ESTADISTICA

1- Estadística Descriptiva

1. 1 – Introducción

El campo de la estadística tiene que ver con la recopilación, organización, análisis y uso de datos para tomar decisiones razonables basadas en tal análisis.
Al recoger datos relativos a las características de un grupo de individuos u objetos, sean alturas y pesos de estudiantes de una universidad o tuercas defectuosas producidas por una fábrica, suele ser imposible o poco práctico observar todo el grupo, en especial si es muy grande. En vez de examinar el grupo entero, llamado ***población*** o universo, se examina una pequeña parte del grupo, llamada ***muestra***.
En muchos problemas estadísticos es necesario utilizar una muestra de observaciones tomadas de la población de interés con objeto de obtener conclusiones sobre ella. A continuación se presenta la definición de algunos términos

Una ***población*** está formada por la totalidad de las observaciones en las cuales se tiene cierto interés.

Una ***muestra*** es un subconjunto de observaciones seleccionada de una población

Si una muestra es representativa de una población, es posible inferir importantes conclusiones sobre la población a partir del análisis de la muestra. La parte de la estadística que trata sobre las condiciones bajo las cuales tal inferencia es válida se llama ***estadística inductiva*** o ***inferencia estadística***. Ya que dicha inferencia no es del todo exacta, el lenguaje de las probabilidades aparecerá al establecer nuestras conclusiones.
La parte de la estadística que estudia la muestra sin inferir alguna conclusión sobre la población es la ***estadística descriptiva***.
En particular la estadística descriptiva trata sobre los métodos para recolectar, organizar y resumir datos.
La estadística descriptiva puede a su vez dividirse en dos grandes áreas: métodos gráficos y métodos numéricos.
En lo referente a la notación, ***n*** representa el número de observaciones en un conjunto de datos, las observaciones están representadas por una variable con subíndice (por ejemplo x_1, x_2, …, x_n). Así la representación de los cinco valores, $n = 5$, de la velocidad de un chip de computadora en MHz medida por un ingeniero, será: $x_1 = 481.5$, $x_2 = 493.7$, $x_3 = 471.8$, $x_4 = 486.4$, $x_5 = 496.2$,

1.2 – Distribución de frecuencias e histogramas

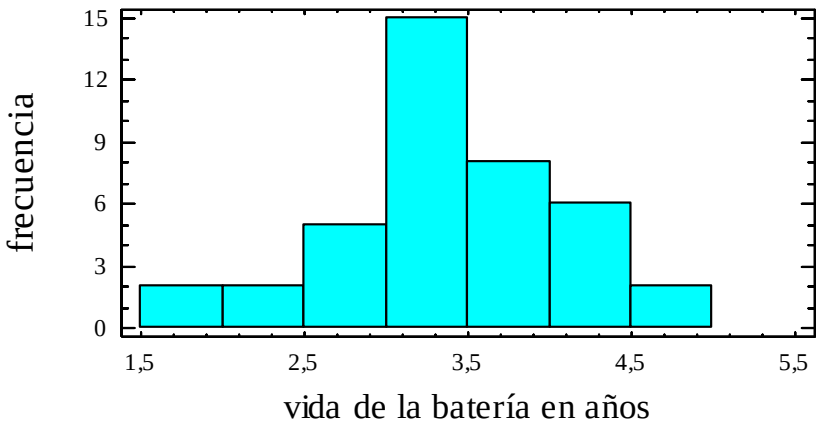
Supongamos que los siguientes datos representan la vida de 40 baterías para automóvil similares, registradas al décimo de año más cercano. Las baterías se garantizan por tres años.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6	3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1	3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5								

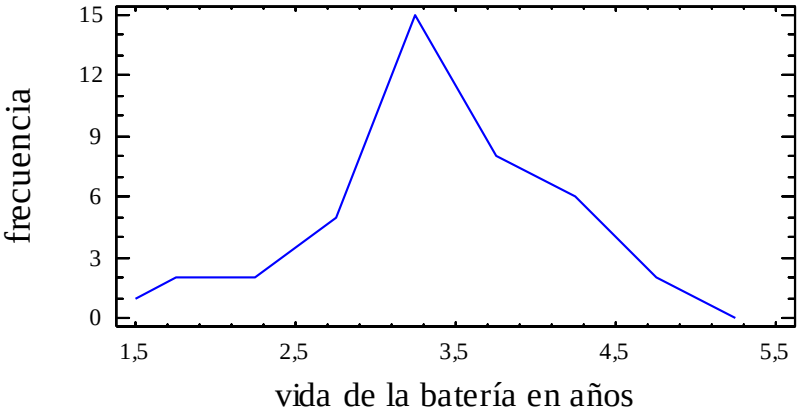
Para organizar los datos buscamos el mínimo y el máximo de la muestra, en este caso el mínimo es 1.6 y el máximo es 4.7
Elegimos un intervalo (a, b) que contenga todos los datos, por ejemplo a = 1.5 y b = 5.0. Dividimos el intervalo (a, b) en subintervalos que pueden ser de igual longitud, pero no necesariamente, y contamos cuántas observaciones caen en cada subintervalo, esa será la ***frecuencia*** del intervalo. Para esto debemos decidir cuántos subintervalos utilizaremos. En general se puede usar la regla de tomar aproximadamente $\sqrt{n}$ subintervalos. Los subintervalos se llaman ***intervalos de clase*** o simplemente ***clases***. Resulta satisfactorio utilizar no menos de 5 clases ni más de 20.
En el ejemplo $\sqrt{40} \approx 6$, entonces 6 o 7 clases será una elección satisfactoria.
Como $b - a = 5.0 - 1.5 = 3.5$, si tomamos r = 7 clases entonces la longitud de cada una sería $(b-a)/r = 0.5$.
Construimos una ***tabla de frecuencias*** de manera tal que, por ejemplo, en el intervalo (1.5 , 2.0) están las observaciones ***mayores*** a 1.5 y ***menores o iguales que*** 2.0.
Los extremos de los intervalos de clase son los ***límites de clase inferior y superior.***

Intervalo de clase	***Marca de clase***	***Frecuencia f***	***Frecuencia relativa***	***Frecuencia acumulada***	***Frecuencia Acumulada relativa***
1.5 – 2.0	1.75	2	0.05	2	0.075
2.0 – 2.5	2.25	2	0.05	4	0.1
2.5 – 3.0	2.75	5	0.125	9	0.225
3.0 – 3.5	3.25	15	0.375	24	0.6
3.5 – 4.0	3.75	8	0.2	32	0.8
4.0 – 4.5	4.25	6	0.15	38	0.95
4.5 – 5.0	4.75	2	0.05	40	1.000

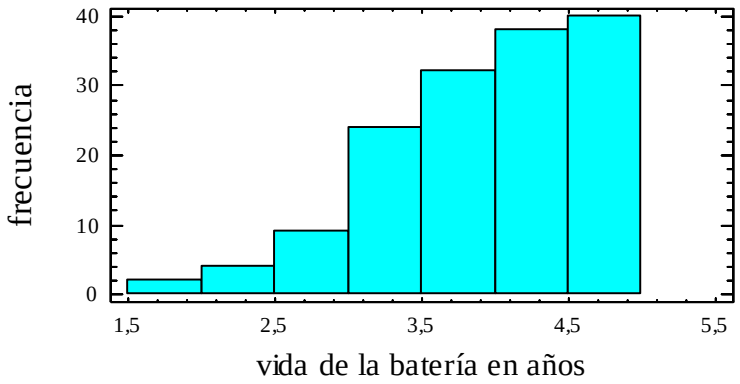
El punto medio de cada clase es la ***marca de clase***. La longitud de cada intervalo de clase es el ***ancho de clase***.
El ***gráfico de la tabla de frecuencias*** es el ***histograma***. Se construye en un sistema de ejes cartesianos. Sobre el eje de absisas se marcan los límites de clase, y en cada clase se construye un rectángulo cuya base es el intervalo de clase y el ***área del mismo*** debe ser proporcional a la ***frecuencia de la clase***. Si los intervalos de clase tienen el ***mismo ancho*** se puede construir cada rectángulo de manera que su ***altura sea igual a la frecuencia*** de la clase correspondiente. Estos histogramas son más fáciles de interpretar.
En la figura siguiente se muestra el histograma referido a la tabla de frecuencias anterior generado por el paquete Statgraphics.



Si se tienen pocos datos los histogramas pueden cambiar de apariencia al variar el número de clases y el ancho de las mismas.
Si se hubiera graficado el ***histograma de frecuencias relativas*** el aspecto sería el mismo con la diferencia de la notación en el eje de ordenadas.
También se puede graficar un ***polígono de frecuencias*** al unir los puntos medios del lado superior de cada rectángulo con segmentos y agregar en los extremos dos clases adicionales de frecuencia cero como indica la siguiente figura



La figura siguiente muestra un ***histograma de frecuencias acumuladas*** disponible en el paquete Statgraphics. En esta gráfica la altura de cada rectángulo representa el número total de observaciones que son menores o iguales al límite superior de la clase respectiva.



Los histogramas son útiles al proporcionar una impresión visual del aspecto que tiene la distribución de las mediciones, así como información sobre la dispersión de los datos.
Al construir una tabla de frecuencias se pierde información, sin embargo esa pérdida de información es a menudo pequeña si se le compara con la concisión y la facilidad de interpretación ganada al utilizar la distribución de frecuencias y el histograma.
Las distribuciones acumuladas también son útiles en la interpretación de datos; por ejemplo en la figura anterior puede leerse de inmediato que existen aproximadamente 25 baterías con duración menor o igual a 3.5 años.

1.3 – Diagrama de tallo y hoja

El diagrama de tallo y hoja es una buena manera de obtener una presentación visual informativa del conjunto de datos $x_1, x_2, \ldots, x_n$, donde cada número x_i está formado al menos por dos dígitos. Para construir un diagrama de este tipo los números x_i se dividen en dos partes: un ***tallo***, formada por uno o más dígitos principales, y una ***hoja***, la cual contiene el resto de los dígitos. Para ilustrar lo anterior consideramos los datos que especifican la vida de 40 baterías para automóvil dados anteriormente. Dividimos cada observación en dos partes de manera que el tallo representa el dígito entero que antecede al decimal, y la hoja corresponde a la parte decimal del número. Por ejemplo, para el número 3.7 el dígito 3 designa el tallo, y el 7 la hoja. Para nuestros datos los cuatro tallos 1, 2, 3 y 4 se listan verticalmente del lado izquierdo de la tabla, en tanto que las hojas se registran en el lado derecho correspondiente del valor del tallo adecuado. Entonces la hoja 6 del número 1.6 se registra enfrente del tallo 1, la hoja 5 del número 2.5 enfrente del tallo 2, y así sucesivamente.

Tallo	*Hoja*
1	6 9
2	2 5 6 6 9
3	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4 5 5 6 7 7 7 8 8 9 9
4	1 1 2 3 4 5 7 7

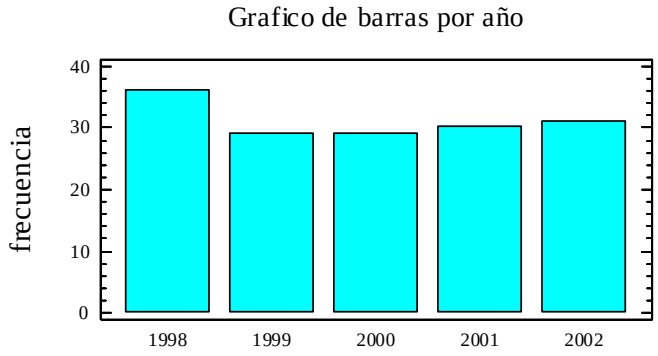
Podríamos aumentar el número de tallos para obtener una forma mas adecuada de la distribución de los datos, para esto escribimos dos veces cada valor del tallo y después registramos las hojas 0, 1, 2, 3 y 4 enfrente del valor del tallo adecuado donde aparezca

por primera vez, y las hojas 5, 6, 7, 8 y 9 enfrente de este mismo valor del tallo donde aparece por segunda vez.
En la tabla siguiente se ilustra el nuevo diagrama de tallo y hoja donde a los tallos que corresponden a las hojas 0 a 4 se les anotó un símbolo *, y al tallo correspondiente a las hojas 5 a 9 se les anotó el símbolo · .

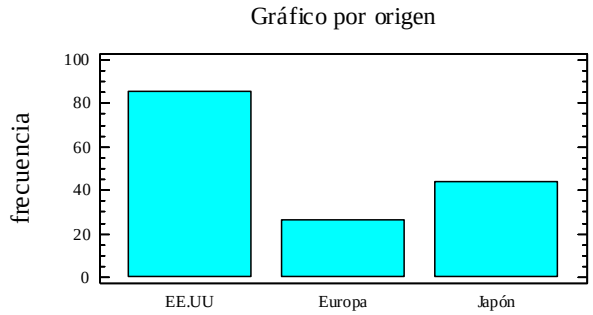
Tallo	***Hoja***
1	6 9
2*	2
2·	5 6 6 9
3*	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4
3·	5 5 6 7 7 7 8 8 9 9
4*	1 1 2 3 4
4·	5 7 7

En cualquier problema específico, se debe decidir cuáles son los valores del tallo adecuados. Se trata de una decisión que se toma algo arbitrariamente, aunque nos guiamos por el tamaño de nuestra muestra. Por lo general se eligen entre 5 y 20 tallos. Cuanto menor es el número de datos disponibles, menor será la elección del número de tallos.

Observación:
Las tablas de frecuencia y los histogramas también pueden emplearse en ***datos cualitativos o categóricos***, es decir la muestra no consiste de valores numéricos ***(datos cuantitativos***) sino que los datos se ordenan en ***categorías*** y se registra cuántas observaciones caen en cada categoría (las categorías pueden ser *masculino , femenino* o *fumador, no fumador* o clasificar según nivel educativo: *primario, secundario, terciario, universitario, ninguno*). Cuando los datos son categóricos las clases se dibujan con el mismo ancho.
Por ejemplo
El gráfico siguiente corresponde a una muestra de 155 autos clasificados según el año de fabricación: 1998, 1999, 2000, 2001 y 2002. Notar que el modelo 1998 es el de mayor frecuencia.



En el siguiente gráfico se clasifican los autos según el origen: americano, europeo o japonés. Notar que hay mayoría de autos americanos.



1.4 – Medidas Descriptivas

1.4.1 – Medidas de Localización

Del mismo modo que las gráficas pueden mejorar la presentación de los datos, las descripciones numéricas también tienen gran valor. Se presentan varias medidas numéricas importantes para describir las características de los datos.
Una característica importante de un conjunto de números es su ***localización*** o ***tendencia central***.

Media

La medida más común de localización o centro de un grupo de datos es el promedio aritmético ordinario o media. Ya que casi siempre se considera a los datos como una muestra, la media aritmética se conoce como ***media muestral.***

Si las observaciones de una muestra de tamaño n son $x_1, x_2, \ldots, x_n$ entonces la ***media muestral*** es

$$\bar{x} = \frac{x_1 + x_2 + \ldots. + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ejemplo:
La media muestral de la vida útil en años de una batería de las 40 observaciones dadas en ejemplos anteriores es

$$\bar{x} = \frac{\sum_{i=1}^{40} x_i}{40} = \frac{2.2 + 4.1 + 3.5 + 4.5 + 3.2 + \ldots. + 3.5}{40} = 3.4125$$

La media tiene como ventaja su fácil cálculo e interpretación, pero tiene como desventaja el hecho de distorsionarse con facilidad ante la presencia de ***valores atípicos*** en los datos. Si en el ejemplo anterior tenemos $x_{40} = 350$ en lugar de 3.5 entonces $\bar{x} = 12.075$ dando así la idea errónea que los datos en su mayor parte se concentran alrededor de 12.075

Mediana

Otra medida de tendencia central es la ***mediana***, o punto donde la muestra se divide en dos partes iguales.

Sean $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ una muestra escrita en orden creciente de magnitud; esto es $x_{(1)}$ denota la observación más pequeña, $x_{(2)}$ la segunda observación más pequeña, ……, y $x_{(n)}$ la observación más grande. Entonces la ***mediana*** $\tilde{x}$ se define como la observación del lugar $\frac{n+1}{2}$ si n es impar, o el promedio de las observaciones de los lugares $\frac{n}{2}$ y $\frac{n}{2}+1$ si n es par. En términos matemáticos

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & n \text{ impar} \\ \dfrac{x_{(n/2)} + x_{((n/2)+1)}}{2} & n \text{ par} \end{cases}$$

La ventaja de la mediana es que los valores extremos no tienen mucha influencia sobre ella.
Ejemplo:
Supóngase que los valores de una muestra son

1, 3, 4, 2, 7, 6, 8

Ordenamos los valores de menor a mayor: 1, 2, 3, 4, 6, 7, 8
Como son $n = 7$ valores y 7 es impar entonces la mediana es el valor del lugar $\frac{7+1}{2} = 4$, es decir $x_{(4)} = 4$
La media muestral es $\bar{x} = 4.4$. Ambas cantidades proporcionan una medida razonable de la tendencia central de los datos.

Si ahora tenemos la muestra 1, 2, 3, 4, 2450, 7, 8, entonces la media muestral es $\bar{x} = 353.6$
En este caso la media no dice mucho con respecto a la tendencia central de la mayor parte de los datos. Sin embargo la mediana sigue siendo $x_{(4)} = 4$, y ésta es una medida de tendencia central más significativa para la mayor parte de las observaciones.

Moda

La ***moda*** es la observación que se presenta con mayor frecuencia en la muestra

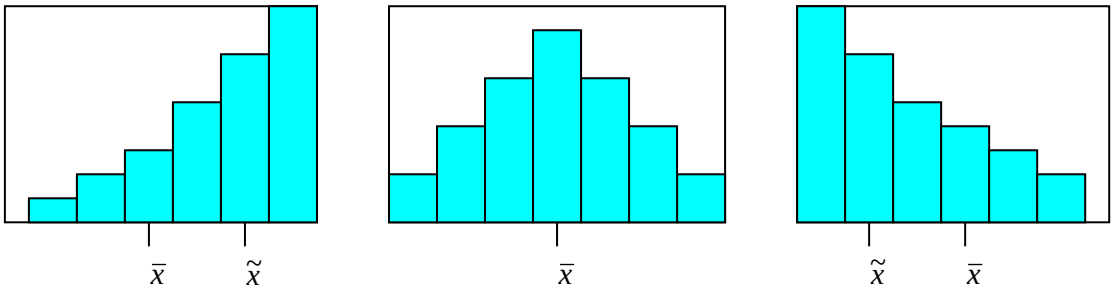
Por ejemplo la moda de los siguientes datos 3, 6, 9, 3, 5, 8, 3, 10, 4, 6, 3, 1
Es 3, porque este valor ocurre 4 veces y ningún otro lo hace con mayor frecuencia.

Puede existir más de una moda. Por ejemplo considérense las observaciones

3, 6, 9, 3, 5, 8, 3, 10, 4, 6, 3, 1, 6, 2, 5, 6

La moda son 3 y 6, ya que ambos valores se presentan cuatro veces., y ningún otro lo hace con mayor frecuencia. Se dice en este caso que los datos son *bimodales*.

Si los datos son simétricos entonces la media y la mediana coinciden. Si además los datos tienen una sola moda entonces la media, la mediana y la moda coinciden.
Si los datos están *sesgados* (esto es, son asimétricos, con una larga cola en uno de los extremos), entonces la media, la mediana y la moda no coinciden. Generalmente se encuentra que moda < mediana < media si la distribución está sesgada a la derecha, mientras que moda > mediana > media si la distribución está sesgada hacia la izquierda.
Eso se representa en la siguiente figura



Sesgo a la izquierda

Simétrica

Sesgo a derecha

Percentiles y cuartiles

La mediana divide los datos de una muestra en dos partes iguales. También es posible dividir los datos en más de dos partes. Cuando se divide un conjunto ordenado de datos en cuatro partes iguales, los puntos de división se conocen como ***cuartiles***. El *primer cuartil* o *cuartil inferior*, q_1, es un valor que tiene aproximadamente la cuarta parte (25%) de las observaciones por debajo de él, y el 75% restante, por encima de él. El *segundo cuartil* , q_2, tiene aproximadamente la mitad (50%) de las observaciones por debajo de él. El segundo cuartil coincide con la mediana. El *tercer cuartil* o *cuartil superior*, q_3, tiene aproximadamente las tres cuartas partes (75%) de las observaciones por debajo de él. Como en el caso de la mediana, es posible que los cuartiles no sean únicos. Por simplicidad en este caso, si más de una observación cumple con la definición se utiliza el promedio de ellas como cuartil.
Ejemplo:
En 20 automóviles elegidos aleatoriamente se tomaron las emisiones de hidrocarburos en velocidad al vacío, en partes por millón (ppm)

141 359 247 940 882 494 306 210 105 880 200 223 188 940 241 190 300 435 241 380

Primero ordenamos los datos de menor a mayor:

105, 141, 188, 190, 200, 210, 223, 241, 241, 247, 300, 306, 359, 380, 435, 494, 880, 882, 940, 940

Buscamos la mediana o segundo cuartil, como $n = 20$ y es número par entonces la mediana es el promedio de la observaciones que se encuentran en los lugares $\frac{n}{2} = 10$ y $\frac{n}{2} + 1 = 11$, es decir $\tilde{x} = q_2 = \frac{247 + 300}{2} = 273.5$

Ahora buscamos el primer cuartil, para esto tomamos las primeras 10 observaciones

105, 141, 188, 190, 200, 210, 223, 241, 241, 247

y de éstas calculamos la mediana, por lo tanto $q_1 = \frac{200 + 210}{2} = 205$

Análogamente, para calcular el tercer cuartil, tomamos las últimas 10 observaciones y calculamos la mediana de éstas

300, 306, 359, 380, 435, 494, 880, 882, 940, 940

$$q_3 = \frac{435 + 494}{2} = 464.5$$

Cuando un conjunto ordenado de datos se divide en cien partes iguales, los puntos de división reciben el nombre de ***percentiles***. En términos matemáticos el $100k$ – ésimo percentil p_k se define:

> El $100k$ – ésimo percentil p_k es un valor tal que al menos el $100k$% de las observaciones son menores o iguales a él, y $100(1-k)$% son mayores o iguales a él.

Notar que $p_{0.25} = q_1$, $p_{0.5} = q_2$, $p_{0.75} = q_3$

Una regla práctica para calcular los percentiles de un conjunto de ***n*** datos es la siguiente:
para calcular p_k hacemos el producto nk

si nk es un número entero i entonces $p_k = \frac{x_{(i)} + x_{(i+1)}}{2}$

si nk no es un número entero entonces tomamos la parte entera de nk : $\lfloor nk \rfloor = i$ y entonces $p_k = x_{(i+1)}$

Con los datos anteriores calculamos $p_{0.88}$
$n = 20$ $k = 0.88$ $nk = 20 \times 0.88 = 17.6$ la parte entera es $\lfloor 17.6 \rfloor = 17$ entonces
$p_{0.88} = x_{18} = 882$

Calculamos $p_{0.10}$:

$n = 20 \quad k = 0.10 \quad nk = 20 \times 0.10 = 2$ entonces

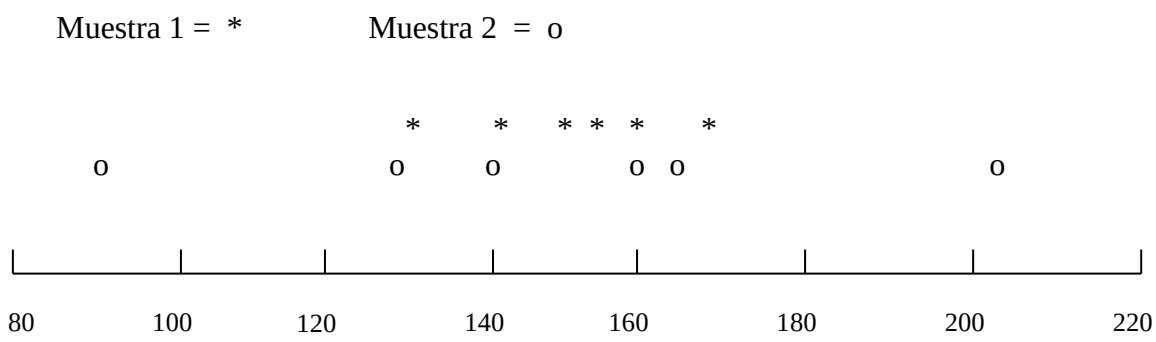
$$p_{0.10} = \frac{x_{(2)} + x_{(3)}}{2} = \frac{141 + 188}{2} = 164.5$$

1.4.2 - **Medidas de variabilidad**

La localización o tendencia central no necesariamente proporciona información suficiente para describir datos de manera adecuada. Por ejemplo, supongamos que tenemos dos muestras de resistencia a la tensión (en psi) de aleación de aluminio-litio:

Muestra 1: 130, 150, 145, 158, 165, 140
Muestra 2: 90, 128, 205, 140, 165, 160

La media en ambas muestras es 148 psi. Sin embargo la dispersión o variabilidad de la muestra 2 es mucho mayor que la de la muestra 1. Para ilustrar esto hacemos para cada muestra un ***diagrama de puntos***:



Rango de la muestra y rango intercuartílico

Una medida muy sencilla de variabilidad es el ***rango de la muestra***, definido como la diferencia entre las observaciones más grande y más pequeña. Es decir

$$rango = \max(x_i) - \min(x_i)$$

Para las muestras anteriores
Muestra 1 ⟶ *rango* = 165-130 = 35

Muestra 2 ⟶ *rango* = 205-90 = 115

Está claro que a mayor rango, mayor variabilidad en los datos.

El rango ignora toda la información que hay en la muestra entre las observaciones más chica y más grande. Por ejemplo las muestras 1, 4, 6, 7, 9 y 1, 5, 5, 5, 9 tienen el mismo rango (rango = 8), Sin embargo en la segunda muestra sólo existe variabilidad en

los valores de los extremos, mientras que en la primera los tres valores intermedios cambian de manera considerable.
Al igual que las observaciones máxima y mínima de una muestra llevan información sobre la variabilidad, el ***rango intercuartílico*** definido como $q_3 - q_1$ puede emplearse como medida de variabilidad.

$$RIC = q_3 - q_1$$

Para los datos de las emisiones de hidrocarburos en velocidad al vacío, el rango intercuartílico es $q_3 - q_1 = 464.5 - 205 = 259.5$.
El rango intercuartílico es menos sensible a los valores extremos de la muestra que el rango muestral.

Varianza muestral y desviación estándar muestral

Si $x_1, x_2, \ldots, x_n$ es una muestra de *n* observaciones, entonces la ***varianza muestral*** es

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$

La ***desviación estándar muestral***, s, es la raíz cuadrada positiva de la varianza muestral

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Por ejemplo, para los datos de las emisiones de hidrocarburos en velocidad al vacío

141 359 247 940 882 494 306 210 105 880 200 223 188 940 241 190 300 435 241 380

la media muestral es $\bar{x} = 395.1$ ppm y la varianza muestral es

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} =$$

$$= \frac{\left(141-395.1\right)^2 + \left(359-395.1\right)^2 + \left(247-395.1\right)^2 + \ldots\ldots + \left(380-395.1\right)^2}{19} = 78998.5 \text{ ppm}^2$$

y la desviación estándar es $s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}} = \sqrt{78998.5} = 281.067 \text{ ppm}$

Observaciones:

1) La varianza muestral está en las unidades de medida de la variable al cuadrado. Por ejemplo si los datos están medidos en metros, las unidades de la varianza son metros al cuadrado. La desviación estándar tiene la propiedad de estar expresada en las mismas unidades de medida de las observaciones.

2) Se puede expresar la varianza muestral como

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}{n-1}$$

pues:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i^2 - 2\,x_i\,\bar{x} + \bar{x}^2)}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\,x_i\,\bar{x} + \sum_{i=1}^{n}\bar{x}^2}{n-1} =$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}n\dfrac{\sum_{i=1}^{n} x_i}{n} + n\,\bar{x}^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\,\bar{x}^2}{n-1} =$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}{n-1}$$

En el ejemplo anterior volvemos a calcular la varianza muestral pero con esta última expresión

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}{n-1} = \frac{4623052 - (20)(395.1)^2}{19} = 78998.5 \text{ ppm}^2$$

Coeficiente de variación

El ***coeficiente de variación muestral*** es

$$cv = \frac{s}{\bar{x}}$$

El coeficiente de variación muestral es útil cuando se compara la variabilidad de dos o más conjuntos de datos que difieren de manera considerable en la magnitud de las observaciones.

Ejemplo: con un micrómetro se realizan mediciones del diámetro de un balero, que tienen una media de 4.03 mm y una desviación estándar de 0.012 mm; con otro micróme-

tro se toman mediciones de la longitud de un tornillo, que tienen una media de 1.76 pulgadas y una desviación estándar de 0.0075 pulgadas.
Los coeficientes de variación son

balero $\longrightarrow$ $cv = \dfrac{0.012}{4.03} = 0.003$

tornillo $\longrightarrow$ $cv = \dfrac{0.0075}{1.76} = 0.004$

en consecuencia, las mediciones hechas con el primer micrómetro tienen una variabilidad relativamente menor que las efectuadas con el otro micrómetro.

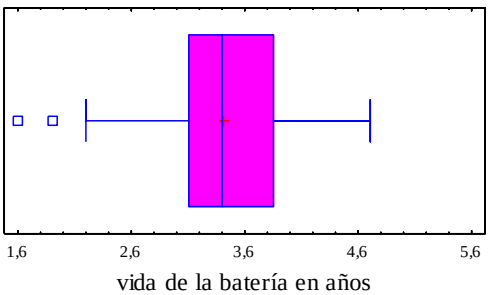
1.5 – Diagramas de caja

El ***diagrama de caja*** es una presentación visual que describe al mismo tiempo varias características importantes de un conjunto de datos, tales como el centro, la dispersión, la desviación de la simetría y la presencia de valores atípicos.
El diagrama de caja presenta los tres cuartiles, y los valores mínimo y máximo de los datos sobre un rectángulo en posición horizontal o vertical. El rectángulo delimita el rango intercuartílico con la arista izquierda ubicada en el primer cuartil, y la arista derecha ubicada en el tercer cuartil. Se dibuja una línea a través del rectángulo en la posición que corresponde al segundo cuartil. De cualquiera de las aristas del rectángulo se extiende una línea, o ***bigote***, que va hacia los valores extremos. Éstas son observaciones que se encuentran entre cero y 1.5 veces el rango intercuartílico a partir de las aristas del rectángulo. Las observaciones que están entre 1.5 y 3 veces el rango untercuartílico a partir de las aristas del rectángulo reciben el nombre de ***valores atípicos***. Las observaciones que están más allá de tres veces el rango intercuartílico a partir de las aristas del rectángulo se conocen como ***valores atípicos extremos***. En ocasiones se emplean diferentes símbolos para identificar los dos tipos de valores atípicos. A veces los diagramas de caja reciben el nombre de ***diagramas de caja y bigotes***.
Ejemplo: volvemos a los datos que representan la vida de 40 baterías para automóvil similares, registradas al décimo de año más cercano.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6	3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1	3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5								

La figura siguiente presenta un diagrama de caja obtenido con el paquete Statgraphics

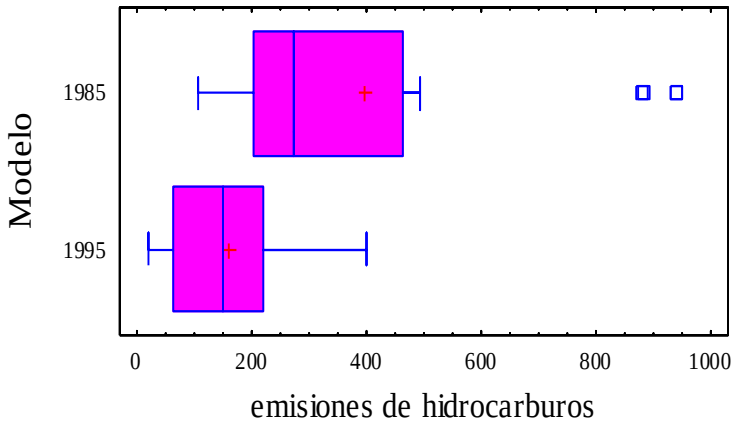
tamaño de muestra = 40
Media = 3.4125
Mediana = 3.4
Mínimo = 1.6
Máximo = 4.7
Rango = 3.1
1º cuartil = 3.1
3º cuartil = 3.85
rango untercuartílico (*RIC*) = 0.75
1.5 *RIC* = 1.125
3 *RIC* = 2.25



El diagrama indica que la distribución de la vida en años de una batería de automóvil es bastante simétrica con respecto al valor central ya que los bigotes izquierdo y derecho, así como la longitud de los rectángulos izquierdo y derecho alrededor de la mediana, son casi los mismos. También se observa la existencia de dos valores atípicos de rango medio en el extremo izquierdo de los datos.

Los diagramas de caja son muy útiles para hacer comparaciones gráficas entre conjuntos de datos, ya que tienen un gran impacto visual y son fáciles de comprender.

Ejemplo: En 20 automóviles elegidos aleatoriamente, se tomaron las emisiones de hidrocarburos en velocidad al vacío, en partes por millón, para modelos de 1985 y 1995.
Modelos 1985:
141 359 247 940 882 494 306 210 105 880 200 223 188 940 241 190 300 435 241 380
Modelos 1995:
140 160 20 20 223 60 20 95 360 70 220 400 217 58 235 380 200 175 85 65
La figura siguiente presenta los diagramas de caja comparativos generados por el paquete Statgraphics



Modelo 1985
Media = 395.1
Mediana = 273.5
Mínimo = 105
Máximo = 940
Rango = 835
1º cuartil = 205
3º cuartil = 464.5
rango untercuartílico (*RIC*) = 259.5

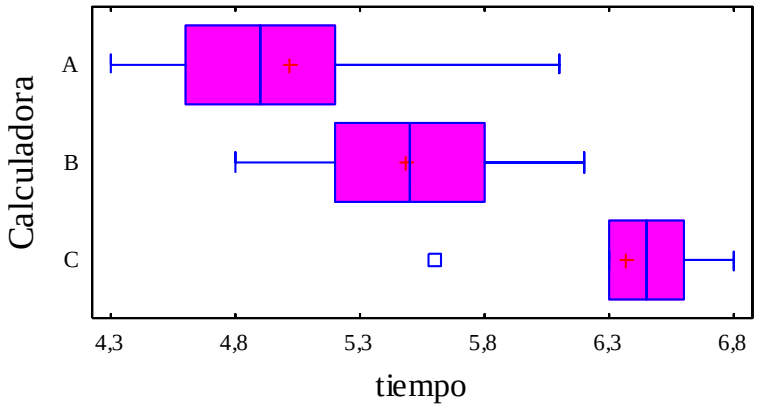
Modelo 1995
Media = 160.15
Mediana = 150
Mínimo = 20
Máximo = 400
Rango = 380
1º cuartil = 62.5
3º cuartil = 221.5
rango untercuartílico (*RIC*) = 159

Se observa que han disminuido las emisiones de hidrocarburos del modelo 1985 al 1995. Además los datos correspondientes al modelo 1985 tienen mayor variabilidad que los del modelo 1995. Para el modelo 1985 hay dos autos con emisiones muy altas.

Ejemplo:
Los siguientes datos representan los tiempos de operación, en horas, para tres tipos de calculadoras científicas de bolsillo, antes de que requieran recarga

Calculadora A:	4.9	6.1	4.3	4.6	5.2		
Calculadora B:	5.5	5.4	6.2	5.8	5.5	5.2	4.8
Calculadora C:	6.4	6.8	5.6	6.5	6.3	6.6	

La figura siguiente presenta los diagramas de caja comparativos generados por el paquete Statgraphics. Queda a cargo del lector comentar las diferencias entre las calculadoras A, B y C.



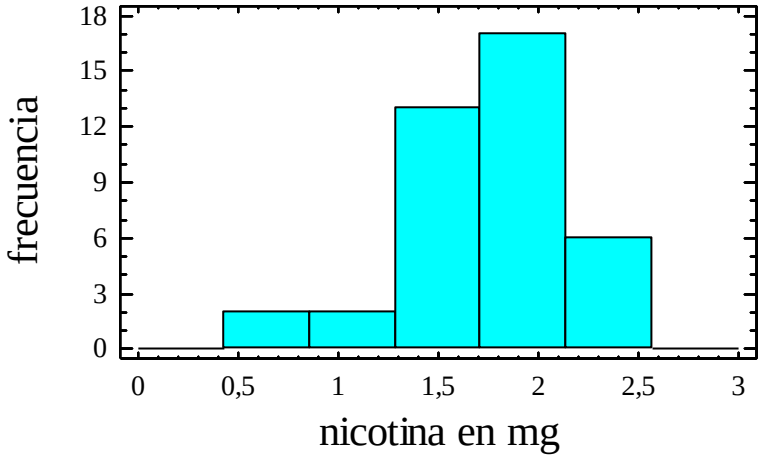
Calculadora	tamaño de muestra	Media	Mediana	Mínimo	Máximo	q1	q3	RIC
A	5	5,02	4,9	4,3	6,1	4,6	5,2	0,6
B	7	5,48571	5,5	4,8	6,2	5,2	5,8	0,6
C	6	6,36667	6,45	5,6	6,8	6,3	6,6	0,3

Ejemplo:
El contenido de nicotina, en miligramos, en 40 cigarrillos de cierta marca se registraron como sigue:

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97	0.85	1.24	1.58	2.03
1.70	2.17	2.55	2.11	1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93	1.40	1.64	2.09	1.75
1.63	2.37	1.75	1.69								

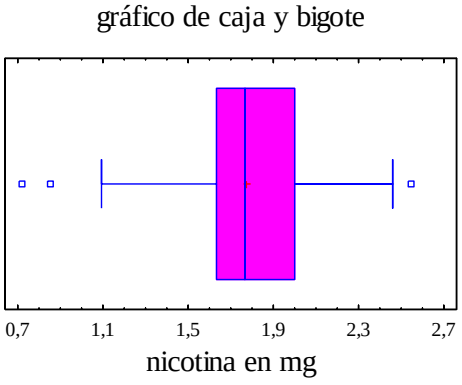
Se muestra la tabla de frecuencias y el histograma de los datos según Statgraphics.
Se tomaron 7 clases, donde el límite inferior de la primera clase es 0 y el límite superior de la última clase es 3.

Clase	limite inferior	limite superior	Marca de clase	frecuencia	frecuencia relativa	frec. acum..	frec.. ac. relativa
1	0,0	0,428571	0,214286	0	0,0000	0	0,0000
2	0,428571	0,857143	0,642857	2	0,0500	2	0,0500
3	0,857143	1,28571	1,07143	2	0,0500	4	0,1000
4	1,28571	1,71429	1,5	13	0,3250	17	0,4250
5	1,71429	2,14286	1,92857	17	0,4250	34	0,8500
6	2,14286	2,57143	2,35714	6	0,1500	40	1,0000
7	2,57143	3,0	2,78571	0	0,0000	40	1,0000



A continuación se dan las medidas descriptivas y el gráfico de caja y bigote:

Tamaño de muestra = 40
Media = 1,77425
Mediana = 1,77
Moda =
Varianza = 0,152456
Desviación estándar = 0,390456
Mínimo = 0,72
Máximo = 2,55
Rango = 1,83
1º cuartil = 1,635
3º cuartil = 2
Rango untercuartílico = 0,365
Coef. de variación = 22,0068%



Practica Nº 1

Estadística Descriptiva

1) Un artículo publicado en *Technometrics* (Vol. 19,1977, pag.425) presenta los datos siguientes sobre el octanaje de varias mezclas de gasolina:

88.5	87.7	83.4	86.7	87.5	91.5	88.6	100.3	96.5	93.3
94.7	91.1	91.0	94.2	87.8	89.9	88.3	87.6	84.3	86.7
84.3	86.7	88.2	90.8	88.3	98.8	94.2	92.7	93.2	91.0

   a) Construya una tabla de frecuencias (absoluta y relativa) y un histograma utilizando 3 clases.
   b) Construya una tabla de frecuencias (absoluta y relativa) y un histograma con 6 clases. Compare la forma del histograma con la que tiene el histograma de la parte a). ¿Los dos histogramas presentan información similar?.

2) Los datos siguientes representan el número de ciclos transcurridos hasta que se presenta una falla en una prueba de piezas de aluminio sujetas a un esfuerzo alternante repetido de 21000 psi, a 18 ciclos por segundo:

1115	1567	1223	1782	1055	798	1016	2100	910	1501
1310	1883	375	1522	1764	1020	1102	1594	1730	1238
1540	1203	2265	1792	1330	865	1605	2023	1102	990

   a) Construya una tabla de frecuencias (absoluta y relativa) y un histograma.
   b) Construya un polígono de frecuencias.
   c) ¿Existe evidencia de que una pieza "sobrevivirá" más allá de los 2000 ciclos?. Justifique su respuesta.

3) Las siguientes mediciones corresponden a las temperaturas de un horno registradas en lotes sucesivos de un proceso de fabricación de semiconductores (las unidades son ºF): 953, 950, 948, 955, 951, 949, 957, 954, 955.
   Calcule:
   a) La media muestral de estos datos.
   b) La mediana muestral de estos datos
   c) ¿En cuánto puede incrementarse la mayor medición de temperatura sin que cambie la mediana muestral?.

4) Se toman ocho mediciones del diámetro interno de los anillos para los pistones del motor de un automóvil. Los datos (en mm) son: 74.001, 74.003, 74.015, 74.000, 74.005, 74.002, 74.005, 74.004.
   a) Encuentre la media y la mediana de estos datos.
   b) Suponga que se elimina la observación más grande (74.015 mm). Calcule la media y la mediana muestrales para los datos restantes. Compare sus resultados con los obtenidos en la parte a).

5) Hallar la media y mediana para los datos de los ejercicios 1) y 2).

6) Los siguientes datos son las temperaturas de unión de los *O-rings* (en grados F), en cada prueba de lanzamiento o de un lanzamiento real, del motor del cohete del transbordador espacial (tomados de *Presidential Commission on the Space Shuttle Challenger Accident*, Vol.1, pags. 129-131): 84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31.
   a) Calcule la media y la mediana muestrales.
   b) Encuentre los cuartiles inferior y superior de la temperatura.
   c) Encuentre los percentiles quinto y noveno de la temperatura.
   d) Elimine la observación más pequeña (31ºF) y vuelva a calcular lo que se pide en los incisos a), b) y c). ¿Qué efecto tiene la eliminación de este punto?.

7) La contaminación de una pastilla de silicio puede afectar de manera importante la calidad de la producción de circuitos integrados. De una muestra de 10 pastillas se obtienen las siguientes concentraciones de oxígeno:
   3.15, 2.68, 4.31, 2.09, 3.82, 2.94, 3.47, 3.39, 2.81, 3.61. Calcule:
   a) La varianza muestral.
   b) La desviación estándar muestral.
   c) El rango de la muestra.

8) Considere los datos para anillos de pistón, del ejercicio 4). Calcule:
   a) La varianza muestral.
   b) La desviación estándar muestral.
   c)El rango de la muestra.
   d)Suponga que se elimina la observación más grande (74.015). Calcule la varianza muestral, la desviación estándar muestral, el rango de la muestra. Compare los resultados con los obtenidos en los incisos anteriores. Para esta medición en particular, ¿ cuán sensibles son la varianza muestral, la desviación muestral y el rango de la muestra?

9) Considere los datos del ejercicio 5).
   a) La varianza muestral.
   b) La desviación estándar muestral.
   c) El rango de la muestra y rango intercuartílico.
   d) Construya un diagrama de caja y discuta sobre la forma de la distribución y la posible presencia de valores atípicos.

10) Se tienen las millas por galón de autos, los que se clasifican según su origen. Se obtienen así las siguientes tres muestras:

Muestra 1, consiste en las millas por galón autos de origen americano:

36.1, 19.9, 19.4, 20.2, 19.2, 20.5, 20.2, 25.1, 20.5, 19.4, 20.6,
20.8, 18.6, 18.1, 19.2, 17.7, 18.1, 17.5, 30, 30.9, 23.2, 23.8, 21.5,

Muestra 2, consiste en las millas por galón de autos de origen europeo:

43.1, 20.3, 17, 21.6, 16.2, 31.5, 31.9, 25.4, 27.2, 37.3, 41.5, 34.3,
44.3, 43.4, 36.4, 30.4, 40.9, 29.8, 35, 33, 34.5, 28.1, 30.7, 36, 44.

Muestra 3, consiste en las millas por galón de autos de origen japonés:

32.8, 39.4, 36.1, 27.5, 27.2, 21.1, 23.9, 29.5, 34.1, 31.8, 38.1,
37.2, 29.8, 31.3, 37, 32.2, 46.6, 40.8, 44.6, 33.8, 32.7, 23.7, 32.4,

a) Hallar media, mediana y moda de cada muestra.
b) Hallar rango, rango intercuartílico, desviación estándar de cada muestra.
c) Hallar el coeficiente de variación para cada muestra.
d) Haga un gráfico de cajas simultáneas.
e) ¿Qué pude concluir sobre el consumo de los autos europeos y japoneses con respecto al de los autos americanos?.

11) Los precios de autos se clasifican según el origen de los mismos, obteniéndose las siguientes tres muestras:

Muestra 1, consiste en el precio de autos de origen americano:

1900, 3300, 3125, 2850, 2800, 3275, 2375, 2275, 2700, 2300,
3300, 2425, 2700, 2425, 3900, 4400, 2525, 3000, 2100, 2250,
3200, 2400, 3925, 3200, 2975, 3150, 3325, 4650, 4850, 5725,

Muestra 2, consiste en el precio de autos de origen europeo:

4475, 5875, 4200, 5450, 3675, 3100, 4675, 2275, 7000, 5900,
3900, 3825, 7975, 14275, 2575, 7000, 5000, 5650, 4600, 8500,
8225, 8550, 5200, 5075, 2400, 15475.

Muestra 3, consiste en el precio de autos de origen japonés:

2200, 2725, 2250, 2975, 2775, 3700, 2975, 3425, 2750, 2750,
3850, 3525, 5500, 4675, 4050, 3975, 3350, 3300, 3925, 3625,
8150, 7250, 4700, 4400, 4000, 3950, 3775, 4475, 3975, 5550,
4650, 5825, 5650, 9475, 8375, 4900, 5100, 6350, 6500, 5425,
4625, 4875, 5075, 7700.

a)Hallar media, mediana y moda de cada muestra.
b)Hallar rango, rango intercuartílico, desviación estándar de cada muestra.
c)Hallar el coeficiente de variación para cada muestra.
d)Haga un gráfico de cajas simultáneas.
e)¿Hay diferencias de precio entre los diferentes orígenes?.

12) Se clasifica el precio de autos según el año del modelo, obteniéndose las siguientes cinco muestras:

Muestra 1, precios de 36 ***autos modelo año 78:***
2400, 1900, 2200, 2725, 2250, 3300, 3125, 2850, 2800, 3275,
2375, 2275, 2700, 2300, 3300, 2425, 2700, 2425, 3900, 4400,

Muestra 2, precios de autos modelo año 79:
3925, 3200, 2975, 3150, 3325, 4650, 4850, 5725, 4025, 5225,
4825, 4100, 4725, 3100, 2750, 2700, 2725, 15475, 9900, 4675,

Muestra 3, precios de autos modelo año 80:
7000, 3850, 2900, 3525, 3625, 3525, 3625, 3700, 5900, 5500,
4675, 4050, 3975, 3350, 3200, 3300, 3900, 3825, 7975, 14275,

Muestra 4, precios de autos modelo año 81:
5100, 5175, 4950, 4550, 4900, 4400, 3600, 4000, 3950, 3775,
4475, 3975, 3450, 3850, 4100, 5650, 4600, 5550, 4650, 5825,

Muestra 5, precios de autos modelo año 82:
5275, 5500, 5175, 5650, 6250, 5650, 5225, 5150, 5200, 4900,
5100, 4350, 4550, 6350, 6500, 5425, 4625, 4875, 5075, 7300,

a)Hallar media, mediana y moda de cada muestra.
b)Hallar rango, rango intercuartílico, desviación estándar de cada muestra.
c)Hallar el coeficiente de variación para cada muestra.
d)Haga un gráfico de cajas simultáneas.
e)¿Ha variado el precio de los autos durante los años considerados?.