ESTADÍSTICA DESCRIPTIVA

CON

Ramos, A. – Villasante, F. – Alvarez, P.

Alcides Ramos Calcina / Fredy H. Villasante Saravia / Teresa P. Alvarez Rozas

ESTADÍSTICA DESCRIPTIVA
con R

Puno - Perú

ESTADÍSTICA DESCRIPTIVA
con R

Autores:

Alcides Ramos Calcina
Fredy Heric Villasante Saravia
Teresa Paola Alvarez Rozas

Docentes de la Facultad de Ingeniería Estadística e Informática
Universidad Nacional del Altiplano

Editado por:
Fredy Heric Villasante Saravia
Jr. Iquitos N° 181, Puno
Puno - Perú

Primera edición digital, enero 2023

Hecho el depósito Legal en la Biblioteca Nacional del Perú

Registro N° 2023-06901

ISBN: 978-612-00-8827-2

Publicación electrónica en: http://repositorio.unap.edu.pe/handle/20.500.14082/20293

PRESENTACIÓN

El libro "Estadística Descriptiva con R" es una guía completa diseñada específicamente para estudiantes universitarios interesados en adquirir habilidades en el análisis de datos utilizando el lenguaje de programación R. Esta segunda edición ha sido actualizada para ofrecer una experiencia de aprendizaje aún más enriquecedora.

En este libro, los lectores encontrarán una introducción detallada a los conceptos fundamentales de la estadística descriptiva, que es una parte esencial en la investigación y toma de decisiones basadas en datos. Se presentan de manera clara y concisa los métodos estadísticos y las técnicas utilizadas para resumir y visualizar datos, así como para identificar patrones y tendencias.

La principal fortaleza de este libro radica en su enfoque práctico, ya que cada concepto se acompaña de ejemplos detallados y ejercicios resueltos utilizando R. Los lectores aprenderán cómo implementar análisis descriptivos utilizando R y cómo interpretar los resultados obtenidos. Además, se exploran diversos librerías y funciones de R diseñados específicamente para el análisis de datos descriptivos.

La segunda edición del libro aborda temas como el análisis exploratorio de datos multivariables. También se ha tenido en cuenta el valioso feedback de los lectores de la primera edición, incorporando mejoras y aclaraciones para una comprensión más sólida de los conceptos estadísticos.

Con su enfoque claro y práctico, "Estadística Descriptiva con R" se posiciona como un recurso invaluable tanto para estudiantes universitarios como para profesores de estadística. Este libro ofrece las herramientas necesarias para comprender y aplicar eficazmente los conceptos de estadística descriptiva, al tiempo que promueve la fluidez en el uso de R como una herramienta poderosa y versátil para el análisis de datos.

En resumen, esta segunda edición del libro "Estadística Descriptiva con R" proporciona una guía práctica y actualizada para el aprendizaje de la estadística descriptiva, integrando el uso de R como una herramienta esencial. Es un recurso imprescindible para aquellos que deseen adquirir una base sólida en el análisis de datos y aprovechar al máximo las capacidades de R en el proceso.

Los autores.

Puno, julio del 2023.

INDICE DE CONTENIDO

Capítulo 1

Introducción a R

1.1 INTRODUCCIÓN

El Lenguaje R funciona de forma muy diferente a otros softwares estadísticos muy conocidos, como el programa SPSS. La mayor diferencia estriba en que, en lugar de utilizar menús de Windows que se manejan de forma sencilla, R utiliza códigos de sintaxis que se necesita escribir y ejecutar para poder llegar a unos resultados. Esto es, R requiere escribir comandos y funciona como lenguaje de programación. Esto puede hacer difícil el primer contacto con R. Sin embargo, una vez te familiarices con la tarea, enseguida descubrirás que es muy sencillo escribir comandos y ejecutarlos para llegar a un resultado concreto, ya sea un análisis de datos o la elaboración de un gráfico (Ximénez & Revuelta, 2022).

Para trabajar con R nos tenemos que acostumbrar a escribir y ejecutar una serie de comandos, lo que nos permitirá trabajar con mucha autonomía y control

sobre los análisis estadísticos concretos que queramos llevar a cabo (Ximénez & Revuelta, 2022).

1.2 QUÉ ES EL LENGUAJE R

R es un ambiente de programación formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones. Además, es gratuito y de código abierto, un Open Source parte del proyecto GNU, como Linux o Mozilla Firefox.

El software R se presentó al mercado en 1993 de la mano de sus creadores Robert Gentleman y Ross Ihaka, que desarrollaron la herramienta en el Departamento de Estadística de la Universidad de Auckland. Sin embargo, la base de sus orígenes se encuentra en el desarrollo del lenguaje S (Ximénez & Revuelta, 2022).

Principales características del lenguaje R

- R es un lenguaje y entorno de programación para análisis estadístico y gráfico.
- Se trata de un proyecto de software libre, resultado de la implementación del Lenguaje S. Probablemente, R es el lenguaje más utilizado en investigación por la comunidad estadística. A esto contribuye la posibilidad de cargar diferentes Librerías o Paquetes con finalidades específicas de cálculo o gráfico.
- R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.
- R se trata de un lenguaje de programación que es muy potente y con una sintaxis muy sencilla e intuitiva de aprender. No se necesitan conocimientos de otros lenguajes de programación para poder usar R.
- R proporciona un abanico de herramientas estadísticas (modelos lineales y no lineales, test estadísticos, algoritmos de clasificación y agrupamiento, etc.) y gráficas.
- R puede integrarse con distintas bases de datos y existen librerías que facilitan su utilización.
- R permite generar gráficos con alta calidad (Ver Figura 1.1 y 1.2).

- R también puede usarse como herramienta de cálculo numérico y puede ser tan eficaz como otras herramientas tales como MATLAB.
- R forma parte de un proyecto colaborativo y abierto. Sus usuarios pueden publicar Librerías (también denominadas paquetes, del inglés package) que extienden su configuración básica. Las librerías están organizadas por temas.

```
# Gráfico 3D
x <- seq(-10,10,length=50)
y <- x
f <- function(x,y){x^2-y^2}
z <- outer(x,y,f)
persp(x,y,z,theta = 30,phi = 30)
```

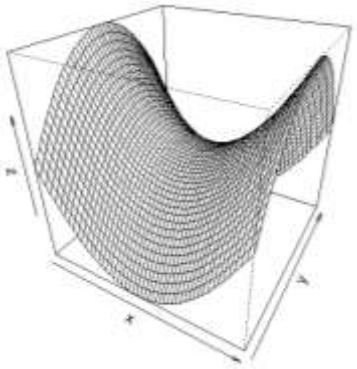


Figura 1.1. Ejemplo de gráfico 3D con R.

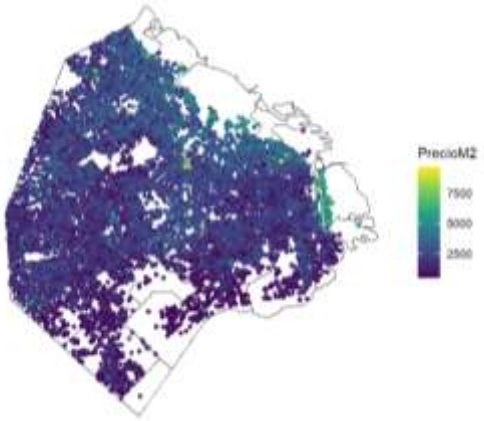


Figura 1.2. Ejemplo de gráfico hecho con R.

1.3 EL ENTORNO DE R

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de:

- Almacenamiento y manipulación efectiva de datos.
- Operadores para cálculo sobre variables indexadas (Arrays), en particular matrices.
- Una amplia, coherente e integrada colección de herramientas para análisis de datos.
- Posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora.
- Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)

El término "entorno" lo caracteriza como un sistema completamente diseñado y coherente, antes que como una agregación incremental de herramientas muy específica se inflexibles, como ocurre frecuentemente con otros programas de análisis de datos (Ximénez & Revuelta, 2022).

1.4 QUÉ ANÁLISIS ESTADÍSTICOS PUEDO HACER CON R

Todos los que el usuario pueda desear. La instalación básica de R incluye procedimientos para prácticamente todos los métodos estadísticos tradicionales: estadística descriptiva, gráficos, inferencia, regresión, análisis de la varianza, etc.

Además, es posible descargar en cualquier momento desde la web de R nuevas librerías o "packages" que incrementan los procedimientos disponibles. En junio de 2014, en la web "oficial" de R (CRAN, Comprehensive R Archive Network), hay 5663 librerías que pueden descargarse libremente (http://www.r-project.org).

Las librerías se encuentran organizadas en grupos (Task Views) en función de su área de aplicación.

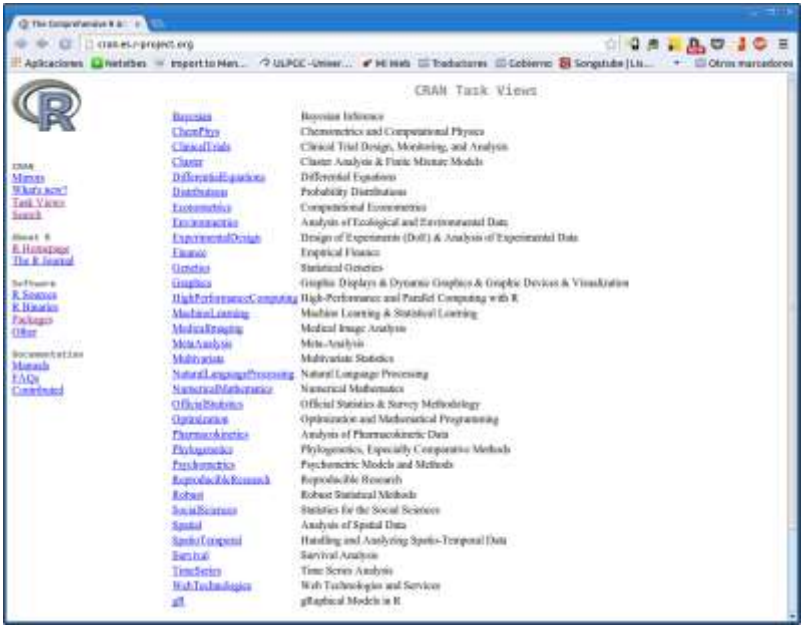


Figura 1.3. Web "oficial" de R (CRAN, Comprehensive R Archive Network)

Hay además numerosas iniciativas orientadas al uso de R en campos específicos. Muchas de ellas cuentan con sus propios repositorios de datos y procedimientos estadísticos implementados en librerías (Santana, 2022):

- *R Open Science*: librerías para el acceso a datos públicos.
- *Bioconductor*: Bioinformática con R, muy orientado a genómica.
- *Fisheries Library*: análisis de pesquerías con R.
- *Rmetrics*: Análisis de mercados financieros con R.
- *R-Geo*: Estadística espacial con R
- *Github*: paquetes en desarrollo, pero listos para su uso.

1.5 CÓMO PUEDO TRABAJAR CON R

Existen múltiples formas de interactuar con R:

1.5.1. Mediante la terminal

Una terminal es una simple ventana del sistema operativo, que arranca R y nos deja el "prompt" esperando por nuestras instrucciones (Santana, 2022):

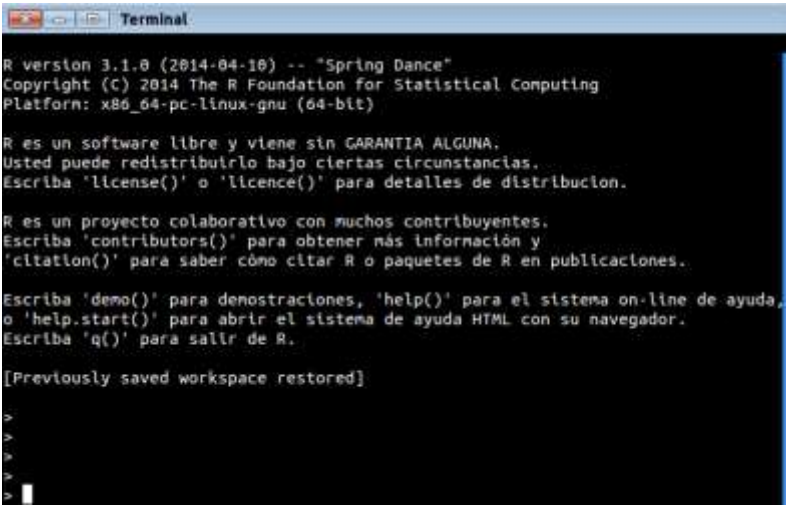


Figura 1.4. Ventana terminal de R.

Obviamente en este tipo de procedimiento no hay menús de ayuda, ni de configuración, ni de edición, etc.

1.5.2. Mediante la GUI (Graphics User Interface) del programa

Las versiones de R para Windows y Mac, cuentan con sus propias interfaces gráficas de usuario que integran una consola de resultados, un editor de código, algún menú de ayuda, y cierta gestión de gráficos.

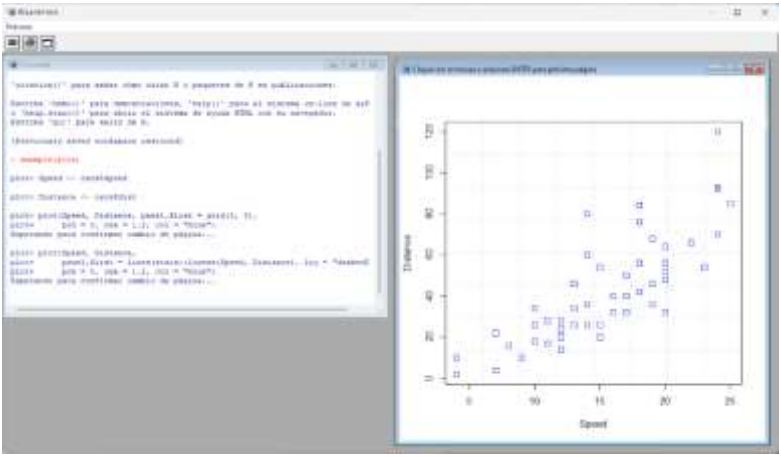


Figura 1.5. GUI de Windows.

1.5.3. Mediante Rstudio

Rstudio es una interfaz gráfica para R desarrollada a comienzos de 2011 por un grupo de informáticos y estadísticos de Boston (EEUU). La versión Open Source puede descargarse libremente desde www.rstudio.com:

Figura 1.6. Pagina web principal de RStudio.

En los tres últimos años ha experimentado un notable desarrollo y se ha convertido, de facto, en la GUI por excelencia para R. Dispone de versiones (idénticas en su funcionamiento) para Linux, Mac y Windows.

La interfaz de RStudio permite un acceso más cómodo a la edición de código, los resultados, los gráficos, la descarga de librerías, los objetos en memoria, etc. Además, permite generar muy fácilmente informes con los procedimientos y resultados de nuestros análisis en varios formatos (html, pdf y word).

Ni Rstudio ni las GUI estándar de R cuentan con ningún menú para el acceso a procedimientos estadísticos. Es el usuario el que debe escribir un "script" (un programa) en el que encadene los comandos necesarios para llevar a cabo el análisis que se propone realizar.

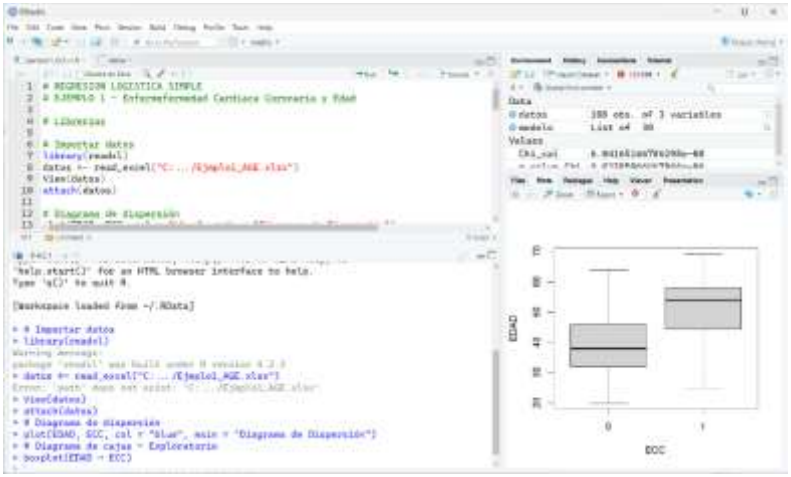


Figura 1.7. Ventana principal de RStudio.

Normalmente un script se estructura en varias secciones (no necesariamente en este orden) (Santana, 2022):

- **Cabecera**: donde se cargan las librerías que se van a utilizar y el usuario define sus propias funciones (si las necesita).
- **Lectura de datos**: se define el directorio de trabajo y se cargan los datos, bien directamente declarándolos en el propio script, o importándolos desde uno o varios archivos externos.
- **Procesamiento de datos**: transformaciones de los datos si se requiere, asignación de etiquetas, identificación de valores perdidos, etc.
- Aplicación de procedimientos estadísticos.
- Generación de informe de resultados.

1.5.4. Mediante sistemas basados en menús

RCommander es una interfaz gráfica desarrollada originalmente por John M. Fox en la Universidad McMaster en Canadá para sus alumnos de psicología, y que rápidamente fue adoptada en otros ámbitos por su facilidad de uso. Esta interfaz cuenta con menús desplegables para el

acceso a los procedimientos estadísticos. RCommander se caracteriza por ser ampliable, esto es, se le pueden añadir menús nuevos o complementar los que ya tiene. Resulta una alternativa interesante para usuarios esporádicos de R que se limiten a hacer análisis estadísticos estándar.

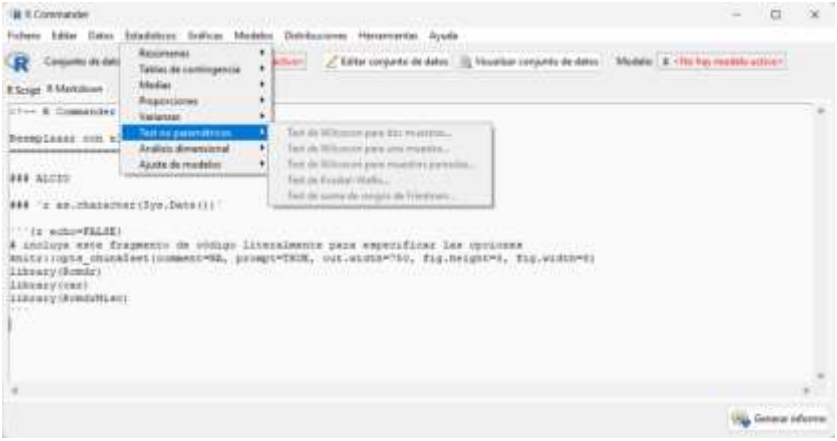


Figura 1.8. Ventana principal de R Commander.

1.6 CÓMO TRABAJA R

1.6.1 Los comandos y el argumento

R es un lenguaje de programación y difiere de otros softwares estadísticos comerciales en que R trabaja mediante comandos. Para ello, utiliza el objeto como entidad básica. Esto es, cualquier expresión evaluada por R tiene como resultado un objeto (Ximénez & Revuelta, 2022).

Los comandos se escriben en el editor de archivos (Ventana 1 de la Figura 1.9). Como resultado, se genera un programa informático (denominado Script), que puede tener cualquier longitud y se ejecuta al pulsar el botón **RUN**. Para la ejecución de los comandos, se necesita tener posicionado el cursor en la línea exacta donde esté escrito el comando. Si queremos ejecutar el programa completo, se puede pulsar **SOURCE**.

Vamos a escribir nuestro primer comando. Por ejemplo, vamos escribir en el terminal:

```
# Ejemplo de ingreso de datos y ploteo
x <- c(10, 12, 14, 15, 11, 12, 19)
y <- c(13, 12, 13, 24, 45, 23, 53)
plot(x,y)
```

Como puede verse en la Figura 1.9, hemos escrito los comandos de la línea 1 a 4 de la Ventana 1. El comando que hemos escrito es **c( )** para el ingreso de valores de la variable x e y, los datos que lleva dentro del paréntesis se denomina argumento (en este caso los datos). Finalmente, se plotea los datos a través del comando **plot( )**. Para ejecutar el comando, es importante dejar el cursor donde hayamos escrito el comando (en este caso en la línea 4), y a continuación pulsamos el botón Run (está en la parte superior derecha de la Ventana 1, en la Figura 1.9, rodeado con un círculo rojo).

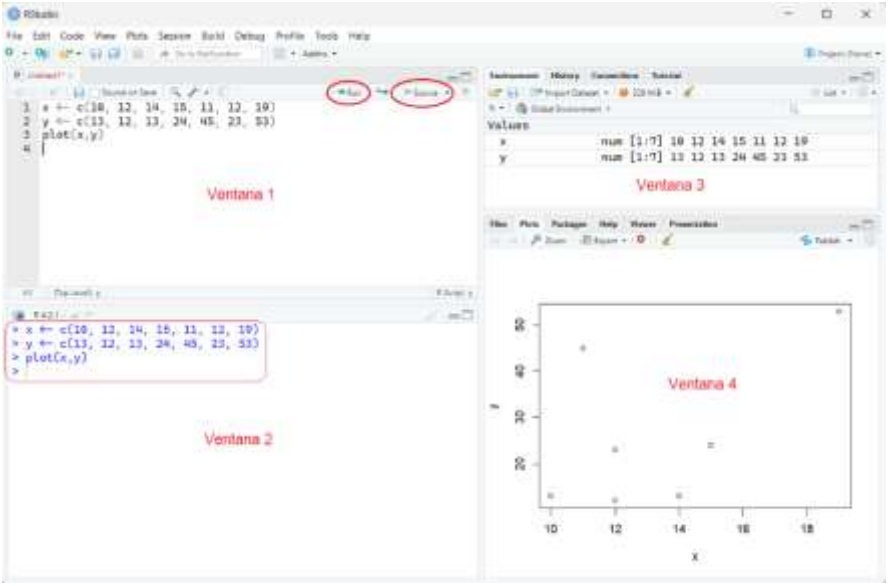


Figura 1.9. Ejemplo de comando Plot.

Uno de los usos más frecuentes de R es como calculadora. Veamos un ejemplo sencillo donde se propone la suma de los valores 3 y 5:

```
3 + 5
[1] 8
```

1.6.2 Crear objetos

Como ya se ha señalado, R es un lenguaje de programación y, como tal, se necesita definir todas las operaciones que queramos llevar a cabo. Para ello, necesitaremos manejar variables denominadas **Objetos**.

Los objetos pueden tener el contenido que queramos. Entre otros:

- Un dato (por ejemplo, el número 2).
- Múltiples datos (por ejemplo, un vector: 1, 2, 3, 4, etc.)
- Nombres y cadenas (“Hola mundo).
- Bases de datos (un fichero SPSS).
- Conjuntos de bases de datos. etc.

Cada objeto recibe un nombre y es guardado en la memoria de R (Ventana 3 de la Figura 1.9) para ser reutilizado posteriormente. Por tanto, un paso importante a realizar implica “dar contenido al objeto”. Por ejemplo, si creamos el objeto x y queremos definir que x = 10, esto sería dar contenido a nuestro objeto (en este caso, un valor numérico) (Ximénez & Revuelta, 2022).

1.6.3 Asignación

Al igual que ocurre con otros lenguajes de programación, R asigna nombres a las operaciones. Esto lo conseguiremos mediante el símbolo “<−”, “−>” o “=”.

Hay que tener en cuenta que R utiliza determinados términos para referirse a algunas funciones por lo que lo mejor es evitar esos nombres a la hora de

las asignaciones, por ejemplo "**c**" se utiliza para crear vectores o "**t**" que calcula la traspuesta de un conjunto de datos (vectores, matrices, dataframe, etc.), pero si nos confundimos no es dramático, ya que podemos arreglarlo (Contreras, 2018).

Vemos un ejemplo.

```
a <- 10
b = 11
suma <- a + b
suma
[1] 21
```

Una vez definidos los objetos a y b, puede definirse el objeto suma que es la operación a + b. Si más adelante queremos utilizar el objeto suma, podremos llamarlo y que forme parte de otro objeto diferente. Podría, como en el ejemplo inferior referido al objeto c, ser el numerador de un cociente:

```
c <- suma/b
c
[1] 1.909091
```

1.6.4 Funciones `ls` y `rm`

La función `ls` saca en pantalla los objetos almacenados en la memoria por el usuario, aunque sólo muestra los nombres de los mismos. Si se quiere listar solo aquellos objetos que contengan un carácter en particular "`ls(pat =)`" y para restringir la lista a aquellos objetos que comienzan con este carácter utilizamos el símbolo exponente "`ls(pat =^`" (Contreras, 2018).

```
ls()
[1] "a"     "b"     "c"     "suma" "x"     "y"
ls(pat = "a")
[1] "a"     "suma"
ls(pat = "^a")
[1] "a"
ls.str()
a :  num 10
b :  num 11
c :  num 1.91
suma :  num 21
x :  num [1:7] 10 12 14 15 11 12 19
y :  num [1:7] 13 12 13 24 45 23 53
```

Tenga en cuenta que es recomendable conocer los elementos que R tiene en memoria, aparte de los que hemos memorizado nosotros. Por ejemplo:

Para borrar objetos almacenados en la memoria, utilizamos la función "`rm()`", por ejemplo `rm(x)` elimina el objeto x; `rm(x; y)` elimina ambos objetos x e y, y "`rm(list = ls( )`" elimina todos los objetos que estén en la memoria. Tenga en cuenta que las opciones mencionadas para la función `ls( )` pueden aplicarse para borrar selectivamente algunos objetos.

```
rm(suma)
ls()
[1] "a" "b" "c" "x" "y"
# Se borro el objeto suma
rm(list = ls())
ls()
character(0)
```

1.6.5 Operaciones con variables

Las operaciones que pueden realizarse con objetos en R son muy diversas. Entre las que nos interesan se encuentran las siguientes:

Operación	Notación
Suma	+
Resta	-
Producto	*
División	/
Exponente	^
Raíz cuadrada	sqrt( )
Producto matrices	%*%

Se tiene algunos ejemplos:

```
x <- 3
y <- 5
s <- x + y
s
[1] 8
r <- y - x; r
[1] 2
p <- s * r; p
[1] 16
q <- sqrt(p); q
[1] 4
```

1.6.6 Vectores

Lo más habitual es que los elementos sean un conjunto de números, aunque también se pueden incluir letras u otros tipos de elementos.

Para crear un objeto o vector que contenga series de números, puede usarse el comando `c()`, que significa conjunto o concatenar. Los elementos incluidos en la serie numérica han de ir separados por comas.

En el ejemplo siguiente se crean dos tipos de series de objetos. Uno con nombres (nomb) y otro con números (x):

```
nomb <- c("Juan", "María")
nomb
[1] "Juan"  "María"
x <- c(1, 3, 5, 7, 9)
x
[1] 1 3 5 7 9
```

Podemos escribir vectores de varias maneras, utilizando la opción ":" (el vector comienza en el primer número suministrado y analiza en el segundo o en un número anterior sin sobrepasarlo, tanto en orden ascendente como descendente).

```
1:10
 [1]  1  2  3  4  5  6  7  8  9 10
10:1
 [1] 10  9  8  7  6  5  4  3  2  1
1.5:9.5
[1] 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5 9.5
c(1:10)
 [1]  1  2  3  4  5  6  7  8  9 10
c(1:5,2,4,-1)
[1]  1  2  3  4  5  2  4 -1
```

Tenemos formas adicionales de crear vectores. Una de las más comunes es utilizando la función "`seq(a,b,c)`", que genera secuencias de números reales, donde el primer elemento indicará el principio de la secuencia, el segundo el final y el tercero el incremento que se debe usar para generar la secuencia.

```
seq(1,10,2)
[1] 1 3 5 7 9
seq(from = 1, to = 10, by = 2)
[1] 1 3 5 7 9
# Si queremos 6 numeros de 1 a 10
seq(from = 1, to = 10, length = 6)
[1]  1.0  2.8  4.6  6.4  8.2 10.0
```

La función "`rep(a; b)`" que crea un vector con "b" elementos idénticos al valor "a".

```
# Repetimos el número 3 diez veces
rep(3,10)
 [1] 3 3 3 3 3 3 3 3 3 3
# Repetimos la secuencia de 1 a 5, dos veces
rep(1:5,2)
 [1] 1 2 3 4 5 1 2 3 4 5
# Repetimos cada elemento de la secuencia 3 veces de 5 en 5
rep(1:3,rep(5,3))
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
```

1.7 EL SCRIPT

Ya hemos hablado del SCRIPT. Es el fichero con extensión *.R donde se escriben los comandos que nos permitirán ejecutar los análisis y/o solicitar gráficos. Los Script siempre se graban desde la Ventana 1 de la Figura 1.9 y contienen el programa o listado de comandos que ejecutaremos para llevar a cabo nuestros análisis.

Es recomendable que este fichero no sea demasiado largo. Y también que anotemos algún comentario que nos permita recordar la operación que hayamos definido. Para añadir comentarios junto a los comandos se usa el símbolo #. A modo de ejemplo, escribiremos el ejemplo anterior con el comentario de lo realizado de la siguiente manera (Ximénez & Revuelta, 2022):

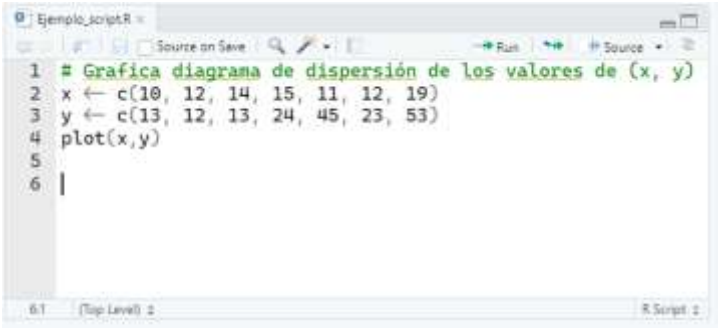


Figura 1.10. Ejemplo de Script en R.

1.8 TRANSFORMACIÓN DE VARIABLES

A continuación, se mostrará como se transforman las variables a través de un sencillo ejemplo.

Ejemplo 1.1
Supongamos que tenemos un fichero con medidas biométricas tales como el peso y la estatura y deseamos calcular el índice de masa corporal de 22 alumnos de la universidad.

Alumno	*Sexo*	*Estatura (cm.)*	*Peso (Kg.)*
1	Mujer	159	49
2	Varón	164	62
3	Mujer	172	65
4	Mujer	167	62
5	Mujer	164	61
6	Mujer	161	67
7	Mujer	168	48
8	Varón	181	74
9	Varón	183	74
10	Mujer	158	50
11	Mujer	156	65
12	Varón	173	64
13	Mujer	168	43
14	Varón	178	74
15	Varón	181	76
16	Varón	182.5	91
17	Varón	176	73
18	Mujer	162	68
19	Mujer	156	52
20	Mujer	162	45
21	Varón	181	80
22	Varón	173	69

Es importante señalar que el Índice de Masa Corporal es un índice del peso de una persona en relación con su altura. A pesar de que no hace distinción entre los componentes grasos y no grasos de la masa corporal total, éste es el método más práctico para evaluar el grado de riesgo asociado con la obesidad (Paco, 2012).

El índice de masa corporal esta dado por la siguiente relación:

$$IMC = \frac{Peso(Kg.)}{[Talla(m)]^2}$$

Los datos del ejemplo serán ingresados a una hoja electrónica o un programa que te permita generar un archivo en formato ***.csv**, el mismo que se caracteriza por su flexibilidad, compatibilidad y tamaño reducido y eso facilitará la importación a cualquier otro programa de análisis de datos, en nuestro caso se usó una hoja electrónica y luego exportado en formato **csv** con delimitación (**;**) pero es posible usar cualquier otro carácter de separación entre los datos de las variables.

R-Studio brinda una interface fácil de usar por lo intuitivo que es, sin embargo, también es posible utilizar código que permita el mismo resultado, con el siguiente:

```
# Librería que permite la importación
library(readr)

# Instrucción que permite importa la data, considerando que el archivo
se encuentra en la siguiente dirección D:/carpeta/ y con nombre de
archivo MedBio.csv
MedBio <- read_delim("D:/carpeta/MedBio.csv",  delim = ";",
escape_double = FALSE, col_types = cols(Alumno = col_number(),
Estatura = col_number(), Peso = col_number()),trim_ws = TRUE)
```

La vista previa de dicha importación tendrá algo similar a:

	Alumno	Sexo	Estatura	Peso
1	1	Mujer	159.0	49
2	2	Varón	164.0	62
3	3	Mujer	172.0	65
4	4	Mujer	167.0	62
5	5	Mujer	164.0	61
6	6	Mujer	161.0	67
7	7	Mujer	168.0	48
8	8	Varón	181.0	74
9	9	Varón	183.0	74
10	10	Mujer	158.0	50
11	11	Mujer	156.0	65
12	12	Varón	173.0	64
13	13	Mujer	168.0	43
14	14	Varón	178.0	74
15	15	Varón	181.0	76

Showing 1 to 16 of 22 entries, 4 total columns

Figura 1.11. Vista de la importación de datos.

Para calcular *IMC*, aplicamos el siguiente script en R:

```
ImasaC <- round(MedBio$Peso/((MedBio$Estatura/100)^2), 2)
ImasaC

# Para incrementar una columna, para el IMC
MedBio <- cbind(MedBio, ImasaC)
MedBio
```

Y obtendremos el siguiente resultado de la nueva tabla modificada:

	Alumno	Sexo	Estatura	Peso	ImasaC
1	1	Mujer	159.0	49	19.38
2	2	Varón	164.0	62	23.05
3	3	Mujer	172.0	65	21.97
4	4	Mujer	167.0	62	22.23
5	5	Mujer	164.0	61	22.68
6	6	Mujer	161.0	67	25.85
7	7	Mujer	168.0	48	17.01
8	8	Varón	181.0	74	22.59

Figura 1.12. Vista de la variable transformada ImasaC.

1.9 RECODIFICACIÓN DE VARIABLES

R permite modificar los valores de las variables recodificándolos. Esto es útil especialmente para añadir o cambiar categorías en una variable.

Siguiendo con el Ejemplo 1.1, vamos a mostrar cómo se recodificaría una variable con medida de escala.

Después de haber calculado el índice de masa corporal, podemos clasificar a los alumnos en tres categorías dependiendo de la puntuación obtenida en dicho índice. Dicha clasificación es la siguiente:

Puntuación en IMC	*Categoría*
Por debajo de 20	Delgado
Entre 20 y 25	Normal
Por encima de 25	Sobrepeso

Para ello aplicaremos el siguiente script:

```
# Caracterización del IMC
# Por debajo de 20     => Delgado
# Entre 20 y 25        => Normal
# Por encima de 25      => Sobrepeso

cIMC <- ifelse(MedBio$ImasaC<20, "Delgado", ifelse(MedBio$ImasaC
>= 20 & MedBio$ImasaC <= 25, "Normal", "Sobrepeso"))
MedBio

# Para incrementar la variable/columna de caracterización
MedBio <- cbind(MedBio, cIMC)
MedBio
```

y el resultado es:

	Alumno	Sexo	Estatura	Peso	ImasaC	cIMC
1	1	Mujer	159.0	49	19.38	Delgado
2	2	Varón	164.0	62	23.05	Normal
3	3	Mujer	172.0	65	21.97	Normal
4	4	Mujer	167.0	62	22.23	Normal
5	5	Mujer	164.0	61	22.68	Normal
6	6	Mujer	161.0	67	25.85	Sobrepeso
7	7	Mujer	168.0	48	17.01	Delgado
8	8	Varón	181.0	74	22.59	Normal
9	9	Varón	183.0	74	22.10	Normal

1.10 SELECCIÓN DE CASOS

Generalmente los análisis a los datos se realizarán sobre todos los casos, pero a veces estaremos interesados en realizar análisis sólo sobre un determinado subconjunto de los casos incluidos en el fichero activo.

R ofrece múltiples opciones para seleccionar casos, la más común es seleccionar casos que cumplan una determinada condición. Por ejemplo si cuando sólo deseamos analizar un determinado subconjunto de la población como por ejemplo la población perteneciente a un determinado género.

Para ello aplicaremos el siguiente script:

```
# Proceso de selección por sexo:
# Si es Mujer => Seleccionada
# SI es Varón => No seleccionado

sIMC <- ifelse(MedBio$Sexo == "Mujer", "Seleccionada", "No
seleccinado")
MedBio <- cbind(MedBio, sIMC)
MedBio
```

Y el resultado es:

	Alumno	Sexo	Estatura	Peso	ImasaC	cIMC	sIMC
1	1	Mujer	159.0	49	19.38	Delgado	Seleccionada
2	2	Varón	164.0	62	23.05	Normal	No seleccinado
3	3	Mujer	172.0	65	21.97	Normal	Seleccionada
4	4	Mujer	167.0	62	22.23	Normal	Seleccionada
5	5	Mujer	164.0	61	22.68	Normal	Seleccionada
6	6	Mujer	161.0	67	25.85	Sobrepeso	Seleccionada
7	7	Mujer	168.0	48	17.01	Delgado	Seleccionada
8	8	Varón	181.0	74	22.59	Normal	No seleccinado
9	9	Varón	183.0	74	22.10	Normal	No seleccinado

Capítulo 2

Aspectos Conceptuales de Estadística

2.1 INTRODUCCIÓN

Actualmente en el mundo académico del análisis de datos R se ha convertido en uno de los programas de Análisis Estadísticos más extendidos debido a que está disponible en diferentes plataformas y de libre disponibilidad con licencia libre.

Trataremos de resumir las características básicas de su funcionamiento mediante el número mínimo posible de conceptos nuevos, teniendo en cuenta la perspectiva del usuario que necesita solamente el manejo de opciones sencillas para el trabajo diario.

Los temas básicos de la estadística descriptivas las enfocaremos a ejemplos y script sencillos con la única finalidad de facilitar la comprensión y manejo del

lenguaje R. Comenzaremos con el desarrollo de algunos de los conceptos básicos

2.2 CONCEPTOS BÁSICOS DE ESTADÍSTICA

2.2.1 Definición de Estadística

El término "Estadística", que tiene origen histórico, hace referencia a una determinada información numérica; esta acepción se encuentra cada día más arraigada en nuestra sociedad debido al abultado conjunto de números y cifras en el que se encuentra inmersa: PBI., índices de precios, tasas de inflación, evolución del paro, cotizaciones bursátiles, accidentes de circulación, porcentajes de votantes, porcentajes de personas que padecen una determinada enfermedad, etc.

Podemos definir la *Estadística* como la ciencia que nos proporciona un conjunto de métodos y procedimientos para recolección, clasificación, análisis e interpretación de datos en forma adecuada para tomar decisiones cuando prevalecen condiciones de incertidumbre. La estadística se clasifica en Estadística Descriptiva y Estadística Inferencial (Sanchez, 2012).

Clasificación de la Estadística (Ibañez, 2001)**:**

- ✓ ***La Estadística Descriptiva*** es parte de la Estadística que se encarga de la recolección, clasificación, presentación y simplificación de los datos (por ejemplo, edad de una población, altura de los estudiantes de una escuela, temperatura en los meses de verano, etc.) tratando de extraer conclusiones sobre el comportamiento de estas variables.
- ✓ ***La Estadística Inferencial*** es la que nos proporciona la teoría necesaria parar inferir o estimar las leyes de una población partiendo de los resultados o conclusiones del análisis de una muestra.

2.2.2 Etapas del Análisis Estadístico

Las diversas fases por las que atraviesa el análisis estadístico son(Vicente, 2016):

a) **Recolección de datos**, que no por ser elemental, está exenta de dificultades e indicaciones que hay que observar, ya que una recogida

mal efectuada puede ocasionar un sesgo de la información y del posterior análisis, por lo que el objeto de la investigación debe plantearse de una manera minuciosa, así como la organización del trabajo de campo necesario para la recolección de datos.

b) **Ordenación y presentación de los datos**, y que suele presentarse mediante unas tablas de simple o de doble entrada.

c) **Resumen de la información**, para tratar de describir las características más relevantes que pueden tener los datos, y que se realiza mediante la determinación de parámetros estadísticos que intentan resumir toda la información que aporte el conjunto de datos.

d) **Análisis estadístico**, a través de métodos facilitados por la Estadística Matemática, para tratar de verificar hipótesis sobre regularidades que pueden detectarse en las etapas previas.

2.2.3 Población y Muestra

Población: conjunto de todos los individuos o elementos que tienen características comunes (personas, objetos, animales, etc.) y que porten información sobre el fenómeno que se estudia. Por ejemplo, si estudiamos el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.

Muestra: dado que no siempre es posible estudiar todos los elementos de la población, ya sea por razones económicas, de rapidez de obtención de la información, o porque los elementos se destruyen en el proceso de la investigación, con frecuencia es necesario examinar sólo una parte de la población, que se denomina *muestra*; para que una muestra sea válida como objeto de estudio, ha de ser representativa de la población, es decir ha de tener las mismas características, en los caracteres estudiados, que la población.

2.2.4 Parámetro y Estadígrafo

Parámetro es un número o medida que describe alguna característica de toda la población, y para determinar su valor, es necesario utilizar la información poblacional completa. Por ejemplo: la media poblacional (μ), varianza

poblacional (σ^2), proporción poblacional (P), etc. y *Estadígrafo* es también un número o medida que se obtiene a partir de los datos muestrales y describe alguna característica de la muestra. Por ejemplo: media muestra ($\bar{x}$), varianza muestral (s^2), proporción poblacional (p), etc.

2.2.5 Variable Estadística

Llamaremos ***variable*** al ***carácter*** objeto de estudio, que puede tomar distintos valores.
Las variables pueden ser cuantitativas o cualitativas, según que tomen, o no, valores cuantificables (Castañeda, 2010).

- ***Variables cualitativas o atributos:*** no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo). Pueden clasificarse en: Nominal y Ordinal.

- ***Variables cuantitativas:*** tienen valor numérico, y estudian caracteres cuantificables (edad, precio de un producto, ingresos anuales). Por su parte, las variables cuantitativas se pueden clasificar en:

  - **Discretas. -** Sólo pueden tomar valores enteros (1, 2, 8, -4, etc.). Por ejemplo: número de hermanos (puede ser 1, 2, 3…, etc. pero, por ejemplo, nunca podrá ser 3,45).
  - **Continuas. -** Pueden tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 80,3 km/h, 94,57 km/h.…etc.

Las variables también se pueden clasificar en:

- ***Variables unidimensionales:*** sólo recogen información sobre una característica (por ejemplo: edad de los alumnos de una clase).
- ***Variables bidimensionales:*** recogen información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase).

- ***Variables pluridimensionales:*** recogen información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

2.2.6 Tipos de Escala

En determinado tipo de estudios, quizá tenga mayor relevancia diferenciar las variables según el tipo de escala utilizada, distinguiendo(Parra, 2014):

- **Escala nominal:** el carácter estudiado se clasifica en categorías no numéricas, sin que puedan establecerse ninguna relación de orden entre ellas, por ejemplo: las profesiones laborales, el estado civil, la ideología política, el sexo, etc.
- **Escala ordinal:** el carácter estudiado es de tipo no numérico, pero se pueden establecer algún tipo de orden entre las distintas categorías. Este es el caso del nivel de estudios (primarios, medios, superiores), los tipos de clases sociales (baja, media, alta), etc.
- **Escala de intervalo:** puede establecerse alguna unidad de medida y cuantificar numéricamente la distancia existente entre dos observaciones. Es la escala cuantitativa, encontrándose en este caso gran número de variables entre ellas como, por ejemplo: salarios, presupuestos, gastos, etc.
- **Escala de proporción:** son aquellas variables en las que además de una unidad de medida, se fija un punto origen, que marca el cero. En este tipo pueden considerarse la edad, el peso, el número de unidades en stock en un inventario, etc.

Capítulo 3

Variables Estadísticas Unidimensionales

3.1 INTRODUCCIÓN

Después de recoger toda la información correspondiente a la investigación, es decir, al agotar todo el trabajo de campo, nuestro escritorio se llena de un cúmulo de datos y cifras desordenadas los cuales, al ser tomados como observaciones individuales, dicen muy poco sobre la población estudiada; es, entonces, tarea del investigador estructurar y ordenar los conjuntos numéricos de los datos obtenidos en la observación de una muestra o población, consignando la información en tablas legibles que denominamos distribuciones de frecuencias, para así poder proceder con más facilidad a su estudio.

3.2 DISTRIBUCIÓN DE FRECUENCIAS

Llamaremos distribución de frecuencias al conjunto de los valores que toma una variable, junto con sus frecuencias correspondientes. Así pues, para determinar una distribución de frecuencias debemos conocer todos los valores x_i de la variable y cualquiera de las columnas de frecuencias (pues el paso de una a otra es inmediato).

Distinguiremos dos tipos fundamentales de distribución de frecuencias: las no agrupadas en intervalos y las agrupadas en intervalos.

3.3 DISTRIBUCIÓN DE FRECUENCIAS NO AGRUPADAS

La distribución de frecuencias no está agrupada en intervalos cuando cada valor de la variable tiene asociado su frecuencia.

Esta forma de representación de datos esta especialmente orientada a ***variables cualitativas*** donde los valores corresponden a categorías de clasificación y ***variables cuantitativas discretas*** donde los valores que toma no son muy diversos y cada uno de ellos se repite muchas veces.

3.3.1 Regla general para construcción de tablas de frecuencia simple

Consideremos una población estadística de n individuos, descrita según un carácter o variable X cuyas modalidades han sido agrupadas en un numero k de clases, que denotamos mediante $x_1, x_2, \ldots, x_k$. Para cada una de las clases x_i, $i = 1, \ldots, k$, introducimos las siguientes magnitudes (Gamarra, 2015):

- ***Frecuencia absoluta de la clase (f_i):*** Es el número de observaciones que presentan una modalidad perteneciente a la clase x_i.

- ***Frecuencia relativa de la clase (hi):*** Es el cociente entre las frecuencias absolutas de dicha clase y el número total de observaciones, es decir:

$$h_i = \frac{f_i}{n}$$

Obsérvese que h_i es el tanto por uno de observaciones que están en la clase x_i.

- ***Frecuencia absoluta acumulada (F_i):*** Se calcula sobre variables cuantitativas o cuasi cuantitativas, y se define:

$$F_i = f_1 + f_2 + f_3 + \ldots + f_i = \sum_{j=1}^{i} f_j$$

- ***Frecuencia relativa acumulada (H_i):*** Se calcula sobre variables cuantitativas o cuasi cuantitativas, y se define:

$$H_i = h_1 + h_2 + h_3 + \ldots + h_i = \sum_{j=1}^{i} h_j$$

- ***Frecuencia relativa porcentual (h_i%):*** Si h_i es el tanto por uno de observaciones que están en la clase x_i, multiplicado por 100% representa el porcentaje de la población que comprende a esa clase.

$$h_i\% = h_i * 100$$

- ***Tabla de frecuencias:***

Modalidad	Frecuencia Absoluta		Frecuencia Relativa		Frec. Rel. Porcentual
	Simple	Acumulada	Simple	Acumulada	
X_i	f_i	F_i	h_i	H_i	h_i%
x_1	f_1	$F_1=f_1$	$h_1 = f_1/n$	$H_1=h_1$	h_1*100
x_2	f_2	$F_2=f_1+f_2$	$h_2 = f_2/n$	$H_2=h_1+h_2$	h_2*100
…	…	…	…	…	…
x_j	f_j	$F_j=f_1+\ldots+f_j$	$h_j = f_j/n$	$H_j=h_1+\ldots+h_j$	h_j*100
…	…	…	…	…	…
x_k	f_k	$F_k=n$	$h_k = f_k/n$	$H_k=1$	h_k*100
TOTAL	n		1		100%

3.3.2 Propiedades de tabla de frecuencias

1. Frecuencias absolutas

   - $\sum_{i=1}^{k} f_i = f_1 + f_2 + \cdots + f_k = n$, Donde: "n" es el total de observaciones.
   - $0 \le f_i \le n, \quad \forall i = 1, 2, ..., k$
   - $f_i \le F_i \le n$
   - $F_k = n$
   - $F_i = f_i + F_{i-1}$

2. Frecuencias relativas

   - $\sum_{i=1}^{k} h_i = h_1 + h_2 + \cdots + h_k = 1$
   - $0 \le h_i \le 1, \quad \forall i = 1, 2, ..., k$
   - $h_i \le H_i \le 1$
   - $H_k = 1$
   - $H_i = h_i + H_{i-1}$

3. La frecuencia absoluta acumulada correspondiente a un valor de la variable se obtiene sumando la frecuencia absoluta acumulada del valor anterior, con la frecuencia absoluta del dato.

Ejemplo 3.1.
Un investigador estaba interesado en estudiar la renta económica (en soles al mes) de un grupo de personas en función de su sexo, edad, y nivel cultural. Los datos de 22 personas que recogió son los siguientes:

Sujeto	*Sexo*	*Edad (en años)*	*Nivel Cultural*	*Renta Económica (S/.)*
1	hombre	21	Bajo	200
2	hombre	27	Muy alto	302
3	hombre	39	Alto	279
4	mujer	25	Bajo	89
5	mujer	47	Normal	95
6	mujer	32	Normal	102
7	mujer	43	Alto	104
8	hombre	34	Muy bajo	312
9	mujer	25	Normal	35
10	hombre	23	Muy bajo	100
11	hombre	33	Muy alto	515
12	hombre	23	Muy bajo	110
13	hombre	34	Normal	287
14	hombre	45	Bajo	198
15	mujer	38	Bajo	99
16	mujer	19	Normal	101
17	mujer	25	Alto	611
18	hombre	39	Normal	209
19	hombre	25	Normal	340
20	hombre	26	Normal	597
21	hombre	21	Bajo	99
22	mujer	35	Muy alto	245

- Introduzca los datos y defina las variables.
- Se desea conocer cuantas mujeres participan en el estudio.
- ¿Cuántas personas tiene nivel cultural Alto?

Para introducir los datos considere las siguientes especificaciones en la definición de las variables, convirtiendo las variables cualitativas de *sexo* y *nivel cultural* en valores numéricos, como se muestra a continuación:

Tabla 3.1
Descripción de variables

VARIABLE	NOMBRE	TIPO	VALORES
Sexo	SEXO	Numérico	1: hombre 2: mujer
Edad	EDAD	Numérico	----
Nivel cultural	NCUL	Numérico	1: Muy bajo 2: Bajo 3: Normal 4: Alto 5: Muy alto
Renta económica	RECO	Numérico	----

En este ejemplo aplicaremos otra forma de importar datos de un archivo ***csv***, el mismo que esta con denominación **RENTA.CSV**, usando el siguiente script:

```
# importar datos desde archivo con extensión .csv
library(readr)
renta <- read_delim("renta.csv", delim = ";", escape_double
= FALSE, col_types = cols(SEXO = col_number(),=
col_number(), NCUL = col_number(),= col_number()), trim_ws =
TRUE)
View(renta)
```

Figura 3.1. Importación de datos.

Luego recategorizaremos valores de algunas variables como SEXO y NCUL a través de:

```
renta$SEXO=factor(renta$SEXO, levels=c(1,2), labels =
c("Hombre","Mujer"))
renta$NCUL=factor(renta$NCUL,levels = c(1,2,3,4,5), labels =
c("Muy bajo","Bajo","Normal","Alto","Muy alto"))

table(renta$SEXO)
table(renta$NCUL)
```

Figura 3.2. Recategorización de las variables SEXO y NCUL.

Los datos resultarán de la siguiente forma:

	SEXO	EDAD	NCUL	RECO
1	Hombre	21	Bajo	200
2	Hombre	27	Muy alto	302
3	Hombre	39	Alto	279
4	Mujer	25	Bajo	89
5	Mujer	47	Normal	95
6	Mujer	32	Normal	102
7	Mujer	43	Alto	104

Figura 3.3. Vista de variables recategorizadas.

Empezaremos a procesar los datos migrados, iniciando con la variable SEXO y poder obtener mediante el siguiente script algunos estadísticos descriptivos importantes:

```
# Instalación de la librería summarytools, que nos facilitará el
cálculo de dichos
# parámetros descriptivos.
install.packages("summarytools")
library(summarytools)

sex <- freq(renta$SEXO, exclude='NA')
sex


RESULTADOS:
Frequencies
renta$SEXO
Type: Factor

               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
------------ ------ --------- -------------- --------- --------------
      Hombre     13     59.09          59.09     59.09          59.09
       Mujer      9     40.91         100.00     40.91         100.00
        <NA>      0                               0.00         100.00
       Total     22    100.00         100.00    100.00         100.00
```

Figura 3.5. Tabla de frecuencias de la variable SEXO

La Figura 3.5 muestra una pequeña tabla de resumen de la variable SEXO, en cuanto al número total de los casos válidos y el número total de casos perdidos (*missings*), esto para poder verificar el correcto ingreso de los mismos. La tabla de frecuencias de la Figura 3.4 nos muestra cinco columnas:

Válidos: Se muestra la etiqueta de cada categoría, en el caso de que este definida.
Frecuencia: Frecuencia absoluta para cada categoría.
Porcentaje: Frecuencia relativa, incluyendo los valores perdidos.
Porcentaje válido: Frecuencia relativa, eliminando de la muestra los valores perdidos (*missings*).
Porcentaje acumulado: Frecuencia relativa acumulada, eliminado valores perdidos.

Para saber cuántas mujeres participan en el estudio solo basta fijarse en la tabla de frecuencias de la variable sexo en la Figura 3.4, en la cual la columna de frecuencias nos indica que hay 9 mujeres ($f_2 = 9$) y que representa el 40.9% ($h_2\% = 40.9$) del total de participantes. Así mismo, se tiene 13 hombres ($f_1 = 13$) que representan el 59.1% (h1%=59.1) del total.

Ahora para mostrar la tabla de frecuencias de la variable NCUL (Nivel Cultural), solicitamos a través del mismo procedimiento anterior:

```
nc <- freq(renta$NCUL)
nc

RESULTADOS:
renta$NCUL
Type: Factor

                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
-------------- ------ --------- -------------- --------- --------------
      Muy bajo      3     13.64          13.64     13.64          13.64
         Bajo      5     22.73          36.36     22.73          36.36
       Normal      8     36.36          72.73     36.36          72.73
         Alto      3     13.64          86.36     13.64          86.36
      Muy alto     3     13.64         100.00     13.64         100.00
         <NA>      0                               0.00         100.00
        Total     22    100.00         100.00    100.00         100.00
```

Figura 3.5. Tabla de frecuencias de la variable RECO

En la tabla de frecuencias de la Figura 3.5 observamos que 3 personas pertenecen a un nivel cultural *Alto*, y representan el 13.6% ($h_4\%$ = 13.6) del total de participantes. También es importante señalar entre los valores que nos muestra la tabla de frecuencias, los porcentajes acumulados, por ejemplo, $H_4\%$ = 86.4, nos indica que el 86.4% de los participantes se encuentran en un nivel cultural de *Muy bajo* a nivel cultural *Alto*.

Ahora también podemos solicitar la tabla de frecuencias para las variables cuantitativas EDAD y RECO, de igual manera que se realizó para las variables cualitativas SEXO y NCUL. Debido a que los valores de la variable RECO son muy diversos y se repiten muy pocas veces, únicamente solicitaremos para ejemplificar la variable EDAD y la atabla de frecuencias de esta variable se muestra.

```
table(renta$EDAD)
ed <- freq(renta$EDAD)
ed

RESULTADOS:
Frequencies
renta$EDAD
Type: Numeric

             Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
---------- ------ --------- -------------- --------- --------------
         19      1      4.55           4.55      4.55           4.55
         21      2      9.09          13.64      9.09          13.64
         23      2      9.09          22.73      9.09          22.73
         25      4     18.18          40.91     18.18          40.91
         26      1      4.55          45.45      4.55          45.45
         27      1      4.55          50.00      4.55          50.00
         32      1      4.55          54.55      4.55          54.55
         33      1      4.55          59.09      4.55          59.09
         34      2      9.09          68.18      9.09          68.18
         35      1      4.55          72.73      4.55          72.73
         38      1      4.55          77.27      4.55          77.27
         39      2      9.09          86.36      9.09          86.36
         43      1      4.55          90.91      4.55          90.91
         45      1      4.55          95.45      4.55          95.45
         47      1      4.55         100.00      4.55         100.00
```

Figura 3.6. Tabla de frecuencias de la variable cuantitativa EDAD

A diferencia de las tablas de frecuencias de las variables cualitativas, en la primera columna nos muestra los valores de la variable asociado a su respectiva frecuencia (columna 2).

Es importante aclarar que frecuentemente, sobre todo en variables de tipo continuo, el número de valores distintos que toma la variable es demasiado alto; y las frecuencias asociadas son muy bajas, como el caso de la variable EDAD. Una tabla de frecuencias construida en estas condiciones, no presenta ninguna utilidad.

3.4 DISTRIBUCIÓN DE FRECUENCIAS AGRUPADAS EN INTERVALOS

Como se analizó antes, usualmente los valores de los datos no permiten un agrupamiento de ellos en una tabla de frecuencias simple, debido a que se encuentran distribuidos a través de todo el recorrido y el número de veces que se repite cada observación no es significativo en todos los casos, y en la mayoría de ellos su frecuencia es baja. Para mayor comodidad en el tratamiento de la información, parece aconsejable agrupar esos valores en intervalos, teniendo en cuenta que lo que ganamos en manejabilidad lo perdemos en información de la distribución.

En la agrupación en intervalos hay que tener en cuenta tres aspectos (Sanchez, 2012):

a) Que el máximo de información se obtiene en la recogida de datos y que ésta se pierde al agrupar en intervalos.
b) Las distribuciones agrupadas en intervalos no se presentan realmente así, sino que es el investigador el que las agrupa para manejar mejor los datos.
c) Al agrupar hay que tener en cuenta las frecuencias.

A continuación, se muestran procedimientos comunes para la creación de intervalos de clase en la presentación de los datos en tablas de frecuencia.

i. **Rango (R):** Es llamado también “recorrido de los datos”. Está definido por:

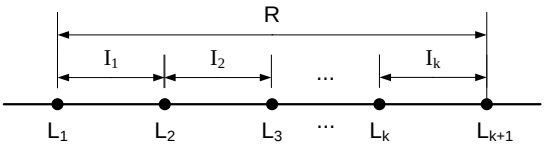
$$R = \text{máx}\{x_i/i=1, 2, \dots, k\} - \text{mín}\{x_i/i=1, 2, \dots, k\}$$

ii. Número de Intervalos de Clase (K): Consiste en dividir el rango en un número de intervalos de clase del mismo tamaño de preferencia.

Por ser los intervalos de clase (I_i) una partición de R se debe cumplir:

a) $I_i \cap I_j = \phi, \quad \forall i \neq j$

b) $\bigcup_{i=1}^{k} I_i = R$



$$I_i = [L_i , L_{i+1}\rangle, \; i = 1, 2, \dots, k$$

El valor entero de K, fundamentalmente, depende del investigador, aunque se recomienda usar los siguientes criterios:

1. $K = 5$, si $n \leq 25$ (n es el tamaño de la muestra)
2. $K = \sqrt{n}$, si $n > 25$
3. Formula de Sturges, cuando $n > 25$

$$K = 1 + 3.22 * \log_{10}(n)$$

Observaciones:

- Si K es un número real con decimales, entonces se recomienda redondear al entero inmediato superior.
- Es recomendable tener entre 5 y 20 intervalos de clase. Por consiguiente, entre más datos se tenga, más intervalos de clase deben considerarse.
- No hay una formula exacta para calcular el número de intervalos de clase, este número es determinado por tentativas y aproximaciones.

iii. Amplitud de Intervalo o Clase (C_i): Llamaremos amplitud del intervalo, a la diferencia entre sus extremos superior e inferior el cual está definido por:

$$C_i = L_i - L_{i-1}$$

Esta amplitud puede ser constante para todos los intervalos, o variable, aunque es más cómodo que sea constante. Entonces podemos utilizar la siguiente formula:

$$C = \frac{R}{K}$$

Observación: Cunado se trabaja con variables cuantitativas discretas, es recomendable redondear a su número inmediato superior.

Cuando un investigador decide agrupar los datos en intervalos se encuentra con dos cuestiones iniciales:

1ª.- ¿Cómo se debe tomar la amplitud, constante o variable?
2ª.- ¿Cuántos intervalos conviene tomar?

La respuesta a estas preguntas depende de la naturaleza del problema, y aunque hay muchas reglas escritas en los textos de estadística, en la práctica nos brindan muy poca ayuda.

iv. Marcas de Clase (X_i): Por último, cabe destacar que tomaremos como representante de cada intervalo su punto medio, que denominaremos marca de clase. Así la marca de clase del intervalo $[L_{i-1}, L_i\rangle$ será:

$$X_i = \frac{L_{i-1} + L_i}{2}, \quad i = 1, 2, 3, \ldots, k.$$

Ejemplo Nº 3.2.

Se desea conocer la distribución de frecuencias de la variable EDAD del Ejemplo Nº 3.1, de acuerdo a los resultados, agrupe la variable en intervalos de clase, si es que es aconsejable.

Como se pudo notar en el ejemplo Nº 3.1, al solicitar el cuadro de frecuencias de la variable EDAD (ver Figura 3.6), esta muestra que el número de valores distinto de la variable es muy alto para un numero pequeño de datos (n = 22) y las frecuencias asociadas son muy bajas, por tanto, es aconsejable agrupar los datos en intervalos de clase.

Primeramente, realizaremos la agrupación de la variable en intervalos de clase con los procedimientos desarrollados anteriormente.

i) Rango o recorrido de los datos:

$$R = \text{máx}\{x_i/i=1, 2, \dots, k\} - \text{mín}\{x_i/i=1, 2, \dots, k\} = 47 - 19 = 28$$

ii) Número de intervalos a través de la fórmula
iii) de Sturges:
m

$$K = 1 + 3.22 * \log_{10}(22) = 5.32 \approx 6$$

iv) Amplitud de intervalo:

$$C = \frac{R}{K} = \frac{28}{6} = 4.67 \approx 5$$

v) Tabla de frecuencias:

I_i	X_i	f_i	F_i	h_i	H_i	h_i%
[19 , 24 ⟩	21.5	5	5	0.227	0.227	22.7
[24 , 29 ⟩	26.5	6	11	0.273	0.500	27.3
[29 , 34 ⟩	31.5	2	13	0.091	0.591	9.10
[34 , 39 ⟩	36.5	4	17	0.182	0.773	18.2
[39 , 44 ⟩	41.5	3	20	0.136	0.909	13.6
[44 , 49 ⟩	46.5	2	22	0.091	1.000	9.10
TOTAL		22		1		100

Para el proceso de cálculo automático de la tabla de distribución de frecuencias para datos agrupados en R aplicaremos el siguiente script:

```
# Cálculo de la distribución de frecuencias, donde:
# f= frecuencia absoluta
# rf= frecuencia relativa
# rf(%) frecuencia relativa porcentual
# cf= frecuencia acumulada
# cf(%)=frecuencia acumulada porcentual

dist <- fdt(renta$EDAD, breaks='Sturges')
dist

RESULTADOS:
     Class limits f   rf rf(%) cf  cf(%)
  [18.81,23.587) 5 0.23 22.73  5  22.73
 [23.587,28.363) 6 0.27 27.27 11  50.00
  [28.363,33.14) 2 0.09  9.09 13  59.09
  [33.14,37.917) 3 0.14 13.64 16  72.73
 [37.917,42.693) 3 0.14 13.64 19  86.36
  [42.693,47.47) 3 0.14 13.64 22 100.00
```

Como podemos observar en los resultados anteriores y en comparación con el método de cálculo manual, la diferencia es mínima además para efectos de interpretación y análisis y en materia de organización de datos es posible identificar en la teoría diferentes métodos bajo la consideración de los objetivos planteados al inicio del análisis.

Capítulo 4

Representaciones Gráficas

4.1 INTRODUCCIÓN

La información proporcionada por las tablas de distribución de frecuencias es bastante completa, pero no todos los lectores alcanzan a comprenderla o no disponen del tiempo suficiente para analizarla. Además, en la experiencia del lector, al comenzar a leer un determinado artículo (científico o no), su vista se dirige primero al título, luego a los gráficos y, finalmente, a las tablas.

De este modo, las representaciones gráficas constituyen uno de los principales y más sencillos métodos de exponer la información, por su capacidad de impactar fácilmente al lector con muy poco esfuerzo por su parte, brindando una información rápida y global de los datos, siendo útiles incluso al

investigador, pues le permiten tener una idea general de los resultados y a veces, sugerir nuevas hipótesis.

R en particular tiene una amplia variedad de gráfico y sus posibilidades de personalización son flexibles mediante el código apropiado, además unas de las librerías más conocidas como **ggplot2** se podrá personalizar gráficos impresionantes.

4.2 TIPOS DE REPRESENTACIONES GRÁFICAS

4.2.1 Gráficos para Variables Cualitativas

Los gráficos más usuales para representar variables cualitativas de tipo nominal u ordinal son los siguientes:

a) Gráfico de barras. - Esta forma de representación gráfica es propia de las distribuciones que tienen muchas observaciones, pero pocos valores distintos de la variable. Se representan en un sistema de ejes de coordenadas cartesianas, en el eje de abscisas los valores de la variable, y en el de ordenadas las frecuencias. Posteriormente, sobre cada valor de la variable se levanta una barra vertical de altura proporcional a la frecuencia, ya sea absoluta o relativa (Sanchez, 2012).

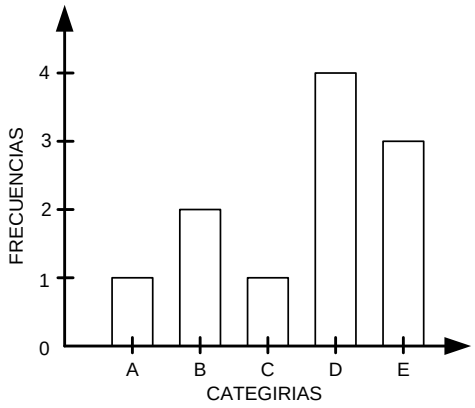


Figura 4.1. Gráfico de Barras para una variable cualitativa

Observación**:** Los gráficos de barras suelen utilizarse también en variables de tipo cuantitativo discretas.

Ejemplo Nº 4.1: Solicitar el gráfico de barras para la variable *Nivel Cultural* (NCUL) del Ejemplo Nº 3.1.

Anteriormente ya solicitamos la tabla de distribución de frecuencias, ahora vamos a completar el análisis realizando un gráfico de barras para la variable NCUL mediante el siguiente script:

```
# Gráfico de barras
barplot(table(renta$NCUL), main="Nivel cultural",
ylab="frecuencia", xlab="Nivel cultural", col=4)
```

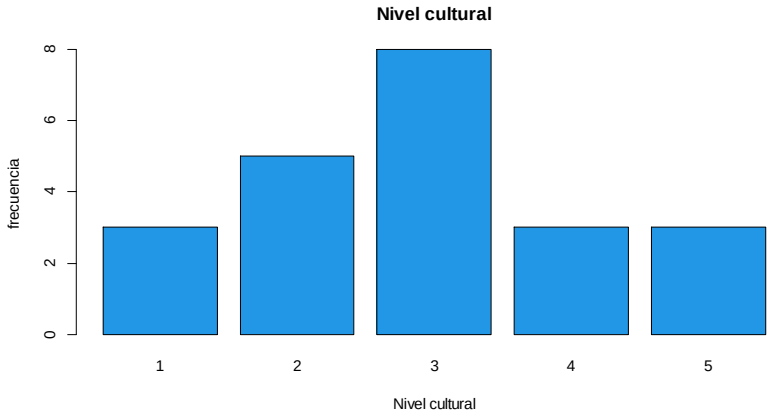


Figura 4.3. Gráfico de frecuencias de la variable NCUL.

b) Gráfico de sectores. - Es la representación gráfica de los datos estadísticos en un círculo, por medio de sectores circulares cuyo ángulo central coincida con la frecuencia absoluta (no se puede utilizar para acumuladas) o relativa del elemento, representando, mediante colores o incluyendo dentro de dicho sector el nombre de la clase o elemento a representar. Vale tanto para frecuencias agrupadas, como no agrupadas.

Previamente hay que calcular los grados que corresponde a cada elemento multiplicando la frecuencia correspondiente a cada dato por el cociente entre 360º y el total de datos:

$$\alpha_i = \frac{f_i}{n} * 360^o, \rightarrow \text{para frecuencias absolutas}$$

$$\alpha_i = h_i * 360^o, \rightarrow \text{para frecuencias relativas}$$

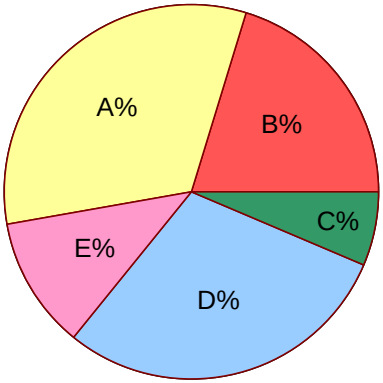


Figura 4.4. Gráfico de Sectores circulares

Observación**:** Los gráficos de sectores circulares se utiliza principalmente para comparar cada valor con el total, es decir, que la suma de los sectores es igual a 1 para frecuencias relativas o 100% para frecuencias relativas porcentuales.

Ejemplo Nº 4.2: Solicitar el gráfico de Sectores para la variable *SEXO* del Ejemplo Nº 3.1, el cual será mediante el script:

```
# Gráfico de sectores
proporciones <- table(renta$SEXO)
etiquetas <- c("Hombre", "Mujer")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
etiquetas <- paste(etiquetas, "%", sep="")
pie(table(renta$SEXO), labels=etiquetas, main="SEXO",
col=rainbow(length(etiquetas)))
       legend("topright", c("Hombre","Mujer"), cex = 0.8, fill
= rainbow(length(etiquetas)))
```

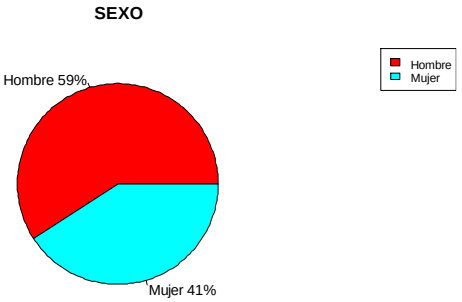


Figura 4.6. Gráfico de sectores de la variable SEXO

Nótese que el gráfico de sectores mostrado tiene ciertas particularidades en relación a la inserción de datos como la leyenda y etiquetas que se logran mediante el código adecuado.

4.2.2 Gráficos para Variables Cuantitativas

Para las variables cuantitativas, consideraremos dos tipos de gráficos, los que para realizarlos usan las frecuencias absolutas o relativas y los otros las frecuencias acumuladas (Bouza, 2017).

a) **Histogramas. -** Se construyen a partir de la tabla de frecuencias, representando sobre cada intervalo, un rectángulo que tiene a este segmento como base. El criterio para calcular la altura de cada rectángulo es el de mantener la proporcionalidad entre las frecuencias absolutas (o relativas) de cada intervalo y el área de los mismos.

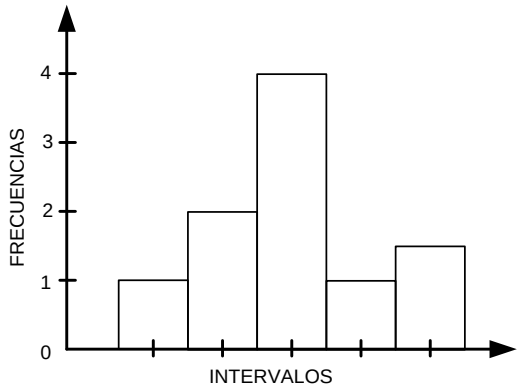


Figura 4.8. Histograma para una variable cuantitativa

b) **Polígono de frecuencias.** - Se construye fácilmente si tenemos representado previamente el histograma, ya que consiste en unir mediante líneas rectas los puntos del histograma que corresponden a las marcas de clase (puntos medios del intervalo).

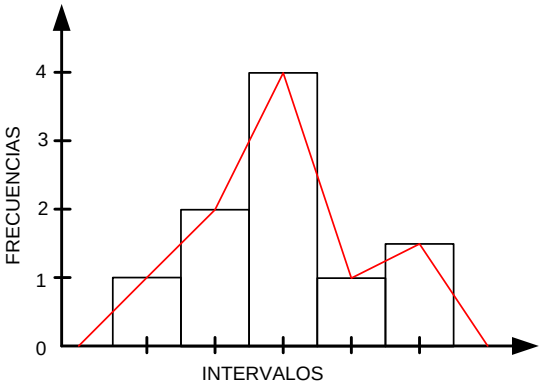


Figura 4.9. Polígono de frecuencias

Ejemplo Nº 4.3: Solicitar el histograma para la variable Renta Económica (*RECO*) del Ejemplo Nº 3.1, el cual lo solucionamos mediante el siguiente script:

```
hist(renta$RECO, main="Histograma", ylab="Frecuencia",
xlab="Renta económica", col=4)
```

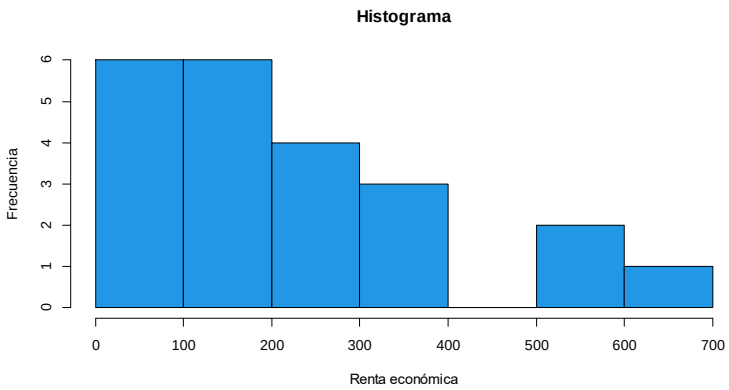


Figura 4.11. Histograma de la variable Renta Económica

En el histograma de la Figura 4.11 podemos observar que el programa R realiza una agrupación automática de los datos, para este caso con una amplitud interválica de 100, iniciando con una reta económica de cero (0) soles hasta 700 soles como máximo. Así mismo, como datos más resaltantes podemos ver que seis personas del grupo reciben una renta económica de S/.100 a S/.200 y no existen personas que perciban como renta la cantidad de S/.400 a S/500.

Capítulo 5

Medidas Descriptivas

5.1 INTRODUCCIÓN

En los capítulos anteriores, nos referimos a la clasificación, ordenación y presentación de datos estadísticos, limitando el análisis de la información a la interpretación porcentual de las distribuciones de frecuencia. No obstante, tras la elaboración de la tabla y su representación gráfica, en la mayoría de las ocasiones resulta más eficaz “condensar” dicha información en algunos números que la expresen de forma clara y concisa.

Los fenómenos en general no suelen ser constantes, por lo que sería necesario que junto a una medida que indique el valor alrededor del cual se agrupan los datos, se asocie una medida que haga referencia a la variabilidad que refleje dicha fluctuación. Por tanto, el objetivo de este capítulo consistirá en definir

algunos tipos de medidas (estadísticos o parámetros) que los sintetizan aún más.

Es decir, dado un grupo de datos organizados en una distribución de frecuencias (o bien una serie de observaciones sin ordenar), pretendemos describirlos mediante dos o tres cantidades sintéticas. En este sentido pueden examinarse varias características, siendo las más comunes(Ximénez & Revuelta, 2022):

- La tendencia central de los datos;
- La dispersión o variación con respecto a este centro;
- Los datos que ocupan ciertas posiciones.
- La simetría de los datos.
- La forma en la que los datos se agrupan.

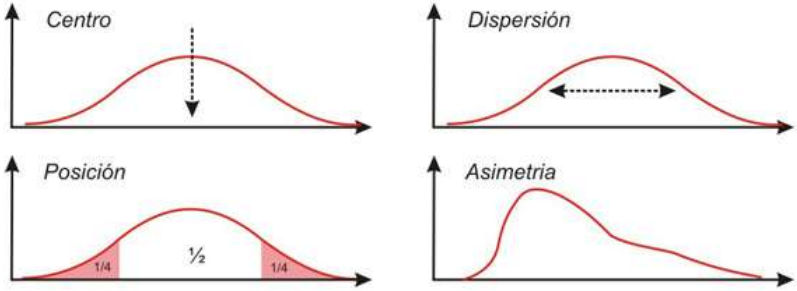


Figura 5.1. Medidas representativas de un conjunto de datos estadísticos

En este capítulo seguiremos trabajando con la opción ***Frecuencias***, el cual nos permite calcular algunos estadísticos básicos de tendencia central, de posición y de dispersión de una variable cuantitativa (numérica). Vamos a completar el análisis de las variable EDAD y RECO (Renta Económica) con algunos de estos estadísticos.

Es importante aclarar que el programa SPSS, realiza los cálculos de las medidas descriptivas para variables cuantitativas (discretas o continuas) con las fórmulas de datos no agrupadas.

5.2 MEDIDAS DE TENDENCIA CENTRAL

El análisis estadístico propiamente dicho, parte de la búsqueda de parámetros sobre los cuales pueda recaer la representación de toda la información. Las medidas de tendencia central, llamadas así porque tienden a localizarse en el centro de la distribución de frecuencias, son de gran importancia en el manejo de las técnicas estadísticas, sin embargo, su interpretación no debe hacerse aisladamente de las medidas de dispersión, ya que la representatividad de ellas está asociada con el grado de concentración de la información.

Las tres medidas más usuales de tendencia central son:

- La media,
- La mediana,
- La moda.

En ciertas ocasiones estos tres estadísticos suelen coincidir, aunque generalmente no es así. Cada uno de ellos presenta ventajas e inconvenientes que precisaremos más adelante.

1.6.1 La Media Aritmética

Habitualmente utilizamos la media aritmética. Cuando, por ejemplo, decimos que un determinado fumador consume una cajetilla de cigarrillos diario, no aseguramos que diariamente deba consumir exactamente los 20 cigarrillos que contiene un paquete, sino que es el resultado de la observación, es decir, dicho sujeto puede consumir 18, un día; 19 otro; 20, 21, 22; pero según nuestro criterio, el número de unidades estará alrededor de 20.

Por tanto, podemos definirla como el valor medio ponderado de la serie de datos. Se pueden calcular la media para datos no agrupados y agrupados.

Datos no agrupados

Se define como la suma de todos los valores de la distribución, dividida por el número total de observaciones. Es decir, sean x_1, x_2, x_3, … x_n valores de la variable X. La media aritmética esta dado por:

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \cdots x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Donde:

n : es el tamaño de la muestra

Datos agrupados

Si designamos por x_i al valor de la variable X, que se repite f_i veces, la media aritmética será:

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots x_k f_k}{n} = \frac{\sum_{i=1}^{k} x_i f_i}{n} = \sum_{i=1}^{k} x_i h_i$$

Donde:

$n = \sum_{i=1}^{k} fi$, tamaño de la muestra

f_i : frecuencia absoluta simple.

h_i : frecuencia relativa simple.

Ventajas y Desventajas de la Media:

Como ventajas de utilizar la media aritmética como un promedio para sintetizar los valores de la variable podemos citar las siguientes:

- Considera todos los valores de la distribución.
- Es siempre calculable (en caso de variable cuantitativa).
- Es única.

Como desventaja de la utilización de la media aritmética cabe mencionar que, a veces, puede dar lugar a conclusiones erróneas, cuando la variable presenta valores muy extremos, que influyen mucho en la media, haciéndola poco representativa.

1.6.2 La Mediana

Es el valor de la distribución que, una vez ordenados los valores de la variable de menor a mayor, deja igual número de frecuencias a su izquierda que a su derecha (un 50% de valores son inferiores y otro 50% son superiores), es decir, el valor que ocupa el lugar central. Puede entenderse también como aquel valor cuya frecuencia absoluta acumulada es n/2.

Sean x_1, x_2, x_3, … x_n valores muestrales de la variable X, tales que: $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_n$. Entonces, la mediana está dada por:

Datos no agrupados

- **Si "n" es impar:**
  Si la distribución está sin agrupar, y hay un número impar de términos, la mediana será el que ocupa la posición central.

$$M_e = x_{\left[\frac{n+1}{2}\right]}$$

- **Si "n" es par:**
  Pero si hay un número par de términos habría dos términos centrales y se toma como mediana la media aritmética de ellos.

$$M_e = \frac{x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}}{2}$$

Datos agrupados

En el caso de variables continuas, las clases vienen dadas por intervalos iguales, y aquí la fórmula de la mediana se complica un poco más (pero no demasiado) y está dado por:

$$M_e = \ell_m + \frac{A_m}{f_m}\left(\frac{n}{2} - F_{m-1}\right)$$

Donde:

ℓ_m : Límite inferior de la clase que contiene a la media.

A_m : Amplitud de clase que contiene a la media.

f_m : Frecuencia absoluta simple de la clase que contiene a la mediana.

F_{m-1} : Frecuencia absoluta acumulada de la clase inmediatamente inferior a la que contiene a la mediana.

n : Tamaño de la muestra.

Ventajas y Desventajas de la Mediana:

Como ventajas de la mediana podemos citar que no está influida por los valores extremos como en el caso de la media, y además tiene sentido en casos de distribuciones en escala ordinal (datos que pueden ser ordenados), siendo la medida más representativa de estos por describir la tendencia central de los mismos.

Como desventaja puede ser la determinación de ésta en los casos de variables agrupadas en intervalos(Sanchez, 2012).

1.6.3 La Moda

Como su nombre lo indica, es el valor más común (de mayor frecuencia dentro de una distribución). Una distribución puede tener una moda y se llama unimodal, dos modas y se llama bimodal, o varias modas y llamarse multimodal. Sin embargo, puede ocurrir que la información no posea moda (distribución amodal).

Datos no agrupados

Para calcular la moda, en el caso que la distribución no esté agrupada, se considera el valor de la variable que más veces se repite en una distribución de frecuencias, es decir, el que tiene mayor frecuencia absoluta.

Datos agrupados

Cuando la información se encuentra agrupada en intervalos de igual tamaño la moda se calcula con la siguiente expresión.

$$M_o = \ell_o + A_o\left(\frac{d_1}{d_1 + d_2}\right)$$

Donde:

ℓ_o : Límite inferior de la clase modal.

A_o : Amplitud de clase modal.

$d_1 = f_m - f_{m-1}$: Diferencia de la clase modal y la frecuencia anterior.

$d_2 = f_m - f_{m+1}$: Diferencia de la clase modal y la frecuencia siguiente.

Ventajas y Desventajas de la Moda:

Como ventajas de la moda cabe citar que cuando la distribución es de escala nominal (no susceptible de ordenación) es la medida más representativa, pues no es posible hacer operaciones con sus observaciones, y por tanto no se pueden calcular las otras medidas. Además, igual que la mediana, no viene influida por los valores extremos de la variable.

Como la desventaja cabe citar el modo de calcularla en los casos de variables agrupadas en intervalos y el hecho de que utiliza un único dato de la distribución(Vicente, 2016).

Relación entre media, mediana y moda

En el caso de distribuciones unimodales, la mediana está con frecuencia comprendida entre la media y la moda (incluso más cerca de la media). En distribuciones que presentan cierta inclinación, es más aconsejable el uso de la mediana. Sin embargo, en estudios relacionados con propósitos estadísticos y de inferencia suele ser más apta la media.

5.3 MEDIDAS DE POSICIÓN

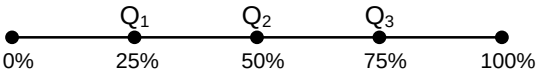
Estos valores no reflejan ninguna tendencia central, sino una posición de la distribución, dividiéndola a ésta en partes iguales. Nos ocuparemos ahora de ciertos parámetros posiciónales muy útiles en la interpretación porcentual de la información.

Las medidas de posición son llamadas también *cuantiles*. Cabe citar entre los de uso más frecuente: cuartiles, deciles y percentiles:

5.3.1 Cuartiles

Son tres valores que dividen a la distribución en cuatro partes iguales (k = 4), estando en cada una de ellas el 25% de sus observaciones. Se indican con Q_i.

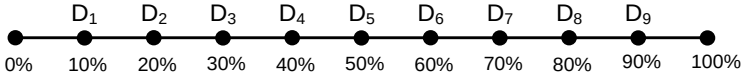
Gráficamente:



5.3.2 Deciles

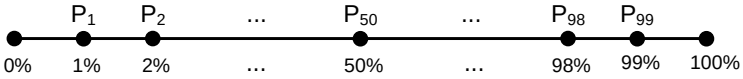
Son nueve valores que dividen a la distribución en diez partes iguales (k = 10), estando en cada una de ellas el 10% de las observaciones. Se indican por D_i.

Gráficamente:



5.3.3 Percentiles

Son noventa y nueve valores que dividen a la distribución en cien partes iguales (k = 100), dejando un 1% de las observaciones entre cada dos de ellos consecutivos. Se nombran por P_i.
Gráficamente:



Hay que tener en cuenta algunas relaciones entre ellos, como son:
$Me = Q_2 = D_5 = P_{50}$
$Q_1 = P_{25}$; $Q_3 = P_{75}$
$D_1 = P_{10}$; $D_2 = P_{20}$; $D_3 = P_{30}$; $D_4 = P_{40}$; $D_6 = P_{60}$

Datos no agrupados

El j-ésimo cuantil C_j es un valor que divide a un conjunto ordenado de datos en dos partes, el C% de ellos con valores inferiores o iguales a C_j y el (100-j)% con valores superiores a C_j. El j-ésimo cuantil particionado en k partes iguales ($C_{j/k}$) se calcula de la siguiente manera:

$$C_{j/k} = X_{\left(\frac{n+1}{k}\right)j} = X_{(E.d)}$$

Donde:

n : tamaño de la muestra
j : j-ésimo cuantil a estimar

Si la expresión $\left(\frac{n+1}{k}\right)j$ que indica la ubicación del estadístico de orden no resulta un valor entero, entonces el Cuantil a estimar será obtenido mediante la siguiente formula:

$$X_{(E.d)} = X_{(E)} + 0.d\left[X_{(E+1)} - X_{(E)}\right]$$

Datos agrupados

Cuando la j-ésima cuantila de la distribución de frecuencias es particionada en k partes iguales, se define:

$$C_{j/k} = \ell_i + \frac{A_i}{f_i}\left(\frac{jn}{k} - F_{i-1}\right);\ j = 1, 2, 3, \ldots, k\text{-}1.$$

Donde:

ℓ_i : Límite inferior de la clase cuantila.
A_i : Amplitud de clase de la clase cuantila.
f_i : Frecuencia absoluta simple de la clase cuantila.
F_{i-1} : Frecuencia absoluta acumulada hasta la clase inmediata anterior a la clase cuantila.
n : Tamaño de la muestra.

Ventajas y Desventajas de las Medidas de Posición:

Las Ventajas y desventajas de las medidas de posición son las mismas de la mediana.

Ejemplo Nº 5.1: Solicitar las medidas de tendencia central y de posición para la variable Renta Económica (RECO) del Ejemplo Nº 3.1. Realice la interpretación de cada medida obtenida.

R nos ofrece una seria de comandos que permiten obtener con los índices de tendencia central que detallamos a continuación:

- **Media:** media aritmética.
- **Mediana:** valor por debajo del cual se encuentra el 50% de los casos.
- **Moda:** valor que más se repite.
- **Suma:** suma de todos los valores.

Al ejecutar el siguiente script aparecerá el valor de los índices mencionados:

```
# Medidas descriptivas
# Resumen
summary(renta$RECO)

# Media aritmética
mean(renta$RECO)

# Mediana
median(renta$RECO)

# Moda
# Calcular la moda
moda <- names(sort(-table(renta$RECO)))[1]
# Imprimir la moda
cat("La moda es:", moda)

# Suma
sum(renta$RECO)
```

Figura 5.3. Tabla de estadísticos de la variable RECO.

- La media o el promedio de la renta económica de las 22 personas es de 228.59 soles, debido a que contamos con datos discretos podemos realizar el redondeo, es decir, la media sería 229 soles.
- El 50% de las personas perciben una renta económica menor e igual a 199 soles (mediana).
- La renta económica más habitual percibido por las personas es de 99 soles (moda).

5.4 MEDIDAS DE DISPERSIÓN O CONCENTRACIÓN

En el análisis estadístico no basta el cálculo e interpretación de las medidas de tendencia central o de posición, ya que, por ejemplo, cuando pretendemos representar toda una información con la media aritmética, no estamos siendo absolutamente fieles a la realidad, pues suelen existir datos extremos inferiores y superiores a la media aritmética, los cuales, en honor a la verdad, no están siendo bien representados por este parámetro.

Así pues, para intentar medir la representatividad de una determinada medida debemos de cuantificar la separación de los valores de la distribución respecto de dicha medida. En ese sentido, resulta necesario que, para completar la información de un promedio (por ejemplo media aritmética), éste vaya acompañado de uno o varios coeficientes que nos midan el grado de dispersión de la distribución de la variable con respecto a él.

Distinguiremos dos tipos de medidas de dispersión: absolutas y relativas.

5.4.1 Medidas de Dispersión Absolutas

5.4.1.1 Recorrido o Rango

Mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el valor más bajo.

$$R = \text{máx}(x_i) - \text{mín}(x_i)$$

Si este recorrido es pequeño respecto al número de datos puede entenderse que existe poca dispersión.

Tiene el inconveniente de que se ve totalmente influenciado por los valores extremos (con los que se calcula).
No utiliza todas las observaciones (sólo dos de ellas); y este aumenta con el número de observaciones, o bien se queda igual. En cualquier caso nunca disminuye.

5.4.1.2 Rango Intercuartílico

Es la diferencia existente entre el tercer y el primer cuartel, y está dado por:

$$R_I = Q_3 - Q_1$$

En esta medida se suprimen el 25% superior e inferior de la distribución, y por lo tanto no se ve influenciado por los valores extremos, y nos indica la longitud del intervalo en el que están el 50% central de los valores.

En algunos casos se utiliza el recorrido semi-intercuartílico que se define como la mitad del recorrido intercuartílico.

$$R_{SI} = \frac{Q_3 - Q_1}{2}$$

5.4.1.3 Desviación Media

Esta medida de dispersión hace referencia a un promedio, cosa que no hacen las anteriores; puede entenderse como la media de las desviaciones de los datos de la variable respecto al promedio utilizado; no obstante, para evitar que las desviaciones positivas queden compensadas por las negativas y que esta desviación media resulte igual a 0, (no presenta dispersión) se utiliza el valor absoluto de la desviación de los datos respecto del promedio.

Así se definirá la desviación media respecto de la media como:

$$D_{\bar{X}} = \sum_{i=1}^{n} \frac{\left| x_i - \bar{X} \right|}{n}, \qquad \text{datos no agrupados}$$

$$D_{\bar{X}} = \sum_{i=1}^{k} \frac{\left|x_i - \bar{X}\right| f_i}{n}, \qquad \text{datos agrupados}$$

También se puede utilizar la desviación media respecto de la mediana como:

$$D_{M_e} = \sum_{i=1}^{n} \frac{\left|x_i - M_e\right|}{n}, \qquad \text{datos no agrupados}$$

$$D_{M_e} = \sum_{i=1}^{k} \frac{\left|x_i - M_e\right| f_i}{n}, \qquad \text{datos agrupados}$$

Las dos nos indicarían la dispersión de los datos respecto del promedio utilizado, en el caso de que ésta fuera grande el promedio sería poco representativo.

5.4.1.4 Varianza

Mide la distancia existente entre los valores de la serie y la media. Se define como la media de los cuadrados de las desviaciones de los valores de la variable respecto de la media aritmética, la varianza muestral se calcula por:

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2}{n-1}, \qquad \text{datos no agrupados}$$

$$S^2 = \frac{\sum_{i=1}^{k} f_i \left(x_i - \bar{X}\right)^2}{n-1}, \qquad \text{datos agrupados}$$

Se utiliza el cuadrado para lograr que todas las desviaciones sean positivas; nos indica la mayor o menor dispersión de los valores de la variable respecto de la media aritmética, y por lo tanto, su representatividad. Es decir, la varianza siempre será mayor que cero. Mientras más se aproxima a cero, más concentrados están los valores de

la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están.

Tiene el inconveniente de no venir expresada en las mismas unidades que la variable, sino en el cuadrado de las mismas, por ello se utiliza más la desviación típica o estándar.

5.4.1.5 Desviación Típica o Estándar

Se define como la raíz cuadrada positiva de la varianza, es decir:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n-1}}, \qquad \text{datos no agrupados}$$

$$S = \sqrt{\frac{\sum_{i=1}^{k} f_i\left(x_i - \overline{X}\right)^2}{n-1}}, \qquad \text{datos agrupados}$$

Al ser la raíz cuadrada de la varianza viene expresada en las mismas unidades que la variable, lo que la hace más apta como medida de dispersión que la varianza, siendo en la actualidad la más utilizada.

5.4.2 Medidas de Dispersión Relativas

Si desearíamos comparar la dispersión de dos distribuciones mediante alguna de las medidas de dispersión halladas antes, no podríamos efectuar tal comparación porque las distribuciones, en general, no vendrán dadas en las mismas unidades y tampoco porque los promedios en general también serán diferentes. Por ello, para poder comparar las dispersiones, es preciso definir medidas de dispersión adimensionales. Entre éstas se encuentra el coeficiente de variación de Pearson.

5.4.2.1 Coeficiente de Variación de Pearson

Generalmente nos interesa establecer comparaciones de la dispersión, entre diferentes muestras que posean distintas magnitudes o unidades de medida.

El coeficiente de variabilidad tiene en cuenta el valor de la media aritmética, para establecer un número relativo, que hace comparable el grado de dispersión entre dos o más variables, y se define como:

$$CV(\%) = \frac{S}{\overline{X}} 100$$

Este coeficiente es adimensional luego permite comparar las dispersiones de dos distribuciones diferentes. Obviamente, a mayor CV menor es la representatividad de media, pues la desviación típica será mayor comparada con la media.

Propiedades:

- Sólo se debe calcular para variables con todos los valores positivos. Todo índice de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva. De ahí que sólo debemos trabajar con variables positivas, para la que tenemos con seguridad que $\overline{X} > 0$.
- Es invariante a cambios de escala. Así por ejemplo el coeficiente de variación de una variable medida en metros es una cantidad adimensional que no cambia si la medición se realiza en centímetros.

Ejemplo Nº 5.2: Solicitar las medidas de dispersión para la variable Renta Económica (RECO) del Ejemplo Nº 3.1. Realice la interpretación de cada medida obtenida.

- **Desviación típica o estándar:** estimación de la variabilidad de las puntuaciones respecto a la media, expresada en las mismas unidades de desviación al cuadrado.

- **Amplitud (Rango):** diferencia entre el valor mínimo y el máximo.
- **Mínimo:** valor más pequeño.
- **Máximo:** valor más grande.
- **Error tipo de la media:** estimación de la variabilidad muestral de la media.

Ejecutamos el procedimiento plasmado en el script y aparecerá el valor de los índices mencionados (ver Figura 5.5):

```
# Medidas de dispersión o concentración
# Varianza
var(renta$RECO)

# Desviación típica
sqrt(var(renta$RECO))

# Rango
range(renta$RECO) # Valores extremos
max(renta$RECO) - min(renta$RECO)
```

Figura 5.5. Tabla de estadísticos de la variable RECO

- La desviación estándar de las rentas económicas percibidas es de 166.752 soles. Entonces podemos señalar que los valores de la renta varían respecto a la media en más o menos S/.166.752.
- La varianza de la renta económica es de 27806.348, este valor es alto lo que nos indica que los datos se están dispersos.
- La amplitud de renta económica es de 576 soles.
- El menor valor de renta percibido es de S/.35 y el que tiene mayor valor es de S/.611.

5.5 MEDIDAS DE FORMA

Para poder conocer una distribución no basta con conocer sus medidas de dispersión y de posición, sino que es necesario, en general, conocer algunos aspectos más de la misma. Las medidas de forma permiten conocer que forma tiene la curva que representa la serie de datos de la muestra.

Dado que la diversidad de comportamientos de las x_i de la distribución se hacía más evidente al realizar la representación gráfica, vamos a tratar de determinar

a continuación más medidas, según la "forma" de la representación; clasificaremos estas medidas en dos grupos: medidas de asimetría y medidas de curtosis o apuntamiento.

5.5.1 Medidas de Asimetría

Tienen por objeto establecer el grado de simetría (o asimetría) de una distribución sin necesidad de realizar la representación gráfica(Sanchez, 2012).

Entenderemos la simetría respecto al eje determinado por la media aritmética, de tal forma que diremos que una distribución es simétrica (no tiene sesgo) cuando los valores de la variable equidistantes de este valor central tengan la misma frecuencia, en caso contrario diremos que es asimétrica, siendo esta asimetría negativa o a izquierda si es más larga la rama de la izquierda, es decir, las frecuencias descienden más lentamente por la izquierda que por la derecha; análogamente llamaremos asimetría positiva o a derechas aquella en que la rama de la derecha es más larga, es decir las frecuencias descienden más lentamente por la derecha que por la izquierda, como se muestra en la Figura 5.6.

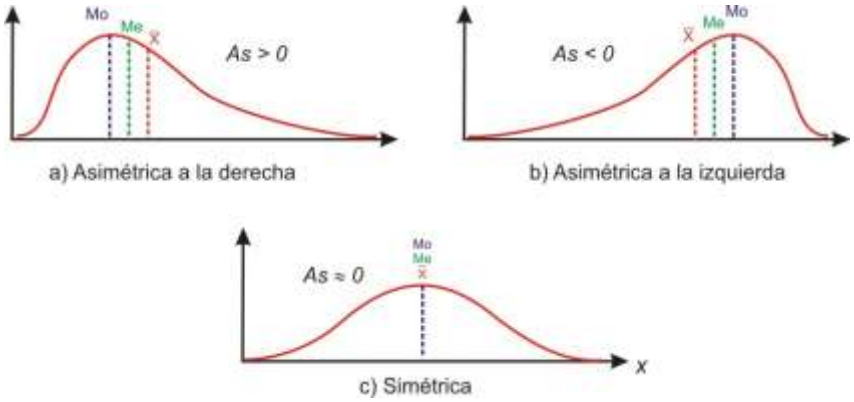


Figura 5.6. Deformación horizontal (asimetría) de la distribución.

Se conoce a toda una familia de estadísticos que ayudan a interpretar la asimetría, denominados índices de asimetría. El más utilizado es el *Coeficiente de Asimetría de Pearson (As)* que definimos a continuación.

$$A_S = \frac{3\left(\overline{X} - M_e\right)}{S}$$

Si: As ≈ 0, ⇒ la distribución es simétrica (Figura 5.2 - c)
As > 0, ⇒ la distribución es sesgado a la derecha (Figura 5.2 - a)
As < 0, ⇒ la distribución es sesgado a la izquierda (Figura 5.2 - b)

De forma general para determinar el coeficiente de asimetría o sesgo en datos no agrupados, podemos utilizar la siguiente formula:

$$\text{Sesgo} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{X}}{S} \right)^3$$

Donde:

$\overline{X}$	: Media aritmética
S	: Desviación estándar
n	: Tamaño de la muestra

5.5.2 Medidas de Apuntamiento

Estas medidas, aplicadas a distribuciones unimodales simétricas o con ligera asimetría, tratan de estudiar la distribución de frecuencias en la zona central, dando lugar a distribuciones muy apuntadas, o poco apuntadas.

Para estudiar el apuntamiento, debemos hacer referencia a una distribución tipo que consideraremos la distribución "Normal"; ésta corresponde a fenómenos muy corrientes en la naturaleza cuya representación gráfica es la campana de Gauss (Vicente, 2016).

Si una distribución tiene mayor apuntamiento que la normal diremos que es a) *leptocúrtica*, si tiene menor apuntamiento que la normal la llamaremos c)

platicúrtica, y a las que tengan igual apuntamiento que la normal las llamaremos b) *mesocúrticas*.

Veamos esto en las figuras siguientes:

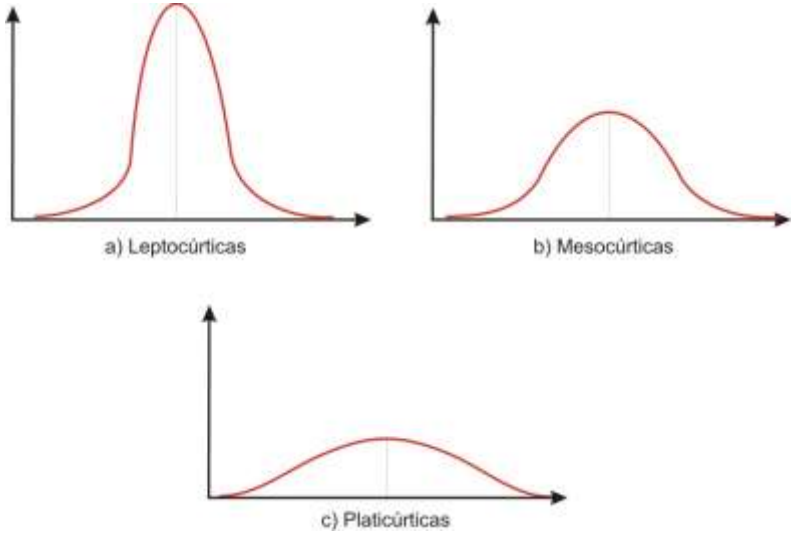


Figura 5.7. Deformación vertical (apuntamiento o curtosis) de la distribución.

El *Coeficiente de Apuntamiento o Curtosis (K)* esta definido por:

$$K = \frac{Q_3 - Q_1}{2\left(P_{90} - P_{10}\right)}$$

Donde:

Q_3 y Q_1 : Cuarteles de tercer y primer orden

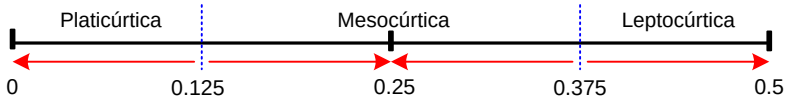
P_{90} y P_{10} : Percentiles de orden 90 y 10

Si: $K = 1/2, \Rightarrow$ la distribución es leptocúrtica

$K = 1/4, \Rightarrow$ la distribución es mesocúrtica

$K = 0, \Rightarrow$ la distribución es platicúrtica

En la práctica podemos utilizar la siguiente relación:



De igual manera podemos determinar el Coeficiente de Curtosis en datos no agrupados, con la siguiente formula:

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{X}}{S} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Donde:

$\overline{X}$: Media aritmética
S : Desviación estándar
n : Tamaño de la muestra

Por tanto:

K = 0, distribución mesocúrtica
K < 0, distribución platicúrtica
K > 0, distribución leptocúrtica

Ejemplo Nº 5.3: Solicitar las medidas de deformación horizontal (asimetría) y deformación vertical (curtosis) para la variable Renta Económica (RECO) del Ejemplo Nº 3.1. Realice la interpretación de cada medida obtenida.

- **Asimetría:** coeficiente de asimetría (deformación horizontal de la distribución).
- **Curtosis:** coeficiente de curtosis (deformación vertical de la distribución).

Al ejecutar el procedimiento del script, aparecerá el valor de los índices mencionados:

```
# install.packages("moments")
library(moments)

# Asimetría de PEARSON
(mean(renta$RECO)-mlv(renta$RECO, method="mfv")) /
mean(renta$RECO)

# Asimetria de Bowley
Q1 <- quantile(renta$RECO, 0.25)
Q2 <- quantile(renta$RECO, 0.50)
Q3 <- quantile(renta$RECO, 0.75)
(Q3+Q1-2*Q2)/(Q1+Q3)

# Coeficiente absoluto de Asimetría
(Q3+Q1-2*Q2) / sd(renta$RECO)

# Coeficiente de Asimetría de Fisher
library(e1071)
skewness(renta$RECO, na.rm=TRUE, type=3)

# coeficiente de apuntamiento o Curtosis
kurtosis(renta$RECO, na.rm=TRUE, type=3)
```

Figura 5.9. Tabla de estadísticos

Ejemplo Nº 5.4: Solicitar los estadísticos básicos para la variable Edad (*EDAD*) del Ejemplo Nº 3.1. Realice la interpretación de cada parámetro obtenido.

Para solicitar los estadísticos básicos para la variable *EDAD* ejecutamos el siguiente script:

```
# Estadístico descriptivos para la variable EDAD
summary(renta$EDAD)
var(renta$EDAD)                # Varianza
sqrt(var(renta$EDAD))          # Desviación estándar
```

Figura 5.12. Tabla de estadísticos descriptivos de la variable Edad.

- De las 22 personas tenemos que la edad mínima es de 19 años y la máxima es de 47 años.
- La media de las edades es de 30.86 ≈ 31 años, debido a que contamos con datos discretos podemos realizar el redondeo. También podemos decir que el promedio de edad de las 22 personas es de 31 años aproximadamente.

- La varianza de la variable EDAD es de 69.933 con una desviación estándar de 8.363, indicándonos estos valores que la distribución de la variable edad esta dispersa con respecto a la media.

Ejemplo Nº 5.5: Un egresado de Estadística de la UNA es contratado por una empresa, el jefe de la Oficina de rentas le encarga que realice un estudio sobre los impuestos que pagan los vecinos en un determinado distrito. Para realizar dicha labor el estadístico elabora un plan de trabajo dentro del cual tiene interés en evaluar dos variables:

NPIS : Número de pisos que tiene la vivienda.
PINP : Pago de impuestos municipales del año anterior (2006)

Los resultados luego de evaluar y ordenar los registros de 36 viviendas elegidas al azar se presentan a continuación:

Nº	*NPIS*	*PINP*	*Nº*	*NPIS*	*PINP*	*Nº*	*NPIS*	*PINP*
1	1	145.1	13	2	216.3	25	3	252.5
2	1	151.0	14	2	225.9	26	3	257.1
3	1	159.0	15	2	227.1	27	3	259.2
4	1	195.6	16	2	231.2	28	3	262.5
5	1	196.9	17	2	234.8	29	3	265.2
6	1	202.6	18	2	238.4	30	3	271.0
7	2	204.9	19	2	239.9	31	3	286.7
8	2	206.1	20	2	241.1	32	4	288.1
9	2	206.5	21	3	242.9	33	4	289.1
10	2	208.0	22	3	244.0	34	4	291.0
11	2	208.0	23	3	247.7	35	4	291.9
12	2	209.3	24	3	249.5	36	4	294.5

a. Solicite la tabla de frecuencias para la variable número de pisos que tiene la vivienda.
b. Elabore los gráficos adecuados para ambas variables.
c. Para la variable NPIS calcule la media, la mediana y la moda.
d. Calcule las medidas de tendencia central, posición (Q_1, D_7, P_{99}) y forma de la variable PINP. Además, halle el coeficiente de variación.

Análisis de la variable Número de Pisos de la Vivienda (NPIS)

Realizaremos el análisis individual de las variables, iniciaremos con la variable Número de Pisos (NPIS), sabemos que esta variable es cuantitativa discreta y, por lo tanto, el análisis es distinto a una variable continua. Podemos solicitar la tabla de frecuencias, el gráfico barras (variables discretas) y las medidas de tendencia central con el siguiente script:

```
#tabla de frecuencias (EDAD)
library(summarytools)
freq(impuesto$NPIS)
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
1	6	16.67	16.67	16.67	16.67
2	14	38.89	55.56	38.89	55.56
3	11	30.56	86.11	30.56	86.11
4	5	13.89	100.00	13.89	100.00
<NA>	0			0.00	100.00
Total	36	100.00	100.00	100.00	100.00

Figura 5.13. Tabla de frecuencias de la variable NPIS.

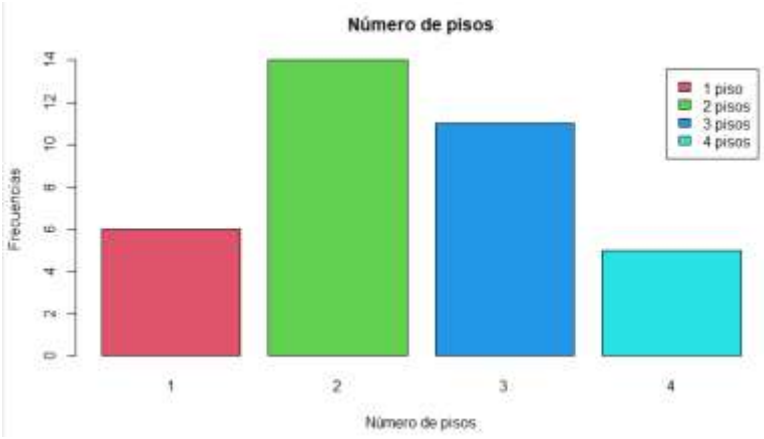


Figura 5.14. Gráfico de barras de la variable NPIS.

```
#Estadísticos descriptivos
round(summary(impuesto$NPIS),2)

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00 2.00     2.00      2.42   3.00     4.00
```

Figura 5.15. Tabla de estadísticos descriptivos de la variable NPIS.

La tabla de frecuencias de la Figura 5.13, como valores más resaltantes podemos ver que el 38.9% de las viviendas estudiadas poseen dos pisos y solamente el 13.9% de ellas poseen 4 pisos. Esto mismo podemos ver más claramente en la Figura 5.14 gráfico de barras del número de viviendas.
En la Figura 5.15 tenemos las medidas de tendencia central solicitados, de lo cual podemos indicar:

- El promedio de número de pisos de las viviendas estudias es de 2.42 ≈ 2.
- El 50% de las viviendas estudiadas tienen 2 o menos pisos.
- De las viviendas estudias la mayoría de ellas posee 2 pisos.

Nótese que las tres medidas son coincidentes en este caso, lo que no lleva a preguntar cuál de ellas es la más representativa, pues cuando las variables son discretas y no agrupadas como este caso, la moda y la mediana son las medidas más representativas.

Análisis de la variable Pago de Impuestos Municipales (PINP)

Ahora realizaremos el análisis de la variable cuantitativa continua PINP, y en este podremos solicitar las medidas de tendencia central, dispersión y forma, así mismo obtendremos el histograma para la distribución aplicando el siguiente script:

```
#Histograma para PINP
hist(impuesto$PINP, labels=TRUE, breaks="Sturges",
main="Histograma", ylab="Frecuencias", xlab="Pago de impuestos")
```

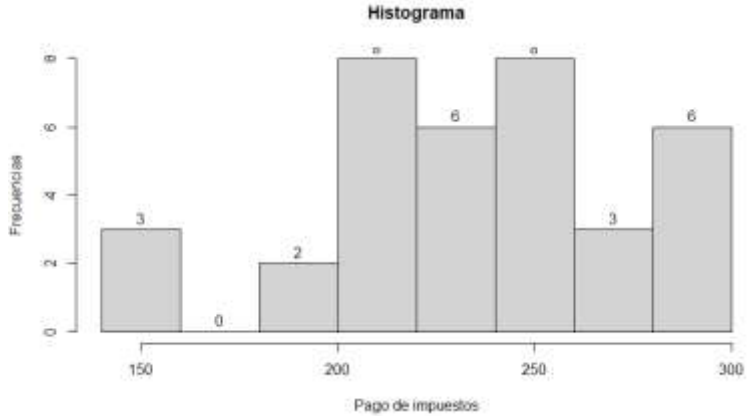


Figura 5.16. Histograma de la variable PINP.

```
summary(impuesto$PINP)
var(impuesto$PINP)
ds <- sqrt(var(impuesto$PINP))
ds
```

Figura 5.17. Tabla de estadísticos de la variable PINP.

El histograma de la Figura 5.16, tenemos que 8 viviendas realizan un pago de impuestos municipales de 200 a 216 aproximadamente igualmente para los realizan un pago de 233 a 250. Así mismo, este gráfico nos da una idea preliminar de la distribución de la variable PINP.

Respecto a los estadísticos que encontramos en la Figura 5.17, podemos señalar lo siguiente:

Medidas de tendencia central:

- El promedio del pago de impuestos municipales de la muestra es de 234.46 soles.
- El 50% de las viviendas pagan impuestos igual a 239.15 soles o menos y el 50% restante realizan un pago mayor a este (mediana).

Medidas de dispersión:

- La desviación estándar de los pagos de impuestos es de 39.05. Entonces podemos señalar que los valores del pago de impuestos varían respecto a la media en más-menos 39.05.
- La varianza del pago de impuestos de 1525.83, este valor no es muy alto con respecto a la media, por tanto, existen sospechas de que la variabilidad de lo los datos sea baja. Esto podemos corroborar con el histograma (Figura 5.16) donde se observa que los datos están alrededor de la media igual a 234.46
- La amplitud del pago de impuestos es de 149.4.
- El menor valor de pago realizado es de S/.145.10 y el que tiene mayor pago es de S/.294.50.

5.6 ALGUNOS GRÁFICOS ADICIONALES

5.6.1 Diagrama de Tallo y Hojas (Stem and Leaf)

El diagrama de tallo y hojas es un diagrama similar al histograma en el sentido de que muestra la distribución de frecuencias de una variable continua. La diferencia fundamental es que se construye utilizando los propios números de los valores de la variable y, a diferencia del histograma, permite recuperar la información original.

Utilizaremos un ejemplo sencillo para ilustrar la construcción de un diagrama de tallo y hojas. Consideremos los siguientes valores temperatura, en grados Fahrenheit:

77, 80, 82, 68, 65, 59, 61, 57, 50, 62, 61, 70, 69, 64, 67, 70, 62, 65, 65, 73, 76, 87, 80, 82, 83, 79, 79, 71, 80, 77

La temperatura mínima es 50 y la máxima es 87. Si hacemos intervalos de amplitud 10, comenzando en el valor 50 tendríamos 4 intervalos con frecuencias 3, 11, 9 y 7 respectivamente. Para estos datos seleccionaremos como tallo la cifra de las decenas y como hojas la cifra de las unidades. Cada tallo será una fila del gráfico y se corresponde con un intervalo de amplitud

10, en cada fila pondremos tantos números como observaciones en el intervalo, cada número escrito son las unidades de la observación correspondiente.

Temperaturas

Frecuencia	**Stem (**Decenas**)**	**Leaf (**Unidades**)**
3	5	079
11	6	11224555789
9	7	001367799
7	8	0002237

Obsérvese que el perfil del gráfico tiene la misma información que el histograma, pero, a diferencia de éste, es posible reconstruir los valores originales de la variable.

5.6.2 Diagrama de Cajas (Box – Plot)

Un box plot (o diagrama de cajas) es un método gráfico inventado por J. Tukey. Para construirlo calculamos primero el primer y el tercer cuartil (Q_1 y Q_3) y la mediana (M_e).

Dibujamos una caja que termine en Q_1 y Q_3 y situamos la mediana dentro de la caja. En el centro de los extremos de la caja añadimos líneas (whiskers) que van hasta los puntos más extremos que no son outliers (valores atípicos), esto es, los valores que están dentro de .15 veces el recorrido intercuartílico de los extremos de la caja. Los puntos que quedan más allá de 1.5 veces el recorrido intercuartílico se dibujan en el gráfico. Si hay varios puntos con el mismo valor, pueden dibujarse uno al lado del otro.

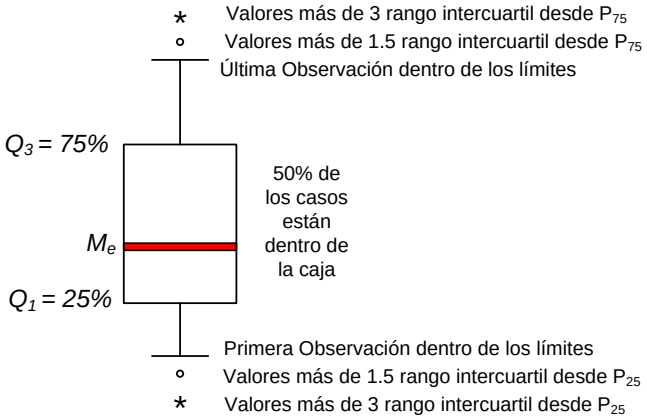


Figura 5.18. Ubicación de los parámetros en el diagrama de cajas.

La utilidad de los boxplots se basa en que permiten, mediante una simple inspección visual, tener una idea aproximada de la tendencia central (a través de la mediana), de la dispersión (a través del recorrido intercuartílico), de la simetría de la distribución (a través de la simetría del gráfico) y de los posibles valores atípicos (Sanchez, 2012).

Permiten, además, la comparación de varios grupos situando varios box-plots en el mismo gráfico, como se muestra en la siguiente figura:

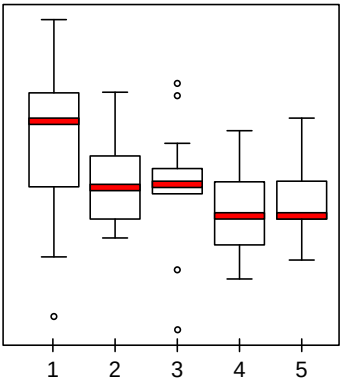


Figura 5.19. Análisis conjunto de diagramas de cajas.

Ejemplo Nº 5.6: Solicitar los grafico de Tallo y hojas y Diagramas de Caja para la variable Pago de Impuestos Municipales (PINP) del Ejemplo Nº 5.5. Realice la interpretación respectiva.

Es importante señalar que la primera y última fila del gráfico Tallo y hojas se utiliza para representar casos extremos (muy alejados del resto), si existen.

Utilizando el siguiente script podremos obtener el gráfico de tallo y hojas:

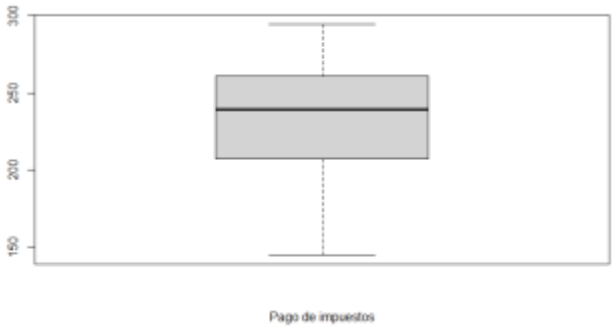


Figura 5.23. Diagrama de Caja de la variable PINP.

El ancho de la caja nos da una idea de la variabilidad de las observaciones y como se ve en la Figura 5.23 este presenta una variabilidad notoria de las observaciones.

De la posición de la mediana (239.15 soles) podemos deducir que la distribución es asimétrica negativa (sesgada a la izquierda), ya que la mediana está ubicada cerca al límite superior de la caja. Ahora encontraremos los valores de los límites:

- Mediana = 239.15
- Cuartel 1 (Q_1) = 206.88
- Cuartel 3 (Q_3) = 261.68
- Rango Intercuartilico (IQR) = 54.80
- Limite inferior: Mediana –(1,5*IQR) = 239.15 – (1.5*54.80) = 156.95. Primera observación dentro de este límite y patilla de caja: 145.1 soles.
- Limite superior: Mediana +(1,5*IQR) = 239.15 + (1.5*54.80) = 321.35. Última observación dentro de este límite y patilla superior de caja: 294.5 soles.
- No existen valores extremos.

Los gráficos de **Diagrama de cajas** son especialmente útiles para comprar la distribución de los valores entre diferentes grupos. En el ejemplo realizaremos la comparación con respecto a la variable Número de Pisos de la Vivienda, se tiene y esto lo lograremos mediante la ejecución del siguiente script:

```
boxplot(impuesto$PINP ~ impuesto$NPIS, data = impuesto, xlab="Número de pisos",
ylab="Pago de impuestos")
```

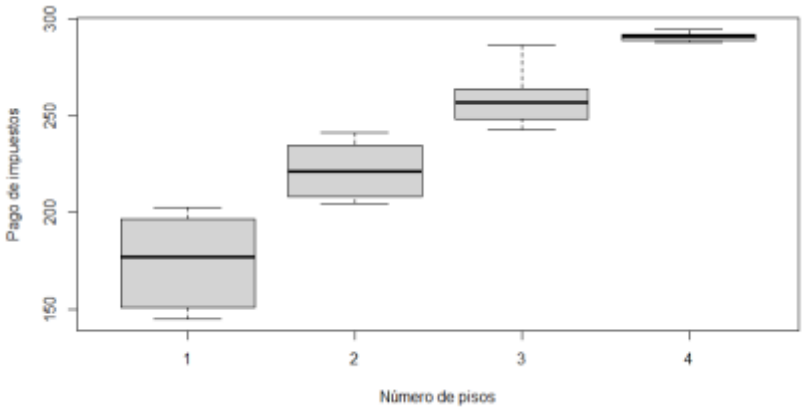


Figura 5.26. Diagrama de cajas conjunto PINP respecto a NPI.

El diagrama de cajas conjunto nos da la idea del comportamiento de las variables Pago de Impuestos Municipales (PINP) dentro de cada grupo de la variable Número de Pisos de la Vivienda (NPIS).

Las viviendas que cuentan con un piso realizan pago de impuestos muy variados (dispersos) a pesar que estas observaciones son muy pocas, a contrario de las viviendas que tienen 4 pisos estas realizan un pago de impuestos casi similares (no dispersos.

Finalmente observamos que a medida que la vivienda incrementa el número de pisos también incrementa el pago de impuestos municipales, pero al contrario la variabilidad de las observaciones dentro de cada grupo va descendiendo.

Capítulo 6

Análisis Exploratorio de Datos (AED)

6.1 INTRODUCCIÓN

La finalidad del Análisis Exploratorio de Datos (AED) es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes.

En este capítulo se va a dar una breve visión general de dicho conjunto de técnicas exponiendo, brevemente, cuál es su finalidad; así mismo, podremos responder a algunas interrogantes tales como (Figueras, 2003):

- ✓ ¿Existe algún tipo de estructura (normalidad, multimodalidad, asimetría, curtosis, linealidad, homogeneidad entre grupos, homocedasticidad, etc.) en los datos que voy a analizar?
- ✓ ¿Existe algún sesgo en los datos recogidos?
- ✓ ¿Hay errores en la codificación de los datos?
- ✓ ¿Cómo se sintetiza y presenta la información contenida en un conjunto de datos?
- ✓ ¿Existen datos atípicos (outliers)? ¿Cuáles son? ¿Cómo tratarlos?
- ✓ ¿Hay datos ausentes (missing)? ¿Tienen algún patrón sistemático? ¿Cómo tratarlos?

Estas interrogantes generalmente se encuentran cuando se realiza el análisis de un conjunto de variables y serán ilustradas con ejemplos.

6.2 ¿QUÉ ES EL ANÁLISIS EXPLORATORIO DE DATOS?

El Análisis Exploratorio de Datos (A.E.D.) es un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas. Para conseguir este objetivo el A.E.D. proporciona métodos sistemáticos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de los mismos, tratamiento y evaluación de datos ausentes (missing), identificación de casos atípicos (outliers) y comprobación de los supuestos subyacentes en la mayor parte de las técnicas multivariantes (normalidad, linealidad, homocedasticidad).

El examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.

6.3 ETAPAS DEL A.E.D.

Para realizar un A.E.D. conviene seguir las siguientes etapas (Figueras, 2003):

a. Preparar los datos para hacerlos accesibles a cualquier técnica estadística.
b. Realizar un examen gráfico de la naturaleza de las variables individuales a analizar y un análisis descriptivo numérico que permita cuantificar algunos aspectos gráficos de los datos.
c. Realizar un examen gráfico de las relaciones entre las variables analizadas y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.
d. Evaluar, si fuera necesario, algunos supuestos básicos subyacentes a muchas técnicas estadísticas como, por ejemplo, la normalidad, linealidad y homocedasticidad.
e. Identificar los posibles casos atípicos (outliers) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
f. Evaluar, si fuera necesario, el impacto potencial que pueden tener los datos ausentes (missing) sobre la representatividad de los datos analizados.

6.4 PREPARACIÓN DE LOS DATOS

El primer paso en un A.E.D. es hacer accesible los datos a cualquier técnica estadística. Ello conlleva la selección del método de entrada (por teclado o importados de un archivo) y codificación de los datos así como la de un paquete estadístico adecuado para procesarlos. Los paquetes estadísticos son conjuntos de programas que implementan diversas técnicas estadísticas en un entorno común. Algunos de los más utilizados son SPSS, SAS, SYSTAT, STATISTICA, STATA, MINITAB, S-PLUS, EVIEWS, STATGRAPHICS, MATLAB y últimamente R y Phyton.

La codificación de los datos depende del tipo de variable. Los paquetes estadísticos existentes en el mercado proporcionan diversas posibilidades (datos tipo cadena, numéricos, nominales, ordinales, etc). En este caso realizaremos la codificación como se trabajó en el capítulo I.
Finalmente, para aumentar la inteligibilidad de los datos almacenados, conviene asociar a la base de datos utilizada, un libro de códigos en el que se detallen los nombres de las variables utilizadas, su tipo y su rango de valores,

su significado, así como las fuentes de donde se han sacado los datos. Todos los paquetes anteriormente citados permiten esta posibilidad.

6.5 ANÁLISIS ESTADÍSTICO UNIDIMENSIONAL

Una vez organizados los datos, el segundo paso de un A.E.D. consiste en realizar un análisis estadístico gráfico y numérico de las variables del problema con el fin de tener una idea inicial de la información contenida en el conjunto de datos así como detectar la existencia de posibles errores en la codificación de los mismos.

El tipo de análisis a realizar depende de la escala de medida de la variable analizada. En la Figura 6.1 se sugieren las representaciones gráficas y resúmenes descriptivos numéricos más aconsejables para realizar dicho análisis. Esto pues no es más que un resumen de todos los capítulos anteriores estudiados hasta aquí. En dicha tabla se sobreentiende que las escalas más informativas pueden utilizar las medidas numéricas y representaciones gráficas de las escalas menos informativas además de las suyas propias (razón, intervalo, ordinal y nominal).

Escala de Medida	Representación Gráfica	Medidas	
		Tendencia Central	Dispersión
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Box-plot	Mediana	Rango Intercuartílico
Intervalo	Histogramas Polígono de frecuencias	Media	Desviación Típica
Razón		Media Geométrica	Coeficiente de Variación

Figura 6.1. Tabla de Medidas Descriptivas Numéricas y Representaciones Graficas aconsejadas en función de la escala de medida de la variable.

Ejemplo Nº 6.1.

Un estudio pretende conocer cuáles son los factores determinantes al momento de evaluar el rendimiento deportivo de los 28 alumnos asistentes a una clase de spinning. Se plantea como primer objetivo fundamental del estudio realizar un análisis exploratorio, mediante técnicas estadísticas.

Los datos se muestran a continuación:

alumno	edad	peso	imc	rendimie	sexo	gym
Alberto	44	90	30,4	72,7%	0	1
Jesús	43	60	22,6	84,7%	0	1
Paco	54	83	28,1	60,2%	0	1
Patricia	50	45	17,6	76,7%	1	1
José	34	90	27,8	72,6%	0	1
Rebeca	26	56	20,6	50,0%	1	1
Carla	35	60	20,8	86,4%	1	1
Blanca	35	53	19,2	94,2%	1	1
Lurdes	30	55	18,4	65,3%	1	1
María	32	63	20,8	67,0%	1	1
Carolina	31	58	20,5	73,8%	1	1
Daniel	27	65	21,2	89,1%	0	1
Feli	51	58	21,3	93,7%	1	2
Javier	35	82	24,0	73,5%	0	2
Paula	24	58	19,6	67,3%	1	2
Manuel	26	70	24,2	82,5%	0	2
Fernando	38	.	.	85,7%	0	2
Helena	29	50	18,4	89,3%	1	2
Pablo	35	85	23,8	64,9%	0	2
Nieves	40	60	22,0	64,5%	1	2
Juan Manuel	42	92	28,4	74,2%	0	2
Raquel	31	73	23,0	69,7%	1	2
Jesús	23	72	23,5	95,4%	0	2
Eduardo	34	78	24,9	64,5%	0	2
Estela	28	72	22,5	80,8%	1	2
Olga	29	75	26,0	73,6%	1	2
David	30	73	23,0	80,0%	0	2
Eugenia	30	65	22,0	100,0%	1	2

Este conjunto de datos contiene las siguientes variables:

- **Alumno:** nombre del deportista – NOMINAL
- **Edad:** edad en años cumplidos del deportista - ESCALA.
- **Peso:** peso del deportista medido en kgs - ESCALA.
- **IMC (índice de masa corporal):** índice que relaciona el peso y la estatura del deportista - ESCALA.

  $$IMC = (Peso/Estatura^2)$$

- **Rendimiento (rendimie):** porcentaje de la capacidad aeróbica máxima alcanzada por el deportista en un minuto, luego de realizados los sprints - ESCALA.
- **Sexo:** género del deportista - NOMINAL.
  0 = "Hombre"
  1 = "Mujer"
- **Gym:** gimnasio al que pertenece el deportista - NOMINAL.
  1 = "Gold"
  2 = "Hercules"

Para efectos de un mejor procesamiento y análisis realizaremos la re codificación mediante el siguiente script:

```
# Asignar los valores originales a la variable sexo
rend$sexo=factor(rend$sexo,levels=c(0,1),labels=c("Hombre","Muj
er"))

# Asignar los valores originales a la variable gym
rend$gym=factor(rend$gym,levels=c(1,2),labels=c("Gold","Hercule
s"))
```

Resultado:

	alumno	edad	peso	imc	rendimie	sexo	gym
1	Alberto	44	90	30.4	72.7	Hombre	Gold
2	Jesús	43	60	22.6	84.7	Hombre	Gold
3	Paco	54	83	28.1	60.2	Hombre	Gold
4	Patricia	50	45	17.6	76.7	Mujer	Gold
5	José	34	90	27.8	72.6	Hombre	Gold
6	Rebeca	26	56	20.6	50.0	Mujer	Gold
7	Carla	35	60	20.8	86.4	Mujer	Gold
8	Blanca	35	53	19.2	94.2	Mujer	Gold

A continuación, realizamos el análisis descriptivo (unidimensional) de las variables. Es recomendable realizar un análisis para variables cualitativas y otro para variables cuantitativas.

Variables Cualitativas

Los datos correspondientes a variables cualitativas de medida *nominal* se agrupan de manera natural en diferentes categorías o clases y se cuenta el número de datos que aparecen en cada una de ellas. Se suelen representar mediante diagrama de barras, sectores o líneas.

A continuación, solicitaremos las tablas de frecuencia y los grafico de sectores circulares para las siguientes variables: *sexo y gym*, y lo obtendremos mediante la ejecución del siguiente script:

```
# Análisis de la variable SEXO
freq(rend$sexo)
          Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
-------- ------ --------- -------------- --------- --------------
 Hombre     13     46.43          46.43     46.43          46.43
  Mujer     15     53.57         100.00     53.57         100.00
   <NA>      0                               0.00         100.00
  Total     28    100.00         100.00    100.00         100.00
```

Figura 6.6. Tabla de frecuencias de la variable SEXO.

Se observa que la mayor parte de los alumnos (53.6%) son mujeres que constituye el valor modal de la distribución de frecuencias, y que la otra parte son hombres (46.4%).

```
# Diagrama de sectores
proporciones <- table(rend$sexo)
etiquetas <- c("Hombre", "Mujer")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
etiquetas <- paste(etiquetas, "%", sep="")

pie(table(rend$sexo), labels=etiquetas, main="SEXO",
col=rainbow(length(etiquetas)))

legend("topright", c("Hombre","Mujer"), cex = 0.8, fill =
rainbow(length(etiquetas)))
```

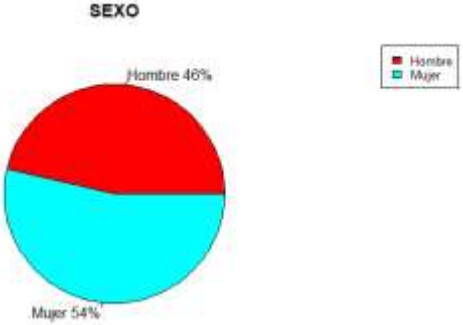


Figura 6.7. Diagrama de sectores circulares de la variable SEXO.

Para obtener los mismos resultados de la variable anterior, aplicaremos el siguiente script para la variable GYM

```
# Análisis de la variable GYM
freq(rend$gym)
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Gold	12	42.86	42.86	42.86	42.86
Hercules	16	57.14	100.00	57.14	100.00
<NA>	0			0.00	100.00
Total	28	100.00	100.00	100.00	100.00

Figura 6.8. Tabla de frecuencias de la variable GYM

Y para el gráfico de sectores ejecutaremos el siguiente script:

```
# Diagrama de sectores
proporciones <- table(rend$gym)
etiquetas <- c("Gold", "Hercules")
pct <- round(proporciones/sum(proporciones)*100)
etiquetas <- paste(etiquetas, pct)
etiquetas <- paste(etiquetas, "%", sep="")

pie(table(rend$sexo), labels=etiquetas, main="Gymnasio",
col=rainbow(length(etiquetas)))

legend("topright", c("Gold","Hercules"), cex = 0.8, fill =
rainbow(length(etiquetas)))
```

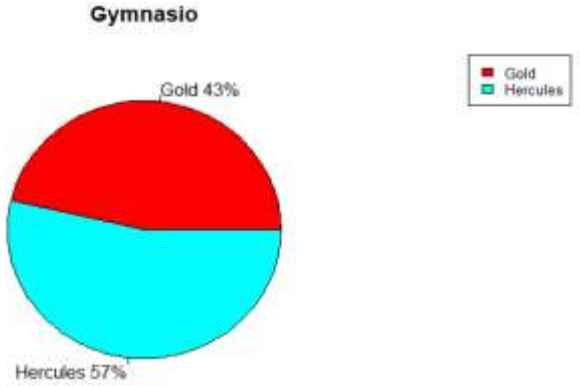


Figura 6.9. Diagrama de sectores circulares de la variable GYM.

Igualmente observamos para la variable gimnasio a la que asiste el alumno (GYM) que el 57% de los alumnos asisten al gimnasio "Hércules" y el 43% de ellos asisten al gimnasio "Gold".

Variables Cualitativas

Las variables cuantitativas discretas con un número pequeño de valores se tratarían de manera similar a las variables cualitativas antes descritas. En el ejemplo tenemos variables discretas, pero no de rango pequeño, por lo cual le realizaremos el análisis como a una variable continua.

Se recomienda obtener la tabla de frecuencias para variables *cuantitativas discretas* de rango pequeño, caso contrario obtener solamente el histograma y los estadísticos descriptivos.

Iniciaremos con gráficos de histogramas de las variables: edad, peso, imc y rendimie.

```
# Para la variable EDAD
hist(rend$edad, prob = TRUE,
     main = "Histograma con curva normal de la variable EDAD",
     ylab = "Frecuencia", xlab="Edad")
x <- seq(min(rend$edad), max(rend$edad), length = 40)
f <- dnorm(x, mean = mean(rend$edad), sd = sd(rend$edad))
lines(x, f, col = "red", lwd = 2)
```

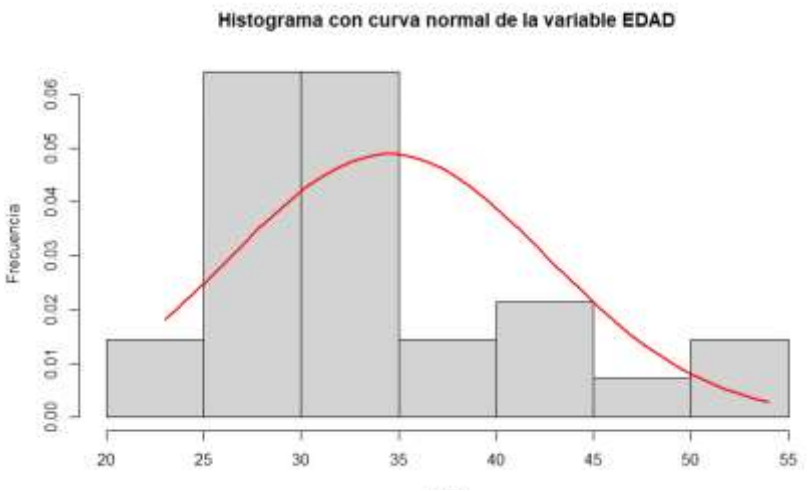


Figura 6.10. Histograma de la variable EDAD

```
# Para la variable PESO
hist(rend$peso, prob = TRUE,
     main = "Histograma con curva normal de la variable PESO",
ylab =
     "Frecuencia", xlab="Peso")
x <- seq(min(rend$peso), max(rend$peso), length = 40)
f <- dnorm(x, mean = mean(rend$peso), sd = sd(rend$peso))
lines(x, f, col = "red", lwd = 2)
```

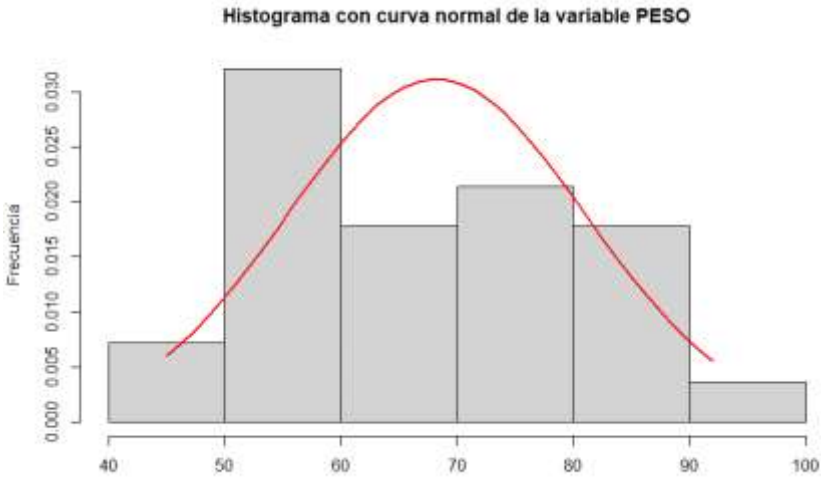


Figura 6.11. Histograma de la variable PESO

```
# Para la variable IMC
hist(rend$imc, prob = TRUE,
     main = "Histograma con curva normal de la variable IMC", ylab
= "Frecuencia", xlab="Índice de masa corporal")
x <- seq(min(rend$imc), max(rend$imc), length = 40)
f <- dnorm(x, mean = mean(rend$imc), sd = sd(rend$imc))
lines(x, f, col = "red", lwd = 2)
```

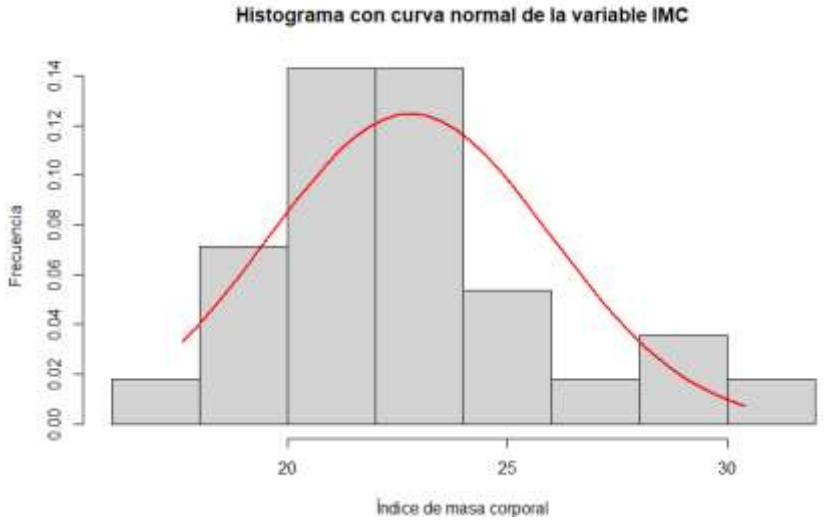


Figura 6.12. Histograma de la variable IMC

```
# Para la variable RENDIMIE
hist(rend$rendimie, prob = TRUE,
     main = "Histograma con curva normal de la variable
RENDIMIE", ylab = "Frecuencia", xlab="Rendimiento (%CAE)")
x <- seq(min(rend$rendimie), max(rend$rendimie), length = 40)
f <- dnorm(x, mean = mean(rend$rendimie), sd = sd(rend$rendimie))
lines(x, f, col = "red", lwd = 2)
```

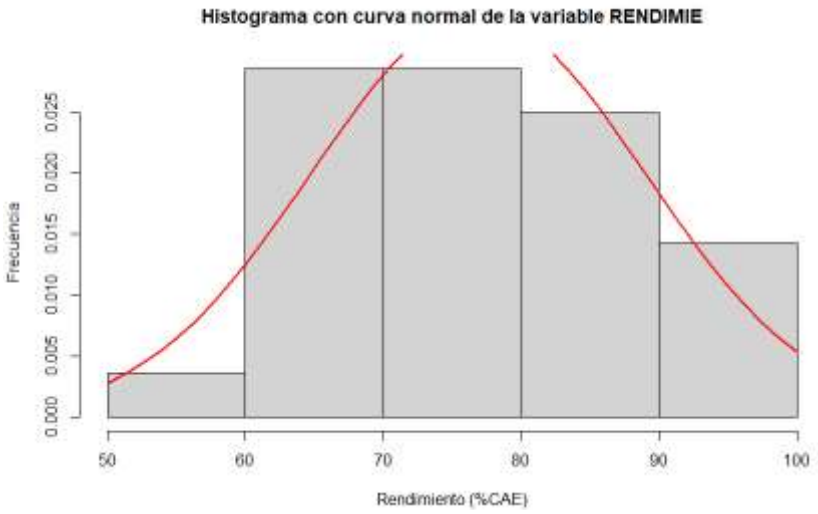


Figura 6.13. Histograma de la variable RENDIMIE

Si la variable analizada es continua o discreta con un elevado número de valores distintos se tabula como una distribución de frecuencias agrupadas y se representa gráficamente mediante histogramas, diagramas de tallos y hojas o box-plots con el fin de estudiar la forma de la distribución y analizar, en particular, la posible existencia de varias modas en la misma que pongan de manifiesto la presencia de diversos grupos homogéneos en la muestra.

El histograma de la Figura 6.13 presenta la forma estilizada, alrededor de un valor central. Este caso presenta un único máximo, por lo que podemos decir que es una distribución unimodal, al igual que el histograma de la Figura 6.15. Pero también observamos claramente que presentan asimetría en la distribución, por lo que sería necesario acompañar a este análisis con diagrama de cajas para ambas variables y encontrar posibles valores extremos que estén afectando a la distribución.

A continuación solicitamos los diagrama de cajas para la variable EDAD e IMC,m mediante el script:

```
# Diagrama de cajas para la variable EDAD
boxplot(rend$edad, horizontal=TRUE)
```

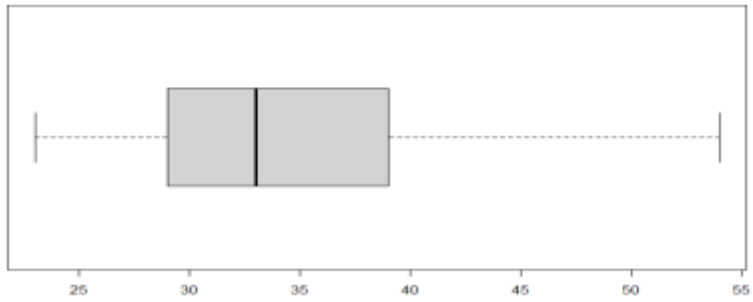


Figura 6.14. Diagrama de cajas de la variable EDAD.

```
# Diagrama de cajas para la variable EDAD
boxplot(rend$imc, horizontal=TRUE)
```

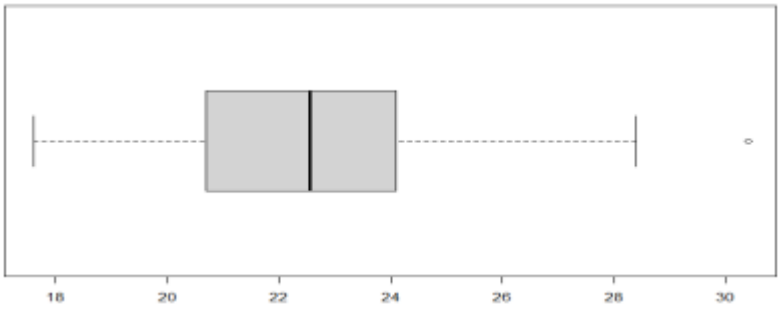


Figura 6.15. Diagrama de cajas de la variable IMC.

La variable IMC también tiene un valor extremo que corresponden a al caso 1, por lo tanto, estos valores pueden ser que influya en la presencia de asimetría en la distribución.

Ahora analizaremos dos casos particulares, la Figura 6.12 y la Figura 6.13 presentan dos máximos o modas – uno a cada lado del centro de simetría – Este patrón aparece cuando los datos responden a una mezcla de dos grupos heterogéneos y, siempre que sea posible, conviene estudiar ambos grupos por separado. Pero estas dos variables parecen tener una distribución simétrica, pues los datos están alrededor de las medidas centrales como se observa en los histogramas respectivos.

Es importante para poder corroborar todo lo señalado anteriormente, realizar el análisis de los estadístico descriptivos, en los cuales encontraremos valores numéricos que nos explicarán con mayor precisión el comportamiento de las distribuciones que están bajo estudio y estos valores se encuentran en la Figura 6.15.

```
# Otros estadísticos
summary(rend)
```

edad	peso	imc	rendimie	sexo	gym
Min. :23.0	Min. :45.00	Min. :17.60	Min. : 50.00	Hombre: 0	Gold : 0
1st Qu.:29.0	1st Qu.:58.00	1st Qu.:20.75	1st Qu.: 67.22	Mujer : 0	Hercules: 0
Median :33.0	Median :67.00	Median :22.55	Median : 74.00	NA's :28	NA's :28
Mean :34.5	Mean :68.21	Mean :22.80	Mean : 76.87		
3rd Qu.:38.5	3rd Qu.:75.75	3rd Qu.:24.05	3rd Qu.: 85.88		
Max. :54.0	Max. :92.00	Max. :30.40	Max. :100.00		

Figura 6.16. Tabla de estadísticos descriptivos.

En la tabla de la Figura 6.16 se muestran los resultados del análisis estadístico de las variables Edad, Peso, Índice de masa corporal y Rendimiento con sus respectivos histogramas y diagrama de cajas mostrados anteriormente.

La edad media de los alumnos es 34.5 ≈ 35 años y su mediana es 33 años, es decir que el 50% de ellos tiene igual o menor edad a esta. De la distribución podemos decir que es asimétrica y unimodal tal y como refleja su histograma (ver Figura 6.13) y diagrama de cajas (ver Figura 6.15) que muestra la existencia de 2 atípicos. Este hecho es responsable del nivel de asimetría de la variable (0.936) y con una deformación vertical normal, es decir, mesocúrtica (0.228)

El peso medio de los alumnos es 68.19 kg. y el 50% de ellos tiene un peso igual o menor a 65 kg. (mediana). Esta distribución es sin embargo, claramente multimodal y simétrica (0.289) tal y como refleja su histograma (ver Figura 6.14) y su curtosis negativa (-0.854).

El promedio del índice de masa corporal de los alumnos es 22.76 y su mediana es 22.47, es decir que el 50% de ellos tiene igual o menor índice de masa corporal a esta. De la distribución podemos decir que es asimétrica (0.635) y unimodal tal y como refleja su histograma (ver Figura 6.15) y diagrama de cajas (ver Figura 6.18) que muestra la existencia de 1 valor extremo. Esto influye en la presencia del nivel de asimetría de la variable y el bajo nivel de curtosis, platcúrtica (0.010)

El rendimiento de alumnos en promedio es 76.9%. y el 50% de ellos tiene un rendimiento igual o menor a 74%. Además en esta distribución existen fuertes

sospechas que sea multimodal y simétrica (0.028) tal y como refleja su histograma (ver Figura 6.16) y el valor negativo de la curtosis (-0.457).

6.6 ESTUDIO DE NORMALIDAD

Muchos métodos estadísticos se basan en la hipótesis de normalidad de la variable objeto de estudio. De hecho, si la falta de normalidad de la variable es suficientemente fuerte, muchos de los contrastes utilizados en los análisis estadístico-inferenciales no son válidos. Incluso aunque las muestras grandes tiendan a disminuir los efectos perniciosos de la no normalidad, el investigador debería evaluar la normalidad de todas las variables incluidas en el análisis.

Existen varios métodos para evaluar la normalidad de un conjunto de datos que pueden dividirse en dos grupos: los métodos gráficos y los contrastes de hipótesis.

6.6.1 Métodos gráficos

El método gráfico univariante más simple para diagnosticar la normalidad es una comprobación visual del histograma que compare los valores de los datos observados con una distribución normal. Aunque atractivo por su simplicidad, este método es problemático para muestras pequeñas, donde la construcción del histograma puede distorsionar la representación visual de tal forma que el análisis sea poco fiable.

Otras posibilidades, también basadas en información gráfica, consisten en realizar diagramas de cuantiles (Q-Q plots). Los diagramas de cuantiles comparan en un sistema de coordenadas cartesianas, los cuantiles muestrales (eje X) con los cuantiles esperados bajo la hipótesis normalidad. Si la distribución de partida es normal dichos diagramas tenderán a ser rectas que pasan por el origen. Cuanto más se desvíen de una recta menos normales serán los datos. En la Figura 6.19 se muestran posibles diagramas de cuantiles según la forma de la distribución de frecuencias.

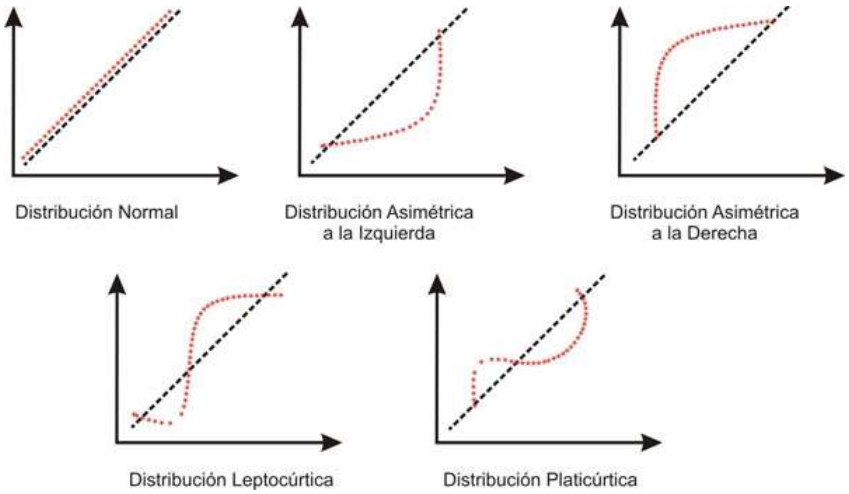


Figura 6.19. Diagrama de cuartiles correspondientes a diferentes distribuciones.

6.6.2 Contrastes de Hipótesis

La segunda de las formas para comprobar la normalidad de una distribución se
efectúa a través de un contraste de hipótesis. No existe un contraste óptimo para probar la hipótesis de normalidad. La razón es que la potencia relativa depende del tamaño muestral y de la verdadera distribución que genera los datos. Desde un punto de vista poco riguroso, el contraste de Shapiro y Wilks es, en términos generales, el más conveniente en muestras pequeñas ($n < 30$), mientras que el contraste de Kolmogorov-Smirnov, en la versión modificada de Lilliefors es adecuado para muestras grandes ($n \geq 30$).

En el test de Kolmogorov-Smirnov la hipótesis nula que se pone a prueba es que los datos proceden de una población con distribución normal frente a una alternativa de que no es así. Este contraste calcula la distancia máxima entre la función de distribución empírica de la muestra y la teórica. Si la distancia calculada es mayor que la encontrada en las tablas, fijado un nivel de significación (α), se rechaza el modelo normal.

Para realizar contrastar la hipótesis, en general se puede proceder de la siguiente manera:

Supongamos que se dispone de una muestra de la población y que, sobre cada individuo de la muestra, se mide una variable continua X. Las pruebas de bondad de ajuste como Kolmogorov-Smirnov y Shapiro y Wilks que se utilizan para contrastar la hipótesis nula (Ho) de que la muestra procede de una población en la que la distribución de X es una distribución Normal. Es decir, la hipótesis nula que se desea es:

$$H_o: F = \text{Distribución Normal} \sim N(0,1)$$

Si el *p*-valor asociado al estadístico de contraste es menor que α (nivel de significancia), se rechaza la hipótesis nula al nivel de significancia α.

Ejemplo Nº 6.2: Realice un estudio de normalidad a la variable Índice de Masa Corporal (IMC) a través del método gráfico y contraste de hipótesis.

Para solicitar estas pruebas para las variables indicadas, ejecutamos en siguiente script:

```
qqnorm(rend$imc, main="Gráfico Q-Q normal de Índice de Masa
Corporal", ylab="Valor observado", xlab="Normal esperado")
qqline(rend$imc)
```

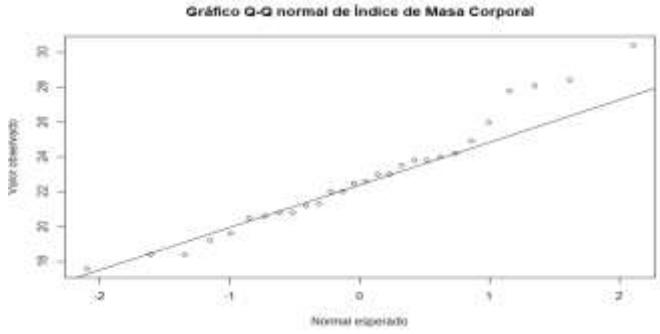


Figura 6.21. Gráfico de Q-Q normal para la variable IMC.

El grafico de la Figura 6.21 se encuentran los valores correspondientes a una distribución normal teórica vienen representados por la recta y los puntos corresponden a las diferentes puntuaciones de los sujetos en la distribución

empírica, es decir, que los 28 alumnos. De acuerdo a los diferentes gráficos de distribución mostrados en la Figura 6.19, los puntos están próximos a la recta, quiere decir que el ajuste es aceptable (tiene distribución normal).

Observación: En ocasiones la falta de normalidad de una variable puede arreglarse mediante una transformación de la misma. Se muestran algunas de las transformaciones más utilizadas.

Transformaciones para conseguir normalidad

Forma de la Distribución	Transformación aconsejada
Asimetría Positiva	Log(X+C)
Asimetría Negativa	Log(C-X)
Leptocurtosis	1/X
Platicurtosis	X^2

6.7 DATOS ATÍPICOS (OUTLIERS)

Los casos atípicos son observaciones con características diferentes de las demás. Este tipo de casos no pueden ser caracterizados categóricamente como benéficos o problemáticos sino que deben ser contemplados en el contexto del análisis y debe evaluarse el tipo de información que pueden proporcionar. Su principal problema radica en que son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos. Por otra parte, aunque diferentes a la mayor parte de la muestra, pueden ser indicativos de las características de un segmento válido de la población y, por consiguiente, una señal de la falta de representatividad de la muestra(Ximénez & Revuelta, 2022).

6.7.1 Tipos de outliers

Los casos atípicos pueden clasificarse en 4 categorías (Ximénez & Revuelta, 2022).

- La primera categoría contiene aquellos casos atípicos que surgen de un error de procedimiento, tales como la entrada de datos o un error de codificación. Estos casos atípicos deberían subsanarse en el filtrado de los datos, y si no se puede, deberían eliminarse del análisis o recodificarse como datos ausentes.

- La segunda clase es la observación que ocurre como consecuencia de un acontecimiento extraordinario. En este caso, el outlier no representa ningún segmento válido de la población y puede ser eliminado del análisis.

- La tercera clase contiene las observaciones cuyos valores caen dentro del rango de las variables observadas pero que son únicas en la combinación de los valores de dichas variables. Estas observaciones deberían ser retenidas en el análisis pero estudiando qué influencia ejercen en los procesos de estimación de los modelos considerados.

- La cuarta y última clase comprende las observaciones extraordinarias para las que el investigador no tiene explicación. En estos casos lo mejor que se puede hacer es replicar el análisis con y sin dichas observaciones con el fin de analizar su influencia sobre los resultados. Si dichas observaciones son influyentes el analista debería reportarlo en sus conclusiones y debería averiguar el porqué de dichas observaciones.

6.7.2 Identificación de outliers

Los casos atípicos pueden identificarse desde una perspectiva univariante o multivariante.

La perspectiva univariante examina la distribución de observaciones para cada
variable, seleccionando como casos atípicos aquellos casos cuyos valores caigan fuera de los rangos de la distribución. La cuestión principal consiste en el establecimiento de un umbral para la designación de caso atípico. Esto se puede hacer gráficamente mediante histogramas o diagramas de caja o bien numéricamente, mediante el cálculo de puntuaciones tipificadas. Para

muestras pequeñas (de 80 o incluso menos observaciones), las pautas sugeridas identifican como atípicos aquellos casos con valores estándar de 2.5 o superiores. Cuando los tamaños muestrales son mayores, las pautas sugieren que el valor umbral sea 3.

6.8 DATOS AUSENTES (MISSING)

Los datos ausentes son algo habitual en el Análisis Multivariante; de hecho, rara es la investigación en la que no aparece este tipo de datos.

En estos casos la ocupación primaria del investigador debe ser determinar las razones que subyacen en el dato ausente buscando entender el proceso principal de esta ausencia para seleccionar el curso de acción más apropiado.

Para ello se debe determinar cuál es el proceso de datos ausentes, entendido como cualquier evento sistemático externo al encuestado (errores en la introducción de datos) o acción por parte del encuestado (tales como rehusar a contestar) que da lugar a la ausencia de datos. En particular, el investigador debe analizar si existe algún patrón no aleatorio en dicho proceso que pueda sesgar los resultados obtenidos debido a la pérdida de representatividad de la muestra analizada.

6.8.1 Tipos de valores ausentes

Se distinguen las dos situaciones siguientes (Ximénez & Revuelta, 2022):

1) **Datos ausentes prescindibles**: son resultado de procesos que se encuentran bajo el control del investigador y pueden ser identificados explícitamente. En estos casos no se necesitan soluciones específicas para la ausencia de datos dado que dicha ausencia es inherente a la técnica usada.
   Ejemplos de estas situaciones son aquellas observaciones de una población que no están incluidas en la muestra o los llamados datos censurados que son observaciones incompletas como consecuencia del proceso de obtención de datos seguido en el análisis.

2) **Datos ausentes no prescindibles**: son resultado de procesos que no se encuentran bajo el control del investigador y/o no pueden ser identificados explícitamente.
   Ejemplos de estas situaciones son los errores en la entrada de datos, la renuncia del encuestado a responder a ciertas cuestiones o respuestas inaplicables. En estos casos se debe analizar si existen o no patrones sistemáticos en el proceso que puedan sesgar los resultados obtenidos.

Si los datos ausentes son no prescindibles conviene, por lo tanto, analizar el grado de aleatoriedad presente en los mismos. Según este grado el proceso de datos ausentes se puede clasificar del siguiente modo (Ximénez & Revuelta, 2022):

1) **Datos ausentes completamente aleatorios (MCAR)**: este es el mayor grado de aleatoriedad y se da cuando los datos ausentes son una muestra aleatoria simple de la muestra sin un proceso subyacente que tiende a sesgar los datos observados. En este caso se podría solucionar el problema sin tener cuenta el impacto de otras variables
2) **Datos ausentes aleatorios (MAR)**: en este caso el patrón de los datos ausentes en una variable Y no es aleatorio sino que depende de otras variables de la muestra X.
   Ahora bien, para cada valor de X, los valores observados de Y sí representan una muestra aleatoria de Y.
   Así, por ejemplo, si X es el sexo del encuestado e Y es su renta, un proceso MAR se tendría si existen más valores ausentes de Y en hombres que en mujeres y, sin embargo, los datos son aleatorios para ambos sexos en el sentido de que, tanto en los hombres como en las mujeres, el patrón de ausentes es completamente aleatorio. Si, además, tampoco existen diferencias por sexos los datos ausentes serían MCAR.
   Si los datos ausentes son MAR cualquier solución al problema deberá tener en cuenta los valores de X dado que afectan al proceso generador de datos ausentes.
3) **Datos ausentes no aleatorios**: en este caso existen patrones sistemáticos en el proceso de datos ausentes y habría que evaluar la magnitud del problema calibrando, en particular, el tamaño de los sesgos introducidos por dichos patrones. Si éstos son grandes habría que

atacar el problema directamente intentando averiguar cuáles son dichos valores.

6.8.2 Localización de datos ausentes

El primer paso en el tratamiento de datos ausentes consiste en evaluar la magnitud del problema. Para ello se comienza analizando el porcentaje de datos ausentes por variables y por casos.

Si existen casos con un alto porcentaje de datos ausentes se deberían excluir del problema. Así mismo si existe una variable con un alto porcentaje de este tipo de casos su exclusión dependerá de la importancia teórica de la misma y la posibilidad de ser reemplazada por variables con un contenido informativo similar.

Como regla general, sin embargo, si dicha variable es dependiente debería ser eliminada ya que cualquier proceso de imputación de valores puede distorsionar la significación estadística y práctica de los modelos estimados para ella.

6.8.3 Diagnóstico de la aleatoriedad en el proceso de datos ausentes

Existen 3 métodos (Ximénez & Revuelta, 2022):

a. Para cada variable Y formar dos grupos (observaciones ausentes y presentes en Y) y aplicar contrastes de comparación de dos muestras para determinar si existen diferencias significativas entre los dos grupos sobre otras variables de interés. Si se encuentran diferencias significativas el proceso de datos ausentes no es aleatorio.
b. Utilizar correlaciones dicotomizadas para evaluar la correlación de los datos ausentes en cualquier par de valores. Estas correlaciones indicarían el grado de asociación entre los valores perdidos sobre cada par de variables. Bajas correlaciones implican aleatoriedad en el par de variables y que los datos ausentes pueden clasificarse como MCAR. En caso contrario son MAR.

c. Realizar contrastes conjuntos de aleatoriedad que determinen si los datos ausentes pueden ser clasificados como MCR. Estos contrastes analizan el patrón de datos ausentes sobre todas las variables y las compara con el patrón esperado para un proceso de datos ausentes aleatorio. Si no se encuentran diferencias significativas el proceso puede clasificarse como MCAR; en caso contrario deben utilizarse los procedimientos a) y b) anteriores para identificar los procesos específicos de datos ausentes que no son aleatorios.

6.8.4 Aproximaciones al tratamiento de datos ausentes

Si se encuentran procesos de datos ausentes MAR o no aleatorios, el investigador debería aplicar sólo el método diseñado específicamente para este proceso. Sólo si el investigador determina que el proceso de ausencia de datos puede clasificarse como MCAR pueden utilizarse las siguientes aproximaciones:

a. Utilizar sólo los casos completos: conveniente si el tamaño muestral no se reduce demasiado
b. Supresión de casos y/o variables con una alta proporción de datos ausentes. Esta supresión deberá basarse en consideraciones teóricas y empíricas. En particular, si algún caso tiene un dato ausente en una variable dependiente, habitualmente excluirlo puesto que cualquier proceso de imputación puede distorsionar los modelos estimados.
   Así mismo una variable independiente con muchos datos ausentes podrá eliminarse si existen otras variables muy similares con datos observados.
c. Imputar valores a los datos ausentes utilizando valores válidos de otras variables y/o casos de la muestra.

6.8.5 Métodos de imputación

Los métodos de imputación pueden ser de tres tipos:

1) Métodos de disponibilidad completa que utilizan toda la información disponible a partir de un subconjunto de casos para generalizar sobre la

muestra entera. Se utilizan habitualmente para estimar medias, varianzas y correlaciones.

2) Métodos de sustitución que estiman valores de reemplazo para los datos ausentes, sobre la base de otra información existente en la muestra. Así se podría sustituir observaciones con datos ausentes por observaciones no maestrales o sustituir dichos datos por la media de los valores observados o mediante regresión sobre otras variables muy relacionadas con aquella a la que le faltan observaciones.
3)) Métodos basados en modelos que construyen explícitamente el mecanismo por el que se producen los datos ausentes y lo estiman por máxima verosimilitud. Entran en esta categoría el algoritmo EM o los procesos de aumento de datos.

BIBLIOGRAFÍA

Bouza, C. (2017). *Elementos de Análisis Estadístico de Datos*.

Castañeda, R. (2010). *Estadística Descriptiva*. https://wwwcastamayor.blogspot.com/

Contreras, J. (2018). *INTRODUCCIÓN A LA PROGRAMACIÓN ESTADÍSTICA CON R PARA PROFESORES* (I).

Figueras, S. (2003). *Análsis Exploratorio de Datos*. https://es.slideshare.net/leningvieralopez/estadistica-aed

Gamarra, G. (2015). *Estadística e investigación*.

Iba, V. (2001). *Estadística básica aplicada a la ganaderia*.

Paco, C. (2012). *Nutrición celular = Calidad de vida*. http://nutricioncelularconomnilife.blogspot.com/2012/07/indice-de-masa-corporal-imc.html

Parra, L. (2014). *estadística Descriptiva: Parte básica*. http://docplayer.es/16540405-Estadistica-descriptiva-1-1-parte-basica.html#show_full_text

Sanchez, J. (2012). *Estadística Descriptiva*. https://issuu.com/jenny_tatiana/docs/documento_de_estadistica

Santana, A. (2022). *Introducción a R*. https://estadistica-dma.ulpgc.es/cursoR4ULPGC/1-presentacion.html

Vicente, J. L. (2016). *Estadística Descriptiva con SPSS*. https://www.yumpu.com/es/document/view/14513264/estadistica-descriptiva-con-spss-13-estadistica#google_vignette

Ximénez, C., & Revuelta, J. (2022). Análisis de datos en Lenguaje R. In *Análisis de datos en Lenguaje R*. UAM Ediciones. https://doi.org/10.15366/9788483448304.dt.107