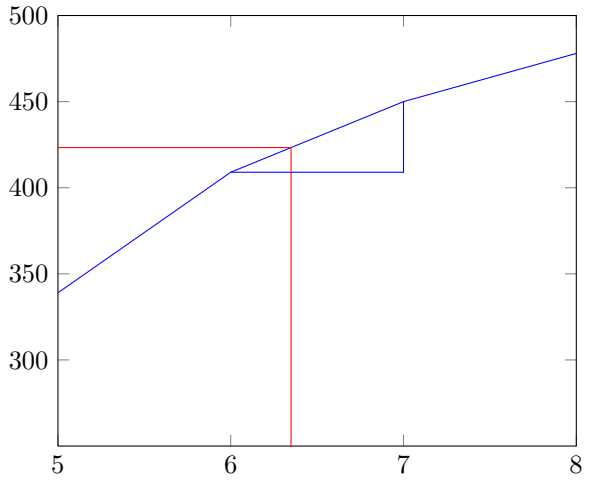
Curva de distribución - Percentil 70



Hacemos $\frac{P_{70}-5}{6-5} = \frac{348.6-339}{409-339}$. Así obtenemos que $P_{70} = 5.137$. Esto es que el 70% de la población tiene una nota inferior a esta, de los cuales un 40% son suspensos. Un 5.137 es la máxima nota para un aprobado.

Para los notables, tenemos que buscar una nota tal que por debajo de ella estén todos los notables, aprobados y suspendidos, es decir, $30 + 40 + 15 = P_{85}$. De nuevo hacemos $P_{85} = \frac{498*85}{100} = 423.3$. Vemos en la tabla que N_7 es la inmediatamente mayor a 423.3. El percentil 85 está en el intervalo $I_7 = (6, 7]$. Vamos a la curva de distribución y hacemos semejanza de triángulos.

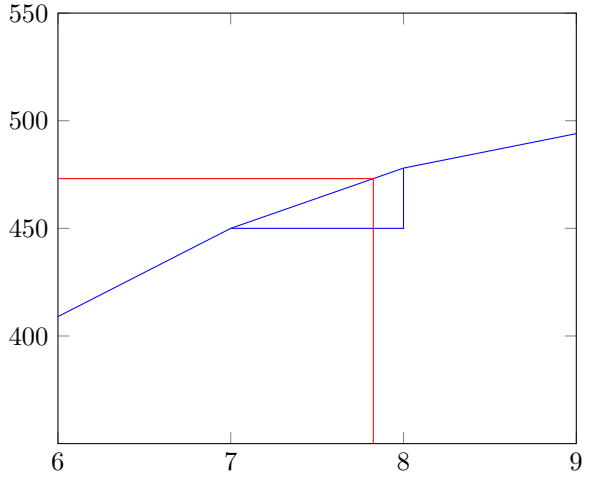
Curva de distribución - Percentil 85



Hacemos $\frac{P_{85}-6}{7-6} = \frac{423.3-409}{450-409}$. Así obtenemos que $P_{85} = 6.349$. Esto es que el 85% de la población tiene una nota inferior a esta, de los cuales un 40% son suspensos y un 30% aprobados. Un 6.349 es la máxima nota para un notable.

Para los sobresalientes, tenemos que buscar una nota tal que por debajo de ella estén todos los sobresalientes, notables, aprobados y suspendidos, es decir, $30 + 40 + 15 + 10 = P_{95}$. De nuevo hacemos $P_{95} = \frac{498*95}{100} = 473.1$. Vemos en la tabla que N_8 es la inmediatamente mayor a 473.1. El percentil 95 está en el intervalo $I_8 = (7, 8]$. Vamos a la curva de distribución y hacemos semejanza de triángulos.

Curva de distribución - Percentil 95



Hacemos $\frac{P_{95}-7}{8-7} = \frac{473.1-450}{478-450}$. Así obtenemos que $P_{95} = 7.825$. Esto es que el 95% de la población tiene una nota inferior a esta, de los cuales un 40% son suspensos, un 30% aprobados y un 10% notables. Un 7.825 es la máxima nota para un sobresaliente.

Para las matrículas, tenemos que buscar una nota tal que por debajo de ella estén todas las matrículas, sobresalientes, notables, aprobados y suspendidos, es decir, todas las notas, $30 + 40 + 15 + 10 + 5 = 100 = P_{100}$. Hacemos $P_{100} = \frac{498*100}{100} = 498$. Vemos en la tabla que N_{10} es 498. Un 10 es la máxima nota para una matrícula.

Problema 9

Se ha medido la altura de 110 jóvenes, obteniendo:

Altura	$(1.55, 1.60]$	$(1.60, 1.70]$	$(1.70, 1.80]$	$(1.80, 1.90]$	$(1.90, 2.00]$
Nº jóvenes	18	31	24	20	17

1. Si se consideran bajos el 3% de los individuos de menor altura, ¿cuál es la altura máxima que pueden alcanzar?
2. Si se consideran altos el 18% de los individuos de mayor altura, ¿cuál es su altura mínima?
3. ¿Qué altura es superada sólo por 1/4 de los jóvenes?
4. Calcular el número de jóvenes cuya altura es superior a 1.75.
5. Calcular la altura máxima de los 11 jóvenes más bajos.
6. Calcular la altura mínima de los 11 jóvenes más altos.

Para hacer todos los cálculos de este ejercicio usaremos la siguiente fórmula:

$$P_r = e_{i-1} + \frac{\frac{nr}{100} - N_{i-1}}{n_i}(e_i - e_{i-1}) \qquad \text{(Percentil)}$$

Donde r es el percentil que queremos calcular, n el tamaño de la población, n_i la frecuencia absoluta, N_i la frecuencia absoluta acumulada y e_i es el final del intervalo i donde encontramos a la cantidad de gente dentro del r% de la población.

Como vemos necesitaremos algunos datos, como la frecuencia absoluta acumulada. Los calcularemos previamente. Aquí están esos datos de forma tabulada:

I	n_i	N_i
(1.55, 1.60]	18	18
(1.60, 1.70]	31	49
(1.70, 1.80]	24	73
(1.80, 1.90]	20	93
(1.90, 2.00]	17	110

Apartado 1

Si se consideran bajos el 3% de los individuos de menor altura, ¿cuál es la altura máxima que pueden alcanzar?

Calculamos P_3. Para ello primero veamos en qué intervalo se encuentra el 3 por ciento de la población con menor altura. Como $0.03 * 110 = 3.3$ sabemos que debemos tomar i como 1. El cálculo quedaría así:

$$\mathrm{P}_3 = 1.55 + \frac{\frac{110 * 3}{100} - 0}{18}(1.6 - 1.55) = 1.559m$$

Por tanto, deducimos que los jóvenes que midan 1.559 metros o menos son el 3% más bajo de la población.

Apartado 2

Si se consideran altos el 18% de los individuos de mayor altura, ¿cuál es su altura mınima?

Nos encontramos en el mismo problema que en el apartado anterior, solo que ahora deberemos calcular P_{82}, pues $100-18=82$. Como $0.82*110=90.2$ sabemos que $i=4$ ya que es el intervalo cuya frecuencia absoluta acumulada es inmediatamente superior. Hacemos los calculos:

$$\mathrm{P}_{82}=1.80+\frac{\frac{110*82}{100}-73}{20}(1.9-1.8)=1.886m$$

Finalmente, se consideran altos los jóvenes que miden 1.886 metros o más.

Apartado 3

¿Qué altura es superada sólo por 1/4 de los jóvenes?

En este caso nos pregunta qué altura es superada por el 25% de la población, o lo que es lo mismo, que altura tiene el 75% más bajo de la población. Para ello calculamos $Q_{75}=P_{75}$. Como $0.75*110=82.5$, por las razones nombradas anteriormente, sabemos que $i=4$. Sustituimos en la expresión:

$$\mathrm{P}_{75}=1.80+\frac{\frac{110*75}{100}-73}{20}(1.9-1.8)=1.847m$$

Deducimos que la altura superada únicamente por un cuarto de la población es 1.847 metros.

Apartado 4

Calcular el número de jóvenes cuya altura es superior a 1.75.

En este apartado nos piden, a fin de cuentas, que hagamos el proceso contrario al que hemos estado haciendo en los apartados anteriores. Antes nos daban r y ahora nos pedían P_r, ahora nos dan P_r y debemos calcular $\frac{nr}{100}$. Para ello usamos la expresión del percentil y despejamos:

$$\mathrm{P}_r=1.75=1.70+\frac{\frac{110*r}{100}-49}{24}(1.80-1.70)$$
$$\frac{nr}{100}=61$$

Como vemos, hay 61 jóvenes que miden 1.75 o menos, por tanto $110-61=49$ jóvenes superan la latura de 1.75 metros.

Apartado 5

Calcular la altura máxima de los 11 jóvenes más bajos.

Primero, esos 11 jóvenes representan el $\frac{11}{110}=0.1$ por ciento de la población, por tanto, debemos calcular $D_1=P_{10}$. Sabemos que los 11 jóvenes más bajos están en el primer intervalo, por lo que $i=1$ y sustituimos:

$$\mathrm{P}_{10}=1.55+\frac{\frac{110*10}{100}-0}{18}(1.60-1.55)=1.581m$$

Luego, la altura máxima de los 11 jóvenes más bajos es 1.581 metros.

Apartado 6

Calcular la altura mínima de los 11 jóvenes más altos.

En este apartado seguiremos el mismo procedimiento que en el apartado anterior. Como los 11 jóvenes representan el 0.1 por ciento de la población, debemos calcular $D_9 = P_{90}$, pues recordemos que son los 11 más altos. Teniendo en cuenta que $110 - 11 = 99$ sabemos que $i = 5$. Procedemos a realizar los cálculos:

$$\mathrm{P}_{90} = 1.90 + \frac{\frac{110 * 90}{100} - 93}{17}(2.00 - 1.90) = 1.935m$$

En conclusión, la altura mínima de los 11 jóvenes más altos es de 1.935 metros.

Problema 10

Realizando una prueba para el estudio del cáncer a 150 personas se obtuvo la siguiente tabla según la edad de los enfermos:

Edad	$(10, 30]$	$(30, 40]$	$(40, 50]$	$(50, 60]$	$(60, 90]$
Nº enfermos	15	22	48	40	25

1. Calcular la edad más común de los individuos estudiados.
2. Calcular la edad mínima y máxima del 30% central de los individuos.
3. Calcular el recorrido intercuartílico y la desviación típica.
4. Calcular e interpretar los valores de los coeficientes de asimetría y curtosis.

Escribamos las tablas con los datos que nos pueden interesar para la resolución del ejercicio:

I_i	n_i	N_i	c_i	h_i	$n_i(x_i - \overline{x})^2$
$(10, 30]$	15	15	20	0.75	12355.55
$(30, 40]$	22	37	35	2.2	4129.18
$(40, 50]$	48	85	45	4.8	657.12
$(50, 60]$	40	125	55	4	1587.6
$(60, 90]$	25	150	75	0.833	17292.25
	150				36021.7

Apartado 1

Se nos pide calcular la edad más común, luego buscamos la moda de la distribución. Dado que los intervalos son de diferente tamaño tenemos que calcular las densidades de frecuencia de cada intervalo. El mayor h_i es 4.8 luego la moda está en el intervalo (40, 50]. Para calcularla tendremos que dibujar el histograma con el intervalo en cuestión y los dos contiguos. Unimos la esquina superior derecha del primer rectángulo con la del segundo, y la esquina superior izquierda del segundo con la del tercero. Ahora, aplicando semejanza de triángulos calculamos la intersección, con lo que obtenemos la moda.

$$\frac{4.8 - 2.2}{4.8 - 4} = \frac{M_O - 40}{50 - M_O}; M_O = 47.647 \approx 48 \text{ años}$$

Apartado 2

Se nos pide calcular el percentil 35 y el 65. Hay que buscar el intervalo asociado al valor inmediatamente superior a $n * p$ en la frecuencia absoluta acumulada. Después, se calcula aplicando semejanza de triángulos :

$$n * 0.35 = 52.5; N_i = 85; I_i = (40, 50]$$
$$\frac{52.5 - 37}{P_{35} - 40} = \frac{85 - 37}{50 - 40} = 43.229 \approx 43 \text{ años.}$$

Luego la edad mínima será aproximadamente 43 años.

$$n * 0.65 = 97.5; N_i = 125; I_i = (50, 60]$$
$$\frac{97.5 - 85}{P_{65} - 50} = \frac{125 - 85}{60 - 50} = 53.125 \approx 53 \text{ años.}$$

Por tanto, la edad máxima será aproximadamente 53 años.

Apartado 3

Para el recorrido intercuartílico necesitaremos los cuartiles 1 y 3, es decir, los percentiles 25 y 75. Procedemos de la misma manera que en el apartado anterior:

$$n * 0.35 = 37.5 \quad N_i = 85 \quad I_i = (40, 50]$$
$$\frac{37.5 - 37}{Q_1 - 40} = \frac{85 - 37}{50 - 40} = 40.104 \approx 40 \text{ años.}$$
$$n * 0.75 = 112.5 \quad N_i = 125 \quad I_i = (50, 60]$$
$$\frac{112.5 - 85}{Q_3 - 50} = \frac{125 - 85}{60 - 50} = 56.875 \approx 57 \text{ años.}$$

Por tanto el recorrido intercuartílico será $Q_3 - Q_1 = 16.771 \approx 17$ años, lo que nos indica que el 50% de la población central se encuentra en un intervalo de unos 17 años.

Para la calcular la desviación típica necesitamos calcular la raíz cuadrada positiva de la varianza. Tenemos que calcular la media del área de los cuadrados de lado la diferencia de cada dato a la media. La media es 48.7. Escribamos los $n_i(x_i - \overline{x})^2$ en la tabla. La varianza será 240.144 años^2. Finalmente la desviación típica será 15.497.

Apartado 4

1. Coeficiente de asimetría de Fisher:

$$\gamma_1(X) = \frac{\mu_3}{\sigma_X^3}; \qquad \mu_3 = m_3 - 3m_2m_1 + 2m_1^3 = \frac{1}{n}\sum_i n_i ci^3 - \frac{3}{n}\sum_i n_i ci^2\overline{x} + 2\overline{x}^3 = 341.256$$

$$\gamma_1(X) = \frac{341.256}{15.4965^3} = 0.0917 > 0 \longrightarrow \text{Asimetría moderada por la derecha.}$$

2. Coeficientes de asimetría de Pearson: Necesitamos la mediana para calcular A*. 150*0.5 = $75 \in I_3$

$$Me = 40 + \frac{75 - 37}{48} * 10 = 47.91\overline{6}$$

$$A_p = \frac{\overline{x} - M_O}{\sigma_X} = 0.06795 \qquad A_p^* = \frac{3(\overline{x} - Me)}{\sigma_X} = 0.1516$$

   Nos indican también una ligera asimetría por la derecha.

3. Coeficiente de curtosis de Fisher:

$$\gamma_2(X) = \frac{\mu_4}{\sigma^4} - 3$$

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 2m_1^4 = 153232.46$$

$$\gamma_2(X) = \frac{153232.46}{15.4965^2} - 3 = -0.3429 < 0$$

Por tanto, la distribución es platicúrtica; presenta menor concentración central de frecuencias que una distribución normal con su media y desviación típica,

4. Coeficiente de curtosis de Kelley:

$$K = \frac{1}{2}\frac{Q_3 - Q_1}{D_9 - D_1} - 0,263$$

Sabemos Q_1 y Q_2, luego calcularemos D_1 y D_9

$1n\alpha = 15 \longrightarrow D_1$ coincide con el extremo superior de I_1, luego será 30 años. $1n\alpha = 135 \longrightarrow D_9 \in I_5$. Aplicamos la fórmula.

$$D_9 = 60 + \frac{135 - 125}{25} * 30 = 72 \text{ años}$$

$$K = -0.06335 < 0$$

Como $K < 0$, la distribución es platicúrtica.