ESTADÍSTICA DESCRIPTIVA

Elaboró: **Oscar Javier Gómez Mora**
Noviembre de 2017
Mail: ojgomez@gmail.com

CONTENIDO

ESTADÍSTICA DESCRIPTIVA

1. Introducción

La Estadística descriptiva es la rama de las matemáticas que comprende la recopilación, tabulación, análisis e interpretación de datos cuantitativos y cualitativos, para tomar decisiones que se requieran a fin de que el comportamiento de los datos se mantenga dentro de los parámetros de control establecidos.

- **Población (N)–** Es el conjunto de todos los elementos de interés para determinado estudio

- **Parámetro –** Es una característica numérica de la población, se identifica con letras griegas (Media = μ, Desviación estándar = σ, Proporción = π, Coeficiente de correlación = ρ)

- **Muestra (n) –** Es una parte de la población, debe ser representativa de la misma.

- **Estadístico –** Es una característica numérica de una muestra, se identifica con letras latinas (Media = $\overline{X}$, Desviación estándar = s, Proporción = p, Coeficiente de correlación = r)

La Estadística descriptiva proporciona un criterio para lograr mejoras, debido a que sus técnicas se pueden usar para describir y comprender la variabilidad. Por ejemplo, consideremos en una caldera de vapor la presión del combustible alimentado y la eficiencia de la caldera, si utilizamos instrumentos de medición con la resolución suficiente, encontraremos que existe variabilidad en esos parámetros, y mediante el uso de técnicas estadísticas podemos realizar mejoras para reducir la variación en rendimiento de la caldera.

Para poder obtener consecuencias y deducciones válidas de los datos de un estadístico, es muy útil contar con información sobre los valores que se agrupan hacia el centro y sobre que tan distanciados o dispersos estén unos respecto a otros. Comenzaremos por definir estas medidas:

La estadística inferencial se refiere a la estimación de parámetros y pruebas de hipótesis acerca de las características de la población en base a los datos obtenidos con una muestra.

2. MEDIDAS DE TENDENCIA CENTRAL Y DISPERSIÓN PARA DATOS SIMPLES.

Medidas de tendencia central

- **Media: ($\bar{x}$)** Es el promedio aritmético de todos los valores que componen el conjunto de datos. Se calcula mediante la siguiente fórmula:

  Para una muestra y para una población se tiene respectivamente:

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

Ejemplo 1: En un equipo de fútbol, una muestra de estaturas de sus integrantes son las siguientes:

1.70,1.79,1.73,1.67,1.60,1.65,1.79,1.84,1.67,1.82, 1.74. Calcule la media.

$$\bar{x} = \sum \frac{xi}{n} = \frac{19}{11} = 1.73$$

- **Mediana: ($\tilde{x}$)** Los datos de "n" observaciones son ordenados del más pequeño al más grande, Si el tamaño de la muestra es "non" la mediana es el valor ordenado en la posición (n+1)/2,
  Cuando el tamaño de la muestra es "par" la mediana es el promedio de los dos valores que se encuentran al centro del conjunto de valores. Se puede calcular mediante:

$$\frac{(n/2) + ([n/2] + 1)}{2}$$

Ejemplo 2: Para el ejemplo anterior ¿cuál es la mediana?

Ordenando los datos de mayor a menor se obtiene:

1.60,1.65,1.67,1.67,1.70,1.73,1.74,1.79,1.79,1.82,1.84;

Como tenemos 11 datos el número es non por lo que (n+1)/2 = 12/2 = 6, buscando el número que ocupa la sexta posición en los datos ordenados encontramos el valor de la mediana $\tilde{x} = 1.73$

- **Media acotada (Truncated Mean):** Determinado porcentaje de los valores más altos y bajos de un conjunto dado de datos son eliminados (tomando números enteros), para los valores restantes se calcula la media.

Ejemplo 3: Para la siguiente serie de datos calcule la media acotada al 20%:

68.7,34.3,97.9,73.4,8.4,42.5,87.9,31.1,33.2,97.7,72.3,54.2,80.6,71.6,82.2,

Como tenemos 11 datos, el 20% de 11 es 2.2, por lo cual eliminamos 2 datos el más bajo y el más alto, ordenado los datos obtenemos:

8.4,31.1,33.2,34.3,42.5,54.2,68.7,71.6,72.3,73.4,80.6,82.2,87.9,97.7,97.9, los valores a eliminar son: 8.4 y 97.9; calculando la media de los datos restantes obtenemos $(\tilde{x}, .20) = 63.82$

Medidas de dispersión

Para comprender el concepto de varianza, supóngase que tenemos los datos siguientes de los cuales queremos saber que tan dispersos están respecto a su media:

2, 3, 4, 5, 6 con media = 20/5 = 4

Si tomamos la suma de diferencias de cada valor respecto a su media y las sumamos se tiene:

(-2) + (-1) + (0) + (1) +(2) = 0

Por lo que tomando diferencias simples no es posible determinar la dispersión de los datos.

Si ahora tomamos esas mismas diferencias al cuadrado y las sumamos se tiene:

4 + 1 + 0 + 1 + 4 = 10

Varianza de los datos

Es una medida que nos ayuda a comprender la variabilidad de los datos, que tan distanciados están de la media

- **Poblacional (σ^2)** Se obtiene dividiendo el valor anterior entre n = 5, o sea el promedio de la suma de las diferencias al cuadrado, tomando n datos.

$$\sigma^2 = \sum \frac{(xi - \bar{x})^2}{n}$$

- **Poblacional (s^2)** Se obtiene dividiendo el valor anterior entre n - 1 = 4, o sea el promedio de la suma de las diferencias al cuadrado, tomando n -1 datos.

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

- **Desviación estándar:** Es la raíz cuadrada de la varianza:

Para el caso de una población $\sigma = \sqrt{\sum \frac{(xi - \bar{x})^2}{n}}$

Para el caso de una muestra $s = \sqrt{\sum \frac{(xi - \bar{x})^2}{n-1}}$

- **Rango (R):** es la diferencia positiva entre el valor mayor y el valor menor de un conjunto de datos. Por ejemplo para el conjunto de datos siguiente: 2.0,2.1,2.4,2.5,2.6,2.8,2.9,2.9,3.0,3.1,3.6,3.8,4.0,4.0

  Su rango es R = 4.0 – 2.0 = 2.0

- **Coeficiente de Variación (CV):** Se utiliza para comparar la dispersión de dos conjuntos de datos que tienen unidades diferentes, ya que representa una medida relativa de dispersión.

$$Coeficiente.de.\mathrm{var}\,iación = CV = \frac{s}{\overline{\overline{X}}}(100)$$

Por ejemplo si la media de tiempos de respuesta es de 78.7 y su desviación estándar es 12.14, el CVt:

$$CV_t = \frac{12.14}{78.7}(100) = 12.05\%$$

Por otra parte si la media de temperaturas es de 10 y su desviación estándar de 2, el CVs de las temperaturas es:

$$CV_s = \frac{2}{10}(100) = 20\%$$

Por tanto la dispersión de las temperaturas es mayor que la de los tiempos de de respuesta, es posible comparar estas dispersiones con el CV aunque los dos conjuntos de datos sean completamente disímbolos.

Ejemplo 4: La resistencia al rompimiento de dos **muestras** de botellas es la siguiente:

Muestra 1:	230	250	245	258	265	240
Muestra 2:	190	228	305	240	265	260

Calcule la desviación estándar para ambas muestras.

Muestra 1:

$\vec{x} = 248$

Suma(Xi - $\bar{x}$ $)^2$ = 790

n - 1 = 5

s = $\sqrt{\frac{790}{5}}$ = 12.56

Rango = 265 – 230 = 35

CV = 12.56/248*100= 5.06%

Muestra 2

$\vec{x} = 248$

Suma(□□Xi - $\bar{x}$ $)^2$ = 7510

n-1 = 5

s = $\sqrt{\frac{7510}{5}}$ = 38.75

Rango = 305 – 190 = 115

CV = 38.75/248*100 = 15.625

Aunque la media en ambas muestras es la misma, la desviación estándar (s), rango y coeficiente de variación, son menores en la muestra 1, por lo cual deducimos que es presenta menor variabilidad.

Ejemplo 5:

Se desea hacer un estudio estadístico de la temperatura del agua, para esto es necesario tomar una muestra y calcular la media, mediana, media acotada al 15%, desviación estándar, rango y coeficiente de variación. Se realizan 14 observaciones arrojando los siguientes resultados en ºC: 2.11, 3.8, 4.0, 4.0, 3.1, 2.9, 2.5, 3.6, 2.0, 2.4, 2.8, 2.6,2.9, 3.0.

1) Calcular la media, mediana, desviación estándar, media acotada al 5%, desviación estándar, rango y coeficiente de variación.

3. OTRAS MEDIDAS DE DISPERSIÓN: PERCENTILES, DECILES Y QUARTILES

Cada conjunto de datos ordenado tiene tres cuartiles que lo dividen en cuatro partes iguales. El primer cuartil es ese valor debajo del cual clasifica el 25% de las observaciones y sobre el cual se encuentra el 75% restante. El segundo cuartil divide a los datos a la mitad similar a la mediana.

Los deciles separan un conjunto de datos ordenado en 10 subconjuntos iguales y los percentiles en 100 partes, la ubicación de un percentil se encuentra en:

$$L_p = (n+1)\frac{P}{100}$$

Donde:

Lp es el sitio del percentil deseado en una serie ordenada
n es el número de observaciones
P es el percentil deseado

Por ejemplo para el conjunto de datos siguiente:

3	10	19	27	34	38	48	56	67	74
4	12	20	29	34	39	48	59	67	74
7	14	21	31	36	43	52	62	69	76
9	15	25	31	37	45	53	63	72	79
10	17	27	34	38	47	56	64	73	80

La localización del percentil 35 se halla en:

$$L_{35} = (50+1)\frac{35}{100} = 17.85$$

O sea que el percentil 35 está al 85% del trayecto comprendido entre la observación 17 que es 29 y la observación 18 que es 31 o sea L35 = 29 + (0.85)(31-29) = 30.7. Por tanto el 35% de las observaciones están por debajo de 30.7 y el 65% restante por encima de 30.7.

De la misma forma los percentiles 25, 50 y 75 proporcionan la localización de los cuartiles Q1, Q2 y Q3 respectivamente.

- **Q1:** es el número que representa al percentil 25 (hay 25% de los datos por debajo de este).

- **Q2 o Mediana:** es el número que representa al percentil 50 (hay 50% de los datos por debajo de este).

- **Q3:** es el número que representa al percentil 75 (hay 75% de los datos por debajo de este).

- **Rango o Recorrido intercuartílico:** es la diferencia entre Q1 y Q3.

DIAGRAMA DE CAJA

Es la representación gráfica de los datos en forma de caja:

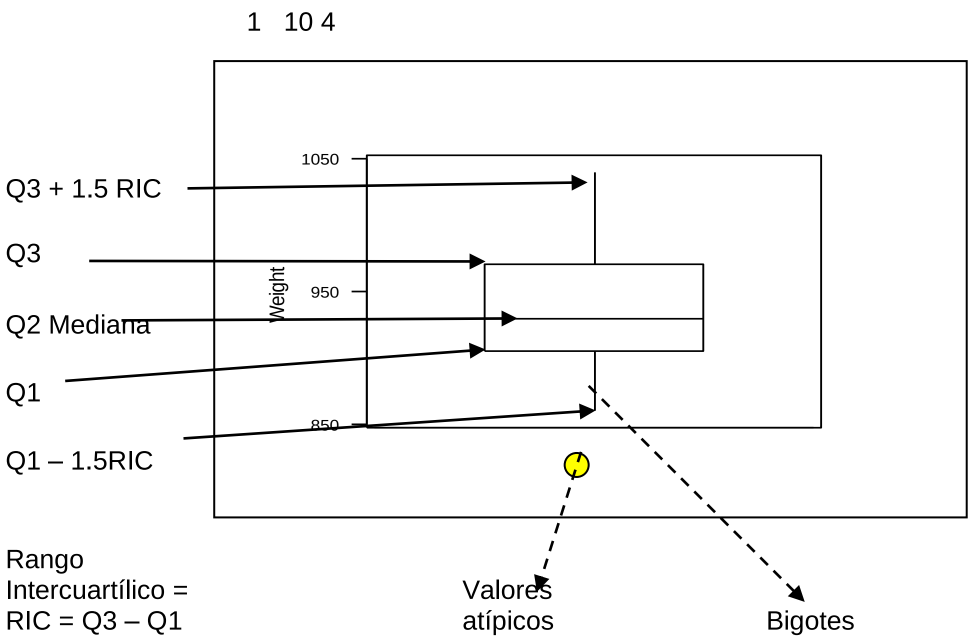


Figura 1. Diagrama de caja con sus cuarteles y bigotes

4. DISTRIBUCIÓN DE FRECUENCIAS E HISTOGRAMAS

Cuando tenemos una cantidad grande de datos es difícil poder analizarlos, a menos que hagamos uso de herramientas que nos permitan hacerlo con mayor facilidad y claridad. El histograma es una de ellas, consiste en un diagrama de barras donde las bases corresponden a los intervalos y las alturas a las frecuencias. Para construir un histograma es necesario tener un mínimo de 50 a 100 datos. Se tienen las siguientes definiciones:

- **Distribución de frecuencias:** es un resumen tabular de un conjunto de datos que muestra el número o frecuencia de artículos en cada una de varias clases que no se traslapan.

- **Frecuencia relativa (f):** Es la frecuencia de la clase dividida entre el total n de datos. Se puede representar en porcentaje.

- **Distribución de frecuencias porcentuales:** es la representación de las frecuencias relativas porcentuales.

- **Frecuencia acumulada (F):** es la acumulación secuencial de las frecuencias de cada clase.

Ejemplo 6

Construir un histograma con la siguiente serie de datos:

2.41	17.87	33.51	38.65	45.70	49.36	55.08	62.53	70.37	81.21
3.34	18.03	33.76	39.02	45.91	49.95	55.23	62.78	71.05	82.37
4.04	18.69	34.58	39.64	46.50	50.02	55.56	62.98	71.14	82.79
4.46	19.94	35.58	40.41	47.09	50.10	55.87	63.03	72.46	83.31
8.46	20.20	35.93	40.58	47.21	50.10	56.04	64.12	72.77	85.83
9.15	20.31	36.08	40.64	47.56	50.72	56.29	64.29	74.03	88.67
11.59	24.19	36.14	43.61	47.93	51.40	58.18	65.44	74.10	89.28
12.73	28.75	36.80	44.06	48.02	51.41	59.03	66.18	76.26	89.58
13.18	30.36	36.92	44.52	48.31	51.77	59.37	66.56	76.69	94.07
15.47	30.63	37.23	45.01	48.55	52.43	59.61	67.45	77.91	94.47
16.20	31.21	37.31	45.08	48.62	53.22	59.81	67.87	78.24	94.60
16.49	32.44	37.64	45.10	48.98	54.28	60.27	69.09	79.35	94.74
17.11	32.89	38.29	45.37	49.33	54.71	61.30	69.86	80.32	96.78

Paso 1: Contar el número de datos **n = 130**

Paso 2: Calcular el rango **R =** Valor mayor – Valor menor, R = 96.78-2.41 = 94.37.
Generalmente los datos no están ordenados por lo cual resulta conveniente ordenarlos de menor a mayor para tener una mejor visualización. En el ejemplo los datos ya han sido previamente ordenados.

Paso 3: Seleccionar el número de columnas, mediante $\sqrt{n}$ = $\sqrt{130} = 11.4 \approx 11$. Por lo cual el histograma se compone de 11 columnas

Paso 4: Calcular el tamaño del intervalo de clase **(C)**, dividiendo el rango entre el número de columnas: **C =** $\frac{94.37}{11} = 8.58 \approx 9$**,** resultando el tamaño del intervalo 9.

- Otra manera de calcular el tamaño del intervalo es el siguiente:

Dividir el valor del rango entre un cierto número de clases (K). La tabla de abajo es una guía que nos muestra para diferentes cantidades de datos el número recomendado de clases a utilizar.

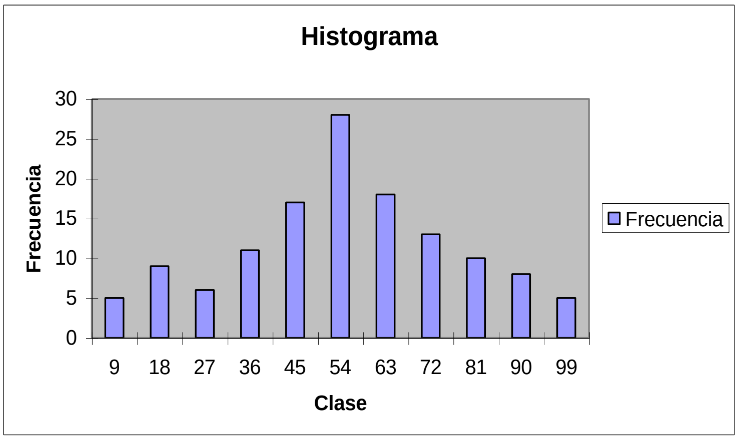
Número de datos (N)	Número de clases (K)
Menos de 50	5 – 7
50 a 100	6 – 10
100 a 250	7 – 12
Más de 250	10 – 20

Paso 5: Calcular los limites de clase de cada intervalo: [0-8], [9-17], etc., considerando que el tamaño del intervalo representa la diferencia entre dos límites de clase adyacentes ya sean inferiores o superiores.

Paso 6: Contar el número de valores que caen en cada intervalo utilizando una **hoja de registro**, de esta manera se obtiene la frecuencia para cada intervalo.

Tabla 1.

Columna	Intervalo	Registro de frecuencias						
1	0 -8	IIIII						5
2	9-17	IIIII	IIII					9
3	18-26	IIIII	I					6
4	27-35	IIIII	IIIII	I				11
5	36-44	IIIII	IIIII	II				17
6	45-53	IIIII	IIIII	IIIII	IIIII	IIIII	III	28
7	54-62	IIIII	IIIII	IIIII	III			18
8	63-71	IIIII	IIIII	III				13
9	72-80	IIIII	IIIII					10
10	81-89	IIIII	III					8
11	90-98	IIIII						5



Paso 7: Basándose en los datos anteriores construya el histograma.

Diagrama de tallo y hojas

Es otra representación de la información, primero se ordenan los dígitos principales a la izquierda de una línea vertical. A la derecha de esta línea se registra el último dígito para cada dato conforme se revisan las observaciones en el orden en que se registraron. Por ejemplo:

Con Minitab: Stat > EDA > Steam and leaf… Indicar columna de datos, increment = 10

```
Stem-and-leaf of Respuest  N  = 50
Leaf Unit = 1.0
```

```
  2     6 89
  8     7 233566
 16     8 01123456
(11)    9 12224556788
 23    10 002466678
 14    11 2355899
  7    12 4678
  3    13 24
  1    14 1
```

5. MEDIDAS DE TENDENCIA CENTRAL Y DE DISPERSIÓN PARA DATOS AGRUPADOS.

- **La media con datos agrupados**: se calcula así:

$$\overline{X}_g = \frac{\sum fM}{n}$$

Donde

f es la frecuencia o número de observaciones en cada clase
M es el punto medio de cada clase, se determina como el valor medio entre los límites de clase.
n es el tamaño de la muestra o la suma de todas las frecuencias de las clases

Ejemplo:

Clase (Presión)	**Frecuencia de clase (días)**	**M**	**fM**	**Frecuencia acumulada F**
50-59	3	54.5	163.5	3
60-69	7	64.5	451.5	10
70-79	18	74.5	1341.0	28
80-89	12	84.5	1014.0	40
90-99	8	94.5	756.0	48
100-109	2	104.5	209.0	50
	50		3935.0	

$$\overline{X}_g = \frac{3935}{50} = 78.7$$

- **Mediana de datos agrupados:**

Primero se identifica la clase donde se encuentra la mediana cuya F es >= n / 2, en este caso la clase de 70 a 79 con punto central de clase = 74.5.

$$Mediana = \tilde{X} = L_{md} + \left[\frac{n/2 - F}{f_{md}}\right](C) = 70 + \left[\frac{50/2 - 10}{18}\right]10 = 78.33\,pasajeros$$

Donde:

Lmd es el límite inferior de la *clase de la mediana cuya F es >= n / 2* o sean (70)
F es la frecuencia acumulada de la clase que antecede a la clase de la mediana (10)
Fmd es la frecuencia de la clase de la mediana (18)
C es el intervalo de clase de la mediana que es la diferencia entre dos límites de clase (10)

- **Moda de datos agrupados:**

Primero se halla la clase que tenga la frecuencia más alta, en este caso la clase 70 a 79.

$$Moda = L_{mo} + \left[\frac{D_a}{D_b + D_a}\right](C) = 70 + \left[\frac{18-7}{(18-12)+(18-7)}\right]10 = 76.47$$

Donde:

Lmo es el límite inferior de la clase modal con la frecuencia más alta (70).
Da es la diferencia entre la frecuencia de la clase modal y la clase que la antecede (18 – 7 = 11)
Db es la diferencia entre la frecuencia de la clase modal y la clase que le sigue (18 – 12 = 6)
C es el intervalo de la clase modal (80 – 70 = 10)

- **Varianza y desviación estándar de datos agrupados:**

$$s^2 = \frac{\sum fM^2 - n\overline{X}^2}{n-1}$$

$$s = \sqrt{s^2}$$

Para los datos anteriores se tiene:

Clase (Presión)	**Frecuencia de clase (días)**	**M**	**fM**	**M²**	**fM²**
50-59	3	54.5	163.5	2790.25	8910.75
60-69	7	64.5	451.5	4160.25	29121.75
70-79	18	74.5	1341.0	5550.25	99904.50
80-89	12	84.5	1014.0	7140.25	85683.00
90-99	8	94.5	756.0	8930.25	71442.00
100-109	2	104.5	209.0	10920.25	21840.50
			3935.0		316902.50

$$\overline{X}_g = \frac{3935}{50} = 78.7$$

$$s^2 = \frac{316902.50 - 50(78.7)^2}{49} = 147.31 \, pasajeros$$

$$s = 12.14 \, pasajeros$$

Con esta información el personal puede tomar sus decisiones

6. USOS FRECUENTES DE LA DESVIACIÓN ESTÁNDAR

- **EL TEOREMA DE TCHEBYSHEV**

Establece que para todo conjunto de datos por lo menos $(1-\frac{1}{K^2})\%$ de las observaciones se encuentran dentro de $\pm$ K desviaciones estándar de la media, con K >= 1.

Por ejemplo si K = $\pm$ 3 desviaciones estándar respecto a la media, se tiene que por lo menos el:

$$(1-\frac{1}{K^2})\% = \left(1-\frac{1}{3^2}\right)\% = 88.89\%$$

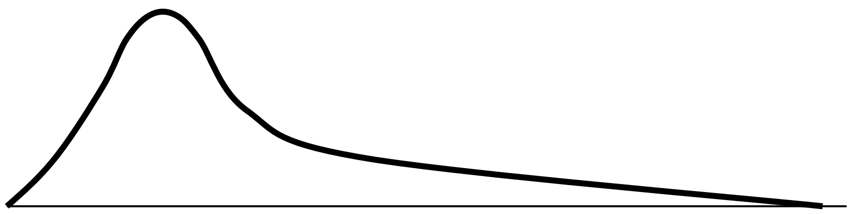
De las observaciones estarán dentro de dicho intervalo.

CASO DE LA DISTRIBUCIÓN NORMAL

68.3% de las observaciones se encuentran dentro de $\pm$ 1 desviación estándar de la media
95.5% de las observaciones se encuentran dentro de $\pm$ 2 desviaciones estándar de la media
99.7% de las observaciones se encuentran dentro de $\pm$ 3 desviaciones estándar de la media

- **SESGO**

En la distribución normal si no es simétrica y tiene una cola más amplia del lado derecho, se dice que existe un sesgo a la derecha y viceversa.



El coeficiente de sesgo o asimetría P se determina como sigue:

$$P = \frac{3(\overline{X} - Mediana)}{s}$$

Si P < 0 los datos están sesgados a la izquierda, si P > 0 están sesgados a la derecha; si P = 0 están distribuidos normalmente.

Para el caso de los datos del ejemplo anterior se tiene:

$$P = \frac{3(78.7 - 78.33)}{12.14} = 0.03$$ Los datos están un poco sesgados hacia la derecha.

Coeficiente de asimetría de Fisher

Otra estimación del sesgo o coeficiente de asimetría se hace a través de momentos estadísticos (diferencias contra la media) como lo sugiere Fisher:

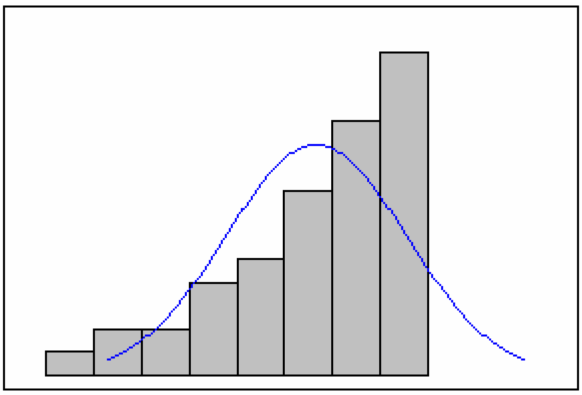
$$M_j = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^j}{n} \qquad j = 1,2,3,4$$

$$Sesgo = \hat{\beta}_1 = \frac{M_3}{M_2^{3/2}} \quad \text{o} \quad \gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(Xi - \overline{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(Xi - \overline{X})^2\right)^{3/2}}$$ Para la distribución normal debe ser 0.

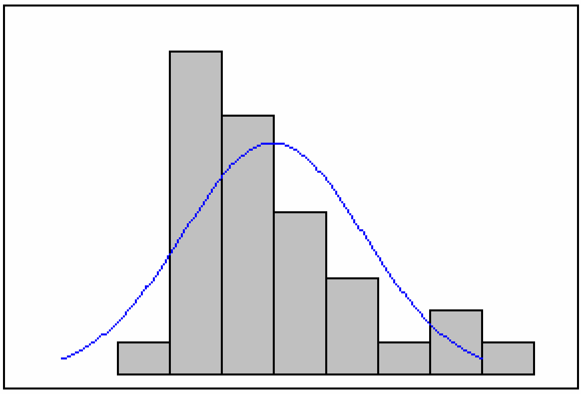
Se puede considerar que una distribución es simétrica si $\gamma 1 = 0$, asimétrica hacia la izquierda con $\gamma 1 < 0$ o hacia la derecha $\gamma 1 > 0$.

Por ejemplo:

Ejemplo de una distribución con sesgo negativo o sesgada hacia la izquierda con Sesgo = -1.01



Ejemplo de una distribución con sesgo positivo o sesgada hacia la derecha con Sesgo = 1.08



- **CURTOSIS**

En la distribución normal si no es acampanada y es más picuda o aplanada de lo normal se dice que tiene una Curtosis diferente de cero que es lo normal, si es mayor es más picuda o más plana al revés.

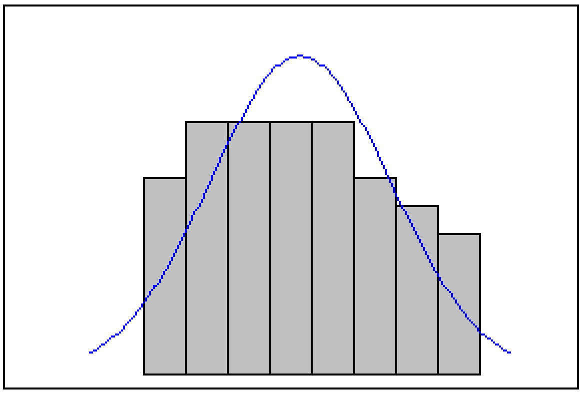
Coeficiente de Curtosis de Fisher

$$Kurtosis = \beta_2 = \frac{M_4}{M_2^2} - 3 \quad \text{o} \quad \gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(Xi-\overline{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(Xi-\overline{X})^2\right)^2} - 3$$

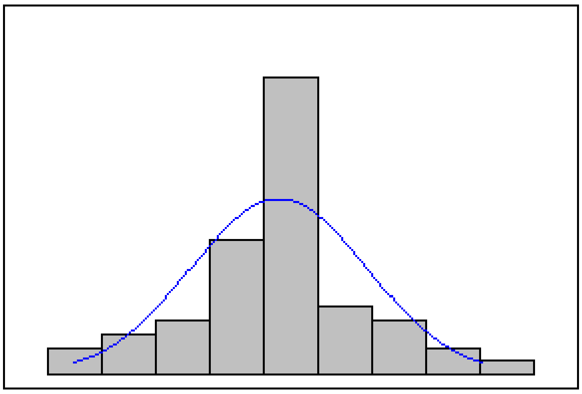
Para la distribución normal debe ser 0.

La distribución es mesocúrtica (plana normal) si $\gamma 2 = 0$, leptocúrtica si $\gamma 2 > 0$ más puntiaguda que la normal o platicúrtica (más plana que la normal) con $\gamma 2 < 0$.

Ejemplo de curva más plana que la normal Curtosis = -1.03



Ejemplo de curva más picuda que la normal Curtosis = 0.76

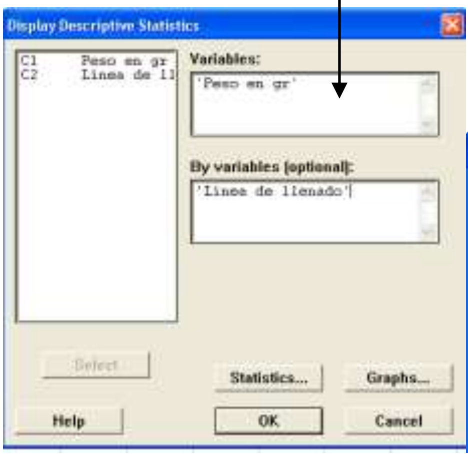


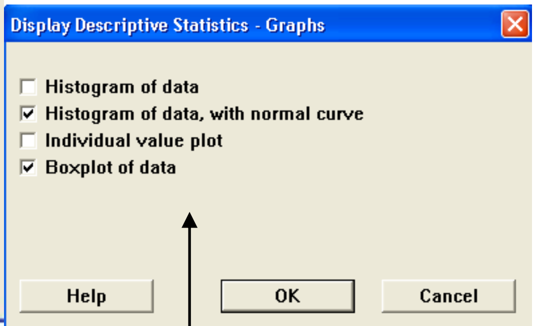
7. USO DE MINITAB y EXCEL

Para la obtención de las estadísticas descriptivas con Minitab las instrucciones son:

- **Stat > Basic statistics > Display descriptive statistics**

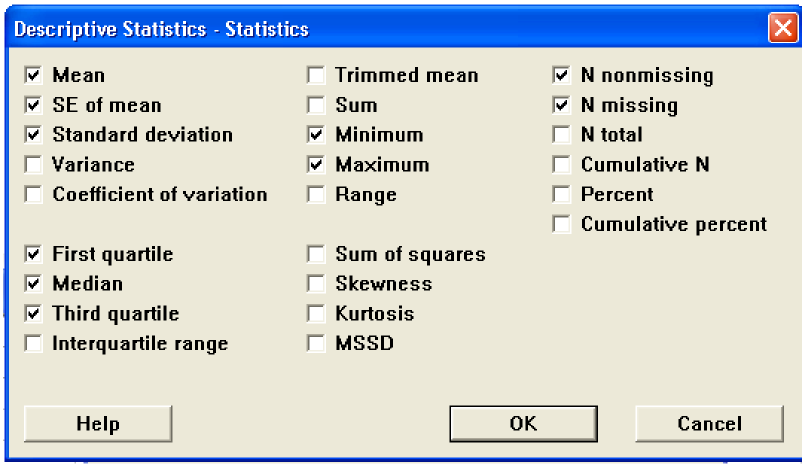
Indicar las variables de las cuales se quieren obtener las estadísticas básicas y la variable categórica si se desean varios grupos.





Seleccionar las gráficas opcionales para los datos: Histograma, diagrama de caja y de puntos.

Seleccionar los estadísticos específicos que se desean obtener:



Los resultados son los siguientes:

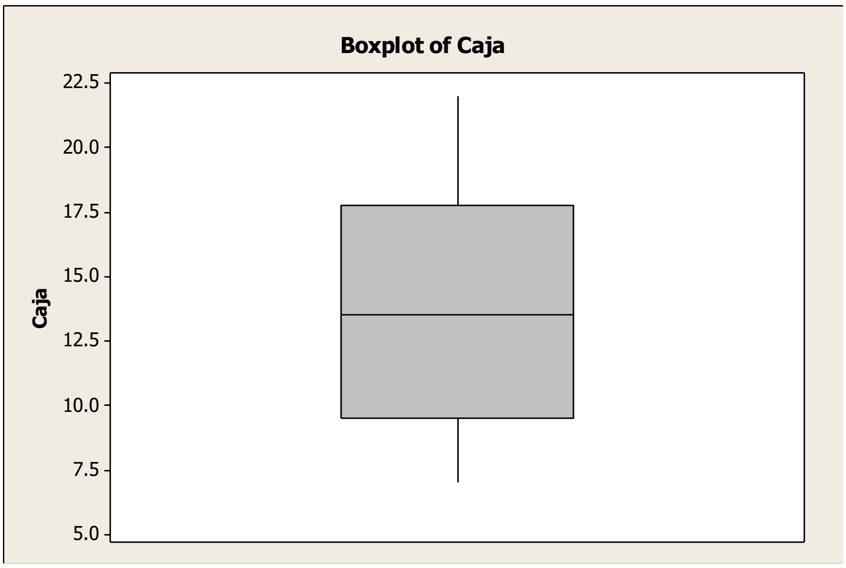
Descriptive Statistics: Peso en gr

```
Variable     Línea       N  N*     Mean  SE Mean  StDev  Minimum      Q1  Median
Peso en gr   1         250   0   3999.6     3.14   49.6   3877.0  3967.8  3999.5
             2         250   0   4085.6     3.32   52.5   3954.0  4048.8  4087.0

Variable     Línea        Q3  Maximum
Peso en gr   1        4040.0   4113.0
             2        4121.5   4202.0
```

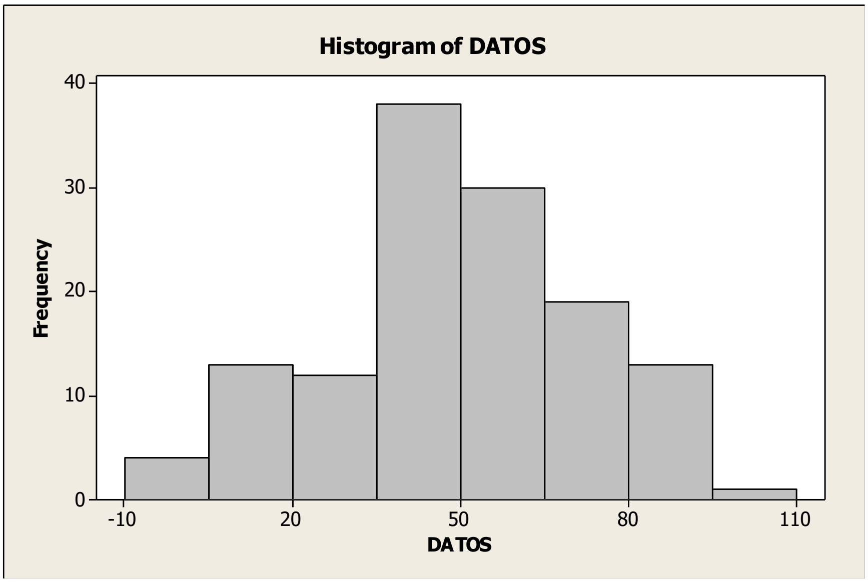
Diagramas de caja en Minitab:

1. Capture datos en la hoja de trabajo: 7 8 9 9 11 12 12 13 14 15 16 17 18 19 20 22
2. Seleccione la opción: Graph> Boxplot
3. Seleccione la variable C1 como se muestra en la pantalla y presione clic en ok
4. A continuación se muestra el diagrama de caja:



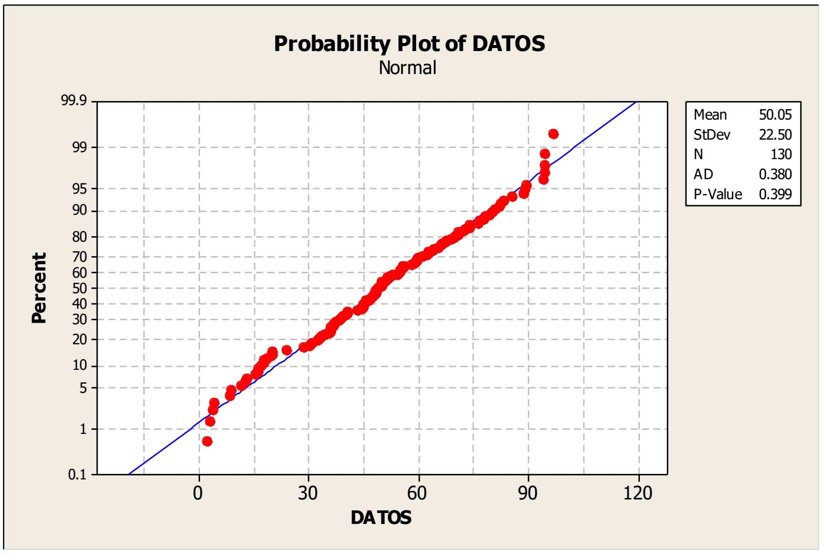
Histograma en Minitab:

1. Capture los datos del ejemplo 6 en la hoja de trabajo:
2. Seleccione la opción: Graph> Histogram (simple)
3. Seleccione la variable C1 como se muestra en la pantalla y presione clic en ok
4. En Options se puede cambiar el número de celdas con Number of intervals (6 – 8)
5. A continuación se muestra el Histograma:



Prueba de normalidad en Minitab:

1. Capture los datos del ejemplo 6 en la hoja de trabajo:
2. Seleccione la opción: Stat > Basic statistics
3. Seleccione la variable C1 como se muestra en la pantalla y presione clic en ok
4. Seleccione la prueba de Anderson Darling
5. A continuación se muestra la grafica normal, si P value > 0.05 los datos son normales.



USO DE EXCEL

1. En el menú **Herramientas** seleccione la opción **Análisis de datos**. Datos de ejemplo 6.
2. Seleccione la opción **Estadística descriptiva**.
3. Seleccione el rango de entrada, estos corresponden a los datos numéricos de la tabla.
4. Seleccione Resumen de estadísticas.
5. En opciones de salida seleccione en Rango de salida, una celda de la hoja de calculo que este en blanco (a partir de está celda serán insertados los resultados).

La hoja mostrará las siguientes medidas estadísticas de los datos presentados:

Columna1	
Media	50.0537692
Error típico	1.9738137
Mediana	49.345
Moda	50.1
Desviación estándar	22.5049388
Varianza de la muestra	506.47227
Curtosis	-0.4466339
Coeficiente de asimetría	-0.0352296
Rango	94.37
Mínimo	2.41
Máximo	96.78
Suma	6506.99
Cuenta	130

8. EJERCICIOS:

1. Las empresas de generación de energía eléctrica están interesadas en los hábitos de consumo de los clientes para obtener pronósticos exactos de las demandas de energía. Una muestra de consumidores de 90 hogares con calefacción de gas arrojó lo siguiente (FURNACE.MTW):

BTU.In_1				
2.97	7.73	9.60	11.12	13.47
4.00	7.87	9.76	11.21	13.60
5.20	7.93	9.82	11.29	13.96
5.56	8.00	9.83	11.43	14.24
5.94	8.26	9.83	11.62	14.35
5.98	8.29	9.84	11.70	15.12
6.35	8.37	9.96	11.70	15.24
6.62	8.47	10.04	12.16	16.06
6.72	8.54	10.21	12.19	16.90
6.78	8.58	10.28	12.28	18.26
6.80	8.61	10.28	12.31	
6.85	8.67	10.30	12.62	
6.94	8.69	10.35	12.69	
7.15	8.81	10.36	12.71	
7.16	9.07	10.40	12.91	
7.23	9.27	10.49	12.92	
7.29	9.37	10.50	13.11	
7.62	9.43	10.64	13.38	
7.62	9.52	10.95	13.42	
7.69	9.58	11.09	13.43	

a) Determinar los estadísticos de tendencia y dispersión

b) Construir un diagrama de caja e histograma

c) Realizar una prueba de normalidad de los datos

d) Establecer conclusiones