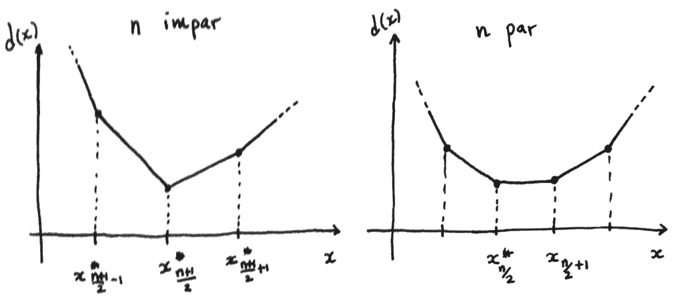
# 1. Estadística descriptiva

### Ejercicio 1

Sean $x_1 = 4$, $x_2 = 3$, $x_3 = 2$, $x_4 = 10$, $x_5 = 6$, $x_6 = 14$ y $x_7 = 10$ los datos proporcionados

1. Se busca $\min_x \sum_{i=1}^{7} |x_i - x|$. Intuitivamente $x$ debe minimizar la distancia a todos los datos y se puede probar que el mínimo se alcanza en la mediana de los datos, es decir $x = 6$. Formalmente, dado que la función a minimizar no es derivable, se estudia el comportamiento de la función. Para esto se ordena la muestra de menor a mayor y resulta que la función es lineal en los intervalos $[x_i*, x*_{i+1}]$. El mínimo depende entonces de la paridad de $n$ (ver siguiente figura)




2. En este caso el valor buscado es la moda de los datos, es decir $x = 10$.
3. Se busca $\min_x \max_{i=1,..,7} |x_i - x|$. En este caso el valor de $x$ resulta ser el rango medio, es decir el promedio del mínimo y el máximo, es decir $x = \frac{2+14}{2} = 8$.
4. Se busca $\min_x |\sum_{i=1}^{7} x_i - x| = |\sum_{i=1}^{7} x_i - 7x|$ que vale cero si $x$ es el promedio de los datos. Por lo tanto el valor que minimiza la multa es $x = 7.14$.
5. Se busca $\min_x \sum_{i=1}^{7} (x_i - x)^2$, como la función a minimizar es derivable basta hallar los extremos y se encuentra que el mínimo se alcanza en $x = 7.14$ el promedio de los datos.

### Ejercicio 2

1. Diagrama de tallos y hojas espalda con espalda

| Hoja Hombres | Tallo | Hoja Mujeres |
|---|---|---|
| 8 2 0 0 | 0 | 0 0 1 2 3 4 6 7 8 8 |
| 4 2 0 0 | 1 | 0 0 0 |
| 8 6 3 0 | 2 | 2 4 |
| 3 | 3 | 0 3 7 9 |
| | 4 | 7 |
| 6 5 5 3 0 | 5 | |
| | 6 | |
| | 7 | 3 |
| 1 | 8 | |
| | 9 | 4 |
| | 10 | |
| | 11 | |
| | 12 | 8 |
| | 13 | |
| | 14 | 6 |

2. Diagrama de tallos y hojas para los datos combinados

| Tallo | Hoja |
|---|---|
| 0 | 0 0 0 0 1 2 2 3 4 6 7 8 8 8 |
| 1 | 0 0 0 0 0 2 4 |
| 2 | 0 2 3 4 6 8 |
| 3 | 0 3 3 7 9 |
| 4 | 7 |
| 5 | 0 3 5 5 6 |
| 6 | |
| 7 | 3 |
| 8 | 1 |
| 9 | 4 |
| 10 | |
| 11 | |
| 12 | 8 |
| 13 | |
| 14 | 6 |

3. La mediana es $m = x^*_{21} = 20$ mientras que el promedio es $\bar{x} = 29.7$. Los datos presentan entonces asimetría hacia la derecha.

**Ejercicio 3**

1. $\bar{x}+\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} x_i + y_i = \overline{x+y}.$

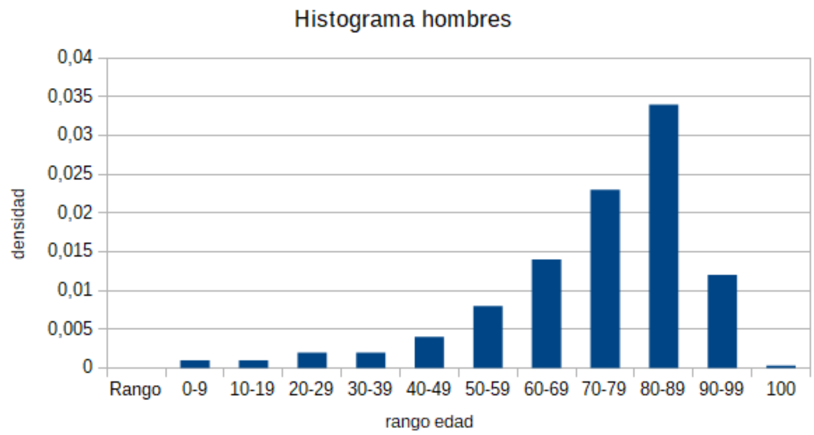
2. $a\bar{x}+b = a\frac{1}{n}\sum_{i=1}^{n} x_i + b = \frac{1}{n}\sum_{i=1}^{n} ax_i + b = \overline{ax} + b.$

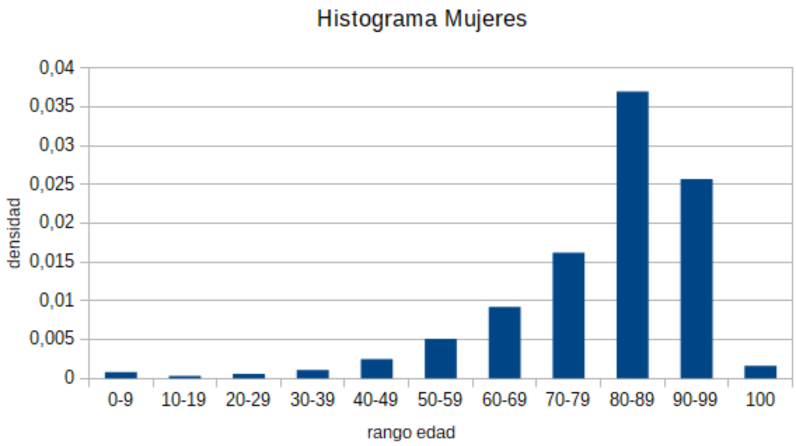
3.
$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\bar{x} + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2
\end{aligned}$$

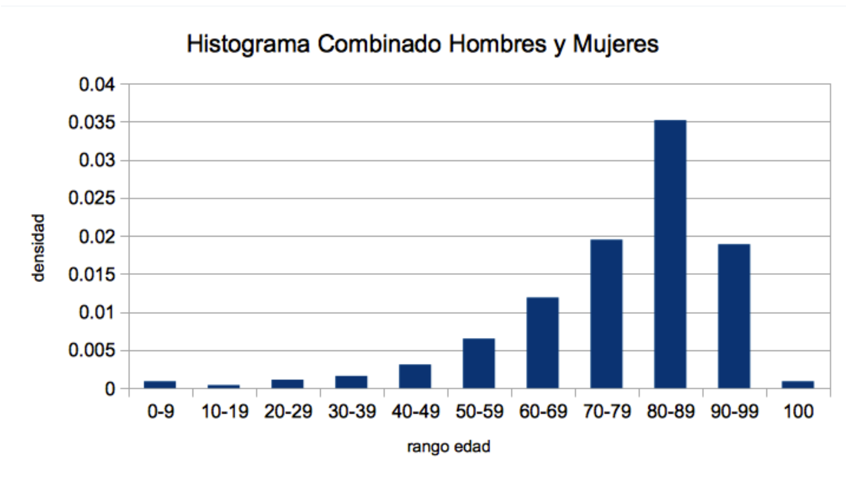
4.
$$\begin{aligned}
\sigma^2_{x+y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i+y_i-\overline{x+y})^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i-\bar{x})+(y_i-\bar{y})]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&= \sigma_x^2 + \sigma_y^2 + \frac{2}{n}\sum_{i=1}^{n} x_iy_i + \bar{x}\bar{y} - x_i\bar{y} - y_i\bar{x} = \sigma_x^2 + \sigma_y^2 + 2[\frac{1}{n}\sum_{i=1}^{n}(x_iy_i) - \bar{x}\bar{y}]
\end{aligned}$$

**Ejercicio 4**

1. Histogramas normalizados (es decir con área igual a 1) para el conjunto de datos de hombres, mujeres y combinados.




2. No es posible calcular el promedio a partir del histograma (o de las frecuencias) sin embargo podemos obtener una cota inferior asumiendo que todos los datos de un intervalo dado coinciden con el extremo izquierdo y analógamente una cota superior utilizando el extremo derecho de cada intervalo. En el caso de los hombres se obtiene que la cota inferior es $\bar{x}_{inf} = 69.17$ y la cota superior es $\bar{x}_{sup} = 76.87$ (para el último intervalo utilizamos 110 como extremo derecho aunque sería mejor utilizar el máiximo de los datos).

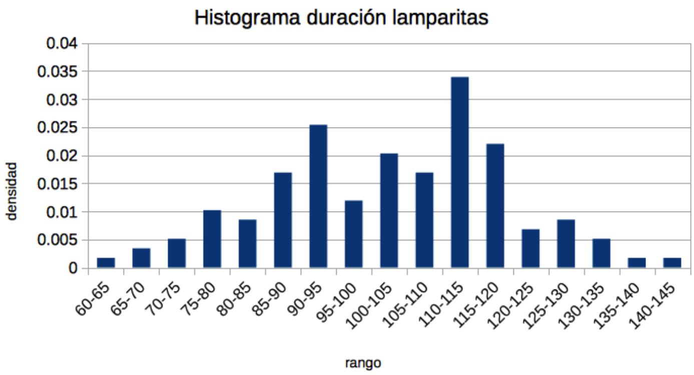
3. Podemos hallar cuál es el intervalo central del histograma (primer intervalo que contiene más del 50 % de los datos) y tomar su punto medio como aproximación de la mediana. Para el caso de los hombres tendríamos que este intervalo es el 75-79 (si observamos los intervalos originales) y por lo tanto la estimación de la mediana resulta en $m = 77.5$. Podemos observar que es mayor que la cota superior dada para el promedio, lo que coincide con la asimetría hacia la izquierda observada en los histogramas.

**Ejercicio 5**

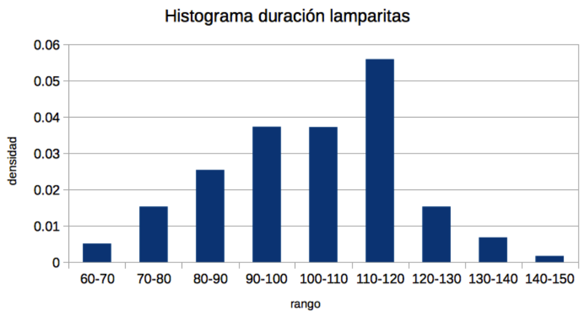
Se tienen en realidad $N = 118$ datos. Construimos tabla de frecuencia e histograma.

| Intervalo | Punto Medio | Frec. Absoluta | Frec. Relativa | Densidad |
|---|---|---|---|---|
| 60 - 65 | 62.5 | 1 | 1/118 | 1/590 |
| 65 - 70 | 67.5 | 2 | 2/118 | 2/590 |
| 70 - 75 | 72.5 | 3 | 3/118 | 3/590 |
| 75 - 80 | 77.5 | 6 | 6/118 | 6/590 |
| 80 - 85 | 82.5 | 5 | 5/118 | 5/590 |
| 85 - 90 | 87.5 | 10 | 10/118 | 10/590 |
| 90 - 95 | 92.5 | 15 | 15/118 | 15/590 |
| 95 - 100 | 97.5 | 7 | 7/118 | 7/590 |
| 100 - 105 | 102.5 | 12 | 12/118 | 12/590 |
| 105 - 110 | 107.5 | 10 | 10/118 | 10/590 |
| 110 - 115 | 112.5 | 20 | 20/118 | 20/590 |
| 115 - 120 | 117.5 | 13 | 13/118 | 13/590 |
| 120 - 125 | 122.5 | 4 | 4/118 | 4/590 |
| 125 - 130 | 127.5 | 5 | 5/118 | 5/590 |
| 130- 135 | 132.5 | 3 | 3/118 | 3/590 |
| 135 - 140 | 137.5 | 1 | 1/118 | 1/590 |
| 140 - 145 | 142.5 | 1 | 1/118 | 1/590 |

Los extremos se toman cerrados a la izquierda y abiertos a la derecha $[a,b)$.



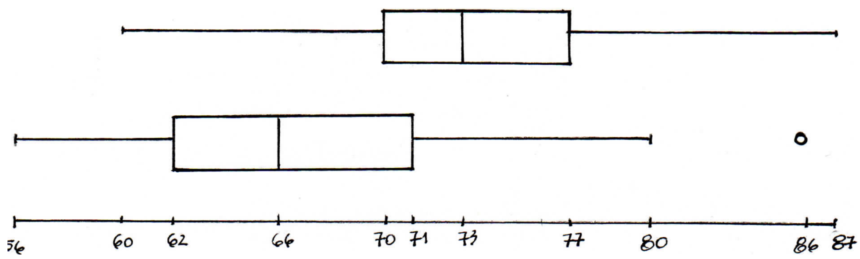

Se puede observar que la cantidad de intervalos parece ser mucha pues los datos quedan muy dispersos. Así que repetimos el histograma con intervalos de largo 10.



**Ejercicio 6**

1. Año 1985: datos necesarios para construir el boxplot
   - $n = 51$.
   - $m = x^*_{26} = 66$
   - $q_i = x^*_{13} = 62$ y $q_s = x^*_{39} = 71$, de donde $RIC = 9$
   - $L_i = q_i - 1.5 \times RIC = 48.5$ y $x_- = 56$. No hay datos atípicos menores que la mediana.
   - $L_s = q_s + 1.5 \times RIC = 84.5$ y $x_+ = 80$. El dato 86 es un dato atípico.
2. Año 1976: datos necesarios para construir el boxplot
   - $n = 51$.
   - $m = x^*_{26} = 73$
   - $q_i = x^*_{13} = 70$ y $q_s = x^*_{39} = 77$, de donde $RIC = 7$
   - $L_i = q_i - 1.5 \times RIC = 59.5$ y $x_- = 60$. No hay datos atípicos menores que la mediana.
   - $L_s = q_s + 1.5 \times RIC = 87.5$ y $x_+ = 87$. No hay datos atípicos mayores que la mediana.

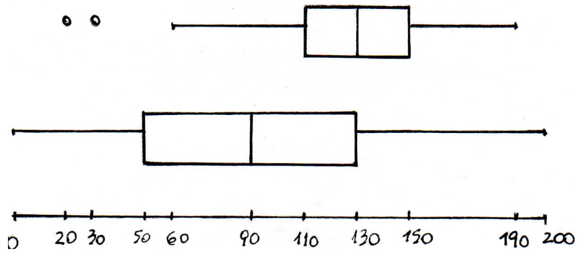
   Comparando los boxplots podemos que en el año 1985 la tasa de ocupación de los hospitales es menor que en el año 1976.




**Ejercicio 7**

1. Estantes superiores: datos necesarios para construir el boxplot:
   - $n = 95$.
   - $m = x^*_{48} = 90$
   - $q_i = x^*_{24} = 50$ y $q_s = x^*_{72} = 130$, de donde $RIC = 80$
   - $L_i = q_i - 1.5 \times RIC = -70$ y $x_- = 0$. No hay datos atípicos menores que la mediana.
   - $L_s = q_s + 1.5 \times RIC = 250$ y $x_+ = 200$. No hay datos atípicos mayores que la mediana.
2. Estantes inferiores: datos necesarios para construir el boxplot:

- $n = 95$.
- $m = x^*_{48} = 130$
- $q_i = x^*_{24} = 110$ y $q_s = x^*_{72} = 150$, de donde $RIC = 40$
- $L_i = q_i - 1.5 \times RIC = 50$ y $x_- = 60$. Los datos 20 y 30 son atípicos.
- $L_s = q_s + 1.5 \times RIC = 210$ y $x_+ = 190$. No hay datos atípicos mayores que la mediana.




**Ejercicio 8**

1. Bloque B.
2. 20 %.
3. 70 %.

**Ejercicio 9**

1. Clase (a) mayor que 50 %, clase (b) menor que 50 % y clase (c) alrededor de 50 %.
2. La clase (b).
3. Entre 90 y 100.

**Ejercicio 10**

El primer histograma corresponde al grupo (C), el segundo corresponde al grupo (A) y el último al grupo (B).

**Ejercicio 11**

El histograma (expresado en este caso en porcentajes) debería tener área igual a 100 % y no la tiene.

**Ejercicio 12**

La altura debería ser 15.

**Ejercicio 13**

En media tiene mayor presión sanguínea el grupo que tiene 4 hijos. No podemos afirmar que se deba a eso sino que puede haber muchos otros factores que influyan que no están tenidos en cuenta al comparar los histogramas.

**Ejercicio 14**

La edad al morir por causas naturales corresponde al histograma (i) (probabilidad concentrada en valores altos, es decir en edades avanzadas).

**Ejercicio 15**

Los dos primeros histogramas son simétricos, por lo que promedio y mediana coinciden. Podemos estimar el promedio en 50 para el primer caso y en 25 para el segundo caso. En el tercer caso tenemos una asimetría hacia la derecha que corresponde a un promedio mayor que la mediana. El promedio se puede estimar en 60 dentro de los valores posibles.

**Ejercicio 16**

Las correspondencias son: (i)-(a), (ii)-(b) y (v)-(c).