Estadística Descriptiva: una introducción

Julio Muñoz

1. Medidas centrales

Una medida central es cualquier valor que se ubique en una posición central dentro de los datos que se examinan. La idea que se pretende que es elegir aquellas medidas que proporcionen información relevante sobre los datos y por ende del fenómeno del que proceden.
La *moda*, el valor *mediana* y la *media aritmética* son las medidas centrales más comúnes. En ocasiones estas tres mediddas coinciden, pero en general no es así. Es útil el cálculo de las tres ya que en conjunto pueden aportar bastante información. La más importante es la media.

1.1. Media aritmética

La media aritmética del conjunto de datos (de la muestra de tamaño n) x_1, x_2, ... , x_n se define como

$$a = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

En ciertas situaciones convendrá combinar distintas medias aritmétics procedentes de distintos conjuntos de datos. Si $a_1, a_2, \ldots, a_k$ son las medias corresponditnes a k muestras de tamaños $n_1, n_2, \ldots, n_k$, se define la media promediada de éstas como

$$\overline{a} = \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i} = \frac{n_1 m_1 + n_2 m_2 + \cdots + n_k m_k}{n_1 + n_2 + \cdots + n_k}$$

1.2. Moda

Recordamos que la moda de una muestra es la observación (u observaciones) que más repite. Definimos el valor modal como aquél cuya frecuencia es superior a la de los valores anterior y posterior a él. La aparición de distintas modas indica en principio heterogeneidad para la población.

Ejemplo La moda de $\{2, 3, 4, 4, 5, 6, 6, 6, 8\}$ es 6, y la de $\{2, 3, 4, 4, 4, 6, 6, 6, 8\}$ es 4 y también lo es 6.

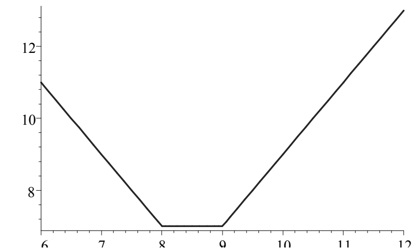
1.3. Valor Mediana

La mediana se define como el valor central de la muestra. Si N es impar y $x_1, ..., x_N$ es la muestra entonces $x_{\frac{N+1}{2}}$es la mediana. Cuando N es par entonces se considera el intervalo, intervalo cuyos extremos son los dos valores centrales de la muestra.
El valor mediana de un conjunto de valores es el que deja la mitad de los valores de la muestra a su derecha y el resto a su izquierda. De otro modo, el valor mediana es x_j si f_j, la frecuencia acumulada hasta el valor x_j es $N/2$. Cuando esto n se consigue entonces se considera el intervalo mediana, intervalo cuyos extremos son los valores centrales de la muestra de frecuencias acumuladas inmediatamente inferior a $N/2$ e inmediatamente superior a $N/2$. En esta situación se suele optar po definir el valor mediana como la media aritmética de estos dos valores centrales.

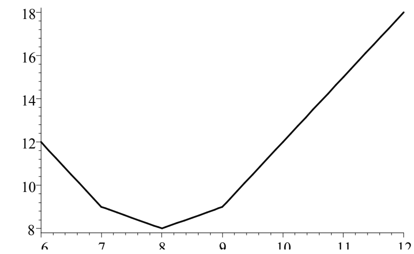
Nota: el valor mediana admite una definición alternativa, es aquel valor M_e que minimiza la función

$$f(M) = \sum_{i=1}^{n} |x_i - M|$$

Ejemplo Si se consideran los datos $\{6, 8, 9, 12\}$ el intervalo mediana es $[6, 8]$ y el valor mediana 7. Si consideramos la muestra $\{6, 7, 8, 9, 12\}$, el valor mediana es 7.



Grfico de $|6 - x| + |8 - x| + |9 - x| + |12 - x|$



Grfico $|6 - x| + |7 - x| + |8 - x| + |9 - x| + |12 - x|$

Véanse los gráficos adjuntos para entender la definición de valor mediana en términos de solución de un problema de minimización.

1.4. Cuantiles

Cuantil es una generalización de valor mediana. Un cuantil de orden $p \in (0, 1)$ es aquel valor de la muestra tal que los valores de la muestra menores o iguales que él acumulan una frecuencia relativa p.

1.5. Media Geométrica

Para el conjunto de datos no negativos $x_1, x_2, \ldots, x_n$ la **media geométricaometric mean** es el producto

$$\sqrt[n]{x_1 x_2 \cdots x_n}$$

Ejemplo Un rectángulo de aristas con longitudes a y b tiene area ab. El cuadrado que tiene el mismo área es el aristas con longitud $\sqrt{ab}$, la media geométrica de a y b.

4×9 rectángulo $\qquad$ 6×6 cuadrado

La media geométrica es bastante útil cuando la razón entre dos datos cualesquiera de los datos se mantiene casi constante, tal y como ocurre en las inversiones de dinero con interés compuesto. Nota: lka media geométrica es siempre menor o igual que la media aritmética.
Sea la siguiente lista correspondiente a los valores de una cuenta al principio de cada año, donde la inversión inicial es deThe table lists the values of the account at the beginning of each year, with initial investment of \$1,000 y la tas de interés compuesto anual es del 10%

Año	Valor
2000	1000.00
2001	1100.00
2002	1210.00
2003	1331.00
2004	1464.10

La media geométrica es

$$\sqrt[5]{1000 \times 1100 \times 1210 \times 1331 \times 1464.1} = 1210.00$$

la cual predice un valor central de los valores. Podemos ahora comparar ésta con la media aritmética:

$$\frac{1000 + 1100 + 1210 + 1331 + 1464.1}{5} = 1221.00$$

1.6. Media Armónica

Para el conjunto de valores no negativos $x_1, x_2, \ldots, x_n$ se define la media armónica con el recíproco de la media de los inversos, como

$$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Suele emplearse esta media para el cálculo de promedios de velocidades cuando las distancias recorridas por las trayctorias es la misma.

Ejemplo Se realiza un trayecto auna velocidad de $70\,\mathrm{mi/h}$ y se lleva a cabo el camino de vuelta, sobre la misma ruta, a una velocidad de $50\,\mathrm{mi/h}$. Entonces la media es

$$\frac{2}{\frac{1}{70} + \frac{1}{50}} = 58.333\,\mathrm{mi/h}$$

1.7. Representación de tallos y hojas

Sean los datos

95	95	100	100	100	100	100	105	105	105
110	110	110	110	110	110	110	110	110	115
115	115	115	115	115	115	115	115	115	115
120	120	120	120	120	120	120	120	125	125
125	125	130	130	130	130	135	135	140	140

Además de los métodos habituales podemos representar los datos mediante el llamado diagrama de tallos y hojas: elegimos dos dígitos representativos para el digrama, las centenas y las decenas, éstos apareceran en la primera columna. Representan los distintos grupos que tienen ése número de decenas. Esto es lo que se llama el tallo. El dígito siguiente, las unidades, se representan a continuación del tallo que le corresponde, y se escribe tantas veces como se repita en la muestra. Esto constituyen las hojas:

$$
\begin{array}{r||l}
09 & 5\ \ 5 \\
10 & 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 5\ \ 5\ \ 5 \\
11 & 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 5\ \ 5\ \ 5\ \ 5\ \ 5\ \ 5\ \ 5\ \ 5\ \ 5\ \ 5\ \ 5 \\
12 & 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 5\ \ 5\ \ 5\ \ 5 \\
13 & 0\ \ 0\ \ 0\ \ 0\ \ 5\ \ 5 \\
14 & 0\ \ 0
\end{array}
$$

2. Medidas de Dispersión

Una medida de dispersión es cualquier indicador que informe sobre cómo de diseminados están los datos con respecto a una medida central. La s medidas de dispersión sirven en cierto modo para validar o despreciar la información la información que proporcione la medida central. Cuanto menor sea la dispersión mejor representará la medida central a la población.
Las medidas de dispersión más comúnes son la varianza y la desviación típica (en algunos textos o programas informáticos se llama desviación estándar). Damos a conocer las medidas de dispersión más elementales.

2.1. Rango o Recorrido

- El **rango** de una muestra es la diferencia entre el mayor y el menor elemento de la muestra.

Ejemplo El rango de la muestra $\{-5, 0, 1, 2, 20\}$ es $20 - (-5) = 25$. Y el de $[0, 0, 0, 1, 2, 30]$ es $30 - 0 = 30$.

2.2. Desviación absoluta media

La desviación absoluta media del conjunto de datos $\{x_1, x_2, ..., x_N\}$ con respecto a una medida centra M_c se define como

$$d = \frac{\sum_{j=1}^{N} |x_j - M_c|}{N}$$

El cálculo también puede hacerse con la ayuda de las frecuencias:

$$d = \frac{\sum_{j=1}^{r} |x_j - M_c| \, n_j}{N}$$

La más común entre las desviaciones absolutas medias es la que se realiza con respecto a la media aritmática, i.e. cuando

$$M_c = a = \frac{\sum_{i=1}^{n} x_i}{n}$$

Téngase en cuenta que el valor absoluto de la diferencia $x - m$, i.e. $|x - m|$, responde a la fórmula

$$|x - m| = \begin{cases} x - m & \text{si} \quad x \geq m \\ m - x & \text{isi} \quad x < m \end{cases}$$

Nota Para cualquier conjunto de datos $x_1, x_2, \ldots, x_N$, la suma de las diferencias con respecto a la media arimética es cero:

$$\sum_{j=1}^{n} (x_j - a) = 0$$

lo que significa que la dispersión sin valor absoluto, con respecto a la media no establece ningún grado de dispersión au cuando de hecho los datos puedan estar tan diseminados como quieran.

2.3. Varianza and Desviación Típica

La **varianza** de una muestra de tamaño N, designada por σ^2, se define como el promedio de los cuadrados de las diferencias entre los valores de la muestra y la media aritmética a:

$$\sigma^2 = \frac{\sum_{j=1}^{n} (x_j - a)^2}{n}$$

o

$$\sigma^2 = \frac{\sum_{j=1}^{r} (x_j - a)^2 n_j}{N}$$

La **desviación típica** de la muestra es la raíz cuadrada de la varianza y se denota mediante σ :

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N} (x_j - \mu)^2}{N}}$$

Covarianza y Desviación Estándar

La covarianza de una muestra de tamaño N y media aritmética a es

$$s^2 = \frac{\sum_{j=1}^{N} (x_j - a)^2}{N - 1}$$

La desviación estándar de la muestra de media aritmética a es

$$s = \sqrt{\frac{\sum_{j=1}^{N} (x_j - a)^2}{N - 1}}$$

La razón por la que se divide por $N-1$ en lugar de N es porque s provee una estimación centrada de la desviación típica σ (en el sentido de que la media aritmética de los valores $\frac{\sum_{j=1}^{N}(x_j-a)^2}{N-1}$ es

σ^2). $N-1$ es el número de grados de libertad de s en el sentido de que si conocemos $N-1$ desviaciones de la muestra con respecto a la media (i.e. se conocen $N-1$ sumandos $x_j - a$) entonces, dado que la media es cero, podemos conocer e N-ésimo. *reference or study.*
Un sencillo cálculo nos lleva a la fórmula

$$s^2 = \frac{N\sum_{j=1}^{N} x_j^2 - \left(\sum_{j=1}^{N} x_j\right)^2}{N\,(N-1)}$$

Tipificación de datos

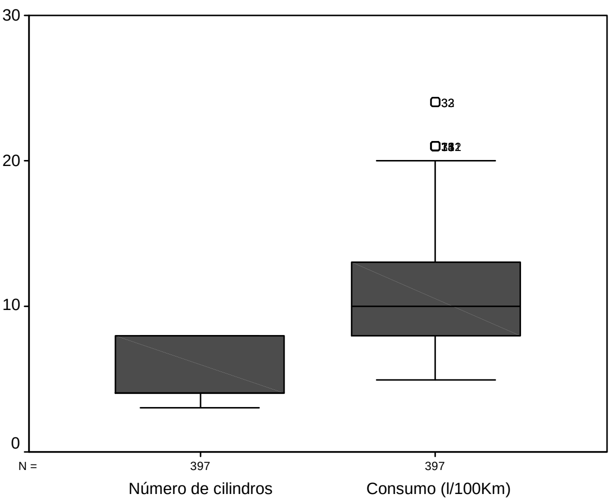
Las desviaciones o medidas expuestas, aparte de informar sobre la dispersión de una muestra con respecto a una medida central, también son útiles para comparar dos muestras o poblaciones distintas (que en principio se expresan en unidades diferentes o cuyas escalas difieren). Para llevar a cabo estas comparaciones es preciso analizar la estandarización de los datos, lo que se conoce con el nombre de tipificación. En lugar de analizar los datos x_i consideramos $\frac{x_i-a}{\sigma}$ (o eventualmente $\frac{x_i-M_c}{\widetilde{\sigma}}$ con M_c una medida central y $\widetilde{\sigma}$ la deviación estándar de la muestra con respecto a esa medida[1]). Por consiguiente si pretendemos comparar dos muestras dadas mediante la realización (u obsevación) de dos estadísticos X e Y hemos de considerar los nuevos estadísticos tipificados

$$Z_X = \frac{X-m}{\sigma}$$

y

$$Z_Y = \frac{Y-m}{\sigma}$$

Ejemplo Considérese el archivo de datos de SPSS **coches.sav** y analicemos las variables número de cilindros y consumo. Un análisis descriptivo sencillo puede resumirse con el siguiente diagrama de cajas en el que se analiza el consumo en función del número de cilindros (se hace el estudio clasificando primero cada coche en una clase, 3, 4, 5, 6 u 8, y en cada clase se realiza la estadística sobre el estadístico consumo) También podemos considerar por separado las variables consumo y número de cilindros:



[1]Podríamos tomar por ejemplo $\widetilde{\sigma} = \sqrt{\frac{\sum_{j=1}^{N}(x_j - M_c)^2}{N-1}}$

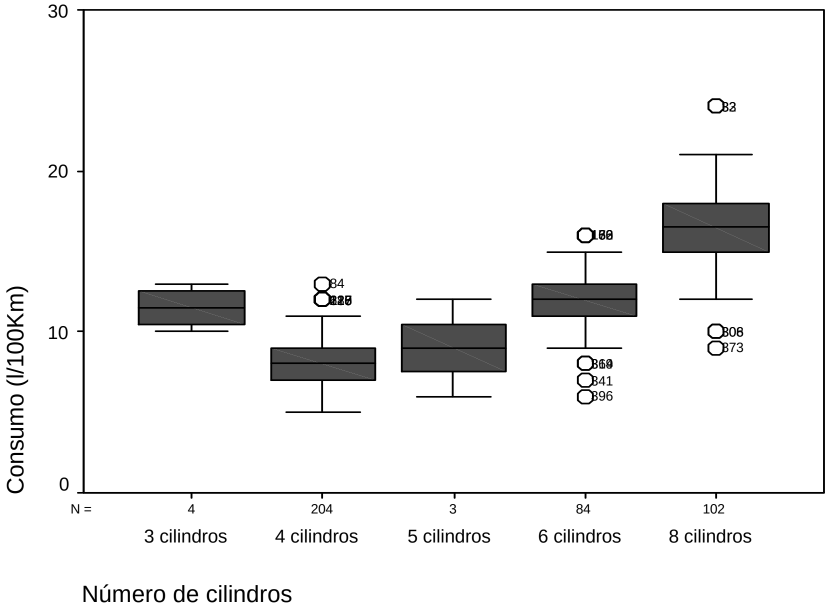
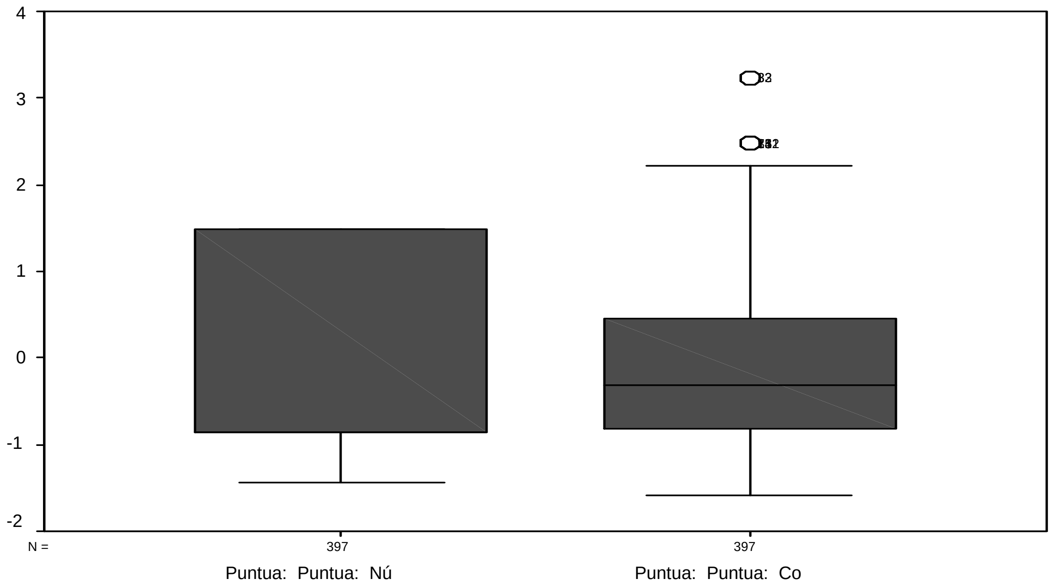


Figure 1:

Con estos diagramas ya podemos emitir algún juicio sobre las variables. Pero si quisiéramos comparar las dispersiones de número de cilindros y consumo habría que tipificarlas. SPSS lo hace de manera automática, crea las variables adimensionales zzconsum y zzcilind (Z_X y Z_Y), cuya representación por cajas es



Esta representación sirve para comparar dos muestras. Pero desde con un punto de vista más preciso podríamos optar por calcular los coeficientes de variación de Pearson (ver más abajo). Para esto hacemos uso SPSS y construimos la siguiente tabla de datos descriptivos:

Estadísticos descriptivos

		Consumo (l/100Km)	Número de cilindros	N válido (según lista)
N	Estadístico	398	405	397
Mínimo	Estadístico	5	3	
Máximo	Estadístico	26	8	
Media	Estadístico	11,23	5,47	
	Error típico	,20	8,50E-02	
Desv. típ.	Estadístico	3,95	1,71	
Varianza	Estadístico	15,572	2,923	
Asimetría	Estadístico	,760	,513	
	Error típico	,122	,121	
Curtosis	Estadístico	,078	-1,402	
	Error típico	,244	,242	

Coeficiente de Variación de Pearson

El **Coeficiente de Variación de Pearson** se define como

$$V = \frac{\sigma}{a}$$

Es una medida de dispersión que expresa la magnitud de la variación con independencia de la magnitud de los datos de la muestra

Ejercicio Compara las dispersiones de las variables consumo y número de cilindros del ejemplo anterior.

2.4. Momentos

Son de especial importancia ya que además de caracterizar a la distribución, sirven para construir medidas centrales y de dispersión.

El **momento de orden** r de la muestra$\{x_1, x_2, \ldots, x_N\}$ se define como

$$\alpha_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - a)^r$$

Del mismo modo se define el momento de orden r de la muestra$\{x_1, x_2, \ldots, x_N\}$ con respecto a a como

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - a)^r$$

Así $m_1 = 0$, $a = \alpha_1$ y $m_2 = \alpha_1^2 + \sigma^2$.

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^r$$

Ejemplo Los momentos de orden 1, 2, 3 y 4 de la muestra $\{2,4,6,8,10,12,14,16,18\} = \{2j\}_{j=1}^{9}$ con repecto a la media aritmética

$$a = \frac{1}{9}\sum_{j=1}^{9} 2j = 10$$

son

$$\tfrac{1}{9}\sum_{i=1}^{9}\left(2i - \tfrac{1}{9}\sum_{j=1}^{9} 2j\right)^{1} = 0 \qquad \tfrac{1}{9}\sum_{i=1}^{9}\left(2i - \tfrac{1}{9}\sum_{j=1}^{9} 2j\right)^{2} = \tfrac{80}{3} \approx 26.67$$

$$\tfrac{1}{9}\sum_{i=1}^{9}\left(2i - \tfrac{1}{9}\sum_{j=1}^{9} 2j\right)^{3} = 0 \qquad \tfrac{1}{9}\sum_{i=1}^{9}\left(2i - \tfrac{1}{9}\sum_{j=1}^{9} 2j\right)^{4} = \tfrac{3776}{3} \approx 1258.7$$

2.5. Correlación

Los métodos de Estadística Descriptiva nos proveen, fundamentalmente, de haerramientas para analizar por separado a estadísticos o muestras. Sin embargo es sumamente interesante tratar de establecer relaciones entre los valore de dos estadísticos (o muestras) distintos. Para poder establecer cómo es la posible relación, la correlación, entre dos estadíticos, a un nivel muy elemental, se desarrollo la regresión lineal.
Para poder comparar las variables X e Y tenemos en cuenta sus muestras, del mismo tamaño, N, yconstruimos el estadístico

$$\rho(X,Y) = \rho = \frac{\sum_{i=1}^{N}(x_i - a_X)(y_j - a_Y)}{N}$$

donde $\overline{x} = a_X$ e $\overline{y} = a_Y$, son las medias aritméticas de X e Y respectivamente, y σ_X y σ_Y son las desviaciones dtípicas. $\rho(X,Y)$ se conoce con el nombre de coeficiente de correlación. Se sabe que $\rho = \rho(X,Y)$ sstisface $-1 \leq \rho \leq 1$, y que $\rho = 0$ indica que no existe correlación lineal. Si se trata de dos variables independientes entonces $\rho = 0$. Se define la recta de regresión linel entre X and Y como

$$Y = aX + b$$

for some constants a and b, donde

$$\begin{aligned} a &= \frac{\rho}{\sigma_X^2} \\ b &= a_Y - a_X\left(\frac{\rho}{\sigma_X^2}\right) \end{aligned}$$

[2]

[2]Se pospone la explicación de la obtención de la recta de regresión lineal.

3. Ejemplos

1. Moda, mediana y Valor Mediana de una distribución discreta con un número impar de datos: $\{2, 2, 7, 4, 5\}$. Idem para una distribución discreta con un número par de datos: $\{2, 3, 0, 2, 4, 5, 6, 5\}$.

2. En 1879, Michelson obtuvo los siguientes valores para la velocidad de la luz en el aire (se dan los resultados restando 299000 a los datos originales en Km/seg): 850, 740, 900, 1070, 930, 850, 950, 980, 980, 880, 1000, 980, 930, 650 y 760.
   En 1882, Newcomb, utilizando otro procedimiento, obtuvo 883, 816, 778, 796, 682, 711, 611, 599 1051, 781, 578, 796, 774, 820, 772. Se pide:

   (a) Diagramas de tallo y hojas para ambas distribuciones

   (b) Medias y desviaciones típicas

   (c) Conclusiones

3. Realizar un estudio descriptivo para la distribución de calificaciones de 180 alumnos que a continuación se relaciona

X_i	n_i	N_i
0	4	4
1	5	9
2	15	24
3	16	40
4	20	60
5	40	100
6	34	134
7	26	160
8	12	172
9	6	178
10	2	180

4. Idem para la siguiente distribución la distribución de calificaciones de 160 alumnos

X_i	n_i	N_i
0	4	4
1	6	10
2	8	18
3	12	30
4	15	45
5	35	80
6	37	117
7	16	133
8	13	146
9	8	154
10	6	160

5. Se estudia el recorrido realizado por 100 animales de una especia, en estado de cautividad. Las observaciones del recorrido se agrupan en intervalos de clase, según la siguiente tabulación final.

$[C_i]$	x_i	n_i	f_i	N_i	F_i
[12.000 − 12.300)	12.150	12	0.12	12	0.12
[12.300 − 12.600)	12.450	20	0.20	32	0.32
[12.600 − 12.900)	12.750	35	0.35	67	0.67
[12.900 − 13.200)	13.050	30	0.30	97	0.97
[13.200 − 13.500]	13.350	3	0.03	100	1.00

Se pide

(a) El intervalo modal y la Moda

(b) Mediana

(c) ¿Cuál es la distancia superada por el 80%& de los animales?

(d) ¿Cuál es la frecuencia de animales que superan los 13.000 metros?

6. Sea la muestra realizada sobre el tamaño de los huevos de cuco (Biometrika, 1902):

Anchura	N^o de huevos
$[13.75, 14.25]$	1
$[14.25, 14.75]$	1
$[14.75, 15.25]$	5
$[15.25, 15.75]$	9
$[15.75, 16.25]$	73
$[16.25, 16.75]$	51
$[16.75, 17.25]$	80
$[17.25, 17.75]$	15
$[17.75, 18.25]$	7
$[18.25, 18.75]$	0
$[18.75, 19.25]$	1

Realizar un estudio descriptivo de la muestra.

7. Un proceso químico se lleva a cabo en tres etapas. En la primera la tasa de reacción es de 10 moles/sg. Y dura 30 segundos, en la segunda es de 20 moles/sg. Y dura 20 sg, y en la tercera es de 30 moles/sg. Y dura 10 sg. Calcular la tasa de reacción media del proceso a través de las siguientes fórmulas:

(a) Media aritmética

(b) Media cuadrática

(c) Media armónica

(d) Media geométrica

8. Dos lombrices de una especie determinada son colocadas en un terrario de 30 mm de longitud eficaz. La primera recorre los 30 mm a una velocidad de 10 mm/minuto y la segunda recorre los 30 mm a una velocidad de 20 mm./minuto ¿Cuál será la velocidad media de avance de esa especie de lombriz? Probar con otros tipos de medias y analizar cuál es la más adecuada.

9. En tres grandes poblaciones de vegetales que crecen en un bosque A, B y C la proporción de individuos infectados por un determinado hongo es del 30, 60 y 10% respectivamente. Se toma al azar una de las 3 poblaciones (las tres son igual de probables), y de ella elegimos a 10 individuos al azar, resultando que 2 de ellos están infectados. ¿A qué población es más probable que pertenezcan?

1. Para el primer caso se tiene que la media y la mediana valen 4, la moda es 2 y el intervalo mediana es $(2, 4)$. En el segundo caso la media es $3.\,375$, la mediana es $\frac{7}{2}$ y el valor mediana es 3.

2. Para la muestra

$$\{850, 740, 900, 1070, 930, 850, 950, 980, 980, 880, 1000, 980, 930, 650, 760\}$$

tenemos media aritmética $a = 896.\,67$ y desviación típica $\sqrt[2]{12524} = 111.\,91$, y para

$$\{883, 816, 778, 796, 682, 711, 611, 599, 1051, 781, 578, 796, 774, 820, 772\}$$

$a = 763.\,2$, y $\sigma = \sqrt[2]{14273} = 119.\,47$.
Parece más adecuado el primer muestreo pues tiene menor desviación. Sin embargo la diferencia es muy reducida. En todo caso una aproximación (medida central para la velocidad de la luz) sería $299000 + 763.2 = 2.\,997\,6 \times 10^5$ km/seg
Los diagramas de tallos y hojas no predicen con clarida cua podría ser la predicción más precisa: ´después de dividir por 10 todos los datos quedan los diagramas siguientes:

5		5	0
6	5	6	810
7	46	7	81877
8	558	8	82002
9	0358883	9	
10	70	10	5

3.

4.

5.

6.

7.

8.

9. Del enunuciado del problema se desprende lo siguiente: si Inf denota estar infectado y A, B y C haber elegido a dichas poblaciones, entonces

$$P(Inf|\,A) = 3/10,\ \ P(Inf|\,B) = 6/10,\ \ P(Inf|\,C) = 1/10,$$

y

$$P(A) = P(B) = P(C) = \frac{1}{3}.$$

Se sabe que S el suceso consistente en la realización de 10 obsevaciones con resultado de 2 infecciones ha tenido lugar. La pregunta es saber cuál de las siguientes cantidades es mayor:

$$P(A|\,S),\ \ P(B|\,S),\ \ P(C|\,S).$$

Calculemos: se sabe que

$$P(A|\,S) = \frac{P(A \cap S)}{P(S)} = \frac{P(S|\,A)P(A)}{P(S|\,A)P(A) + P(S|\,B)P(B) + P(S|\,C)P(C)},$$

ahora bien, $P(S|\,A)$ es la probabilidad de que hayan tenido lugar 2 infecciones de 10 individuos sabiendo que la probabilidad de infección en A es $3/10$. Esta probabilidad no es otra cosa que el valor de la masa de un binomial de parámetros $p = 3/10$ y $n = 10$; análogamente para las otras probabilidades condicionadas. Por tanto

$$P(S|\,A) = \binom{10}{2} \left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^8$$

y así

$$P(A|\,S) = \frac{\binom{10}{2} \left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^8}{\binom{10}{2} \left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^8 + \binom{10}{2} \left(\frac{6}{10}\right)^2 \left(\frac{4}{10}\right)^8 + \binom{10}{2} \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^8}.$$

Puesto que el denominador que aparece en el cálculo de $P(A|\,S)$, $P(B|\,S)$ y $P(C|\,S)$ es el mismo, basta con comparar

$$\left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^8, \left(\frac{6}{10}\right)^2 \left(\frac{4}{10}\right)^8, \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^8.$$

El valor más grande es el correspondiente a $P(A|\,S)$ [3]. La población con más probabilidad de haber sido elegida es A.

[3] Para cerciorarse de ello basta con esbozar la gráfica de la función $\psi(p) = p^2(1-p)^8$ y notar que el máximo se da para $p = 0.2$. Los valores en 0.6 y 0.1 están más lejos de este máximo que 0.3.