Medidas de Tendencia Central y de Variabilidad

Contenidos

- Medidas descriptivas de forma: curtosis y asimetría
- Medidas de tendencia central: media, mediana y moda
- Medidas de dispersión: rango, varianza y desviación estándar. Coeficiente de variación
- Percentiles
- Diagrama de caja

MEDIDAS DE TENDENCIA CENTRAL

Al trabajar con histogramas y polígonos de frecuencias, vimos que las distribución de los datos pueden adoptar varias formas. En algunas distribuciones los datos tienden a agruparse más en una parte de la distribución que en otra. Comenzaremos a analizar las distribuciones con el objeto de obtener medidas descriptivas numéricas llamadas ***estadísticas,*** que nos ayuden en el análisis de las características de los datos. Dos de estas características son de particular importancia para los responsables de tomar decisiones: *la tendencia central y la dispersión*

MEDIDAS DE TENDENCIA CENTRAL: Moda, mediana y media

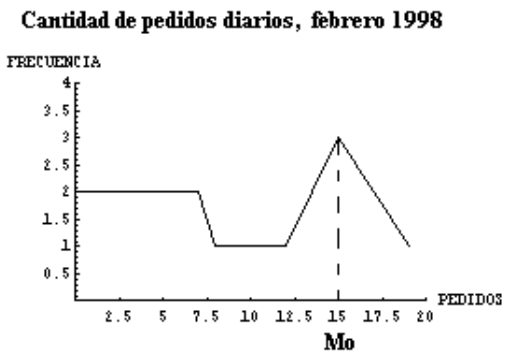
Tendencia central : La tendencia central se refiere al punto medio de una distribución. Las medidas de tendencia central se denominan medidas de posición.

> ***Moda:***
> *es el valor que más se repite en un conjunto de datos.*

Ejemplo 1: Los siguientes datos representan la cantidad de pedidos diarios recibidos en un período de 20 días, ordenados en orden ascendente

0	0	1	1	2	2	4	4	5	5
6	6	7	7	8	12	**15**	**15**	**15**	19

Mo = 15 **La cantidad de pedidos diarios que más se repite es 15**



Fte: Empresa NN. 2009

Autores: Liliana Marconi / Adriana D´Amelio

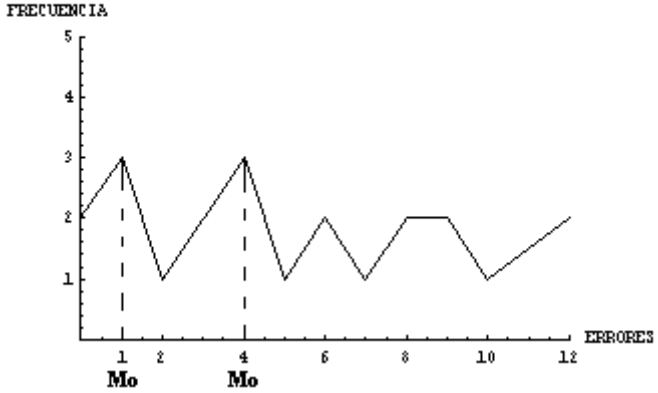
Ejemplo 2: La cantidad de errores de facturación por día en un período de 20 días, ordenados en orden ascendente es

0	0	1	1	1	2	4	4	4	5
6	6	7	8	8	9	9	10	12	12

Esta distribución tiene 2 modas. Se la llama *distribución bimodal* .

Mo = 1 y Mo = 4

Errores de facturación por día, febrero 1998



Fte: Empresa NN. 2009

Cálculo de la moda para datos agrupados

Si los datos están agrupados en una distribución de frecuencias, se selecciona el intervalo de clase que tiene mayor frecuencia llamado ***clase modal***.

Para determinar un solo valor de este intervalo para la moda utilizamos la siguiente ecuación:

$$Mo = L_{Mo} + \left(\frac{d_1}{d_1 + d_2} \right).h$$

M_o Moda

L_{Mo} Límite inferior de la clase modal

d_1 frecuencia de la clase modal menos la frecuencia de la clase anterior a ella ($d_1 = f_i - f_{i-1}$)

d_2 frecuencia de la clase modal menos la frecuencia de la clase posterior a ella ($d_2 = f_i - f_{i+1}$)

h amplitud del intervalo de clase

Ejemplo 3: La edad de los jubilados encuestados en Mendoza en noviembre del 2008

EDAD	m_i	f_i	f_{ri}	$f_{ri\%}$	F_i	F_{ri}	$F_{ri\%}$
[50,60)	55	10	0,20	20	10	0,20	20
[60, 70)	65	18	0,36	36	28	0,56	56
[70, 80)	75	14	0,28	28	42	0,84	84
[80, 90)	85	6	0,12	12	48	0,96	96
[90,100)	95	2	0,04	4	50	1	100

La clase modal es [60, 70) , ya que es la que presenta la mayor frecuencia

$L_{Mo} = 60 \qquad f_i = 18 \qquad f_{i-1} = 10 \qquad f_{i+1} = 14 \qquad h = 10$

$d_1 = f_i - f_{i-1}$ **=18-10 =8** $\qquad d_2 = f_i - f_{i+1}$ **= 18-14=4**

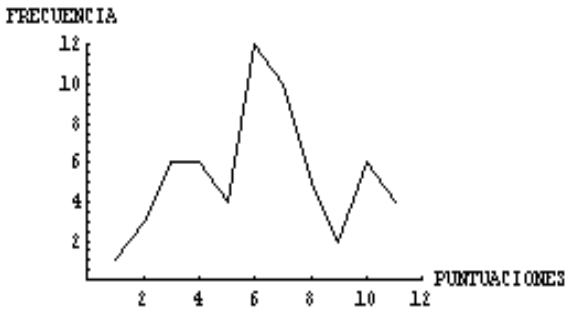
$$Mo = 60 + \left(\frac{8}{8+4}\right).10 = 66{,}66$$

- **La edad que más se repite es 66,66 años**

<u>VENTAJAS Y DESVENTAJAS DE LA MODA</u>

- **Se puede utilizar para datos cualitativos nominales u ordinales y para datos cuantitativos**
- **No se ve afectada por los valores extremos**
- **Se puede utilizar cuando la distribución de frecuencias tenga clases abiertas**
- **Cuando todas las puntuaciones de un grupo tienen la misma frecuencia, se dice que <u>no tiene moda</u>**
- **Si un conjunto de datos contiene 2 puntuaciones adyacentes con la misma frecuencia común (mayor que cualquier otra), la moda es el promedio de las 2 puntuaciones adyacentes Ej. (0,1,1,2,2,2,3,3,3,4,5) tiene Mo=2,5**
- **Si en un conjunto de datos hay dos que no son adyacentes con la misma frecuencia mayor que las demás, es una distribución bimodal. Conjuntos muy numerosos se denominan bimodales cuando presentan un polígono de frecuencias con 2 lomos, aún cuando las frecuencias en los 2 picos no sean exactamente iguales. Estas ligeras distorsiones de la definición están permitidas porque el término bimodal es muy conveniente y en último término es descriptivo. Una distinción conveniente puede hacerse entre la *moda mayor* y la *moda menor*. Por ejemplo en el gráfico siguiente, la moda mayor es 6 y las menores son 3,5 y 10**

Puntuaciones obtenidas en un examen de aptitudes



Fte: Elaboración propia. 2009

Mediana:
es el valor que divide al conjunto ordenado de datos, en dos subconjuntos con la misma cantidad de elementos. La mitad de los datos son menores que la mediana y la otra mitad son mayores

En general, vamos a representar un conjunto de n datos como $x_1,\ x_2,\ x_3, \ldots, x_n$

Si los datos están ordenados, los indicaremos $x_{(1)},\ x_{(2)},\ x_{(3)}, \ldots, x_{(n)}$

donde el subíndice encerrado entre paréntesis indica el orden o ubicación en el conjunto ordenado

Se presentan dos situaciones:

- Número impar de datos: La mediana es el dato que está en la posición $\frac{n+1}{2}$

$$Me = \tilde{m} = \tilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

Sea el conjunto ordenado de datos:

2	3	5	6	8
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$

$$Me = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{5+1}{2}\right)} = x_{(3)} = 5$$

- **La mitad de las observaciones son menores o iguales que 5 y la otra mitad son mayores o iguales que 5.**

- Número par de datos: Es el promedio entre los dos datos centrales.

$$Me = \tilde{m} = \tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

2	3	5	6	8	9
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$

$$Me = \frac{x_{\left(\frac{6}{2}\right)} + x_{\left(\frac{6}{2}+1\right)}}{2} = \frac{x_{(3)} + x_{(3+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{5+6}{2} = 5{,}5$$

- **La mitad de las observaciones son menores o iguales que 5,5 y la otra mitad son mayores o iguales que 5,5.**

Autores: Liliana Marconi / Adriana D´Amelio

Cálculo de la mediana para datos agrupados

Si los datos están agrupados en una distribución de frecuencias, se selecciona el intervalo de clase que contiene a la mediana llamado ***clase mediana***. Para ello, debemos determinar la frecuencia acumulada absoluta que contenga al elemento número $\frac{n+1}{2}$. El valor de este intervalo para la mediana se calcula utilizando la siguiente ecuación:

$$Me = \tilde{m} = \tilde{x} = L_m + \left(\frac{\frac{n+1}{2} - F_{i-1}}{f_i} \right).h$$

M_e **Mediana**
L_m **Límite inferior de la clase mediana**
n **cantidad de datos**
F_{i-1} **frecuencia acumulada absoluta de la clase anterior al intervalo mediana**
f_i **frecuencia absoluta de la clase mediana**
h **amplitud del intervalo de clase**

Ejemplo (Continuación): La edad de los residentes en un complejo de viviendas tiene la siguiente distribución:

EDAD	m_i	f_i	f_{ri}	$f_{ri\%}$	F_i	F_{ri}	$F_{ri\%}$
[50,60)	55	10	0,20	20	10	0,20	20
[60, 70)	65	18	0,36	36	28	0,56	56
[70, 80)	75	14	0,28	28	42	0,84	84
[80, 90)	85	6	0,12	12	48	0,96	96
[90,100)	95	2	0,04	4	50	1	100

La clase mediana es la que contenga el elemento en la posición $\frac{50+1}{2}$, es decir en la posición 25,5 . Buscamos en la frecuencia acumulada F_i y vemos que se halla en el intervalo [60, 70)

$L_{Me} = 60 \qquad F_{i-1} = 10 \qquad f_i = 18 \qquad h = 5$

$$Me = 60 + \left(\frac{25{,}5 - 10}{18} \right).10 = 68{,}61$$

INTERPRETE: .....

VENTAJAS Y DESVENTAJAS DE LA MEDIANA

- **Se puede utilizar para datos cualitativos ordinales y para datos cuantitativos**

- **No se ve afectada por los valores extremos. Esta es la propiedad más importante que tiene.**

- **Se puede utilizar cuando la distribución de frecuencias tiene clases abiertas, a menos que la mediana caiga en una de las clases abiertas**

- **Si hay un gran número de datos, el tener que ordenarlos para hallar la mediana insume esfuerzo y tiempo.**

Media o media aritmética:
Es el promedio de los datos

- **Una** muestra **con n (minúscula) observaciones, tiene una media $\bar{x}$ (que se denomina estadística)**
- **Una** población **con N (mayúscula) elementos tiene una media μ (que se denomina parámetro)**

Cálculo de la media para datos no agrupados

$$\mu = \frac{\sum x}{N} \qquad\qquad \bar{x} = \frac{\sum x}{n}$$

Vemos que es la suma de las observaciones divididas el total de datos. Cuando calculamos la media de la población, dividimos por la cantidad de datos de la población N y cuando se calcula la media muestral por n

Ejemplo: El Departamento de Acción Social ofrece un estímulo especial a aquellas agrupaciones en las que la edad promedio de los niños que asisten está por debajo de 9 años. Si los siguientes datos corresponden a las edades de los niños que acuden de manera regular al Centro ¿calificará éste para el estímulo?

8 5 9 10 9 12 7 12 13 7 8

$$\bar{x} = \frac{\sum x}{n} = \frac{8+5+9+10+9+12+7+12+13+7+8}{11} = 9{,}09$$

Interpretación: .....
.....

Cálculo de la media para datos agrupados

Para calcular la media para datos agrupados, primero calculamos el punto medio de cada clase (marca de clase m_i). Después multiplicamos cada punto medio por la frecuencia absoluta de cada intervalo

$$\bar{x} = \frac{\sum m_i . f_i}{n}$$

Una manera de hacer los cálculos es utilizando la siguiente tabla:

EDAD	m_i	f_i	$\mathbf{m_i . f_i}$
[50,60)	55	10	**550**
[60, 70)	65	18	**1170**
[70, 80)	75	14	**1050**
[80, 90)	85	6	**510**
[90,100)	95	2	**190**
Total		**50**	**3470**

$$\bar{x} = \frac{3470}{50} = 69{,}4$$

La edad promedio es de 69,4 años

VENTAJAS Y DESVENTAJAS DE LA MEDIA

- **Se trata de un concepto familiar e intuitivamente claro**
- **Cada conjunto de datos tiene una media y es única**
- **Es útil para llevar a cabo procedimientos estadísticos como la comparación de medias de varios conjuntos de datos. En estadística inferencial es la medida de tendencia central que tiene mejores propiedades**
- **Aunque la media es confiable en el sentido de que toma en cuenta todos los valores del conjunto de datos, puede verse afectada por valores extremos que no son representativos del resto de los datos. La media puede malinterpretarse si los datos no forman un conjunto homogéneo.**
- **No se puede calcular la media si la distribución de frecuencias tiene clases abiertas**

COMPARACIÓN ENTRE LA MEDIA, LA MEDIANA Y LA MODA

- **Las distribuciones simétricas tienen el mismo valor para la media, la mediana y la moda.**
- **En una distribución con sesgo positivo, la moda se halla en el punto más alto de la distribución, la mediana está hacia la derecha de la moda y la media más a la derecha. Es decir Mo < Me < $\bar{x}$**
- **En una distribución con sesgo negativo, la moda es el punto más alto, la mediana está a la izquierda de la moda y la media está a la izquierda de la mediana. Es decir, $\bar{x}$ < Me < Mo**
- **Cuando la población tiene una distribución sesgada, con frecuencia la mediana resulta ser la mejor medida de posición, debido a que está siempre entre la media y la moda. La mediana no se ve altamente influida por la frecuencia de aparición de un solo valor como es el caso de la moda, ni se distorsiona con la presencia de valores extremos como la media.**
- **La selección de la media, la mediana o la moda, depende de la aplicación. Por ejemplo, se habla del salario promedio (media); el precio mediano de una casa nueva**

puede ser una estadística más útil para personas que se mudan a un nuevo vecindario (si hay una o dos crestas que distorsionan la media). Y mientras que la familia promedio conste de 1,7 niños, tiene más sentido para los diseñadores de automóviles pensar en la familia modal, con dos niños.

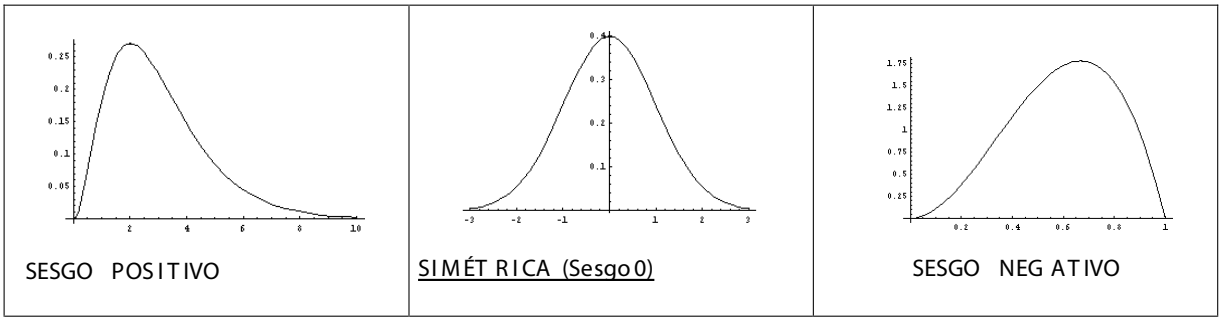
MEDIDAS DE VARIABILIDAD

Dispersión: La dispersión se refiere a la extensión de los datos, es decir al grado en que las observaciones se distribuyen (o se separan).

Existen otras dos características de los conjuntos de datos que proporcionan información útil: el sesgo y la curtosis.

Sesgo (skewness): Las curvas que representan un conjunto de datos pueden ser simétricas o sesgadas. Las curvas simétricas tienen una forma tal que una línea vertical que pase por el punto más alto de la curva, divide al área de ésta en dos partes iguales. Si los valores se concentran en un extremo se dice sesgada. Una curva tiene sesgo positivo cuando los valores van disminuyendo lentamente hacia el extremo derecho de la escala y sesgo negativo en caso contrario.

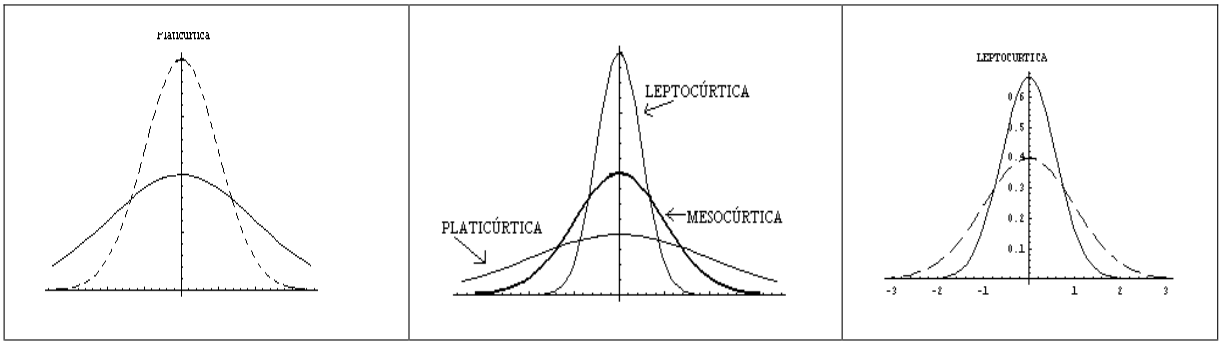
? El sesgo es una medida de la asimetría de la curva. En general es un valor que va de -3 a 3. Una curva simétrica toma el valor 0.



SESGO POSITIVO | SIMÉTRICA (Sesgo 0) | SESGO NEGATIVO

Curtosis (Kurtosis): Nos da una idea de la agudeza (o lo plano) de la distribución de frecuencias. Una curva normal (es el patrón con el que se compara la curtosis de otras curvas) tiene curtosis 0. Esta curva se llama mesocúrtica. Si la curtosis es mayor que 0, la curva es más empinada que la anterior y se denomina leptocúrtica (Lepto, del griego, "empinado" o "estrecho"). Si la curtosis es menor que 0, es relativamente plana y se denomina platicúrtica ("plano", "ancho")

(En el gráfico la curva punteada es la curva normal (mesocúrtica))



Autores: Liliana Marconi / Adriana D´Amelio

MEDIDAS DE DISPERSIÓN

Las medidas de dispersión son útiles porque:

Nos proporcionan información adicional que nos permite juzgar la confiabilidad de nuestra medida de tendencia central. Si los datos están muy dispersos la posición central es menos representativa de los datos, como un todo, que cuando estos se agrupan más estrechamente alrededor de la media.

Ya que existen problemas característicos de distribuciones muy dispersas, debemos ser capaces de distinguir que presentan esa dispersión antes de abordar los problemas

Nos permiten comparar varias muestras con promedios parecidos

Los analistas financieros están preocupados por la dispersión de las ganancias de una empresa que van desde valores muy grandes a valores negativos. Esto indica un riesgo mayor para los accionistas y para los acreedores. De manera similar los expertos en control de calidad, analizan los niveles de calidad de un producto

RANGO:

Es la diferencia entre el mayor y el menor de los valores Observados

$$R = x_{(n)} - x_{(1)}$$

Siendo $x_{(n)}$ la observación mayor y $x_{(1)}$ la observación Menor

- **El rango es fácil de entender y de encontrar, pero su utilidad como medida de dispersión es limitada. Como sólo toma en cuenta el valor más alto y el valor más bajo ignora la naturaleza de la variación entre todas las demás observaciones, y se ve muy influido por los valores extremos.**
- **Debido a que considera sólo dos valores tiene muchas posibilidades de cambiar drásticamente de una muestra a otra en una población dada.**
- **Las distribuciones de extremo abierto no tienen rango.**

VARIANZA Y DESVIACIÓN ESTÁNDAR

Las descripciones más comprensibles de la dispersión son aquellas que tratan con la desviación promedio con respecto a alguna medida de tendencia central. Veremos dos medidas que nos dan una distancia promedio con respecto a la media de la distribución: *varianza y desviación estándar.*

VARIANZA DE LA POBLACIÓN:

Es el promedio de las distancias al cuadrado que van de las observaciones a la media

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} = \frac{\sum x^2}{N} - \mu^2$$

σ^2 : Varianza de la población
x : Elemento u observación
μ : Media de la población
N : Número total de elementos de la población

DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN:

Es la raíz cuadrada de la varianza

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} = \sqrt{\frac{\sum x^2}{N} - \mu^2}$$

Para calcular la varianza de la población, dividimos la suma de las distancias al cuadrado entre la media y cada elemento de la población. Al elevar al cuadrado cada una de las distancias, logramos que todos los números que aparecen sean positivos y, al mismo tiempo asignamos más peso a las desviaciones más grandes. Las unidades de la varianza están elevadas al cuadrado (pesos al cuadrado, unidades al cuadradro, etc.) lo que hace que no sean claras o fáciles de interpretar.

La desviación estándar, que es la raíz positiva de la varianza, se mide en la misma unidad que la variable, y su interpretación es **" en promedio los valores se alejan de la media en ..... unidades"**

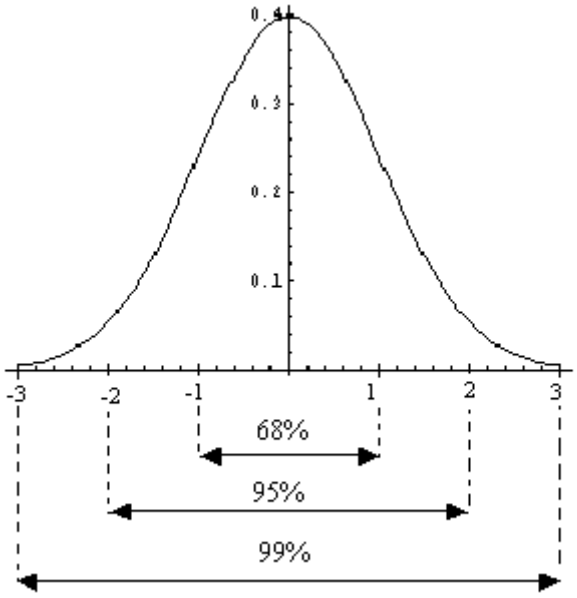
Aplicación de la desviación estándar poblacional

La desviación estándar nos permite determinar, con un buen grado de precisión, dónde están localizados los valores de una distribución de frecuencias con relación a la media.

Para curvas cualesquiera, el ***teorema de Chebyshev*** asegura que al menos el 75% de los valores caen dentro de $\pm 2\sigma$ (2 desviaciones estándar) a partir de la media μ, y al menos el 89% de los valores caen dentro de $\pm 3\sigma$.

Se puede medir con más precisión el porcentaje de observaciones que caen dentro de un rango específico de ***curvas simétricas con forma de campana (regla empírica):***

1. Aproximadamente 68% de las observaciones cae dentro de $\pm 1\sigma$
2. Aproximadamente 95% de las observaciones cae dentro de $\pm 2\sigma$
3. Aproximadamente 99% de las observaciones cae dentro de $\pm 3\sigma$



En el gráfico interpretamos el 0 como μ, y los números como unidades de σ. Por ejemplo, 1 es $\mu+\sigma$; -1 es $\mu-\sigma$; 2 es $\mu+2\sigma$; etc.

Cálculo de la varianza y la desviación estándar utilizando datos agrupados

$$\sigma^2 = \frac{\sum (m_i - \mu).f_i}{N} = \frac{\sum m_i^2.f_i}{N} - \mu^2$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (m_i - \mu).f_i}{N}} = \sqrt{\frac{\sum m_i^2.f_i}{N} - \mu^2}$$

σ^2 : Varianza de la población
σ : Desviación estándar de la población
f_i : frecuencia absoluta de la clase i
m_i : marca de clase de la clase i
μ : media de la población
N : tamaño de la población

VARIANZA Y DESVIACIÓN ESTÁNDAR MUESTRAL

Para calcular la varianza y la desviación estándar muestral se utilizan las mismas fórmulas que las poblacionales, sustituyendo μ con $\bar{x}$ y N con $n-1$.

La utilización de $n-1$ en lugar de n se verá con más detalle más adelante.

Las expresiones para el cálculo de la varianza y desviación estándar muestral son:

DATOS SIN AGRUPAR

VARIANZA MUESTRAL:

$$s^2 = \frac{\sum \left(x - \bar{x}\right)^2}{n-1} = \frac{\sum x^2}{n-1} - \frac{n.\bar{x}^2}{n-1}$$

s^2 : Varianza de la muestra
x : Elemento u observación
$\bar{x}$: Media de la muestra
n : Número de elementos de la muestra

DESVIACIÓN ESTÁNDAR MUESTRAL:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum \left(x - \bar{x}\right)^2}{n-1}} = \sqrt{\frac{\sum x^2}{n-1} - \frac{n.\bar{x}^2}{n-1}}$$

DATOS AGRUPADOS

VARIANZA MUESTRAL:

$$s^2 = \frac{\sum (m_i - \bar{x}).f_i}{n-1}$$

DESVIACIÓN ESTÁNDAR MUESTRAL:

$$s = \sqrt{\frac{\sum (m_i - \bar{x}).f_i}{n-1}}$$

s^2 : Varianza de la muestra
s : Desviación estándar de la muestra
f_i : frecuencia absoluta de la clase i
m_i : marca de clase de la clase i
$\bar{x}$: media de la muestra
n : tamaño de la muestra

Ejemplo: Los siguientes datos representan una muestra de la cantidad de pedidos diarios entregados :

17 25 28 27 16 21 20 22 18 23

a) Hallar el rango, la varianza y la desviación estándar e interpretar.
b) Hallar el porcentaje de observaciones que están alrededor de la media a una distancia de 2 desviaciones estándar. Comparar con el teorema de Chebyshev y con la regla empírica
   a) Para hallar el rango ordenamos el conjunto de mayor a menor

16 17 18 20 21 22 23 25 27 28

R = $x_{(10)} - x_{(1)}$ = 28 - 16 = 12 **La diferencia entre el mayor y el menor valor observado es 12**

Para el cálculo de la varianza conviene realizar un cuadro:

x (1)	$\bar{x}$ (2)	$x - \bar{x}$ (3)	$(x - \bar{x})^2$ (4)	x^2 $(1)^2$
16	21,7	-5,7	32,49	256
17	21,7	-4,7	22,09	289
18	21,7	-3,7	13,69	324
20	21,7	-1,7	2,89	400
21	21,7	-0,7	0,49	441
22	21,7	0,3	0,09	484
23	21,7	1,3	1,69	529
25	21,7	3,3	10,89	625
27	21,7	5,3	28,09	729
28	21,7	6,3	39,69	784
$\sum x = 217$			$\sum (x - \bar{x})^2 = 152,1$	$\sum x^2 = 4861$

1) $$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{152,1}{10 - 1} = 16,9$$

$s = \sqrt{s^2} = 4,11$ **En promedio, la cantidad de pedidos se separa de la media, en 4,11 (pedidos).**

2) $$s^2 = \frac{\sum x^2}{n - 1} - \frac{n.\bar{x}^2}{n - 1} = \frac{4861}{9} - \frac{10.(21,7)^2}{9} = \frac{152,1}{9} = 16,9$$

b) $(\bar{x} - 2s; \bar{x} + 2s) = (21,7 - 8,22; 21,7 + 8,22) = (13,48; 28,92)$

Todos los valores de la variable caen en este intervalo o sea el 100%

Según Chebyshev: al menos el 75% de los valores caen en ese intervalo, por lo tanto se verifica

Según la regla empírica: aproximadamente el 95% de las observaciones caen en dicho intervalo, (el 100% es un valor bastante cercano)

COEFICIENTE DE VARIACIÓN:

La desviación estándar es una medida absoluta de la dispersión que expresa la variación en las mismas unidades que los datos originales. Pero no puede ser la única base para la comparación de dos distribuciones. Por ejemplo si tenemos una desviación estándar de 10 y una media de 5, los valores varían en una cantidad que es el doble de la media. Si por otro lado tenemos una desviación estándar de 10 con una media de 5000, la variación respecto a la media es insignificante.

Lo que necesitamos es una medida relativa que nos proporcione una estimación de la magnitud de la desviación respecto de la magnitud de la media.

El **coeficiente de variación** es una medida relativa de dispersión que expresa a la desviación estándar como un porcentaje de la media

$CV = \frac{\sigma}{\mu}.100\%$ **en la población**	$CV = \frac{s}{\bar{x}}.100\%$ **en la muestra**

Se lo utiliza en la comparación de variación de dos o más grupos.

Ejemplo: Se pretende comparar el desempeño en ventas de 3 vendedores. Los resultados siguientes dan los promedios de puntajes obtenidos en los cinco años pasados por la concreción de los objetivos

A	88	68	89	92	103
B	76	88	90	86	79
C	104	88	118	88	123

$\bar{x}_A = 88$ $\quad s_A = 12,67$ $\quad CV = \frac{12,67}{88}.100\% = 14,4\%$

$\bar{x}_B = 83,8$ $\quad s_B = 6,02$ $\quad CV = \frac{6,02}{83,8}.100\% = 7,18\%$

$\bar{x}_C = 104,2$ $\quad s_C = 16,35$ $\quad CV = \frac{16,35}{104,2}.100\% = 15,69\%$

Vemos que el vendedor C tiene la mayor variabilidad, mientras que el B tiene la menor. El desempeño de C parece ser mejor si analizamos la media, pero hay que tener en cuenta que también tiene la mayor variabilidad en la concreción de los objetivos.

PERCENTILES

Un percentil aporta información acerca de la dispersión de los datos en el intervalo que va del menor al mayor valor de los datos. En los conjuntos de datos que no tienen muchos valores repetidos, el percentil p divide e los datos en dos partes. Cerca del p porciento de las observaciones tienen valores menores que el percentil p y aproximadamente (100-p) por ciento de las observaciones tienen valores mayores o iguales que este valor.
Definición:
El percentil p es un valor tal que por lo menos p porciento de las observaciones son menores o iguales que este valor y por lo menos (100-p) por ciento de las restantes son mayores o iguales que ese valor.

Cálculo del percentil:
Paso 1. Ordenar los datos de menor a mayor en orden ascendente.
Paso2. Calcular el índice i

$$i=\left(\frac{p}{100}\right)n$$

donde ***p*** es el percentil deseado y ***n*** el número de observaciones.

Paso 3. (a) Si no es un número entero, debe redondearse al primer entero mayor que ***i*** denotando la posición del percentil ***p.***
(b) Si es un número entero, el percentil p es el promedio de los valores en las posiciones **i** e **i+1**

Ejemplo:
Se tiene los primeros sueldos de 12 egresados en Administración.
Ordenados son:

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

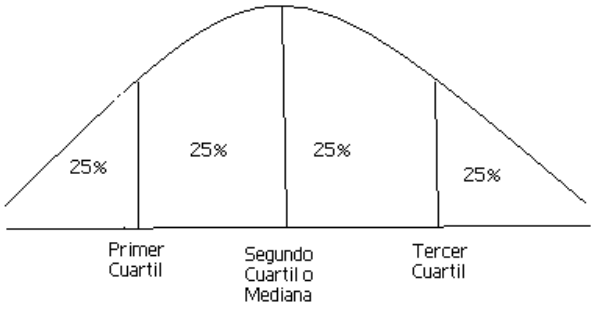
Paso 2: $$i=\left(\frac{p}{100}\right)n=\left(\frac{85}{100}\right)12=10.2$$

Paso 3. Como ***i*** no es un número entero se debe redondear al primer entero mayor que es 11.
Es decir el percentil 85 se encuentra en la posición 11. Este es 3730

CUARTILES

Con frecuencia es conveniente dividir los datos en cuatro partes, así cada una contiene el 25% de los datos. A los puntos de división se los llama cuartiles :

Q_1= primer cuartil o percentil 25
Q_2= segundo cuartil o percentil 50
Q_1= tercer cuartil o percentil 75



Rango intercuartílico (RIC) es también una medida importante a tener en cuenta, es la diferencia entre el tercer y primer cuartel

$$\mathbf{RIC}= Q_3- Q_1$$

Nos indica el 50 % de las observaciones centrales

DIAGRAMA DE CAJA Y BIGOTES

Un diagrama de caja es un resúmen gráfico de los datos con base en el resumen de cinco números .
La clave para elaborar un diagrama de cajas está en calcular Q_1, Q_3 ***y la mediana o*** Q_2.
También hay que calcular el **RIC**= $Q_3- Q_1$
Pasos para dibujar el diagrama de cajas:

1. Se dibuja una caja cuyos extremos se localicen en el primer y tercer cuartel. En nuestros datos de salarios $Q_1=3465$ ***y*** $Q_3= 3600$. Significa que la caja contiene el 50% de los datos centrales .

2. En el punto dónde se localiza la mediana (3505) se traza una línea horizontal o vertical según se represente la caja en posición vertical u horizontal respectivamente. Si se quieren comparar dos poblaciones a veces también se representa la media dentro de la caja.

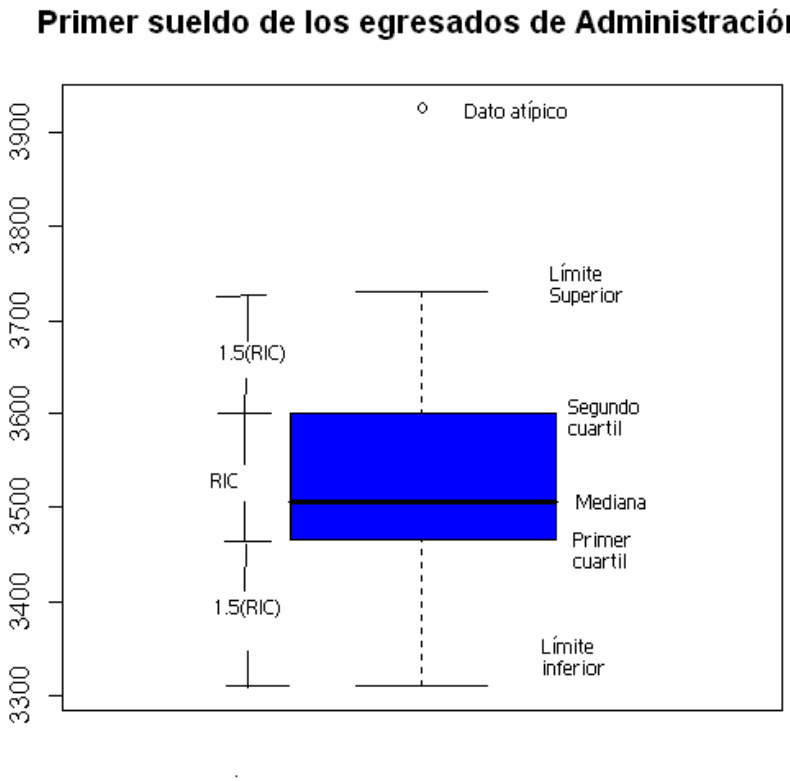
3.Usando el rango intercuartílico **RIC**= $Q_3- Q_1$ se localizan los límites. En un diagrama de caja los límites se encuentran en 1,5*(RIC) abajo del Q_1 y 1,5(RIC) arriba del Q_3 . En el caso de los salarios el **RIC**= $Q_3- Q_1$ = 3600-3465=135. por lo tanto los límites son
L_i=3465 - 1,5*(RIC)= 3465-1,5*135 = 3262,5
L_s= 3600+ 1,5*(RIC)= 3600+1,5*135=3802,5
Los datos que quedan fuera de estos límites se consideran observaciones atípicas.

4. A las líneas punteadas se las llama bigotes . Los bigotes van desde los extremos de la caja hasta los valores menor y mayor de los correspondientes a los límites inferior y superior encontrados en el paso 3.Por lo tanto los bigotes terminan en los salarios cuyos valores son 3310 y 3730.

5. Por último con un círculo o asterisco se identifica la observación atípica 3925.

Actividad con R

```
> sueldo<-c(3310,3355,3450,3480,3480,3490,3520,3540,3550,3650,3730,3925)
> boxplot(sueldo, main="Primer sueldo de los egresados de Administración", col="blue")
```



Este gráfico no se puede realizar con Excel.
Para obtener todas las medidas juntas usando R se utiliza el comando summary.

```
Summary(sueldo)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   3310    3472    3505    3540    3575    3925
```

Microsoft Excel -

Para datos sin agrupar en el caso de la edad de los jubilados encuestados se colocan en una columna y luego en el menú herramientas se busca análisis de datos estadística descriptiva se marca el rango de las celdas y se le pide resumen de estadísticas aceptar y larga

Edad de los jubilados encuestados en Mendoza en noviembre del 2008.

	Columna1
Media	68,42
Error típico	1,47277054
Mediana	65,5

Moda	65
Desviación estándar	10,4140604
Varianza de la muestra	108,452653
Curtosis	-0,6706671
Coeficiente de asimetría	0,43071849
Rango	40
Mínimo	53
Máximo	93
Suma	3421
Cuenta	50

Ejemplo de los salarios de los egresados de Administración:

Columna1	
Media	3540
Error típico	47,8198957
Mediana	3505
Moda	3480
Desviación estándar	165,652978
Varianza de la muestra	27440,9091
Curtosis	1,71888364
Coeficiente de asimetría	1,09110869
Rango	615
Mínimo	3310
Máximo	3925
Suma	42480
Cuenta	12