CAPITULO III

ORGANIZACION DE DATOS

3.1 ORGANIZACION DE DATOS CUALITATIVOS

CUADRO DE FRECUENCIAS.

Se recomienda realizar la tabla o cuadro de frecuencias.

Ejemplo: A 40 alumnos que habían desaprobado un curso en el semestre anterior, se les consultó que curso fué el que desaprobaron; las respuestas fueron las siguientes:

```
desaprobados <- c("Cálculo II", "Cálculo II", "Cálculo I", "Algebra I",
"Estadística", "Estadística", "Cálculo II", "Biología", "Química",
"Cálculo I", "Estadística", "Cálculo I", "Estadística", "Algebra I",
"Algebra I", "Física", "Cálculo I", "Algebra I", "Estadística",
"Cálculo II", "Algebra I", "Algebra I", "Cálculo I", "Cálculo I",
"Estadística", "Cálculo II", "Cálculo II", "Cálculo II", "Estadística",
"Cálculo I", "Estadística", "Genética", "Procesos", "Agrometría",
"Estadística", "Cálculo I", "Bioquímica", "Cálculo II", "Cálculo I",
"Cálculo I")

tabla1 <- table(desaprobados)

tabla2 <- round(tabla1*100/sum(tabla1),1)

tabla3<-
data.frame(alumnos=cbind(tabla1)[,1],porcentaje=cbind(tabla2)[,1])

tabla4 <- subset(tabla3,tabla3$alumnos>1)

otros <- subset(tabla3,tabla3$alumnos==1)

adicional<-data.frame(row.names="Otros",alumnos=

sum(otros$alumnos),porcentaje=sum(otros$porcentaje))

final<-rbind(tabla4,adicional)
final
```

	alumnos	porcentaje
Algebra I	6	15.4
Cálculo I	9	23.1
Cálculo II	8	20.5
Estadística	9	23.1
Otros	7	18.2

CUADRO DE FRECUENCIAS

Cuadro 1. Distribución de alumnos desaprobados en un curso en el semestre 89-I.

Curso	Nro. de Alumnos	% de Alumnos
Cálculo II	8	20
Cálculo I	10	25
Estadística	9	22.5
Algebra I	6	15
Otros cursos	7	17.5
Total	40	100

En el caso de que se trate con variables de tipo cualitativo jerárquico, los valores de la variable deben de colocarse ordenadamente de mayor a menor.

REPRESENTACIONES GRAFICAS

Se recomienda la utilización de gráfico de barras o circular.

GRAFICO DE BARRAS

Existen los gráficos de barras horizontales y verticales.

- Se hace uso de los ejes cartesianos
- Las barras son de ancho iguales
- Las barras están igualmente espaciadas
- Pueden representarse tridimensionalmente
- Existen las barras simples y compuestas
- Se grafican los porcentajes

El gráfico de barras horizontales y verticales de los datos del cuadro 1, se muestra a continuación:

```
par(mar=c(4,8,8,4),cex=0.8)

barplot(final$porcentaje,name=row.names(final),col=colors()[21],horiz=T,las=2)
title(main="Graf. 1. Barra Horizontal: Porcentaje\nde desaprobados")

barplot(final$porcentaje,name=row.names(final),col=colors()[45])
title(main="Graf. 2. Barra Vertical: Porcentaje\nde desaprobados")
```

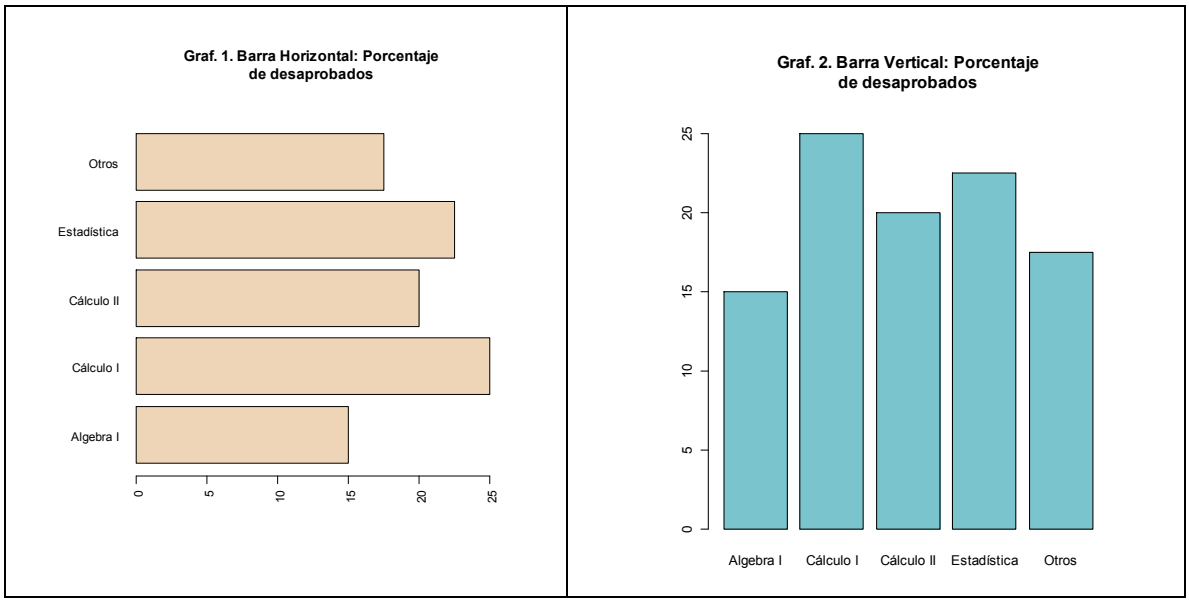
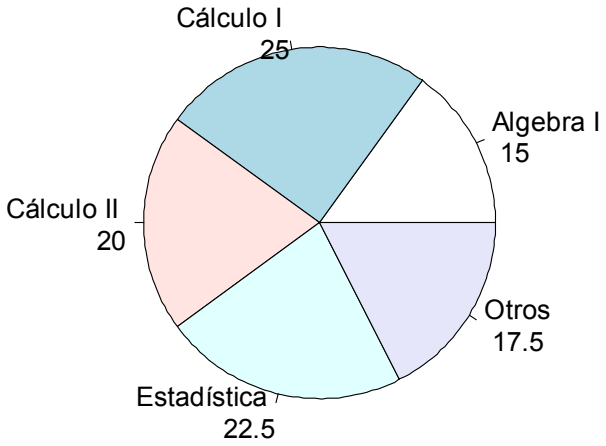


GRAFICO CIRCULAR.- Mediante áreas proporcionales de sectores circulares se representan los porcentajes de cada valor de la variable.

- Tener presente que : °G = % x 3.6

```
par(mar=c(2,3,2,2),cex=1.5)
pie(final$porcentaje,labels=paste(rownames(final),"\n",final$porcentaje))
```



3.2 ORGANIZACION DE LOS DATOS CUANTITATIVOS CONTINUOS

Se debe organizar o clasificar los datos en una TABLA DE DISTRIBUCION DE FRECUENCIAS.

Una tabla de distribución de frecuencias comprende frecuencias absolutas y relativas para intervalos que cubren toda la amplitud de datos.

TABLA DE DISTRIBUCION DE FRECUENCIA (T.D.F)

CONSTRUCCION.- Se ilustrará la construcción de la T.D.F. mediante el siguiente ejemplo:

```
library(agricolae)
data(genxenv)
rdto <- subset(genxenv$YLD,genxenv$ENV==2)
rdto <- round(rdto,1)
```

corresponde al rendimiento de 50 genotipos de papa del Banco de Germoplasma de CIP. Equivalente a tonedadas por Hectarea.

17.2	13.5	17.7	20.1	13.7	16.6	18	18.9	14.4	13.5
19.3	15.4	17.6	21.6	19	12.8	15.5	17.9	19.7	9.9
17.2	15.7	16	18.6	12.1	10.2	18.8	17.8	14.2	13.9
9.9	17.1	18.4	14.2	15.7	13.9	22.8	19.3	13.9	18.8
14.8	20	17.4	13.1	17.2	14.4	17.5	26.1	18.7	17

```
> sturges.freq(rdto)
$maximum
[1] 26.1

$minimum
[1] 9.9

$amplitude
[1] 16.2

$classes
[1] 7

$interval
[1] 2.4

$breaks
[1]  9.9 12.3 14.7 17.1 19.5 21.9 24.3 26.7
```

```
HIST(rdto)
rdto <- round(subset(genxenv$YLD,genxenv$ENV==2),1)
HIST(rdto)
rdto
rdto
history()
```

Para construir una T.D.F. se debe seguir los siguientes pasos:

1. Determinar la amplitud o rango.

$A = X_i$ máx - X_i min

max(rdto) - min(rdto)

A = 26.1 - 9.9 = 16.2

2. Determinar el número de intervalos de clase (k) para la tabla.

- Se recomienda que este comprendido entre 5 y 20
- Existen diversas reglas
- Se recomienda la regla de STURGES

k = 1 + 3.. log n ; n = número de datos.
k = 1 + 3.3* log(50) = 6.6 $\approx$ 7, recomendable 6 (parte entera)

- El tamaño puede ser ajustado despues si se tiene frecuencias bajas (menos del 5%).

Para este caso se utilizara 7 clases

3. Determinar el tamaño de los intervalos de clase (TIC)

TIC = A/k
TIC = 16.2 / 7 = 2.31 $\approx$ 2.4

El redondeo es por exceso y hasta el número de decimales que tienen los datos.

Ejemplo se redondeo por exceso a 1 decimal.

4.51 → 4.6
3.03 → 3.1

4. Determinar los límites de los intervalos de clase. Se debe tener en cuenta los siguientes aspectos:

- Se emplearán límites semiabiertos del tipo [,). La observación menor X_i min = Limite inferior del 1er. intervalo.
- El límite inferior de un intervalo es igual al límite superior del intervalo anterior.
- Los intervalos deben contener a todos los datos.

9.9 12.3 14.7 17.1 19.5 21.9 24.3 26.7

Clase	Inf	Sup
1	9.9	12.3
2	12.3	14.7
3	14.7	17.1
4	17.1	19.5
5	19.5	21.9
6	21.9	24.3
7	24.3	26.7

5. Realizar el conteo que consiste en asignar cada observación al intervalo correspondiente.

Clase	Inf	Sup	Conteo
1	9.9	12.3	\\\\
2	12.3	14.7	\\\\\ \\\\\ \\
3	14.7	17.1	\\\\\ \\\
4	17.1	19.5	\\\\\ \\\\\ \\\\\ \\\\\
5	19.5	21.9	\\\\
6	21.9	24.3	\
7	24.3	26.7	\

6. Determinar las frecuencias absolutas de cada intervalo (f_i) que resulta del consolidado del conteo.

Las frecuencia absoluta del intervalo "i" , se expresa como f_i e indica el número de observaciones que son mayores o iguales que su límite inferior pero menores que su límite superior.

Clase	Inf	Sup	f_i
1	9.9	12.3	4
2	12.3	14.7	12
3	14.7	17.1	8
4	17.1	19.5	20
5	19.5	21.9	4
6	21.9	24.3	1
7	24.3	26.7	1

Se verifica que:

$$\sum_{i=1}^{k} f_i = n$$

$f_4 = 20$; indica que existen 20 genotipos cuyo rendimiento esta entre 17.1 y 19.5 toneladas por hectarea.

7. Determinar las frecuencias relativas de cada intervalo (fr_i). se define como: $fr_i = f_i / n$

La frecuencia relativa del intervalo "i" fr_i, expresado porcentualmente, indica el porcentaje de observaciones que son mayores o iguales que su límite inferior pero menores que su límite superior.

Clase	Inf	Sup	f_i
1	9.9	12.3	0.08
2	12.3	14.7	0.24
3	14.7	17.1	0.16
4	17.1	19.5	0.40
5	19.5	21.9	0.08
6	21.9	24.3	0.02
7	24.3	26.7	0.02

Verifica que:

$$\sum_{i=1}^{k} fr_i = 1$$

$fr_4 = 0.40$, (40%) significa que el 40% de los genotipos tienen rendimientos mayores o iguales a 17.1 ton/ha pero menores que 19.5 ton/ha.

8.- Determinar las frecuencias acumuladas absolutas para cada intervalo (F_i)

Los valores de cada F_i se obtienen mediante la suma de todas las frecuencias absolutas anteriores al intervalo "i" más la frecuencia absoluta del mismo intervalo "i".

$$F_i = \sum_{j=1}^{i} f_j$$

Clase	Inf	Sup	F_i
1	9.9	12.3	4
2	12.3	14.7	16
3	14.7	17.1	24
4	17.1	19.5	44
5	19.5	21.9	48
6	21.9	24.3	49
7	24.3	26.7	50

Se observa que:

$F_1 = f_1$
$F_i = F_{i-1} + f_i$
para i=2,3,..k
$F_k = n$

La frecuencia acumulada absoluta F_i del intervalo "i", indica el número de observaciones que son menores que su límite superior.
F_4 = 44, indica que existen 44 genotipos cuyo rendimiento es menor a 19.5 ton/ha.

9.- Determinar las frecuencias acumuladas relativas de cada intervalo (Fr_i).

Se define como $Fr_i = \sum_{i=j}^{i} fr_j$

La frecuencia acumulada relativa Fr_i del intervalo "i", indica el porcentaje de observaciones con valores menores que su límite superior.

Estas frecuencias cumplen:

$Fr_1 = fr_1$

$Fr_i = Fr_{i-1} + fr_i$ para i=2,3,..k

$Fr_i = F_i/n$

Fr_4 = 44/50 = 0.88 (88%) expresa que el 88% de los genotipos tienen rendimientos inferiores a 19.5 ton/ha

10.- Determinar la marca de clase o punto medio de cada intervalo (X'_i). Se obtiene mediante la semisuma de los límites inferior y superior del intervalo. No se redondea los valores obtenidos.

$$X'_i = \frac{LI_i + LS_i}{2}$$

Se verifica que :

$X'_+ = X'_i + TIC$

$LS_i = X'_i + TIC/2$
$LI_i = X'_i - TIC/2$

El cuadro completo de distribución de frecuencia se muestra en el cuadro 2.

Cuadro 2. Distribucion de frecuencias del rendimiento de los genotipos de papa

Clase	Inf	Sup	MC	fi	fri	Fi	Fri
1	9.9	12.3	11.1	4	0.08	4	0.08
2	12.3	14.7	13.5	12	0.24	16	0.32
3	14.7	17.1	15.9	8	0.16	24	0.48
4	17.1	19.5	18.3	20	0.40	44	0.88
5	19.5	21.9	20.7	4	0.08	48	0.96
6	21.9	24.3	23.1	1	0.02	49	0.98
7	24.3	26.7	25.5	1	0.02	50	1.00

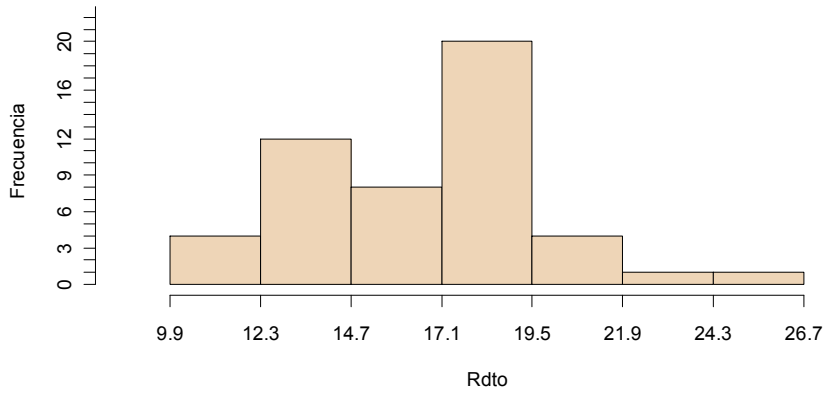
REPRESENTACIONES GRAFICAS

HISTOGRAMA DE FRECUENCIAS

Mediante barras paralelas y adyacentes muestra comparativamente las frecuencias absolutas y relativas.

En el eje X los intervalos de clase y en el eje Y las frecuencias.

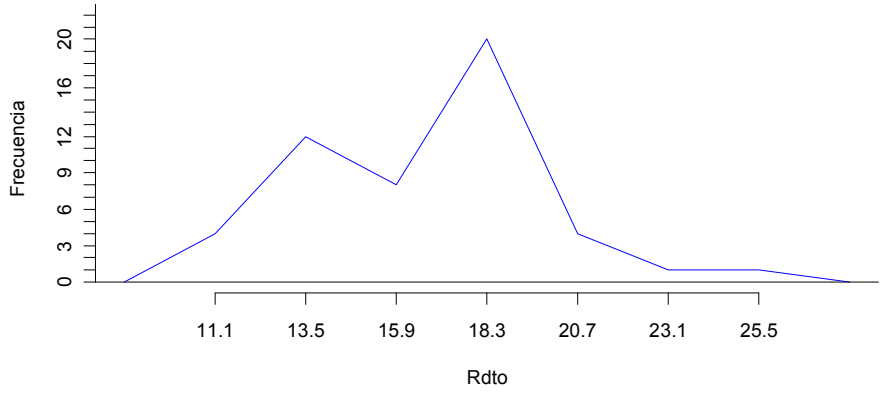
Graf. 4. Distribución de frecuencias absolutas



POLIGONO DE FRECUENCIAS

Muestra la variación de las frecuencias absolutas o relativas al pasar de un intervalo a otro. Puede graficarse simultáneamente con el histograma de frecuencias. En el eje X se ubica la marca de clase y en el eje Y las frecuencias.

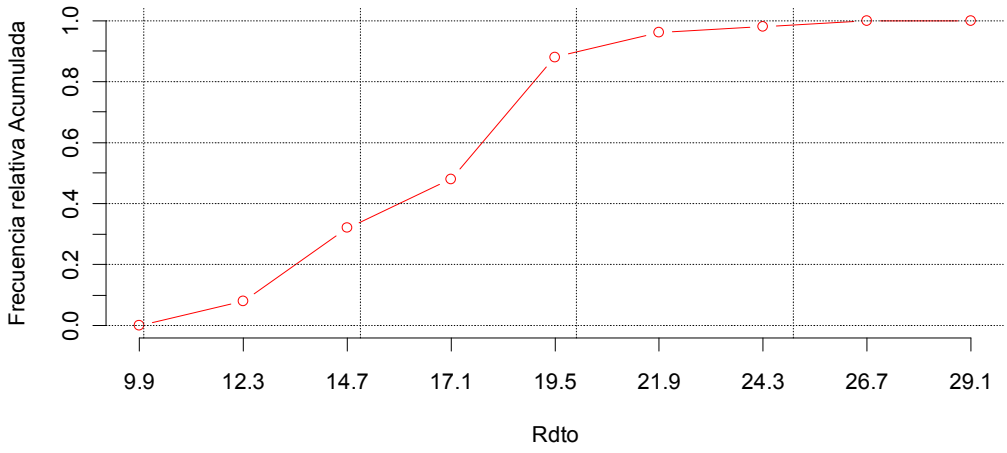
Graf. 5. Poligono de frecuencias



OJIVA

Muestra el comportamiento de las frecuencias acumuladas absolutas o relativas de todos los intervalos de clase. Para el grafico, se utiliza los limites superiores de clase para el eje X y las frecuencias absolutas o relativas acumulativas en el eje Y.

Graf. 6. Ojiva del rendimiento de genotipos de papa



3.3 ORGANIZACION DE DATOS CUANTITATIVOS DISCRETOS

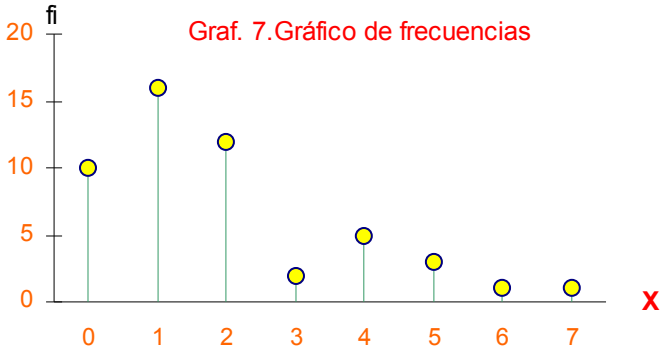
Es muy similar al caso de los datos continuos, en la tabla de distribución de frecuencias se cambia los intervalos de clase por los valores de la variable, la inerpretación se hará teniendo en cuenta esta relación.

Ejemplo

A 50 madres de familia se les preguntó respecto al número de veces semanal que incluyen carne de res en su menú del dia, las respuestas fueron las siguientes:

2, 2, 1, 1, 3, 4, 6, 7, 0, 0, 0, 1, 1, 1, 2, 2, 1, 0, 0, 0, 0, 5, 5, 1, 2, 2, 1, 1, 1, 2, 1, 3, 4, 4, 4, 1, 2, 1, 1, 1, 2, 2, 2, 4, 5, 0, 0, 0, 2

Variable X : Número de veces por semana que se consume carne de res.



Cuadro 3. Tabla de distribución de frecuencias del consumo semanal de carne de res.

X	fi	fri%	Fi	Fri%
0	10	20	10	20
1	16	32	26	52
2	12	24	38	76
3	2	4	40	80
4	5	10	45	90
5	3	6	48	96
6	1	2	49	98
7	1	2	50	100

3.4 INTERPOLACION DE LA OJIVA

La ojiva permite responder directamente preguntas cuando está incluído algún límite superior de los I.C. Si esto no ocurre es necesario realizar la interpolación cuya validéz es correcta bajo el supuesto de que el incremento de las frecuencias a lo largo de cada intervalo permanece constante

Ejemplos (ref cuadro 2)

1. ¿ Cuántos genotipos tienen un rendimiento de por lo menos 19.5 ?

   Resp. 4 + 1+ 1 = 6 genotipos

2. Estimar el porcentaje de genotipos cuyo rendimiento es inferior a 15

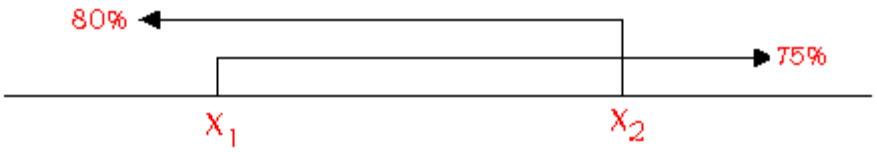
   Hasta 14.7 se tiene 32%
   15 – 14.7 = 0.3

   si a 2.4 le corresponde 16% en la categoria 3
   a 0.3 cuanto le corresponde.....2 = 0.3*16/2.4

   El porcentaje de genotipos con rendimiento inferior a 15 es: 32 + 2 = 34%

3. ¿Entre qué valores debe estar comprendido el rendimiento de un genotipo para estar considerado en el intercepto de los siguientes conjuntos ?.

   A = { EL 75% de los genotipos con mayores rendimientos }
   B = { EL 80% de los genotipos con menores rendimientos }



   Resolver ..

CAPITULO IV

MEDIDAS DE TENDENCIA CENTRAL

Son medidas estadísticas calculadas con la información de una muestra o una población, que tienden a localizar el centro de la distribución de datos. Son valores representativos de un conjunto de datos, pudiendo ser:

- Valores estadísticos, si se calculan con la informacion de la muestra.
- Parametros, si se calculan con la información de una población.

Las principales medidas son: Media o promedio aritmético, mediana, moda. Otras medidas de tendencia central son la media geométrica y media armónica

MEDIDAS DE TENDENCIA CENTRAL PARA DATOS NO AGRUPADOS

MEDIA O PROMEDIO ARITMETICO.- Es el cociente entre la suma de las observaciones de una muestra (X_1, X_2, ...X_n) o de una población (X_1, X_2, ...X_N) y el número de observaciones con que cuenta la muestra (n) o la población (N).

Muestra	Población
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

La diferencia de las observaciones respecto a la media muestral, para cualquier elemento X_i se llama "desviación respecto a la media" y esta dado por:

$(X_i - \bar{X})$

PROPIEDADES:

1. La suma de las desviaciones respecto al promedio es igual a cero.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

2. La suma de cuadrados de las desviaciones respecto al promedio es mínima.

$\sum_{i=1}^{n}(x_i - \bar{x})^2$, es un valor mínimo

Es decir:

$\sum_{i=1}^{n}(x_i-\bar{x})^2 < \sum_{i=1}^{n}(x_i-h)^2$, Para todo "h" diferente del promedio

3. La media o promedio aritmético es un valor típico, si se sustituye cada observación por la media, la suma total de observaciones no varía.

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

$$\bar{x} + \bar{x} + ... + \bar{x} = \sum_{i=1}^{n} \bar{x} = n\bar{x}$$

ambos resultados son iguales.

4. La media esta afectada por valores extremos (pequeños o grandes).

5. Si se tiene dos muestras o subpoblaciones de tamaño n_1 y n_2, entonces la media o promedio del total se calcula por:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n_1+n_2} x_i}{n_1 + n_2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

donde: $n = n_1 + n_2$

Para K constante.

6. Si $Y_i = X_i \pm K$, entonces $\bar{y} = \bar{x} \pm K$

7. Si $Y_i = K X_i$, entonces $\bar{y} = K\,\bar{x}$

8. Si $Y_i = X_i/K$, $K \neq 0$, entonces $\bar{y} = \bar{x}/K$

MEDIA PONDERADA.- Cuando cada observación está asociada a un peso o ponderación, que mide la importancia relativa de dicha observación. Si w_1, w_2, ... w_k son los pesos asociados a los valores de una variable X, de elementos x_1, x_2, ... x_k
entonces la media ponderada esta dado por:

$$\bar{x}_p = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i}$$

MEDIANA (me).- Es una medida de tendencia central y divide al conjunto de observaciones, previamente ordenadas de acuerdo a su magnitud (ascendente o descendente), en dos grupos, de tal modo que el 50% de las observaciones son menores y el otro 50% son mayores que el valor de la mediana.

Para determinar la mediana, se observa si n = número de observaciones es par o impar, según este resultado se tiene:

para n=impar

$m= (n+1)/2 \quad me = X_m$

Para n=par

$m= n/2 \qquad me= (X_m + X_{m+1})/2$

El valor de "m" indica la posición de la mediana.

Propiedades:

1. La suma en valor absoluto de las desviaciones respecto a la mediana es mínima.

$$\sum_{i=1}^{n} |X_i - me|, \text{ es mínima}$$

$$\text{es decir: } \sum_{i=1}^{n} |X_i - me| < \sum_{i=1}^{n} |X_i - h|; \quad \text{para todo } h \neq me$$

2. La mediana no está afectada por los términos extremos (pequeños o grandes).

3. Si $Y_i = X_i \pm K$ (K=constante), entonces : $me_y = me_x \pm K$

4. Si $Y_i = K X_i$ (K=constante), entonces : $me_y = K\, me_x$

5. Si $Y_i = X_i/K$ (K=constante), entonces : $me_y = me_x/K$

MODA (mo).- Es el valor que se presenta con mayor frecuencia en el conjunto de observaciones.

mo = Observación con mayor frecuencia.

Propiedades.

1. Si $Y_i = X_i \pm K$ (K=constante), entonces: $mo_y = mo_x \pm K$

2. Si $Y_i = K X_i$ (K=constante), entonces: $mo_y = K\, mo_x$

3. Si $Y_i = X_i/K$ (K=constante), entonces: $mo_y = mo_x/K$

CALCULO DE MEDIDAS DE TENDENCIA CENTRAL PARA DATOS AGRUPADOS

DATOS DISCRETOS .- Las observaciones son valores discretos, definido en un conjunto finito de elementos, este conjunto define las clases o categorias.

Clase	X_i	f_I	F_i
1	X_1	f_1	F_1
2	X_2	f_2	F_2
.			
.			
K	X_k	f_k	F_I

donde:

f_i: Frecuencia absoluta
de la clase i

F_i: Frecuencia acumulada
absoluta de la clase i

k : número de clases

Tamaño de la muestra o número de observaciones: $n = \sum_{i=1}^{k} f_i$

$$\text{PROMEDIO } \bar{x} = \frac{\sum_{i=1}^{k} X_i f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} X_i f_i}{n}$$

MEDIANA (me) .- Como "n" es el tamaño de la muestra o total de observaciones, primero se determina la clase mediana. La clase mediana (m),

es aquella clase donde en la columna de las F_i acumuló o superó el 50% de los datos.

me = X_m, sólo si n es par y F_m =n/2, entonces me=(X_m+X_{m+1})/2

MODA (mo) Corresponde a la observación de la variable X, con más frecuencia f_i

Si X_4 tiene la mayor frecuencia, es decir f_4 es mayor que cualquier f_i , entonces:

mo = X_4

Ejemplo: En una encuesta de 60 productores de maiz, tomados al azar, se les preguntó por el número de peones contratados para la presente campaña agrícola.

X_i : Número de peones

f_i : número de productores

el cuadro de frecuencias, resumen de la encuesta fue:

Clase	X_i	f_i	F_i
1	0	5	5
2	2	10	15
3	3	15	30
4	5	10	40
5	6	20	60
		60	

n = 60
k = 5

Promedio:

$$\bar{x} = \frac{(0)(5)+(2)(10)+(3)(15)+(5)(10)+(6)(20)}{60} = 3.92$$

Interpretación: el promedio de peones contratados para la presente campaña agrícola es de 3.92

Mediana. El 50% de 60 es 30. según la tabla hasta la clase 3 se tiene 30, esto significa que esta en el límite, es decir la mediana podría ser me = 3 o me = 5, por lo tanto, la mediana será un promedio de ambos:

me = (3+5)/2 = 4

Interpretación: El 50% de los productores contrataron a 4 ó menos peones.

Moda. La clase de mayor frecuencia es la quinta, es decir f_5
con valor 20, entonces la moda corresponde a $X_5 = 6$, la moda será mo = 6

Interpretación: El número de peones contratados más frecuente por los porductores es de 6 peones.

DATOS CONTINUOS

Tabla de distribución de frecuencias

Clase	Intervalos de clase [I.C.]	Marca de clase X'_i	Frec. Abs. f_i	Frec. Relat. fr_i	Frec. Acum. Abs. F_i	Frec. Acum. Relativa Fr_i
1	LI_1- LS_1	X'_1	f_i	fr_1	F_1	Fr_1
2	LI_2- LS_2	X'_2	f_i	fr_2	F_2	Fr_2
.						
.						
.						
K	LI_k- LS_k	X'_k	f_i	fr_k	F_k	Fr_k
			n	1.00		

MEDIA O PROMEDIO

$$\overline{X} = \frac{\sum_{i=1}^{k} x_i' f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i' f_i}{n} = \sum_{i=1}^{k} x_i' fr_i$$

donde : X'_i = Marca de la clase i
f_i = Frecuencia absoluta de la clase i

MEDIANA (me)

$$me = LI_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} TIC$$

donde : i = clase mediana, posición (n+1)/2.
La clase mediana es el intervalo de clase donde en la columna de las F_i acumuló o superó el 50% de los datos.

LI_i = Límite inferior de la clase mediana.
F_{i-1}= Frecuencia acumulada absoluta de la clase anterior a la clase mediana.
f_i = Frecuencia absoluta de la clase mediana.

MODA (mo)

$$mo = LI_i + \frac{d_1}{d_1 + d_2} TIC$$

donde: i = clase modal. La clase modal es identificado por la frecuencia absoluta (f_i) más alta.

LI_i = Límite inferior de la clase modal.
d_1 = Diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta anterior
($f_i - f_{i-1}$).
d_2 = Diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta siguiente
($f_i - f_{i+1}$).

Ejemplo: Considere el cuadro 2 distribución de frecuencia del rendimiento de los genotipos (capitulo II).

La tabla de frecuencia resulta:

Clase	Inf	Sup	MC	fi	fri	Fi	Fri
1	9.9	12.3	11.1	4	0.08	4	0.08
2	12.3	14.7	13.5	12	0.24	16	0.32
3	14.7	17.1	15.9	8	0.16	24	0.48
4	17.1	19.5	18.3	20	0.40	44	0.88
5	19.5	21.9	20.7	4	0.08	48	0.96
6	21.9	24.3	23.1	1	0.02	49	0.98
7	24.3	26.7	25.5	1	0.02	50	1.00

TIC = 2.4, k=7 y n=50

Promedio:

$$\bar{x} = \frac{11.1(4) + 13.5(12) + ... + 25.5(1)}{50} = 16.62$$

también, con las frecuencias relativas:

$$\bar{x} = 11.1(0.08) + ... + 25.5(0.02) = 16.62$$

Interpretación: El rendimiento promedio por genotipo es de 16.62 ton/ha.

MEDIANA: Posición (n+1)/2 = (50+1)/2 = 25.5

Clase mediana : i=4, F_4 =44, dado que F_3 = 24

$$me = 17.1 + \frac{25-24}{20} 2.4 = 17.22$$

Interpretación: El 50% de las personas tienen un ingreso mensual de $101.9 o menos, mientras que el otro 50% tienen ingresos mensuales mayores a $101.9

MODA: La frecuencia más alta es f_4 = 20; la clase modal es i=4.

$d_1 = f_4 - f_3 = 20-8 = 12$
$d_2 = f_4 - f_5 = 20-4 = 16$

$$mo = 17.1 + \frac{12}{12+16} 2.4 = 18.1285$$

Interpretación: El rendimiento mas frecuente de los genotipos es de 18.12 ton/ha. en el intervalo 17.1 y 19.5

En el caso que haya más de una frecuencia modal, se debe calcular una moda por cada frecuencia modal que exista.

MEDIDAS DE TENDENCIA CENTRAL EN DATOS CUALITATIVOS

Cuando el conjunto de observaciones son datos cualitativos no es posible realizar el cálculo de la media y la mediana, sólo es posible hallar la observación con mayor frecuencia, esto es la MODA.

Ejemplo:

Tipo de enfermedad	fr_i	%
Cardivasculares	0.15	15
Gastrointestinales	0.35	35
Vias respiratorias	0.25	25
Otras afecciones	0.25	25
		100

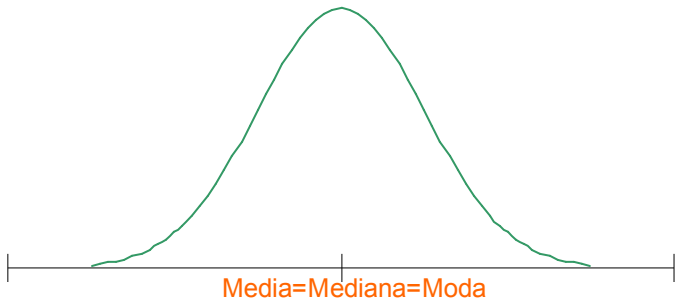
MODA: mo = Gastrointestinales.

CARACTERISTICAS DE ALGUNAS DISTRIBUCIONES

Distribuciones simétricas monomodales

Es una distribución cuyos valores de tendencia central son iguales o aproximadamente iguales. es decir:

Graf. 8. Distribución simétrica

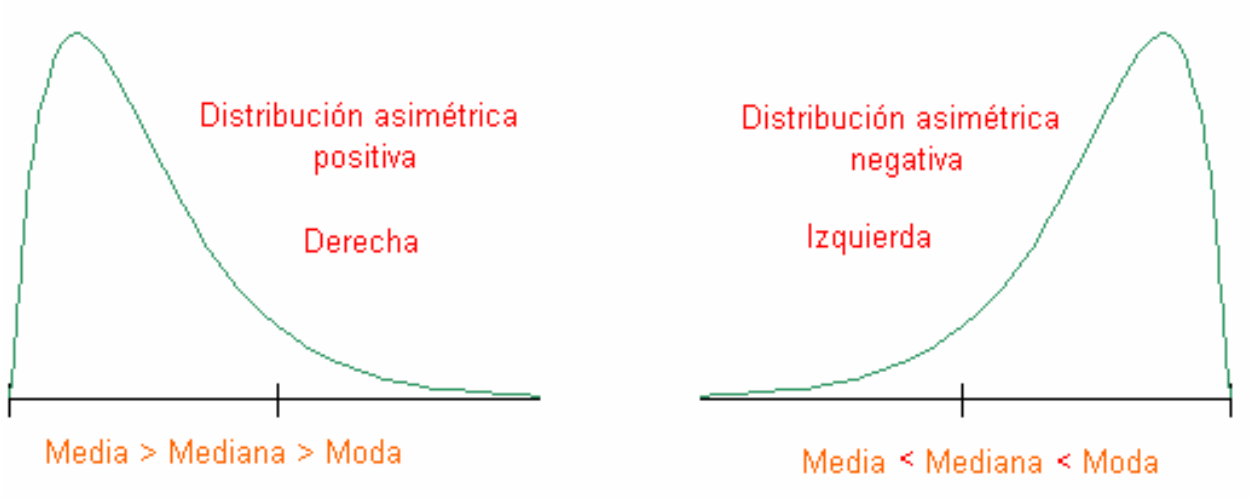


$\mu \approx Me \approx Mo$

μ, Me, Mo son la media, mediana y moda, parámetros de la población que representan las medidas de centralización.

Distribuciones asimétricas monomodales

La asimetría puede ser a la izquierda (negativa) o a la derecha (positiva), como se muestra:



PERCENTILES: Es el punto (Pp) que divide al conjunto de datos en p% menor o igual que el valor percentil y (1-p)% mayores que el valor percentil.

$$Pp = Li_i + \frac{np - F_{i-1}}{f_i} TIC, \qquad 0<p<1$$

donde: i = Clase percentil.
La clase percentil es el intervalo de clase donde se supera por primera vez los (np) datos, $F_i > np$, o también el primer intervalo de clase que satisface $Fr_i > p$

LI_i = Límite inferior de la clase percentil
F_{i-1} = Frecuencia acumulada absoluta de la clase anterior a la clase percentil.
f_i = Frecuencia absoluta de la clase percentil.

Ejemplo: De la tabla de distribución de frecuencias de los rendimiento de genotipos. Hallar $P_{0.25}$

np = (50)(0.25) = 12.5

i = 2; puesto que $F_2 = 16 > np=12.5$; tambien $Fr_2 = 0.32 > 0.25$

F_2 =16 no satisface, en forma equivalente $Fr_1 = 0.08$

Entonces:

$$P_{0.25} = 12.3 + \frac{50(0.25) - 4}{12} 2.4 = 14$$

Interpretación: EL 25% de los genotdipos de la muestra tienen un rendimiento menor o igual a 14 ton/ha, mientras que el 75% tienen un rendimiento mayor de 14 ton/ha.

Para el caso de datos no agrupados:

1. Se ordena los datos
2. Si son n datos, se cacula la posicion posicion = p*(n+1)
3. Si el valor es entero, entonces el dato correspondiente es el percentil.
4. Si el valor es decimal, se toma los dos datos que encierra el percentil, se calcula el equivalente de la parte decimal en el intervalo y se suma al menor de los dos valores.

```
x<- sort(rdto)
n<- length(x)    # n= 50
x
9.9  9.9 10.2 12.1 12.8 13.1 13.5 13.5 13.7 13.9 13.9 13.9 14.2 14.2
14.4 14.4 14.8 15.4 15.5 15.7 15.7 16.0 16.6 17.0 17.1 17.2 17.2
17.2 17.4 17.5 17.6 17.7 17.8 17.9 18.0 18.4 18.6 18.7 18.8 18.8 18.9
19.0 19.3 19.3 19.7 20.0 20.1 21.6 22.8 26.1

Posicion = 0.25(n+1) = 12.75, el valor esta entre 13.9 y 14.2.
Entonces: P0.25 = 13.9 + 0.75(14.2-13.9) = 14.125
```

```
> quantile(rdto,0.25,type=6)
   25%
14.125     ← R reporta igual valor que Minitab y SPSS.
```

Nota: si la posición es entero, entonces el valor que le correspone es el percentil.

Tallos y Hojas

Es una forma de visualizar la distribución de los datos, cuando estos no son muchos y puede realizarse manualmente. En el programa R se tiene la función stem para este proceso.

Considere el siguiente objeto:

```
> x<- c(58,4,73,68,82,60,1,69,60,36,15,6,63,56,86,62,68,48,3,89)
> stem(x,scale=2)

  The decimal point is 1 digit(s) to the right of the |

  0 | 1346
  1 | 5
  2 |
  3 | 6
  4 | 8
  5 | 68
  6 | 0023889
  7 | 3
  8 | 269
```

Para descifrar este resultado, ordene los valores de “x”

```
> sort(x)
 [1]  1  3  4  6 15 36 48 56 58 60 60 62 63 68 68 69 73 82 86 89
```

Se puede observar grupos de valores:

```
grupo 0 : 1, 3, 4, 6
grupo 1 : 15
grupo 3 : 36
grupo 4 : 48
...
grupo 8 : 82, 86, 89
```

Mediante SAS.

```
Stem Leaf                     #  Boxplot
   8 269                      3     |
   7 3                        1     |
   6 0023889                  7  +-----+
   5 68                       2  |  +  |
   4 8                        1  |     |
   3 6                        1  |     |
   2                             +-----+
   1 5                        1     |
   0 1346                     4     |
```

Aplicar al caso del rendimiento de genotipos:

```
> sort(rdto)
9.9  9.9 10.2 12.1 12.8 13.1 13.5 13.5 13.7 13.9 13.9 13.9
14.2 14.2 14.4 14.4 14.8 15.4 15.5 15.7 15.7 16.0 16.6 17.0
17.1 17.2 17.2 17.2 17.4 17.5 17.6 17.7 17.8 17.9 18.0 18.4
18.6 18.7 18.8 18.8 18.9 19.0 19.3 19.3 19.7 20.0 20.1 21.6
22.8 26.1

> stem(rdto,scale=2)

  The decimal point is at the |

   9 | 99
  10 | 2
  11 |
  12 | 18
  13 | 1557999
  14 | 22448
  15 | 4577
  16 | 06
  17 | 01222456789
  18 | 0467889
  19 | 0337
  20 | 01
  21 | 6
  22 | 8
  23 |
  24 |
  25 |
  26 | 1
```

En el grupo de 9 tiene dos valores, sus decimales son 9 y 9
En el grupo 10 un solo valor y es 2
...
Grupo 26 tiene un solo valor y es 1.

Boxplot, diagrama de caja o caja de Tukey

Este ha sido un aporte fundamental realizado por Tukey (1977). Es un gráfico simple, ya que se realiza básicamente con cinco números, pero poderoso. Se observa de una forma clara la distribución de los datos y sus principales características. Permite compara diversos conjuntos de datos simultáneamente. Como herramienta visual se puede utilizar para ilustrar los datos, para estudiar simetría, para estudiar las colas, y supuestos sobre la distribución, también se puede usar para comparar diferentes poblaciones. Este gráfico contiene un rectángulo, usualmente orientado con el sistema de coordenadas tal que el eje vertical tiene la misma escala del conjunto de datos. La parte superior y la inferior del rectángulo coinciden con el tercer y primer cuartil de los datos. Esta caja se divide con una línea horizontal a nivel de la mediana. Se define un "paso" como 1.5 veces el rango intercuartil, y una línea vertical (un bigote) se extiende desde la mitad de la parte superior de la caja hasta la mayor observación de los datos si se encuentran dentro de un paso. Igual se hace en la parte inferior de la caja. Las observaciones que caigan mas allá de estas líneas son mostradas individualmente como valores extremos.

La definición de los cuartiles puede variar y otras definiciones del paso son planteadas por otros autores (Frigge et al., 1989).

Propiedades del grafico de caja

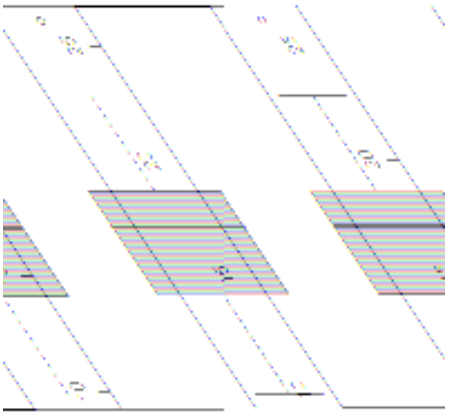
1. Cinco números de resumen de los datos son representados gráficamente de tal forma que proporciona información acerca de la localización, la dispersión, el sesgo y las colas del conjunto de datos que se aprecia de una sola mirada. La localización está representada en la línea que corta la caja y representa la mediana (que está dentro de la caja), la dispersión está dada por la altura de la caja, como por la distancia entre los extremos de los bigotes. El sesgo se observa en la desviación que exista entre la línea de la mediana con relación al centro de la caja, y también la relación entre las longitudes de los bigotes. Las colas se pueden apreciar por la longitud de los bigotes con relación a la altura de la caja, y también por las observaciones que se marcan explícitamente.
2. El gráfico de caja contiene información detallada sobre las observaciones de las colas.
3. La grafica de caja es fácil de calcular, dibujar e interpretar.

Existen muchas variaciones de este grafico, las cuales tratan de involucrar otras características de los datos que en un momento dado puedan ser de interés para el investigador, por ejemplo, a veces se utilizan muescas en la caja para comparar la localización de diferentes muestras y ver si la diferencia es significativa desde el punto de vista estadístico. Otros ponen una marquilla para ubicar la media aritmética, otros deforman la caja para obtener más claridad

acerca de la distribución, por ejemplo Benjamini, (1988) crea el grafico “vaso”, en el cual se involucran conceptos de estimación de densidades. Zani, Riani y Corbellini (1998) presentan una generalización del gráfico de caja a dos dimensiones.

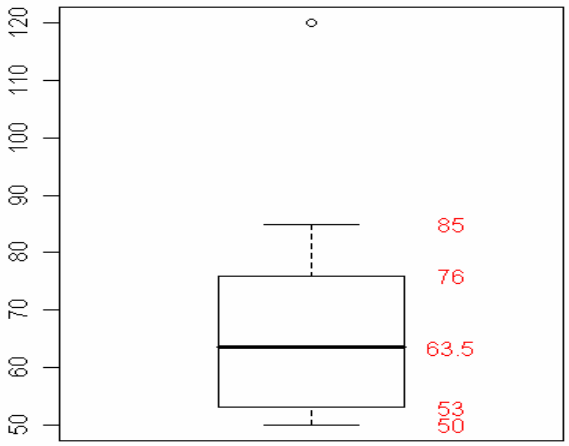
```
Con R. Para rendimientos de los
genotipos de papa (ton/ha.)

> G<-boxplot(rdto,col="yellow")
> G
$stats
      [,1]
[1,]  9.90
[2,] 14.20
[3,] 17.15
[4,] 18.70
[5,] 22.80
```



Para el caso de los pesos: 50, 52, 53, 54, 63, 64, 75, 76, 85, 120

```
Con R. Para Pesos
> boxplot(pesos)
> quantile(pesos, 0.25, type=1)
53
> quantile(pesos, 0.5, type=1)
63.5
> quantile(pesos, 0.75, type=1)
76
ric = 76-53 = 23
ISI = 76+1.5*23 = 110.5
ISS= 53-1.5*23 = 18.5
Valor mas alto < 110.5 es 85
Valor mas bajo > 18.5  es 50
```




```
> stem(pesos,scale=2)
  The decimal point is 1 digit(s) to the right of the |
   5 | 0234
   6 | 34
   7 | 56
   8 | 5
   9 |
  10 |
  11 |
  12 | 0
```

CAPITULO V

MEDIDAS DE VARIABILIDAD

Son medidas estadísticas que permiten conocer el grado de homogeneidad o heterogeneidad de un conjunto de datos, evaluando la dispersión que presentan entre ellos. Estas medidas son:

Medidas de variabilidad absoluta.- Aquellas que presentan unidades de medida:

Rango: R r
Variancia: σ^2 S^2
Desviación estándar: σ S

Medidas de variabilidad relativa.- Aquellas que no presentan unidades de medida.

Coeficiente de variabilidad CV cv

RANGO.- Es la diferencia entre la observación de mayor y menor valor.

RANGO = Observación mayor - Observación menor

VARIANCIA.- Es una medida de dispersión absoluta de las observaciones, esta dada por la suma de las diferencias cuadráticas de las observaciones respecto a su promedio, y dividido por el total de observaciones.

Variancia muestral S^2:

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

Variancia poblacional σ^2:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} X_i^2 - N\mu^2}{N} = \frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2$$

DESVIACION ESTANDAR.- Es la raíz cuadrada de la variancia.

La variancia y desviación estándar se utilizan para comparar dos conjuntos de datos expresados en las mismas unidades y cuyos valores medios sean aproximadamente iguales.

Ejemplo: Se desea comparar los ingresos mensuales del año 1989 de dos empresas.

Empresa A: $\mu_A = 28{,}000$ $\sigma^2_A = 2500$

Empresa B: $\mu_B = 30{,}000$ $\sigma^2_B = 5000$

Entonces de puede afirmar que los ingresos mensuales del año 1989, han sido más variables para la empresa B que los de la empresa A ($\sigma^2_A < \sigma^2_B$)

COEFICIENTE DE VARIABILIDAD.- Es una medida de variabilidad que no presenta unidades y que expresa el número de veces que la desviación estandar contiene a la media. Esta medida estadística se utiliza para comparar conjuntos de datos que tienen diferentes unidades o cuyos valores medios son muy diferentes.

Muestral: $CV = \frac{S}{\bar{x}}100\%$ Poblacional: $CV = \frac{\sigma}{\mu}100\%$

Estos valores se expresan en porcentaje.

CALCULOS DE MEDIDAS DE VARIABILIDAD PARA DATOS AGRUPADOS

RANGO: $R \approx LS_k - LI_1$

Variancia Muestral:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i' - \bar{x})^2 f_i}{n-1} = \frac{\sum_{i=1}^{N} X_i'^2 f_i - n\bar{x}^2}{n-1}$$

X'_i es la marca de clase, f_i la frecuencia absoluta

Variancia poblacional σ^2:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i' - \mu)^2 f_i}{N} = \frac{\sum_{i=1}^{N} X_i'^2 f_i - N\mu^2}{N} = \frac{\sum_{i=1}^{N} X_i'^2 f_i}{N} - \mu^2$$

DESVIACION ESTANDAR:

Muestral: $S = \sqrt{S^2}$, Poblacional: $\sigma = \sqrt{\sigma^2}$

Ejemplo: considerando la tabla de distribución de frecuencias de ingresos mensuales.

Clase	Inf	Sup	X'_i	f_i	$X'_i f_i$	$X_i'^2 f_i$
1	9.9	12.3	11.1	4	44.4	492.84
2	12.3	14.7	13.5	12	162	2187
3	14.7	17.1	15.9	8	127.2	2022.5
4	17.1	19.5	18.3	20	366	6697.8
5	19.5	21.9	20.7	4	82.8	1714
6	21.9	24.3	23.1	1	23.1	533.61
	24.3	26.7	25.5	1	25.5	650.25
					831.0	14298.94

Rango: r = 26.1 – 9.9 = 16.2

$$Variancia: \quad S^2 = \frac{14297.94 - \frac{831^2}{50}}{49} = 9.933$$

Desviación estándar: S = 3.15

Interpretación: Los rendimientos una dispersion respecto de su promedio (16.62) de 3.15 ton/ha.

Coeficiente de variación: cv = 3.15/16.62 = 0.1895

Interpretación: Los rendimientos de los genotipos presentan una variabilidad relativa de 18.95%

COEFICIENTE DE ASIMETRIA.- Son medidas que indican la existencia o no de valores extremos (superior o inferior) que presenta una distribución de datos.

Coeficiente de Asimetría de Pearson.- Determina la asimetria de la distribución de los datos:

$$Skp = \frac{3(\overline{X} - me)}{S}$$

Teóricamente Skp varía de -3 a +3, comunmente los valores de Skp fluctuan entre -1 a +1.

Skp cercano o igual a cero, la distribución se considera simétrica.

A medida que se va alejando del valor cero, la distribución va siendo mas asimétrica, asi:

Skp cercano a +1, la distribución es asimétrica hacia la derecha.

Skp cercano a -1, la distribución es asimétrica hacia la izquierda.

Ejemplo: Considerando la distribucion de frecuencia de los rendimientos de los genotipos tratado en cada caso, se tiene:

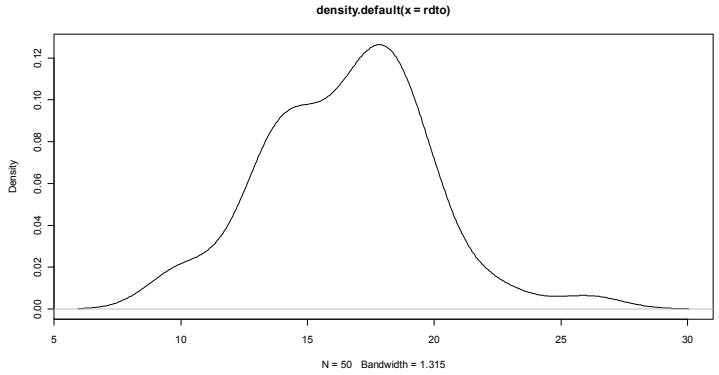
$$S_{kp} = \frac{3(16.62 - 17.22)}{3.15} = -0.5714286$$

Interpretación: La distribución de los rendimientos es ligeramente asimetría hacia la izquierda, pero esta se debe considera simetrica.

En R, se puede calcular con los datos sin agrupar, la funcion esta en agricolae.

library(agricolae)
skewness(rdto)
0.18 ← Este valor es igual al calculado por Minitab. SPSS y SAS

plot(density(rdto)



Rango Intercuatil. Es una medida de variacion que excluye todo valor extremo hasta un 25% superior e inferior.

RIC = $P_{0.75} - P_{0.25}$

En el caso del rendimiento con R:

```
quantile(rdto,0.75,type=6)-quantile(rdto,0.25,type=6)
18.725-14.125 = 4.6
```