3. ESTADÍSTICA DESCRIPTIVA

La estadística descriptiva es una parte de la estadística que se dedica al ordenamiento y tratamiento de la información para su presentación por medio de tablas y de representaciones gráficas, así como a la obtención de algunos parámetros útiles para la explicación de la información. En este contexto, R brinda muchas alternativas para calcular medidas descriptivas de una población o muestra; a continuación se presentan algunas de estas.

3.1 TABLA DE DISTRIBUCIÓN DE FRECUENCIAS

Se entiende como el agrupamiento de datos en categorías, el cual muestra el número de observaciones en cada categoría mutuamente excluyente. Cada una de estas categorías es llamada intervalo de clase; los intervalos de clase usados en la distribución de frecuencias deben ser iguales. Para determinar la amplitud de un intervalo de clase se utiliza la fórmula int=(valor más alto – valor más bajo)/número de clases. En R es posible construir la tabla de distribuciones de frecuencias de la siguiente manera:

```
>
> y=c(15,23.7,19.7,15.4,18.3,23,14.2,20.8,13.5,20.7,17.4,18.6,12.9,20.3,13.7)
>
> range(y) # valor mínimo y valor máximo
[1] 12.9 23.7
>
> rango=23.7-12.9
> rango
[1] 10.8
>
> n=6 # número de clases
>
> amplitud=rango/n # amplitud de clase
> amplitud
[1] 1.8
>
> table(cut(y,breaks=seq(12,24,2),rigth=TRUE))

(12,14] (14,16] (16,18] (18,20] (20,22] (22,24]
      3       3       1       3       3       2
>
```

Imagen 41. Salida R para creación de tabla de frecuencias

En el ejemplo anterior se hace necesario calcular primero la amplitud de clase, y de acuerdo con esta se puede llegar a modificar el inicio de la primera clase y el final de la última clase, mediante la instrucción **breaks**; la instrucción **rigth=TRUE** indica que el intervalo es cerrado a la derecha. Si se necesita determinar la tabla de frecuencias acumuladas, el comando **table** se escribe dentro del comando **cumsum()**, como se muestra a continuación.

```
>
> table(cut(y,breaks=seq(12,24,2),rigth=TRUE))

(12,14] (14,16] (16,18] (18,20] (20,22] (22,24]
      3       3       1       3       3       2
>
> cumsum(table(cut(y,breaks=seq(12,24,2),rigth=TRUE)))

(12,14] (14,16] (16,18] (18,20] (20,22] (22,24]
      3       6       7      10      13      15
```

Imagen 42. Salida R para creación de tabla de frecuencias acumuladas

Si el interés está en ver las tablas de frecuencia relativa y frecuencia relativa acumulada, entonces las tablas obtenidas con los comandos **table y cumsum** se dividen entre el tamaño del vector de datos; a continuación un ejemplo:

```
>
> table(cut(y,breaks=seq(12,24,2),rigth=TRUE)) # tabla de frecuencias

(12,14] (14,16] (16,18] (18,20] (20,22] (22,24]
      3       3       1       3       3       2
>
> table(cut(y,breaks=seq(12,24,2),rigth=TRUE))/length(y) # Frec Relativa

   (12,14]    (14,16]    (16,18]    (18,20]    (20,22]    (22,24]
0.20000000 0.20000000 0.06666667 0.20000000 0.20000000 0.13333333

>
> cumsum(table(cut(y,breaks=seq(12,24,2),rigth=TRUE))/length(y))# F R A
  (12,14]   (14,16]   (16,18]   (18,20]   (20,22]   (22,24]
0.2000000 0.4000000 0.4666667 0.6666667 0.8666667 1.0000000
>
```

Imagen 43. Salida R para creación de tabla de frecuencias relativas

3.2 FUNCIONES GRÁFICAS BÁSICAS PARA EL ANÁLISIS EXPLORATORIO DE DATOS

Los métodos gráficos proporcionan al investigador un conjunto de formas sencillas para examinar tanto las variables de manera individual como las relaciones entre ellas. Los métodos gráficos se distinguen según la cantidad de variables que se analizan; a continuación se presentan algunos gráficos para el análisis exploratorio de datos y su correspondiente algoritmo para ser trabajado en R.

3.2.1 Diagrama de sectores. Utilizado para variables de tipo cualitativas (también llamado circular). Se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa.

Ejemplo: En una encuesta realizada a 10 personas sobre la preferencia que tienen hacia algún deporte, los resultados son los siguientes: 4 optaron por fútbol; 2, por básquet; 3, por voleibol, y 1, por tenis; de acuerdo con los datos, la fracción del diagrama que le corresponde a cada deporte es: fútbol = 0.4, básquet = 0.3, voleibol = 0.3 y tenis = 0.1; con esta información se construyen los vectores necesarios para realizar el diagrama en R.

> X = c(0.4,0.3,0.3,0.1)
> **nombres** = c("Fútbol - 40%","Básquet - 30%","Voleibol - 30%","Tenis - 10%")

El primer vector se refiere a la fracción del diagrama por deporte, y el segundo vector contiene el nombre del deporte y su respectivo porcentaje; con esto se procede a realizar el gráfico en R así:

```
>
>  X=c(0.4,0.3,0.3,0.1)
> nombres=c("Futbol - 40%","Basket - 30%","Voleibol - 30%","tenis - 10%")
>  pie(X,  labels=nombres)
>
```

Imagen 44. Salida R para la creación del diagrama de sectores

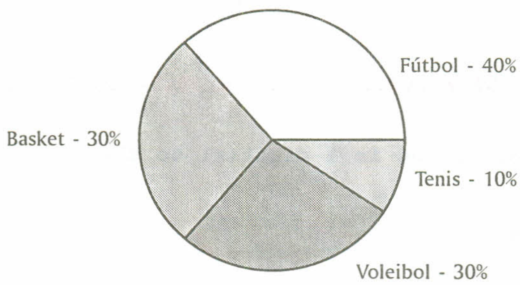

Imagen 45. Salida R diagrama de sectores

3.2.2 Diagrama de barras. Se utiliza para representar los caracteres cualitativos y cuantitativos discretos. En el eje horizontal, o eje de abscisas, se representan los datos o modalidades; en el eje vertical, o de ordenadas, se representan las frecuencias de cada dato o modalidad. Las frecuencias pueden ser absolutas, relativas y relativas acumuladas. Teniendo en cuenta los mismos datos del ejemplo de preferencias de deporte utilizado en el diagrama de sectores, el siguiente comando permite realizar el diagrama de barras en R:

barplot(X, names.arg = nombres)

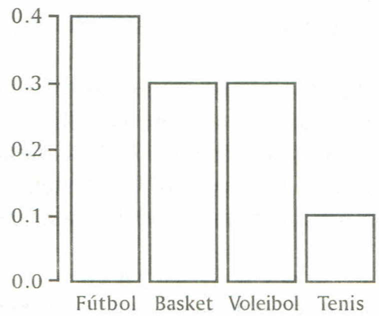

Imagen 46. Salida R diagrama de barras

X se refiere a la fracción de cada deporte, y **names.arg**, al nombre de la fracción en el vector del ejemplo anterior.

3.2.3 Diagrama de tallos y hojas. Utilizado para variables de tipo numérico; uno de los objetivos es descubrir un patrón de comportamiento de los datos, es decir, qué distribución de probabilidad pueden seguir los datos. Es aplicable para valores formados por al menos dos cifras, y su principio es que cada número se divide en dos partes, una llamada "Tallo", y la otra, "ramas u hojas". En R, el comando que permite realizar este diagrama es **stem(vector)**.

Ejemplo: Considerar los números 65, 57, 79, 69, 53, 63, 71, 81, 64, 85, 72, 59, 90, 51, 68. Los tallos serán las decenas, y las ramas serán las unidades; la instrucción en R es:

```
>
> T=c(65,57,79,69,53,71,81,64,85,72,59,90,51,68)
> stem(T)

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

5 | 1379
6 | 34589
7 | 129
8 | 15
9 | 0
```

Imagen 47. Salida R diagrama de tallos y hojas

3.2.4 Diagrama de Caja. Utilizado para variables de tipo numérico. Es un gráfico representativo de las distribuciones de un conjunto de datos en cuya construcción se usan cinco medidas descriptivas de estos, a saber: mediana, primer cuartil, tercer cuartil, valor máximo y valor mínimo; permite identificar de forma individual observaciones que se alejan del resto de los datos; a estas observaciones se les conoce como valores atípicos. El comando que permite realizar este gráfico es **boxplot(vector)**.

Se tiene por ejemplo un vector G cuyos datos son: >G=c(29, 78, 48, 29, 30, 44, 72, 73, 46, 82, 84, 71, 75, 84, 45, 45, 47, 35, 33, 54, 56, 33, 62, 63, 64, 36)

>boxplot(G)

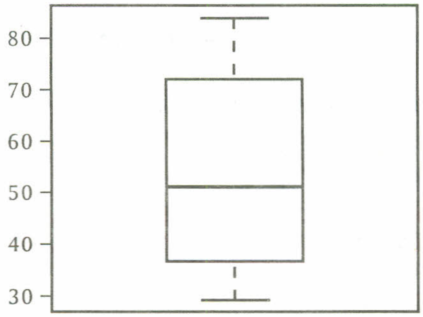

Imagen 48. Salida R diagrama de caja

3.2.5 Histograma. Es el gráfico estadístico que se utiliza para representar datos continuos cuando vienen agrupados en intervalos. Sobre cada uno de estos intervalos se levanta una franja tan ancha como el intervalo y de forma que su área sea proporcional a su frecuencia. El comando que permite realizar un histograma en R es **hist(vector)**.

Con los datos del diagrama de caja, el histograma correspondiente se obtiene así:

>hist(G)

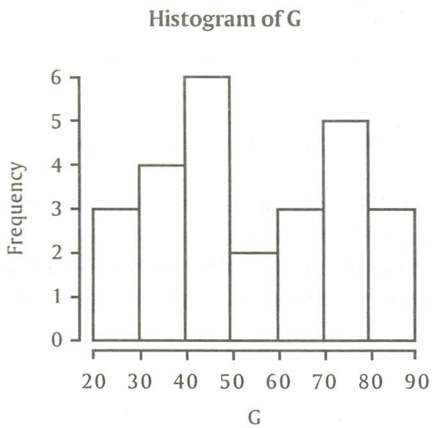

Imagen 49. Salida R histograma

3.2.6 Dispersograma. Gráfico bidimensional usado para variables cuantitativas. Consiste en dos ejes perpendiculares; en cada uno de ellos se ubican los valores de cada una de las variables.

Ejemplo: Los siguientes datos representan las calificaciones de matemáticas para una muestra aleatoria de 12 alumnos de primer grado de cierta universidad, junto con sus calificaciones de una prueba de inteligencia que se les aplicó cuando aún eran alumnos del último año de bachillerato:

Calificación prueba de inteligencia (x) = 65, 50, 55, 65, 55, 70, 65, 70, 55, 70, 50, 55.

Calificación prueba de matemáticas (y) = 85, 74, 76, 90, 85, 87, 94, 98, 81, 91, 76, 74.

A continuación la sintaxis para realizar el diagrama de dispersión de estos conjuntos de datos y su respectivo gráfico.

```
>
> X=c(65,50,55,65,55,70,65,70,55,70,50,55)
> Y=c(85,74,76,90,85,87,94,98,81,91,76,74)
>
> plot(X,Y)
```

Imagen 50. Salida R creación de dispersograma

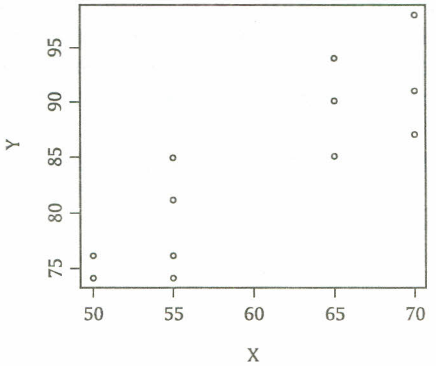

Imagen 51. Salida R dispersograma

Las funciones gráficas tienen posibilidades para ser modificadas, hasta ahora solo se ha definido el comando básico para realizar el gráfico con una instrucción. Para incluir opciones en las gráficas, estas se escriben después del nombre del conjunto de datos; la estructura por utilizar es **función-gráfica(datos, opción = parámetro)**; la tabla siguiente muestra algunas de estas opciones:

Tabla 3. Opciones gráficas

Opción	Descripción
main = "título"	Título principal; debe ser de tipo carácter
sub ="subtítulo"	Subtítulo (escrito en letra más pequeña)
xlab=", ylab="	Títulos en los ejes: deben ser variables de tipo carácter
xlim=, ylim=	Especifica los límites inferiores y superiores de los ejes
axes = TRUE	Si es FALSE no dibuja los ejes ni la caja del gráfico
col = "color"	Le da un color específico a los puntos o a las líneas

El tipo de gráfico también es posible determinarlo para los dispersogramas, de la siguiente forma:

Tabla 4. Opciones gráficas para el dispersograma

Opción	Descripción
type = "p"	Puntos
type = "*l*"	Líneas
type = "b"	Puntos conectados por líneas
type = "o"	Igual al anterior, pero las líneas están sobre los puntos
type = "h"	Líneas verticales
type = "s"	Escaleras, los datos se representan como la parte superior de las líneas verticales
type = "S"	Escaleras, los datos se representan como la parte inferior de las líneas verticales

También es posible cambiar los puntos por otra figura (como una de las que se aprecian en la imagen 52); para esto se introduce la opción **pch** en el comando **plot**(), cada figura aparecerá de acuerdo con el número que se especifique después del comando **pch=número**. Las figuras y su correspondiente número son:

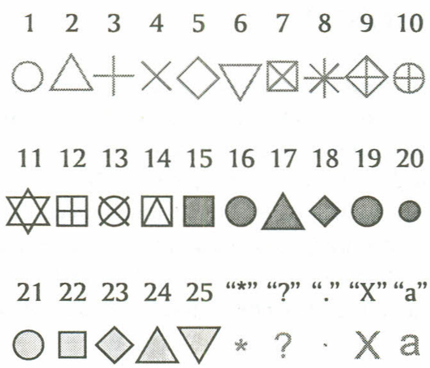

Imagen 52. Opciones para el argumento pch

Con el fin de visualizar algunas de las figuras mencionadas, se presenta un diagrama de dispersión empleando la figura de diamante.

Ejemplo: El supervisor de mantenimiento de una línea de autobuses cree que existe una relación entre el costo anual de mantenimiento de las unidades y los años que llevan de operación. Considera que si tal relación existe podrá hacer un mejor pronóstico de presupuesto. Los datos tomados por el supervisor sobre 15 autobuses de la empresa se muestran a continuación:

x = c(8, 5, 3, 9, 11, 2, 1, 8, 12, 4, 7, 10, 6, 3, 9)
y = c(8.6, 6.8, 4.7, 7, 11, 2.2, 3.2, 6.5, 10.5, 5.6, 6.8, 10.8, 6.2, 5, 8)

```
> plot (x,y,main="Diagrama de dispersión",
+ xlab="Tiempo de operación",ylab="Costo de mantenimiento",
+ col="red",col.main="blue",pch=18)
```

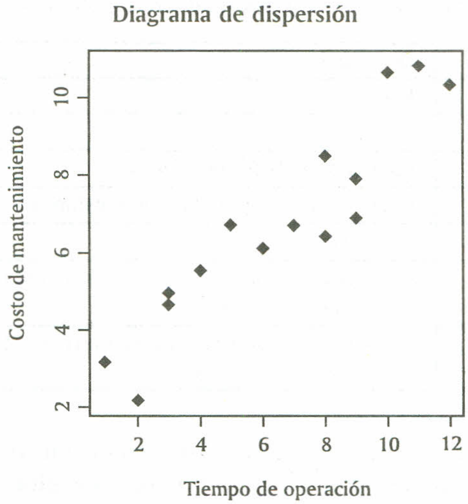

Imagen 53. Salida R argumentos gráficos

3.3 MEDIDAS DE TENDENCIA CENTRAL

Al describir grupos de observaciones, con frecuencia se desea describir el grupo con un solo número; desde luego, no se usará el valor más elevado ni el valor más pequeño como único representante, ya que solo representan los extremos y no los valores que generalmente tienen mayor ocurrencia en una población; sería más adecuado buscar un valor central. Las medidas que describen un valor típico en un grupo de observaciones suelen llamarse medidas de tendencia central. Es importante tener en cuenta que estas medidas se aplican a grupos más que a individuos. Un promedio es una característica de grupo, no individual. En los ejemplos para las medidas de tendencia central se utilizará el siguiente conjunto de datos localizados en el vector W:

```
>
> W=c(24,30,32,35,48,16,14,15,21,32,30,25,26,25,30,16,15,17,21,20,30)
>
```

Imagen 54. Salida R vector de trabajo

3.3.1 Media aritmética. La medida de tendencia central más utilizada que se puede elegir es el valor obtenido sumando las observaciones y dividiendo esta suma por el número de observaciones que hay en el grupo; en R, el comando que calcula directamente la media aritmética es: **mean(vector)**. La media aritmética en R se calcula como se muestra a continuación:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$	`>`
	`> mean(W)`
	`[1] 24.85714`
	`>`
	`> media=sum(W)/length(W)`
	`> media`
	`[1] 24.85714`
	`>`

Ecuación 1 e imagen 55. Salida R para la media aritmética

3.3.2 Mediana. Definida como el valor de la variable que deja el mismo número de datos antes y después que él.

```
>
>
median(W)
[1] 25
>
```

Imagen 56. Salida R para la mediana

3.3.3 Moda. Es el dato que más se repite en un conjunto de datos. Si existen dos datos que se repiten un número igual de veces, entonces, el conjunto será bimodal. Al utilizar el comando **table(vector)** se construye una tabla donde se aprecian cada uno de los valores diferentes de la variable y su frecuencia. Considere el vector "y":

```
>
>
y=c(1,5,6,8,2,4,1,2,2,3,2,6,2)
> table(y)
y
1 2 3 4 5 6 8
2 5 1 1 1 2 1
>
```

Imagen 57. Salida R para la moda

En los datos anteriores se observa que el valor que más se repite es el 2, pues su frecuencia es 5; por lo tanto, este valor es la moda para este conjunto de datos.

3.4 MEDIDAS DE DISPERSIÓN

Se llaman medidas de dispersión a aquellas que permiten expresar la distancia de los valores de la variable a un cierto valor central, o que permiten identificar la concentración de los datos en un cierto sector del recorrido de la variable. Se trata del coeficiente para variables cuantitativas.

3.4.1 Varianza. Es el valor obtenido de sumar los cuadrados de las desviaciones de cada uno de los datos respecto a la media y dividir esta suma por el número de observaciones menos uno; en

R, el comando que calcula directamente la varianza es **var(vector)**, o bien, se puede programar así, teniendo en cuenta el vector de datos **G** dado en los ejemplos sobre tipos de gráficas, se tiene:

$Var(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$	`>`
	`> Var=sum(c(G-mean(G))^2)/(length(G)-1)`
	`> var`
	`[1] 344.8185`
	`>`
	`> var(G)`
	`[1] 344.8185`
	`> .`

Ecuación 2 e imagen 58. Salida R para la varianza

3.4.2 Desviación estándar. Es la raíz cuadrada de la varianza; el comando **sd(G)** hace el cálculo directo de la desviación estándar sobre un vector numérico.

$desvest = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$	`>`
	`> sqrt(var(G))`
	`[1] 18.56929`
	`>`
	`> sd(G)`
	`[1] 18.56929`
	`>`

Ecuación 3 e imagen 59. Salida R para la desviación estándar

3.4.3 Cuantiles. Se usan con frecuencia en los datos para dividir las poblaciones en grupos. Por ejemplo, se puede utilizar el primer cuantil para determinar cuál valor deja un 25 por ciento de datos por debajo de él, esto se observa a continuación:

```
>
> quantile(G)
   0%   25%   50%   75%  100%
29.00 38.00 51.00 71.75 84.00
>
```

Imagen 60. Salida R para los cuantiles

Si el interés recae en calcular los percentiles de la variable, se utiliza la función

quantile(variable,seq(valor inicial,valor final, incremento))

Luego los percentiles para el vector G son:

```
>
> quantile(G,seq(0.1,0.9,0.1))
  10%   20%   30%   40%   50%   60%   70%   80%   90% 
 31.5  35.0  44.5  46.0  51.0  62.0  67.5  73.0  80.0 
>
```

Imagen 61. Salida R para los percentiles

El comando **summary** permite calcular directamente y a la vez algunas de las medidas de tendencia central y de dispersión como: media, mediana, primer cuartil, tercer cuartil, valor mínimo y valor máximo de un conjunto de datos. Así como se muestra:

```
>
> x = c(8,5,3,9,11,2,1,8,12,4,7,10,6,3,9)
>
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.500   7.000 6.533   9.000 12.000 
>
```

Imagen 62. Salida R para obtener resumen general

3.5 MEDIDAS DE ASIMETRÍA

Comparan la forma que tiene la representación gráfica, bien sea el histograma o el diagrama de barras de la distribución, con la distribución normal.

3.5.1 Sesgo. Diremos que una distribución es simétrica cuando su mediana, su moda y su media aritmética coinciden. El sesgo mide la simetría de la distribución de un conjunto de datos; puede ser negativo, cero o positivo. Una fórmula para calcular este coeficiente de simetría es la siguiente:

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

Ecuación 4. Sesgo

El algoritmo de la fórmula anterior se puede escribir como sigue:

```
>
> x=c(29,78,48,29,30,44,72,73,46,82,84,71,75)
> n=length(x) # tamaño del vector
> n
[1] 13
> s=sd(x)# desviación estándar
> s
[1] 21.27837
> M=mean(x) # media
> M
[1] 58.53846
> sesgo=(n/((n-1)*(n-2)))*sum(((x-M)/s)^3)
> sesgo
[1] -0.3307138
>
```

Imagen 63. Salida R para el cálculo del sesgo

3.5.2 Curtosis. Mide la mayor o menor cantidad de datos que se agrupan en torno a la moda. Se definen tres tipos de distribuciones según su grado de curtosis: **distribución mesocúrtica**, presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal); **distribución leptocúrtica**, presenta un elevado grado de concentración alrededor de los valores centrales de la variable; **distribución platicúrtica**, presenta un reducido grado de concentración alrededor de los valores centrales de la variable. Una fórmula para calcular la curtosis de un conjunto de datos es la siguiente:

$$curtosis=\left[\frac{n \times (n+1)}{(n-1) \times (n-2) \times (n-3)} \sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})}{s}\right)^4\right]-\frac{3 \times (n-1)^2}{(n-2) \times (n-3)}$$

Ecuación 5. Curtosis

Y el algoritmo para la curtosis está dado por:

```
>
> x=c(29,78,48,29,30,44,72,73,46,82,84,71,75)
> n=length(x) # tamaño del vector
> n
[1] 13
> s=sd(x)# desviación estándar
> s
[1] 21.27837
> M=mean(x) # media
> M
[1] 58.53846
> curtosis=((n*(n+1))/((n-1)*(n-2)*(n-3)))*sum(((x-M)/s)^4)-((3*(n-1)^2)/((n-2)*(n-3)))
> curtosis
[1] -1.702062
>
```

Imagen 64. Salida R para el cálculo de la curtosis

Las fórmulas anteriores son un poco engorrosas, pero se muestran con el fin de verificar que, en la mayoría de los casos, fórmulas dispendiosas son posibles de programar en R. El sesgo y la curtosis son posibles de calcular directamente luego de cargar el paquete **moments**, mediante los siguientes comandos:

skewness(vector de datos) para calcular el sesgo
kurtosis(vector de datos) para calcular la curtosis

3.6 MEDIDAS DE ASOCIACIÓN

3.6.1 Covarianza. Es una medida de la intensidad de cierta asociación estadística entre dos variables. Como se menciona, es necesario contar con dos variables para calcular la covarianza. Los comandos utilizados para calcularla son cov(x, y) o var(x, y).

Ejemplo: En el examen de una asignatura que consta de parte teórica y parte práctica, las calificaciones de nueve alumnos se muestran en la tabla siguiente. Determinar la covarianza entre la prueba teórica y la prueba práctica.

Teórica (x)	5	7	6	9	3	1	2	4	6
Práctica (y)	6	5	8	6	4	2	1	3	7

```
>
> x=C(5,7,6,9,3,1,2,4,6)
> y=C(6,5,8,6,4,2,1,3,7)
> cov(x,y)
[1] 4.541667
>
>
> x=C(5,7,6,9,3,1,2,4,6)
> y=C(6,5,8,6,4,2,1,3,7)
> var(x,y)
[1] 4.541667
>
```

Imagen 65. Salida R para el cálculo de la covarianza

Si en lugar de tener dos vectores se tiene una matriz, este comando calcula la covarianza entre columnas de la matriz.

3.6.2 Correlación. Cuando dos fenómenos sociales, físicos o biológicos crecen o decrecen de forma simultánea y proporcional debido a factores externos, se dice que los fenómenos están positivamente correlacionados. Si uno crece en la misma proporción que el otro decrece, los dos fenómenos están negativamente correlacionados. El grado de correlación se calcula mediante un coeficiente de correlación aplicado a los datos de ambos fenómenos. Una correlación positiva perfecta tiene un coeficiente igual a 1, y para una correlación negativa perfecta es -1. La ausencia de correlación da como coeficiente 0.

Ejemplo: Teniendo en cuenta los datos del ejercicio del examen, la correlación entre estas dos variables es:

```
>
> x=c(5,7,6,9,3,1,2,4,6)
> y=c(6,5,8,6,4,2,1,3,7)
> cor(x,y)
[1] 0.7628535
>
```

Imagen 66. Salida R para el cálculo de la correlación

De la anterior medida se puede observar que se presenta una correlación positiva alta entre estas dos variables.

3.7 EJERCICIOS

3.7.1 En un estudio sobre la preferencia de bebidas calientes en una empresa se preguntó a 20 empleados su bebida favorita; los resultados son los siguientes: café (C), té (T) y aromática (A).

C	T	C	A	A	T	C	A	C	C
A	C	T	C	C	A	A	T	C	C

Determine los porcentajes para cada bebida y construya un diagrama de barras o un diagrama de sectores.

3.7.2 Al finalizar un curso de estadística descriptiva las notas de los 16 estudiantes son las siguientes:

4.1	2.0	3.6	4.5	3.1	3.2	3.6	2.8
3.1	4.0	3.5	3.1	4.7	3.1	4.2	3.9

a. Construir la tabla de frecuencias acorde con la situación y su respectivo histograma.
b. Calcular las medidas de tendencia central (media, moda y mediana).
c. Analice la dispersión de los datos.
d. Analice la simetría de la distribución de las notas.

3.7.3 En la empresa CARS se cree que la inversión realizada en atención al cliente está relacionada con el nivel de ventas de cada mes; de darse esta relación las directivas estarían dispuestas a incrementar el porcentaje de inversión en este aspecto. Los datos tomados por un supervisor en los últimos 10 meses se muestran a continuación:

Mes	1	2	3	4	5	6	7	8	9	10
Inversión en dólares	20	50	50	80	90	85	100	105	260	300
Ventas en dólares	1300	1500	1400	1550	1700	1900	1800	1850	1900	1800

a. Construya el dispersograma para estas dos variables.
b. Calcule la covarianza y la correlación para la inversión y las ventas de la empresa.
c. Concluya respecto al gráfico y las medidas calculadas anteriormente.