Percentiles, Deciles y Cuartiles

Además de las medidas de tendencia central, dispersión y forma, también existen algunas medidas interesantes de posición que se utilizan al resumir y analizar las características o propiedades de grandes colecciones de datos.

1. Percentiles

Los *percentiles* son aquellos valores que dividen a los datos *ordenados* de forma creciente, en cien partes iguales. Existen noventa y nueve percentiles que se denotan por P_1, P_2, ... , P_{99}. Entre dos percentiles consecutivos se encuentra el 1% de los datos. Así, por ejemplo, entre los percentiles P_{10} y P_{20} se encuentran 10% de los datos.

Para denotar un percentil cualquiera usamos P_h, donde $h = 1, 2, 3, \ldots, 99$.

Definición de Percentil

El percentil P_h de una colección de datos que previamente han sido ordenados (de forma creciente), es un valor tal que como máximo el h% de los datos son menores que él, y también como máximo un (100-h)% de los datos son mayores que él.

Como en el caso de la mediana, si dos valores consecutivos del conjunto de datos cumplen con la definición anterior, se conviene en tomar como percentil al promedio de ellos dos.

Ejemplo:

Suponga que los pesos de ocho personas (en Kg) son: 52, 97, 108, 63, 90, 74, 86, 73. Hallar lo percentiles: P_{20}, P_{50} y P_{80}.

En primer lugar se deben ordenar de forma creciente los datos:

52 63 73 74 86 90 97 108

El P_{20} es el valor tal que el 20% de los datos, es decir el 20% de 8 = 1,6 datos, como máximo son menores que él, y también como máximo el 80% de 8 = 6,4 datos son mayores que él.

Observe que el valor 63 cumple con estas condiciones.

Por tanto, P_{20} = 63 Kg.

Ahora, en el cálculo de P_{50} se observa que existen dos valores 74 y 86, que cumplen con la definición.

De esta manera, P_{50} = (74 + 86) / 2 = 80 Kg.

Para estos datos, P_{80} tiene como máximo 6,4 datos por debajo de él y a lo sumo 1,6 datos por encima.

El valor 97 satisface esto, así P_{80} = 97 Kg.

Nótese que ni el valor 90 ni 108 cumple con las condiciones.

Por ejemplo, el valor 90 tiene cinco datos por debajo que cumple con lo que se exige pero por encima tiene a dos datos (el 25% de los datos), lo que no satisface los requerimientos para ser percentil 80.

2. Deciles

Los *Deciles* son los valores que dividen a los datos ordenados (de forma creciente) en diez partes iguales.

Existen nueve deciles que se denotarán por D_1, D_2, ... , D_9.

Entre dos deciles consecutivos se encuentra un 10% de los datos.

3. Cuartiles

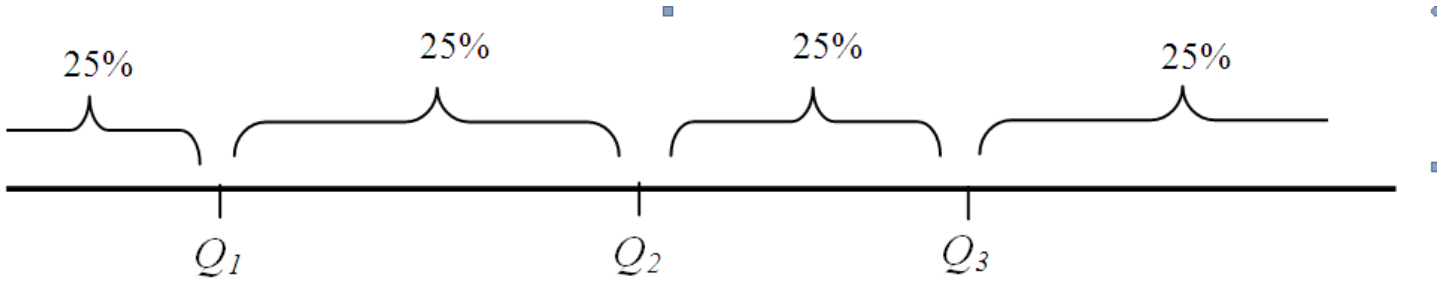
Los *cuartiles* son los valores que dividen a una colección de datos que previamente han sido ordenados en forma creciente, en cuatro partes iguales.

De esta manera, existen tres cuartiles que se denotan Q_1, Q_2 y Q_3.

Entre dos cuartiles consecutivos se encuentra un 25% de los datos.

Por debajo de Q_1, se encuentra un 25% de los datos y por encima un 75%;

Por debajo del cuartil tres, se encuentra un 75% de los datos y por encima de él existe un 25% de los datos.



Relaciones entre los cuartiles deciles y percentiles:

$$Q_1 = P_{25}$$
$$Q_2 = D_5 = P_{50} = Md$$
$$Q_3 = P_{75}$$
$$D_1 = P_{10}$$
$$D_2 = P_{20}$$
$$\vdots$$
$$D_9 = P_{90}$$

Nota:

A los cuartiles, deciles y percentiles en general se les denominan *cuantiles*

Ejercicio:

Para los datos no agrupados de estatura, calcular e interpretar: los cuartiles, el decil tres y el percentil diez.

Uso de percentiles como medidas de dispersión

Los percentiles también son utilizados como indicadores de la dispersión de los datos.

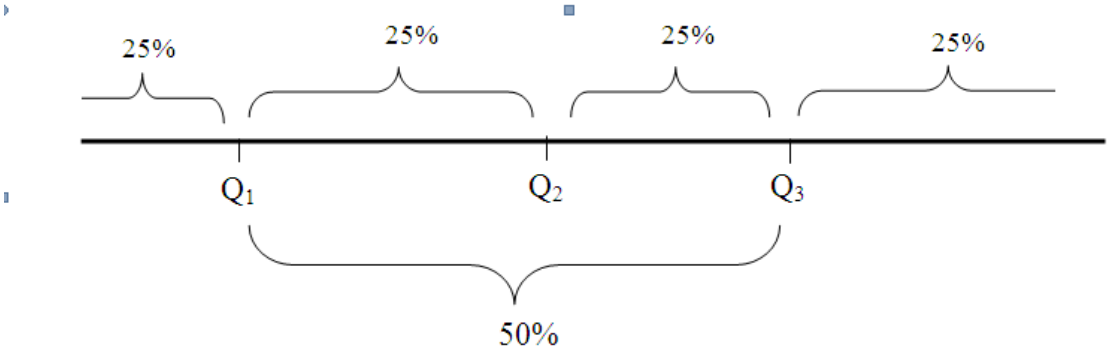
Con ellos se construyen algunas medidas de dispersión:

Recorrido Intercuartil

El *recorrido intercuartil*, viene dado por:

$$RQ = Q_3 - Q_1$$

Esta medida refleja la dispersión de la parte central de la distribución ya que toma en cuenta al 50% de los datos del centro de la distribución:



Desviación Cuartil ó Recorrido Semi-Intercuartil

La *desviación cuartil* se obtiene mediante la siguiente expresión:

$$Q = \frac{Q_3 - Q_1}{2}$$

Si se calcula $Md \pm Q$

se obtiene un intervalo que contiene aproximadamente el 50% de los datos.

Nota

Fácilmente puede notarse que las dos medidas anteriores no toman en cuenta a todos los datos

Esto puede representar una seria desventaja ya que es posible que por debajo de Q_1 o por encima de Q_3, los datos se encuentren muy concentrados o muy dispersos y el efecto sobre RQ y Q será el mismo.

Aunque por otro lado, y por la misma razón, el recorrido intercuartil y la desviación cuartil no son afectados por valores atípicos.

Recorrido Percentil

Es una medida basada en la misma idea que el *RQ*, la cual viene dada por:

$$RP = P_{90} - P_{10}$$

Este indicador refleja el 80% de los datos ubicados en la parte central de la distribución

Ejercicio:

Para los datos correspondientes a las variables peso e ingreso hallar:

a. *RQ*
b. *RP*
c. El intervalo que contiene aproximadamente el 50% de los datos de la parte central de la distribución.

Medidas de Forma

Existen indicadores que cuantifican la asimetría y el apuntamiento de una distribución.

Estos son de utilidad cuando no se dispone del gráfico o para confirmar las conclusiones obtenidas gráficamente.

Tanto las medidas de asimetría como las de apuntamiento son indicadores relativos ya que no vienen expresados en alguna unidad de medida.

Medidas de Asimetría

Los resultados que se discutirán se refieren a distribuciones unimodales:

a. ***Coeficiente de Asimetría de Pearson***

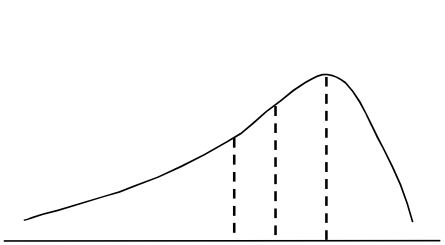
Este indicador se basa en la relación existente entre la media y la mediana:

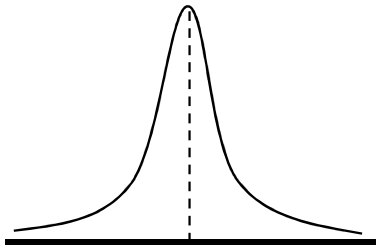
$$ASP = \frac{3(\overline{x} - Md)}{S_*}$$

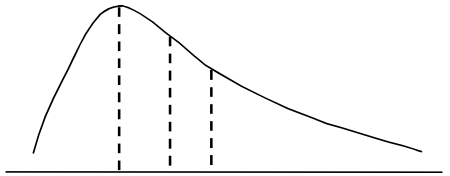
Obsérvese que si la distribución es:

- Simétrica => ASP = 0, ya que en este caso $\overline{x} = Md$
- Asimétrica por la derecha => ASP > 0, debido a que $\overline{x} > Md$
- Asimétrica por la izquierda => ASP < 0, porque $\overline{x} < Md$

El coeficiente de asimetría de Pearson toma valores en el intervalo (-3, 3)







Coeficiente de Asimetría de Fisher

Se denota por γ_1 y viene dado por:

$$\gamma_1 = \frac{\left(\frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{n} \right)}{S_*^3}$$

El coeficiente γ_1 está basado en la media aritmética e indica de que lado las diferencias respecto de éstas son mayores.

Su interpretación es similar a la del coeficiente de Asimetría de Pearson

Medidas de Apuntamiento o Curtosis

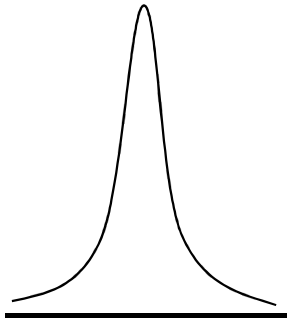
Estas medidas indican el grado de apuntamiento o achatamiento del gráfico.

La medición del apuntamiento de un gráfico se hace tomando como referencia la curva normal (curva de campana o curva de Gauss).

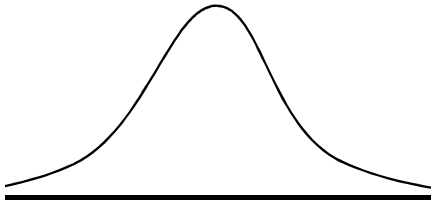
Por tanto, el gráfico al que se le desee medir el apuntamiento, debe ser al menos aproximadamente simétrico.

A la curva normal se le llama mesocúrtica, si es más puntiaguda se le llama leptocúrtica y si es más achatada platicúrtica.

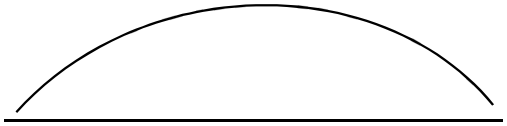
Los indicadores de curtosis, miden el nivel de concentración de datos en la región central.



Leptocúrtica



Mesocúrtica



Platicúrtica

Coeficiente de Pearson

El coeficiente β_2 de Pearson es el más utilizado de las medidas de apuntamiento y viene dado por:

$$\beta_2 = \frac{\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n} \right)}{S_*^4}$$

- Si la curva es normal (mesocúrtica) , β_2 = 3
- Si la curva es leptocúrtica , β_2 > 3
- Si la curva es platicúrtica , β_2 < 3

Medidas Descriptivas Numéricas

Ejercicio:

Calcule e interprete los coeficientes de asimetría ASP y γ_1 para los datos correspondientes a las variables peso, estatura, ingreso y número de hermanos.

Ejercicio:

Calcule e interprete el coeficiente β_2 de Pearson para para los datos correspondientes a las variables peso, estatura y número de hermanos.

Diagrama de caja

El diagrama de tallo y hoja y el histograma proporcionan una impresión visual general del conjunto de datos, mientras que las cantidades numéricas tales como $\overline{x}$ o S brindan información sobre una sola característica de los datos.

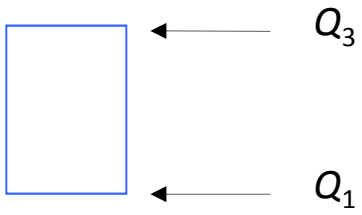
El ***diagrama de caja*** es una presentación visual que describe al mismo tiempo varias características importantes de un conjunto de datos, tales como el centro, la dispersión, la simetría o asimetría y la identificación de observaciones atípicas.

El diagrama de caja representa los tres cuartiles, y los valores mínimo y máximo de los datos sobre un rectángulo (caja), alineado horizontal o verticalmente.

Diagrama de caja

Construcción:

1. El rectángulo delimita el rango intercuartílico con la arista izquierda (o inferior) ubicada en el primer cuartil Q_1, y la arista derecha (o superior) en el tercer cuartil Q_3.



2. Se dibuja una línea a través del rectángulo en la posición que corresponde al segundo cuartil (que es igual al percentil 50 o a la mediana), Q_2 = Md.

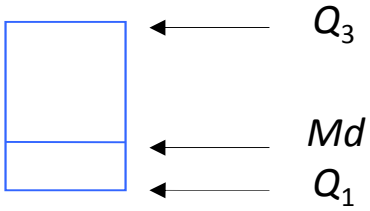
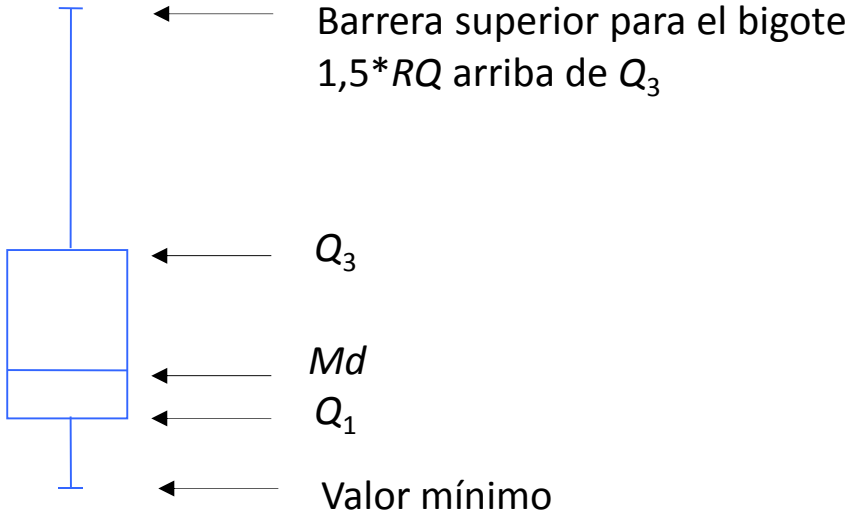


Diagrama de caja

3. De cualquiera de las aristas del rectángulo se extiende una línea, o *bigote*, que va hacia los valores extremos (valor mínimo y valor máximo). Estas son observaciones que se encuentran entre cero y 1.5 veces el rango intercuartílico a partir de las aristas del rectángulo.



A veces, los diagramas de caja reciben el nombre de *diagramas de caja y bigotes.*

Nótese que el rectángulo o caja representa el 50% de los datos que particularmente están ubicados en la zona central de la distribución.

La caja representa el cuerpo de la distribución y los bigotes sus colas.

Diagrama de caja

4. Las observaciones que están entre 1.5 y 3 veces el rango intercuartílico a partir de las aristas del rectángulo reciben el nombre de *valores atípicos*.

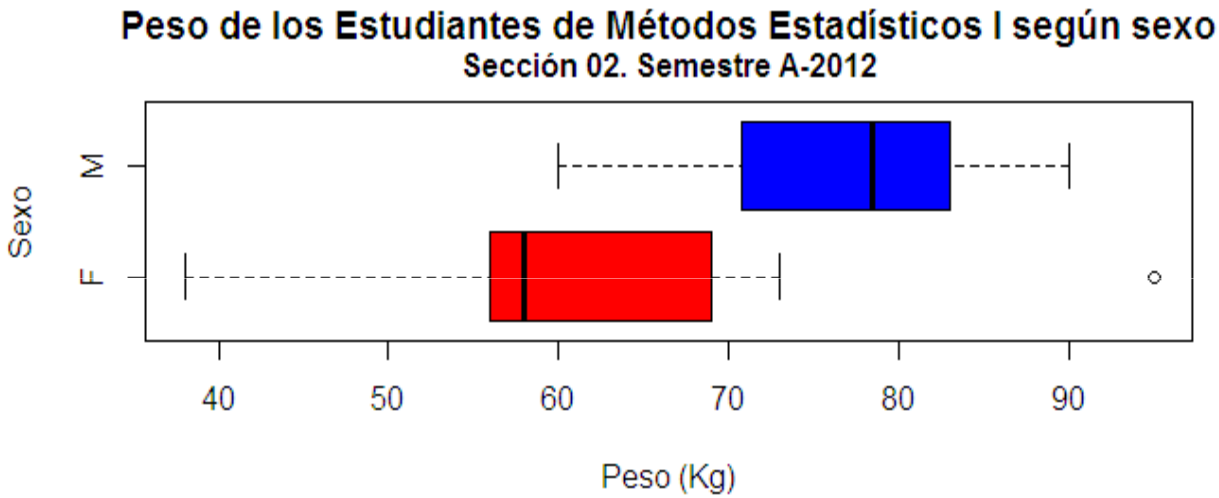
Las observaciones que están más allá de tres veces el rango intercuartílico a partir de las aristas del rectángulo se conocen como *valores atípicos extremos*.

En ocasiones se emplean diferentes símbolos (como círculos vacíos o llenos), para identificar los dos tipos de valores atípicos.

Diagrama de caja

Diagrama de caja

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



Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas-FACES-ULA. Abril 2012

Esta figura presenta los diagramas de caja comparativos para la variable *peso* de los estudiantes de Métodos Estadísticos I clasificados por *sexo*.

Diagrama de caja

El análisis de este diagrama revela que el peso de los varones es, en general, mayor que el de las hembras.

También se observa que la variabilidad de los pesos de las hembras es mayor a la de los varones.

Sin embargo, la variabilidad en la parte central de la distribución de los pesos tanto de las féminas como de los masculinos es muy similar.

Se observa la existencia de un valor atípico en la distribución de las mujeres, que es un peso muy alto (el valor máximo de todos los pesos) en comparación a los pesos del resto de las estudiantes.

La distribución del peso de los varones es asimétrico por la izquierda mientras que las chicas presentan una distribución asimétrica por la derecha influenciada por el valor atípico.

Diagrama de caja

Comparación entre diagramas de caja e histogramas del mismo conjunto de datos.

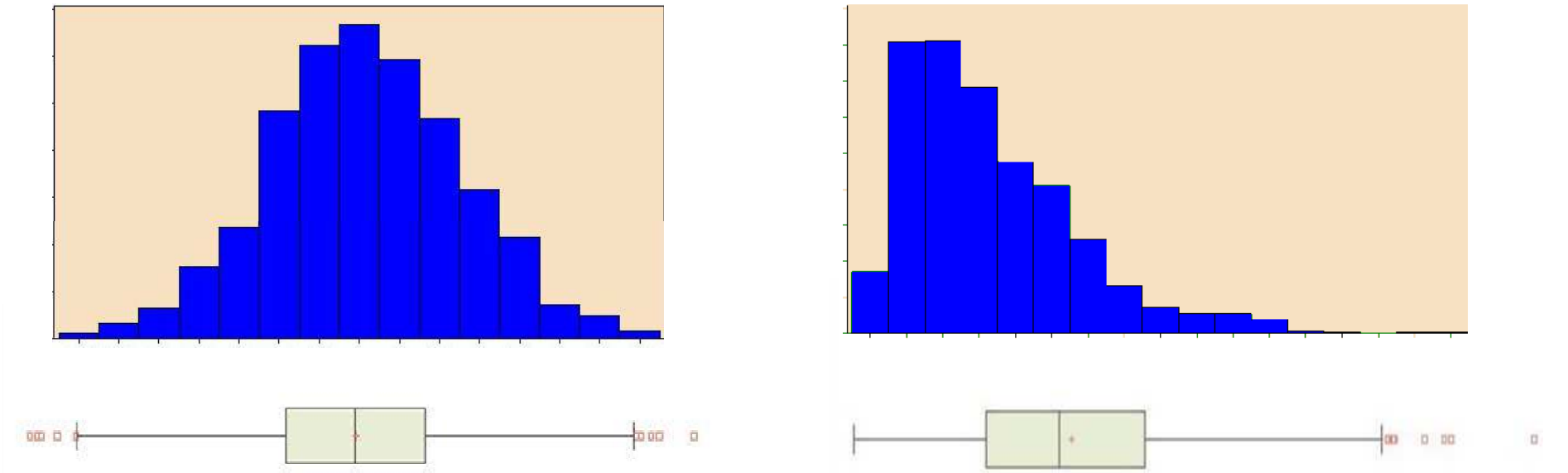
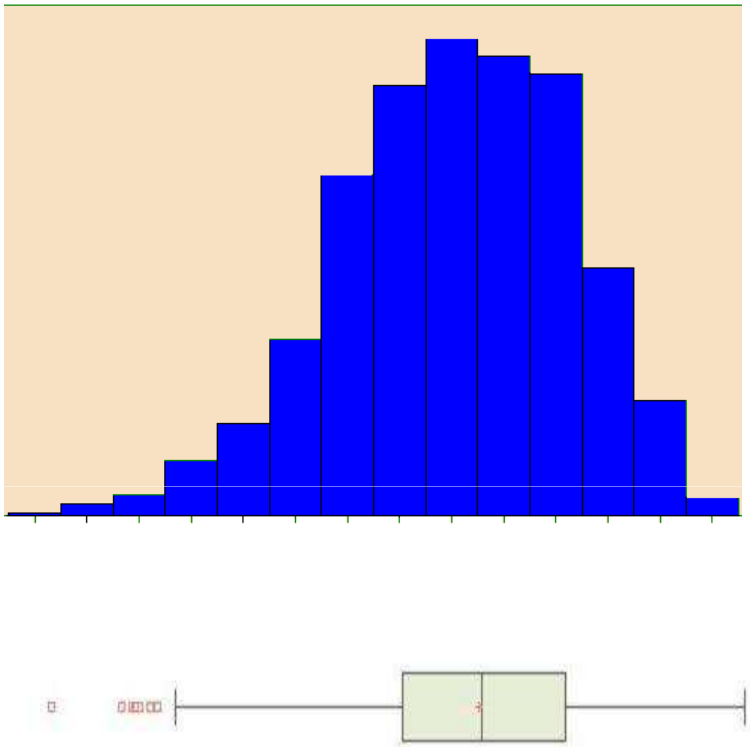


Diagrama de caja



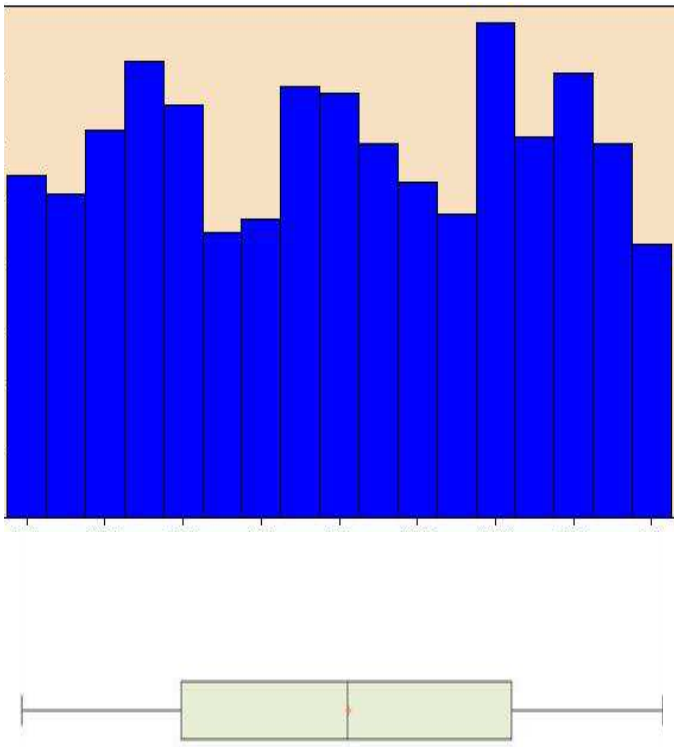


Diagrama de caja

Ejercicio

1. Construya y analice el diagrama de caja para la variable peso de los datos usados en clase.
2. Construya un diagrama de caja de la variable peso para el sexo femenino y otro para el sexo masculino. Compare estos dos diagramas.