Estadística Descriptiva con R. Gráficos avanzados y aplicaciones

Autores:

Manuel Antonio Meneses Freire, Ph.D
Lourdes del Carmen Zuñiga Lema, M.Sc
Silvia Mariana Haro Rivera, M.Sc

GCPI Unach

Estadística Descriptiva con R.

Gráficos avanzados y aplicaciones

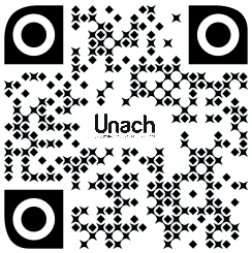
GCPI Unach

Título de la obra: **ESTADÍSTICA DESCRIPTIVA CON R. GRÁFICOS AVANZADOS Y APLICACIONES**

Nombres de los autores: Manuel Antonio Meneses Freire, Lourdes del Cármen Zuñiga Lema, Silvia Mariana Haro Rivera.

Ediciones: Universidad Nacional de Chimborazo (UNACH)
Diseño Gráfico: UNACH
Primera edición – julio 2021
Riobamba - Ecuador

ISBN: 978-9942-835-54-3
ISBN: 978-9942-835-55-0 (DIGITAL)
Registro Biblioteca Nacional
Depósito legal: EN TRÁMITE
DOI: https://doi.org/10.37135/u.editorial.05.35



Estadística Descriptiva con R. Gráficos avanzados y aplicaciones

Filiación Autores:

Manuel Antonio Meneses Freire, Ph.D
Universidad Nacional de Chimborazo
ameneses@unach.edu.ec

Lourdes del Carmen Zuñiga Lema, M.Sc
Escuela Superior Politécnica de Chimborazo
lzuniga@espoch.edu.ec

Silvia Mariana Haro Rivera, M.Sc
Escuela Superior Politécnica de Chimborazo
s_haro@espoch.edu.ec

GCPI Unach

Agradecimiento

En primer lugar, agradecemos al ser sublime nuestro Padre Celestial que hizo posible la realización de este libro. En segundo lugar, agradecemos a nuestros padres, hijos: David, Paolo, Kerly y Rafita quienes son nuestro apoyo y motor para seguir adelante. Silvita plasma su agradecimiento a Dios por ser la luz y guía en su vida. A la familia fuente de inspiración, y razón que empuja a salir adelante; y en especial, su gratitud a Luly y Toñito por la oportunidad de poder llamarlos amigos.

Índice

Prólogo

"Estadística Descriptiva con R. Gráficos avanzados y aplicaciones"

Esta obra está dirigida y diseñada para un curso de *Estadística Descriptiva* que se explican en la asignatura de Estadística a estudiantes de Ingeniería y Ciencias, y se imparten en Instituciones de Educación Superior. No se pretende ser exhaustivos con esta publicación, sino más bien, presentar un compendio que ha sido el fruto conjunto de varios miembros del área de Matemática y Estadística.

La Ciencia y la Ingeniería necesitan de la *Estadística Descriptiva* para analizar y representar sus bases de datos, tanto gráficamente como analíticamente. En la práctica, esta representación se realiza mediante *software*, comercial (su licencia de uso es pagada) o libre (no necesita licencia y se puede descargar de internet de forma gratuita). El presente libro titulado: *Estadística Descriptiva con R. Gráficos avanzados y aplicaciones*, utiliza el *software* estadístico R de uso libre, que es una ventaja frente a otros textos de Estadística que usan *software* comercial con licencia.

A continuación, se relata el contenido del libro enfocado al uso del *software* R.

En el capítulo 1, se dan los primeros pasos con R, desde la descarga del internet, la instalación, el uso de la consola y de los scripts. En el capítulo 2, se desarrolla la parte teórica de la *Estadística Descriptiva*, los elementos y herramientas necesarias para el análisis y la descripción de las variables estadísticas.

En el capítulo 3, se estudian las variables cualitativas, la frecuencia absoluta con la función *table*, el diagrama de barras con la función *barplot* y el gráfico de sectores con la función *pie* del código en R.

En el capítulo 4, se estudian las variables cuantitativas discretas, la frecuencia absoluta con la función *table*, la frecuencia relativa con *prop.table*, las frecuencias acumuladas con *cumsum*, el diagrama de barras con la función *barplot* y el gráfico de sectores con la función *pie* del código en R.

En el capítulo 5, se estudian las variables continuas, su discretización con la función *cut*, la frecuencia absoluta con la función *table*, la frecuencia relativa con *prop.table*, las frecuencias acumuladas con *cumsum*, el histograma con la función *hist* y el diagrama de caja con *boxplot*. También se calculan las medidas de posición: media muestral (*mean*), mediana (*median*) y cuantiles (*quantile*), y las medidas de dispersión: varianza (*var*) y desviación típica (*sd*), rango o rango intercuartílico (*range*) y coeficiente de variación.

En el capítulo 6, los gráficos que se desarrollan se clasifican como avanzados porque en su contexto tanto gráfico como en código del *software* R necesitan más detalle con respecto a los que se han visto en los capítulos anteriores, así como la utilización de funciones de otras librerías o paquetes diferentes a los básicos que vienen por defecto instalados.

En el anexo del libro se tiene paso a paso la instalación de las nuevas librerías en la consola, así como también las ayudas necesarias del paquete y de sus funciones. También se encuentra el código en R para la lectura de una base de datos con extensión *.txt* con la función *read.table*.

Introducción

En la actualidad, la demanda de bases de datos es muy elevada, debido a que existe bastante información de investigaciones en áreas de Ingeniería, Salud, Educación, Ciencias Políticas y en otras en general. Estas bases necesitan un análisis detallado para determinar sus características y cualidades relevantes las que se deben representar de forma resumida y clara.

El presente libro titulado, *Estadística Descriptiva con R. Gráficos avanzados y aplicaciones*, es una alternativa para realizar el análisis de datos utilizando el *software* estadístico R de libre acceso en internet. El desarrollo de este libro está dividido en seis capítulos, los anexos y la bibliografía; a continuación se realiza su descripción. En el primer capítulo, se realiza los primeros pasos que hay que hacer para entrar en confianza con los códigos, ventanas e íconos del *software* R. Este *software* es de libre acceso, que se puede obtener de la página de internet https://www.r-project.org/_y elegir el *cran mirror* más próximo para descargarlo de la forma más rápida. También se desarrolla varios ejemplos y gráficos sencillos, así como la manera de utilizar un *script* y la consola.

En el segundo capítulo, se desarrolla la parte teórica de la Estadística, sus elementos como la población y los parámetros, la muestra y los estimadores, así como también las variables estadísticas y su clasificación (Gutierrez, 2104; Spiegel, 1976).

En el tercer capítulo, se empieza con el estudio de las variables cualitativas, que representan cualidades no medibles, su representación en tablas de frecuencias y las gráficas se realizan mediante diagramas de barras y de sectores, con aspecto en forma bidimensional y tridimensional (Mendenhall, 2010).

En el cuarto capítulo, se estudia las variables cuantitativas discretas, su característica principal es tener un número finito o infinito numerable de valores distintos. A demás en estas variables, sus posibles valores pueden ser ordenados, a diferencia de las variables cualitativas. También las frecuencias absolutas y acumuladas se representan gráficamente mediante diagramas de barras y de sectores (Miller, 1999; Walpole, 2012).

En el quinto capítulo, se presenta las variables continuas con característica principal, ser medibles, las que toman tantos posibles valores como número de observaciones. El resumen estadístico de estas variables se realiza en clases o agrupaciones de datos, con frecuencias absolutas, relativas, frecuencias absolutas acumuladas y frecuencias relativas acumuladas. La representación gráfica se realiza en histogramas. También se calculan las medidas de posición (media muestral, mediana y cuantiles) y las de dispersión (varianza y desviación típica, rango o rango intercuartílico y coeficiente de variación). Un gráfico importante y muy útil para estudiar la distribución de los datos es el diagrama de cajas, donde se observan la existencia o no de datos atípicos para posteriormente ser analizados (Ross, 2005).

En el sexto capítulo, los gráficos que se desarrollan se clasifican como avanzados porque en su contexto tanto gráfico como en código del *software* R, necesita más detalle en comparación a los que se ha visto en los capítulos anteriores, así como la utilización de funciones de otras librerías o paquetes (Fox, 2018; Fox & Weisberg, 2011; Kampstra, 2008; Lemon, 2006).

El objetivo fundamental de este libro es facilitar herramientas y técnicas a toda la comunidad que desee introducirse en la *Estadística Descriptiva* a través del *software* R. Procuramos dar la ayuda necesaria para entender, tanto el lenguaje de programación

como los contenidos de la *Estadística Descriptiva* desarrollados a lo largo de los seis capítulos, y que, los ejemplos propuestos aporten a la comprensión de los contenidos.

También se añaden los anexos como ayuda para una mejor utilización del *software* R (R Core Team, 2017).

Con respecto a los diferentes conjuntos de datos con los que hemos pretendido ilustrar los contenidos de este libro, algunos son conjuntos de datos clásicos, como *titanic*, *iris* entre otros, disponibles en R. Todos ellos se encuentran en el siguiente enlace: http://graduados.unach.edu.ec/estadistica-descriptiva-con-R, de esta manera se podrán reproducir las practicas a lo largo de los capítulos. Y, otros conjuntos de datos han sido simulados en R.

Capítulo 1

Primeros pasos con el software estadístico R

1.1 ¿Qué es R?

R es un lenguaje y entorno de programación, cuya característica principal está enfocada al análisis estadístico para la manipulación de datos, su cálculo y desarrollo gráfico. R está basado en el lenguaje de programación S. En su momento el lenguaje S evoluciono por un lado hacia R, con la singularidad de que es un *software* de uso libre con licencia GNU (*General Public Licenc*: libertad que tienen los usuarios para ejecutar, copiar, distribuir, estudiar, cambiar, actualizar y mejorar el *software*), por otro lado, hacia el *software* comercial S-plus.

En la actualidad ningún otro programa reúne las condiciones de madurez, manejabilidad, cobertura, accesibilidad y cantidad de recursos que tiene R, además dispone de una gran comunidad de desarrolladores/usuarios y soporte técnico de calidad, que se dedican constantemente a mejorar, ampliar sus funcionalidades y capacidades del programa.

Las funciones de R se agrupan en bibliotecas o paquetes (packages, libraries), los que contienen las funciones más habituales, se incluyen por defecto en la distribución e instalación de R, llamadas bibliotecas estándar y otros muchos paquetes desarrollados por investigadores de todo el mundo, se encuentran disponibles a través de Internet en la Comprehensive R Archive Network (CRAN), que abarcan diversos campos como aplicaciones financieras, fiabilidad de materiales, salud, wavelets, análisis de datos espaciales, etc.
R se compila y ejecuta en plataformas LINUX, Windows y Mac OS X. Este libro trata fundamentalmente de la interacción de R bajo Microsoft Windows. Si interactúa con las plataformas LINUX o Mac OS X necesitará realizar algunos pequeños cambios.

El programa para la instalación está ubicado en la dirección URL, https://cran.r-project.org/ . Allí observaremos, en la parte superior un enlace para la instalación de R en Linux, Mac OS X y Windows.

Para Windows debemos seguir el enlace al subdirectorio *base*. Para el caso de Mac OS X, el enlace nos lleva a una página donde encontramos el archivo de instalación de la última versión. Existen varias formas de instalar R en distribuciones de Linux; de forma general, primero es necesario añadir << CRAN >> a la lista de repositorios, luego configurar el nuevo repositorio adecuadamente e instalar el r-base.

1.2 Interfaz de R

Existen diferentes entornos gráficos que ayudan y facilitan la interacción con el usuario, denominados GUI (Graphical Users Interface). Entre los más conocidos en R tenemos RGui, (disponible solo para Windows), TinnR (disponible para Windows y MAC OS X), RKward (Linux), RStudio entre muchos otros. Los ejemplos de este libro han sido elaborados con RGui en Windows. La Figura 1.1 muestra el aspecto de la interfaz RGui, disponible para Windows:

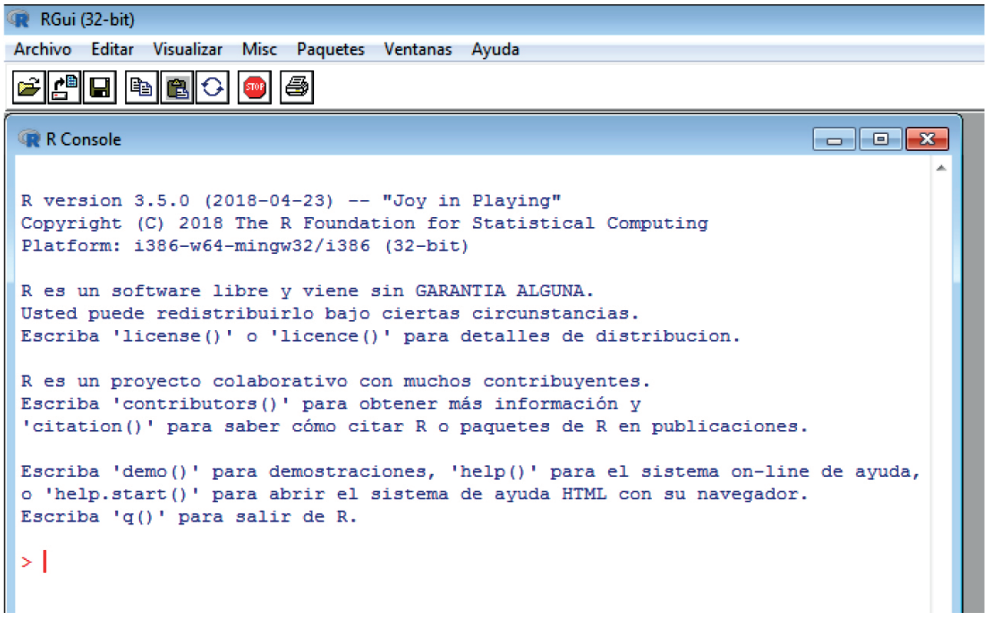


Figura 1.1. Interfaz RGui. Elaboración propia al ejecutar el software R.

RGui se instala automáticamente con el paquete *base* de Windows. El resto de GUI necesitan una instalación específica. Se encuentran gratuitamente en el internet.

Estas GUI, disponen de menús desplegables, cuya complejidad va desde sencillas posibilidades de edición, apertura de ficheros, cargar paquetes, etc. como es el caso de RGui, hasta correcciones de sintaxis de instrucciones, menús desplegables para importar/exportar datos, gráficos y múltiples técnicas estadísticas como son el caso de *R Commander* y *Rkward*.

1.3 Comandos y conceptos básicos

En este apartado describiremos solamente la utilidad específica de cada comando, y si se desea consultar la sintaxis precisa de cada uno de ellos, R dispone de un sistema de ayuda que se puede invocar desde el menú que aparece al iniciar el programa (opción **Help**).

Case sensitivity: el lenguaje de R distingue entre mayúsculas y minúsculas, por lo que *A* y *a* son variables diferentes, Los comandos están separados por punto y coma, ; , o por una nueva línea. Si un comando no está completo al final de la línea, R dará un mensaje diferente, por ejemplo + en la segunda línea y las siguientes, y continuará leyendo la entrada hasta que el comando esté completo sintácticamente.

Robustez: R es un lenguaje robusto, es decir, intenta en lo posible, no dar mensajes de errores. Esto en general, es cómodo, pero se debe tener cuidado, pues al momento de ejecutar un comando, es posible que no obtengamos ningún error, sin embargo, R no puede estar haciendo lo que nosotros pretendemos.

Importar códigos: el comando *source ("comandos.R")* permite introducir comandos procedentes de archivos. Es útil para cargar funciones creadas por el usuario o para ejecutar macros y scripts.

Exportar: el comando *sink("archivo")*, permite que los comandos posteriores se almacenen en *"archivo"*. Para devolver la salida a la pantalla se usa nuevamente el comando *sink( )*. Usado principalmente para guardar salidas de datos en archivos de texto.

Funciones: para consultar el código de una función en R, es suficiente teclear el nombre en la línea de comandos.

Workspace: el comando *getwd( )* devuelve el directorio de trabajo, en cambio el comando *setwd( )* permite modificarlo. Al iniciar el programa R se abre de forma automática un espacio de trabajo, en él se almacena todas las ordenes usadas durante esa sesión. En cualquier instante se puede guardar todos los objetos del espacio de trabajo como un archivo con extensión "*.R" (menú archivo). Como recomendación y norma general, las funciones y objeto de datos hay que guardarlo en archivos propios; y esto evita tener objetos innecesarios en el "workspace".

Librerías: al iniciar R se cargan por defecto librerías básicas, pero a veces es necesario cargar otras librerías consideradas no básicas. Esto se puede hacer por medio del comando *library(nombre)*.

Listados: el comando *ls( )* entrega un listado de los objetos que hay actualmente en el espacio de trabajo; para búsquedas de objetos en el espacio de trabajo comúnmente se utilizan los comandos *apropos( )* y *find( )*.

Borrado: el comando *rm(objetos)* elimina uno o varios objetos del espacio de trabajo.

Historial de comandos: mediante las flechas del teclado se accede a los últimos comandos ejecutados. Por otro lado, el comando *history( )* nos devuelve un archivo tipo texto con últimos comandos ejecutados.

1.4 Objetos y operaciones básicas

En R se puede trabajar con varios tipos de objetos; algunos son de tipo estándar como en cualquier lenguaje de programación y otros objetos son específicamente de R, orientados con propósitos estadísticos.

Vectores: son la estructura de datos más sencilla con la que trabaja R. Se utiliza "<-" para hacer asignaciones, en general no se realizan asignaciones con "=" (esto da lugar a errores lógicos).

```
x <- c(1,3,5,7)
```

Este comando proporciona un vector formado por los 4 primeros números impares. El código *c()* sirve para concatenar objetos, aunque es muy cómoda para crear vectores. En el ejemplo anterior el código x[2] nos devuelve el valor 3.

```
x <- numeric(10)
```

Este comando proporciona un vector de 10 componentes, todos sus elementos 0. También podemos definir vectores de tipo: character, double, en otros.

```
x<-1:15; y<-15:1
```

Este comando proporciona vectores *x* e *y* con los 15 primeros números naturales en forma creciente y decreciente respectivamente.

seq(1,9,by=2)

Devuelve los 5 primeros números impares (va del 1 al 9 con un incremento de 2).

rep(2,10)

Este comando proporciona un vector de 10 componentes, todas ellas con el valor 2.

x<-c("a","e","i","o","u") Nos devolvería un vector de caracteres con las 5 vocales.

Todos los elementos de un vector deben ser del mismo tipo, R admite dos excepciones, los valores NA (not avaliable) y NaN (not a number). Estos dos valores son importantes en el análisis estadístico: NA porque a veces hay valores perdidos en algunos campos de una muestra y NaN porque es el resultado de alguna indeterminación. R permite realizar operación con vectores, aunque sus componentes tengan NA o NaN.

Operaciones elementales con vectores: los operadores básicos son +, -, *, /, y ^ para la potencia; estos operadores funcionan sobre los vectores componente a componente. Otras funciones básicas en R son log, exp, sin, cos, tan, sqrt entre otros. Además, *sum(x)* devuelve la suma de los elementos de x, *prod(x)* su producto, *mean(x)* su media, *var(x)* su quasivarianza.

R tiene dos valores lógicos: TRUE y FALSE. En cambio los operadores lógicos son <, <=, >, >=, == para igual, '!=' para distinto, '&' se usa para indicar intersección ("y"), '|'

indica disyunción ("o") y '!' indica la negación. A veces de un vector dado, nos interesa realizar un tipo de filtro para quedarnos con los elementos de este que verifican una cierta propiedad. Por ejemplo, con los comandos:

x<-c(1,-1,2,-2); x>0 Obtenemos el vector: TRUE FALSE TRUE FALSE

x<-c(1,-1,2,-2); y<-x[x>0] Obtenemos el vector y, cuyos componentes son los números 1 y 2 (los positivos de x).

En ocasiones nos interesa quitar los valores NA y NaN de un vector, esto se realiza con el comando x<-x[!is.na(x)]. En cambio, para quitar sólo los NaN se puede usar el comando x<-x[!is.nan(x)]. De forma análoga, x[is.nan(x)]<-0 reemplaza todos los NaN del vector por 0. El comando which(is.nan(x)) devuelve las posiciones de los elementos de vector que toman el valor NaN.

Arrays y matrices: las matrices son una extensión natural de los vectores. Un array o una matriz se puede definir por ejemplo como:

x <- array(1:20,dim=c(4,5)) genera un array de 4 filas y 5 columnas cuyos componentes son los números del 1 al 20. En R el array se llena por columnas (al contrario que en el lenguaje C), es decir, los componentes de la primera columna del array x serían los números del 1 al 4, la segunda los números del 5 al 8 y así sucesivamente. Lo más importante a la hora de definir un array es especificar las dimensiones. En el ejemplo anterior, el código x[3,2] nos devuelve como resultado el segundo elemento de la tercera fila de x, el código x[,1] devuelve la 1era columna y x[3,] la tercera fila.

Operaciones elementales con arrays y matrices: definida una matriz A, la función *t(A)* calcula la traspuesta de A. Dado

un array de dimensión cualquiera, con la función aperm() nos permite obtener trasposiciones del mismo (intercambiar los índices en la matriz, es decir permutando los índices de columnas, filas). Los códigos *nrow(A)* y *ncol(A)* devuelve el número de filas y columnas de A respectivamente. Definidas dos matrices A y B, el producto de ambas se hace mediante el operador %*%. Si A y B tienen la misma dimensión, el código A*B devuelve el producto componente a componente de sus elementos, no el producto matricial. Por otro lado, la función *solve* aplicada sobre la matriz A calcula su inversa.

Listas y data frames: por último, tenemos dos tipos de datos todavía más generales: las listas y los data frames.

La lista es un objeto cuyas componentes pueden ser arrays, listas, Data Frames, variables lógicas entre otros y cualquier combinación de estos (permite almacenar juntos datos de distinta naturaleza como nombre, edad, ingresos, trabajos anteriores, etc.). Los distintos elementos de la lista no han de ser necesariamente del mismo tipo. Por ejemplo:

Lst <- list(name="Fred", wife="Mary", no.children=3, child.ages=c(4,7,9)) este código devuelve:

```
> Lst
$name
[1] "Fred"
$wife
[1] "Mary"
$no.children
[1] 3
$child.ages
[1] 4 7 9
```

una lista formada por 4 objetos (name, wife, no.children, child.ages), dos variables de tipo carácter (name, wife), un valor numérico (3) y un vector de tres componentes (4, 7, 9). Para referenciar a cada objeto de la lista se hace con Lst[[i]], de este modo Lst[[1]]="Fred" y Lst[[4]][3]=9. Cuando se trabaja con listas es más cómodo usar los nombres de cada objeto de la lista, esto es Lst$wife="Mary". Para saber los nombres de los objetos que componen la lista se usa names(Lst).

Los data.frames (campos de datos) son el objeto más habitual para el almacenamiento de datos. Un data.frame considera que cada fila representa a un individuo de una muestra y el correspondiente valor para cada columna se corresponde con la medición de alguna variable para ese individuo (fila-individuo, columna-variable). Por ejemplo si tenemos 200 alumnos y las notas numéricas de examenes de cada uno de ellos junto con la calificación final (tipo carácter), esto se pondría en un data.frame de 200 filas con una columna por cada examen y otra columna más para la calificación final. A veces resulta tedioso trabajar con un data.frame o una lista cuando tienen con muchos campos anidados (muchos subniveles). En el ejemplo del caso de la lista antes definida se usaba el código Lst$wife para referirse al segundo objeto de la lista, Lst$name al primero y así con los demás objetos. Esta notación puede hacerse menos pesada mediante el comando attach(Lst), este comando hace que Lst "suba un nivel en el workspace" y a partir de ese momento podremos utilizar, el código, child.ages en vez de Lst$child.ages para acceder al vector de las edades. Al utilizar este comando hay que tener cuidado de que en el espacio de trabajo no existan ya objetos con el mismo nombre de alguno de los objetos/columnas de la lista/data.frame. Para volver a la situación anterior se usa el comando dettach(Lst). El comando search()

permitirá ver el número de librerías cargadas en el espacio de trabajo, pero también permite mirar los objetos a los que se ha "subido un nivel" con attach.

1.5 Procedimientos gráficos

R dispone de varias funciones creadas y preparadas para la representación gráfica de datos. La más común es la función *plot*, esta tiene muchas variantes que dependen del tipo de datos que se tome como argumento. La más destacada es *plot(x,y)*, esta función representa un diagrama de puntos de *y* frente a *x*. También tenemos otras funciones como *hist* para graficar histogramas, o *persp* para gráficas tridimensionales (superficies).

Los parámetros más comunes para la mayoría de gráficos son:

type: tipo de gráfico a realizar, el más destacado es type="p" para representar gráfica de puntos (opción por defecto), type="l" para representar líneas, type="b" para representar los puntos unidos por líneas.

lines: permite sobreponer o añadir nuevas gráficas en una gráfica ya existente.

points: permite añadir puntos.

legend: permite añadir una nueva leyenda.

text: permite añadir texto.

pch: indica la forma en que se dibujarán los puntos.

lty: Indica la forma en que se dibujarán las líneas.

lwd: ancho de las líneas.

col: color a usar en el gráfico

font: fuente a usar en el texto.

Para una descripción detallada y completa de estos y otros parámetros invocamos la función help(par)

1.6 Importando datos

R es un entorno pensado para el análisis de datos, por lo que es importante saber cómo importar datos externos (datos de SPSS, Excel y archivos de texto). En este libro se trabaja con archivos de datos tipo texto. A continuación, conoceremos cómo se importa datos con extensión *.txt*:

> *.txt*: si un conjunto de datos viene guardado como un archivo *.txt* se lee con la función *read.table* que nos devuelve un data.frame. Entre sus argumentos se puede definir ciertas especificaciones como si tiene encabezado o no entre otros más.

1.7 Empezando a trabajar con R

Para limpiar la consola pulsamos *ctrl + L* y luego empezamos con el ejemplo de la sesión: tras teclear el texto se debe pulsar *enter:*

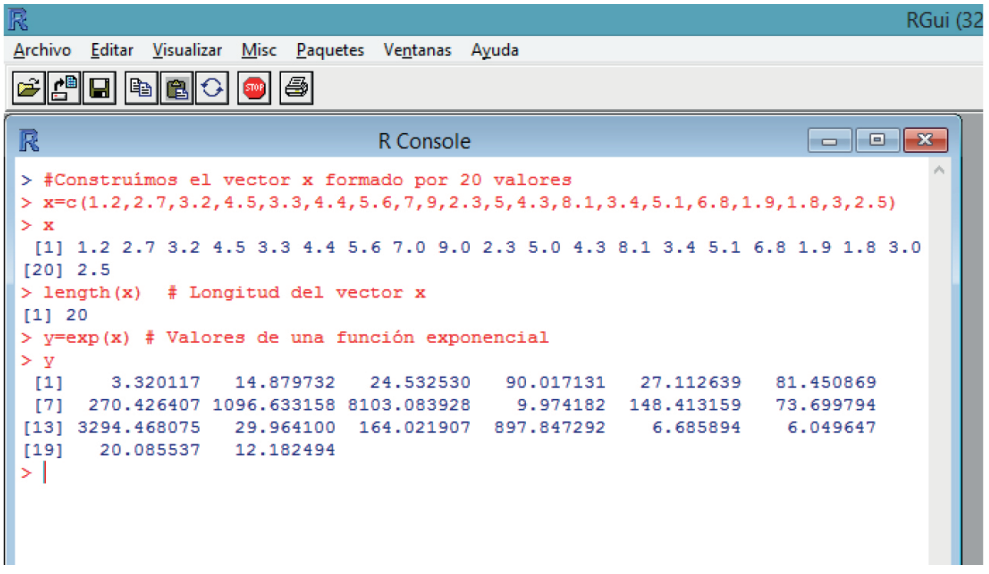


Figura 1.2. Ejemplos en la consola de R . Elaboración propia al realizar los ejemplos en la consola.

1.8 Gráficos en la consola de R

En el *software* R se puede realizar varios gráficos estadísticos entre estos hacemos un par de ejemplos: en la Figura 1.3 se observa la manera de ingresar los valores o variables en la consola para graficar la función raíz cuadrada.

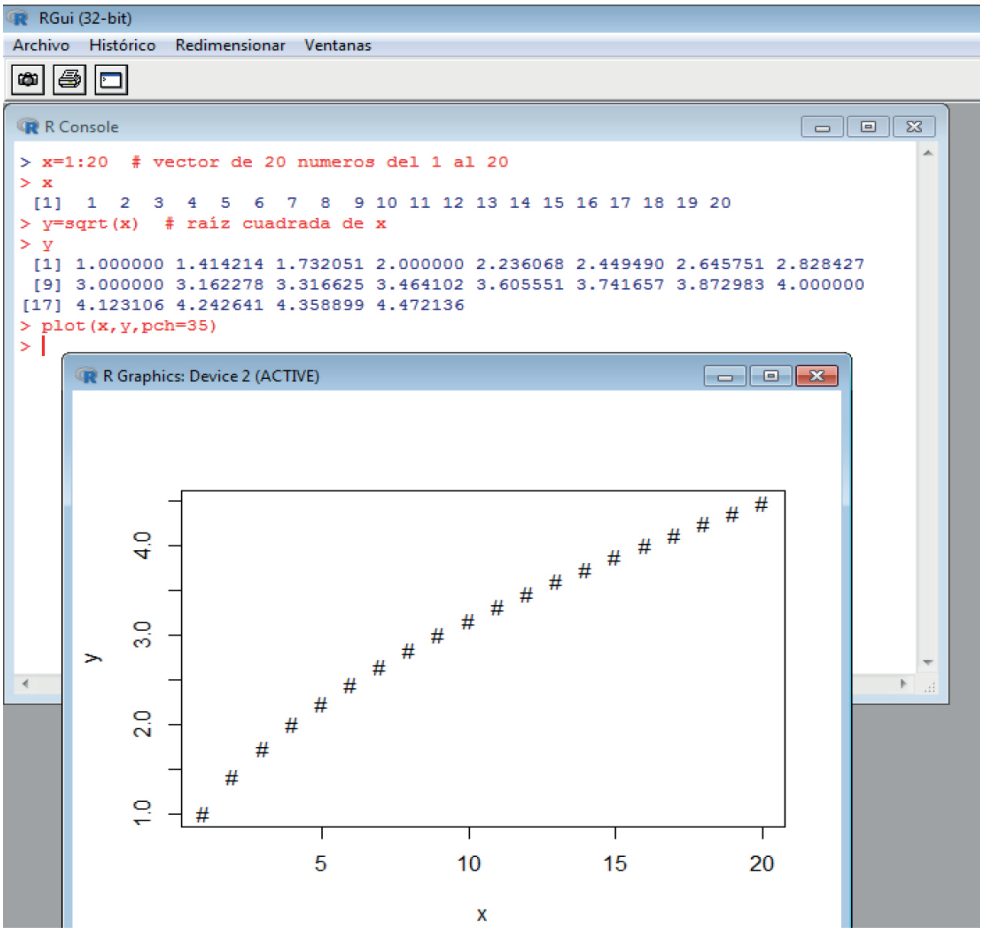


Figura 1.3. Ejemplos del plot en la consola de R. Elaboración propia al realizar los ejemplos en la consola.

También en la Figura 1.4 se realiza un histograma simple de números aleatorios generada con la función *rnorm* de la distribución normal de media 0 y desviación típica 1:

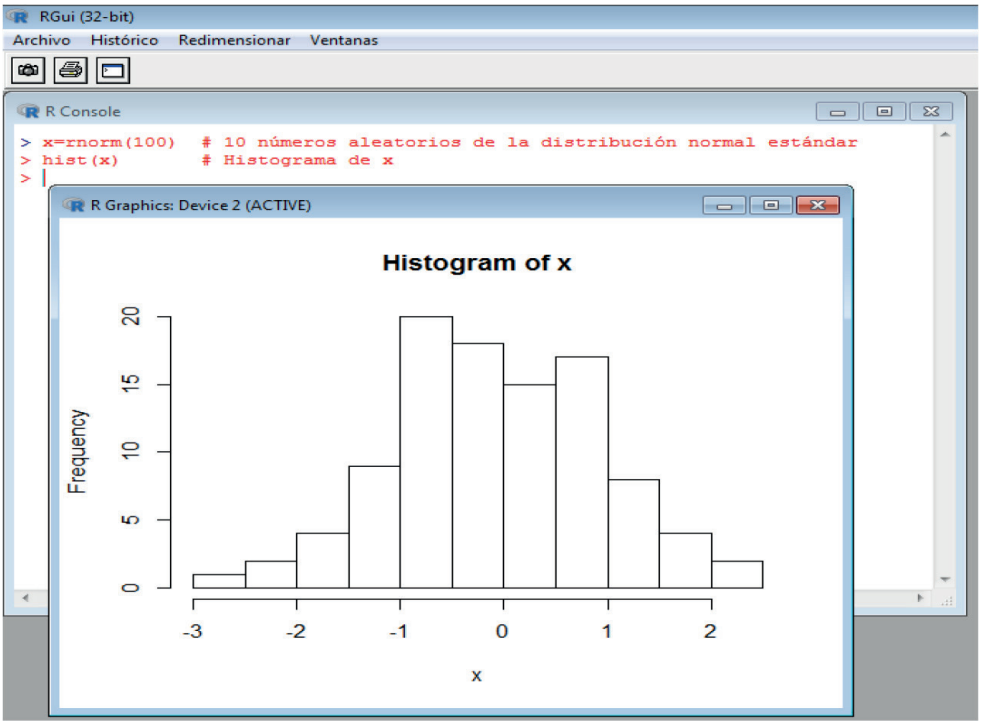


Figura 1.4. Ejemplos del hist en la consola de R. Elaboración propia al realizar los ejemplos en la consola.

1.9 Trabajar con scripts en R

R es un *software* libre que también da la posibilidad de trabajar con scripts para enviar a correr las líneas de código sobre la consola, la manera de utilizar un script se observan paso a paso en las siguientes gráficas:

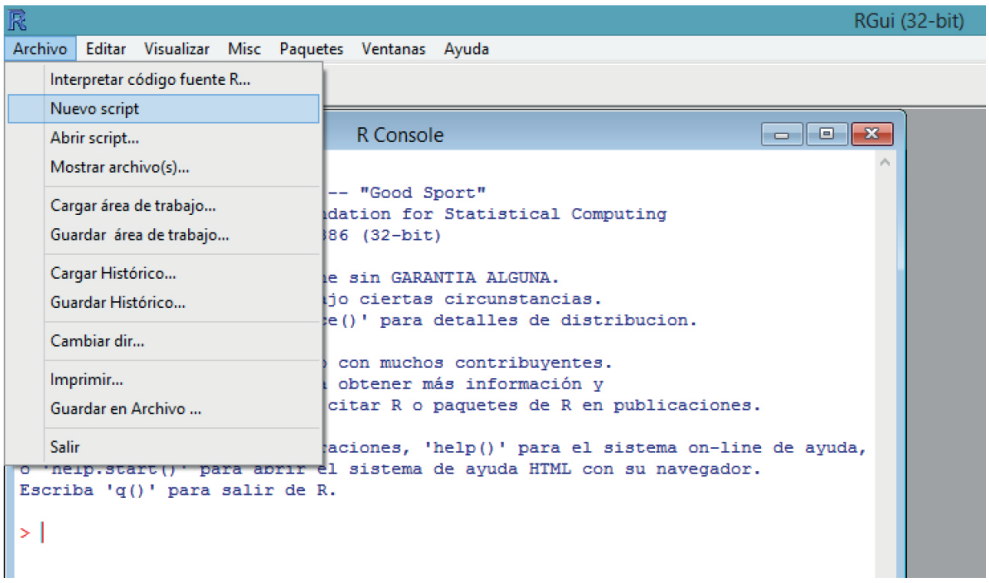


Figura 1.5. Guía para trabajar con scripts en R. Elaboración propia para abrir un nuevo script.

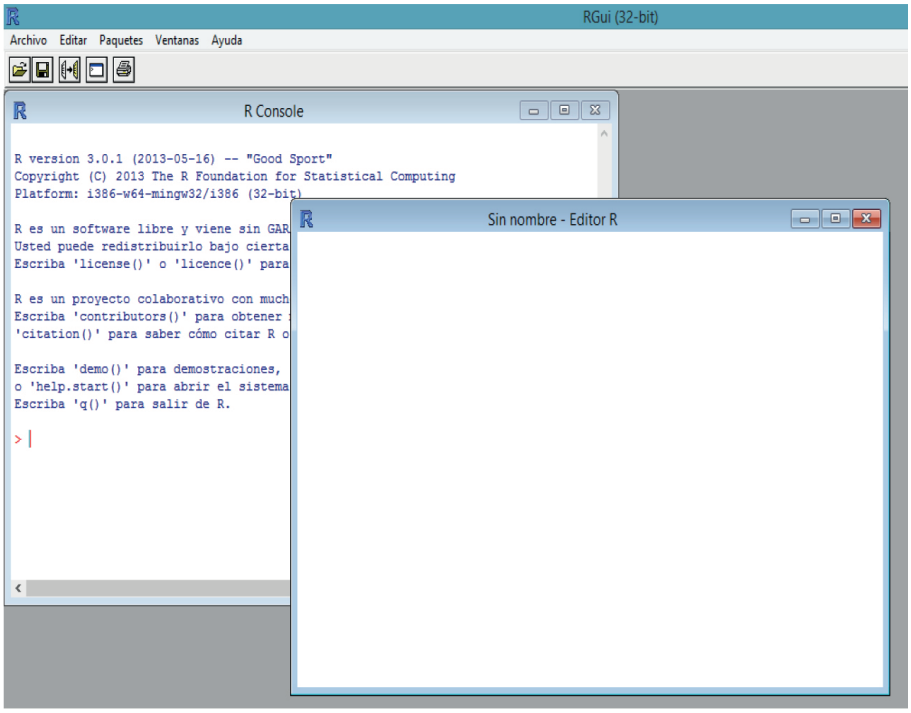


Figura 1.6. Script nuevo en R. Elaboración propia para abrir una ventana de editor.

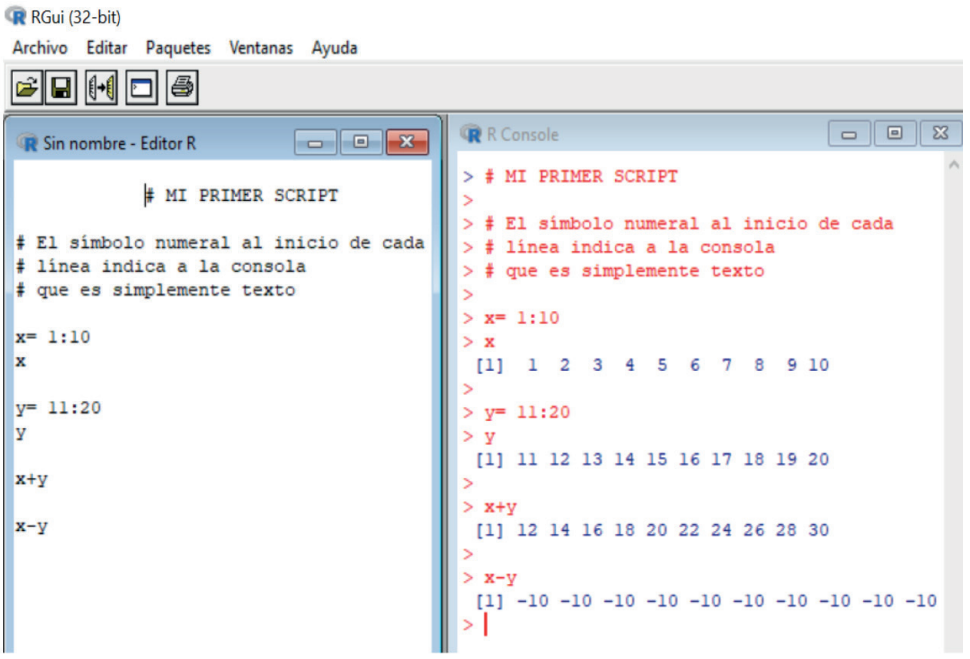


Figura 1.7. Script y la consola con ventanas de forma vertical. Elaboración propia para abrir una nueva ventana y la consola en forma vertical.

En la Figura 1.7, se observa el texto escrito en la ventana izquierda (script) y cada línea es corrida en la ventana derecha (consola) con el ícono que está justo debajo de la palabra Editar, indicando que el símbolo numeral # al inicio de cada línea corresponde un comentario de texto. Además, *x* e *y* son variables numéricas que se pueden sumar, restar y otras operaciones matemáticas.

Capítulo 2

Fundamentos de la Estadística

Introducción a la estadística

Todo lo que tiene que ver con recolección, procesamiento, análisis e interpretación de datos numéricos pertenece al dominio de la estadística.

La estadística juega un rol importante en el mejoramiento de la calidad de cualquier producto o servicio.

Se puede decir, en términos generales, que la estadística se aplica dentro de la organización de una empresa, en las siguientes áreas: producción, finanzas, contabilidad, personal y mercados.

2.1 ¿Por qué estudiar estadística?

Las respuestas dadas por el análisis estadístico pueden sentar las bases para tomar decisiones o elegir acciones. Los funcionarios de la ciudad, por ejemplo desean conocer si el nivel de plomo en el suministro de agua está dentro de los estándares de seguridad. Puesto que no toda el agua puede verificarse, las respuestas deben basarse en la información parcial de las muestras de agua que se recolectan para tal propósito.

Cuando se busca información, las ideas estadísticas sugieren un proceso de recolección típico con cuatro pasos fundamentales.

1. Establecer metas definidas con claridad para la investigación
2. Elaborar un plan de cuáles datos recolectar y cómo recabarlos
3. Aplicar métodos estadísticos adecuados para extraer información a partir de los datos
4. Interpretar la información y extraer conclusiones

Se trata de pasos indispensables que ofrecerán un marco de referencia siempre que se desarrollen las ideas clave de la estadística. El razonamiento y los métodos estadísticos le ayudarán a volverse eficiente para obtener información y obtener conclusiones útiles.

2.2 Estadística moderna

El origen de la estadística está en dos áreas de interés que, en la superficie, tienen poco en común: los juegos de azar y lo que ahora se conoce como ciencia política. Los estudios de probabilidad a mediados del siglo XVIII, motivados en gran medida por el interés en los juegos de azar, condujo al tratamiento matemático de los errores de medición y a la teoría que ahora forma los cimientos de la estadística. En el mismo siglo el interés en la descripción numérica de las unidades políticas (ciudades, provincias, poblados, etc.) llevo a lo que ahora se conoce como estadística descriptiva**.** Al principio, la estadística descriptiva consistía simplemente en la presentación de datos en tablas y gráficas; en la actualidad incluye el resumen de datos mediante descripciones numéricas y gráficas. En décadas recientes, el crecimiento de la estadística se vio en casi cualquier rama de actividad importante, cuya característica más importante en crecimiento ha sido el cambio en el énfasis: de la estadística descriptiva a la inferencia estadística. La inferencia estadística se ocupa de la generalización basada en datos muestrales; se aplica a problemas como la estimación de la emisión promedio de contaminantes de un motor a partir de corridas de prueba, el hecho de probar la afirmación de un fabricante sobre la base de mediciones realizadas a muestras de su producto, entre otros.

Cuando alguien hace una inferencia estadística, es decir, una inferencia que va más allá de la información contenida en un conjunto de datos, siempre debe proceder con cautela. Uno habrá de decidir cuidadosamente cuán lejos hay que ir en la generalización a partir

de cierto conjunto de datos, ya sea que tales generalizaciones sean en absoluto razonables o justificables, o bien que sea aconsejable esperar hasta que existan más datos, etc. De hecho, algunos de los problemas más importantes de la inferencia estadística tienen que ver con la valoración de los riesgos y las consecuencias a las que uno estaría expuesto al realizar generalizaciones a partir de datos muestrales. Esto incluye una valoración de las probabilidades de tomar decisiones equivocadas, así como la posibilidad de hacer predicciones incorrectas y la obtener estimaciones que no reflejan de manera adecuada la situación real.

2.3 Estadística e Ingeniería

Hay pocas áreas donde la influencia del crecimiento reciente de la ingeniería se haya sentido con mayor fuerza que en la ingeniería y la administración industrial. De hecho, sería muy difícil sobreestimar las contribuciones de la estadística para resolver problemas de producción, del uso efectivo de materiales y la mano de obra, de la investigación básica y del desarrollo de nuevos productos. Como en otras ciencias, la estadística se ha convertido en una herramienta vital para los ingenieros. Les permite entender fenómenos sujetos a variación y predecirlos de manera efectiva o controlarlos.

2.4 El rol del científico y del ingeniero en el mejoramiento de la calidad

En la última mitad del siglo pasado e inicios del presente, Estados Unidos se encontró a sí mismo en un mercado mundial cada vez más competitivo. La competencia alentó una revolución internacional en el mejoramiento de la calidad. Las enseñanzas e ideas de W. Edwards Deming (1900-1993) fueron útiles en el rejuvenecimiento de la industria japonesa. Él destacó que la industria estadounidense, con la finalidad de sobrevivir, debería movilizarse con un compromiso continuo por el mejoramiento de la

calidad. Desde el diseño hasta la producción, los procesos necesitan mejorarse de forma continua. El ingeniero y el científico, con sus conocimientos técnicos y armados con habilidades estadísticas básicas en recolección de datos y presentaciones gráficas, podrían ser los principales actores en el logro de dicha meta.

El mejoramiento de la calidad se basa en la filosofía de "hacerlo bien la primera vez". Más aún, uno no debería estar contento con cualquier proceso o producto, más bien tiene que seguir buscando formas de mejorarlo.

2.5 Algunos conceptos necesarios

2.5.1 Unidad (o elemento): una sola entidad, por lo general, un objeto o una persona, cuyas características son de interés

2.5.2 Población de unidades: colección completa de unidades, acerca de la cual se busca información

2.5.3 Características (o caracteres): corresponden a ciertos rasgos, cualidades o propiedades de las unidades determinadas que constituyen la población. Algunos caracteres son mesurables y se describen numéricamente, por tal motivo se denominan caracteres o *variables cuantitativas*, (estatura, peso, ingreso, valor, producción, etc.). Otros se expresan mediante palabras por no ser mesurables pero si cuantificadas, (profesión, cargo, marcas, calidad, etc.), se denominan caracteres o *variables cualitativos* (*atributos*).

2.5.4 Población estadística (o sólo población): es el conjunto de todas las mediciones (o registros de algún rasgo de calidad) correspondientes a cada unidad en toda la población de unidades acerca de la cual se busca información. En la Tabla 2.1 se observan algunos ejemplos de poblaciones, unidades y la característica de la variable que puede ser estudiada.

Tabla 2.1
Ejemplos de poblaciones, unidades y variables

Población	Unidad	Variables/características
Todos los estudiantes actualmente inscritos en la Universidad Nacional de Chimborazo	Estudiantes	Promedio Número de créditos Horas de trabajo por semana Especialidad Diestro/zurdo
Todas las placas de circuito impreso indispensable para armar una computadora, fabricadas durante un mes	Tarjeta	Tipo de defectos Número de defectos Ubicación de defectos
Todos los restaurantes de comida rápida en la ciudad de Riobamba	Restaurante	Número de empleados Número de asientos Contrata/no contrata
Todos los libros en la biblioteca de la Unach.	Libro	Costo de sustitución Frecuencia de salida Reparaciones necesarias

Elaboración propia.

El objetivo de la *Estadística Descriptiva* es la toma de información sobre los elementos de un cierto colectivo llamado población.

2.5.5 Muestras de una población: una muestra de una población estadística es el subconjunto de mediciones que realmente se recolectan en el curso de una investigación. Las unidades se seleccionan aleatoriamente, es decir, todos los elementos que componen la población tienen la misma posibilidad de ser seleccionados. Para que la muestra sea representativa de la población se requiere que las unidades sean seleccionadas al azar, ya sea utilizando el sorteo, tablas de números aleatorios, selección sistemática o cualquier otro método al azar.

Si la muestra coincide con la población, es decir se toma información sobre cada uno de los individuos de la población, la muestra se denomina censo.

Las dificultades para realizar un censo (población infinita, dificultad de acceso a todos los individuos, coste económico, tiempo necesario, etc.) hacen que en muchas situaciones sea preferible el muestreo. En este caso, las técnicas de Inferencia Estadística permitirán obtener resultados de toda la población a partir de los obtenidos en la muestra. En la Figura 2.1 se observa la población y una muestra de puntos que indican sus unidades.

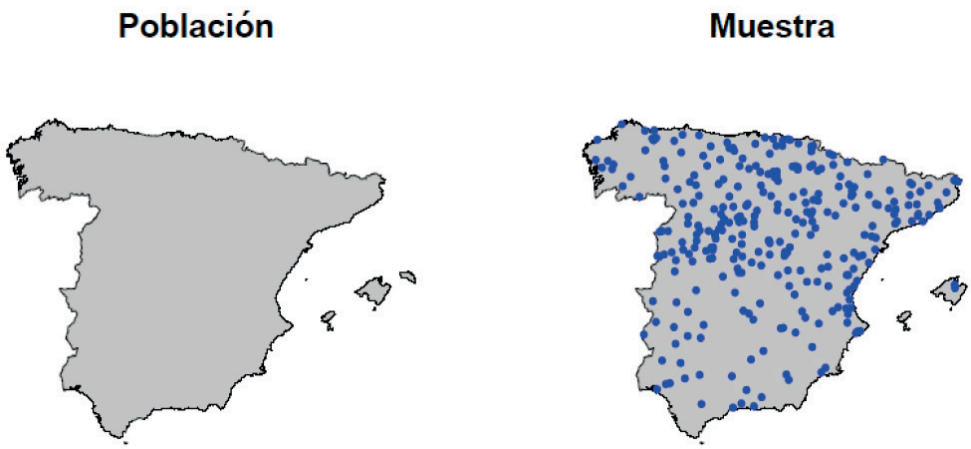


Figura 2.1. Noción gráfica de población y muestra. Elaboración propia para visualizar de forma general la población y la muestra.

2.5.6 Parámetros: son todas aquellas medidas que describen numéricamente la característica de una población. También se les denomina *valor verdadero*, ya que una característica poblacional tendrá un solo parámetro (media, varianza, etc.). Sin embargo, una población puede tener varias características y, por tanto varios parámetros.

2.5.7 Estimadores: la descripción numérica de una característica correspondiente a la muestra, se le denomina *estimador o estadígrafo*. De una población se puede obtener M número de muestras posibles y en cada una de ellas se puede cuantificar la característica, obteniéndose, por lo general, valores diferentes para cada muestra, a pesar de ser utilizado el mismo estimador o medida.

También se conoce como *estimador puntual* si se trata de un promedio, varianza, proporción, etc. Como por lo general, existe una diferencia entre el estimado y el parámetro, denominado *error*, es aconsejable utilizar el estimador por intervalos, dentro del cual deberá estar el parámetro con cierto margen de error.

2.5.8 Variable estadística: cuando se desea estudiar a los individuos de una población se acostumbra a obtener una muestra y anotar información acerca de un conjunto de características.

Ejemplos:

- X = "edad de la población"
- X = "nivel de estudios"
- X = " de hijos"

Tipos de variables: dependiendo de la naturaleza de los valores distinguimos los siguientes tipos de variables estadísticas:

Cualitativas: los valores son cualidades no medibles. Ejemplos: sexo, nacionalidad, marca de un computador,
Asimismo, las *variables cualitativas* se clasifican en:

1. **Nominales:** cuando los datos se agrupan sin ninguna jerarquía entre sí. Ejemplos: nombres de personas, de establecimientos, raza, grupos sanguíneos, estado civil.

2. **Jerárquicas (u ordinales):** cuando los datos poseen un orden, secuencia o progresión natural esperable. Ejemplos: grados de desnutrición, respuesta a un tratamiento, nivel socioeconómico, intensidad de consumo de alcohol, días de la semana.

Cuantitativas: los valores son cantidades numéricas. Ejemplos: edad, peso, duración de una pieza. Asimismo, las *variables cuantitativas* se clasifican en:

1. **Discretas:** número finito o infinito numerable de valores distintos. Ejemplos: número de hijos, número de llamadas a una central de teléfono.

2. **Continuas:** toman infinitos valores en un intervalo de la recta real. Ejemplos: peso, tiempo de respuesta de un servidor.

2.6 Estadística descriptiva: El objetivo de la estadística descriptiva es proporcionar procedimientos para:

- organizar,
- resumir,
- presentar gráficamente y
- analizar información

contenida en una muestra $X_1, \ldots, X_n$ de n individuos de una variable de interés X.

Capítulo 3

Variables cualitativas

Empezaremos con el estudio de las variables cualitativas que representan cualidades no medibles.

Ejemplo – Titanic

El fichero titanic.txt recoge información de 2201 pasajeros del naufragio del buque Titanic, en las cuatro variables cualitativas: clase, sexo, edad y superviviente.

- clase: primera, segunda, tercera, tripulación.
- sexo: hombre, mujer.
- edad: variable binaria con posibles valores: niño, adulto.
- superviviente: si, no.

Código R

```
datos <- read.table("titanic.txt", header=T)
head(datos)

        # salida de la consola
  clase      sexo  edad superviente
1 tercera hombre niño        no
2 tercera hombre niño        no
3 tercera hombre niño        no
4 tercera hombre niño        no
5 tercera hombre niño        no
6 tercera hombre niño        no

dim(datos)

        # salida de la consola
[1] 2201    4
```

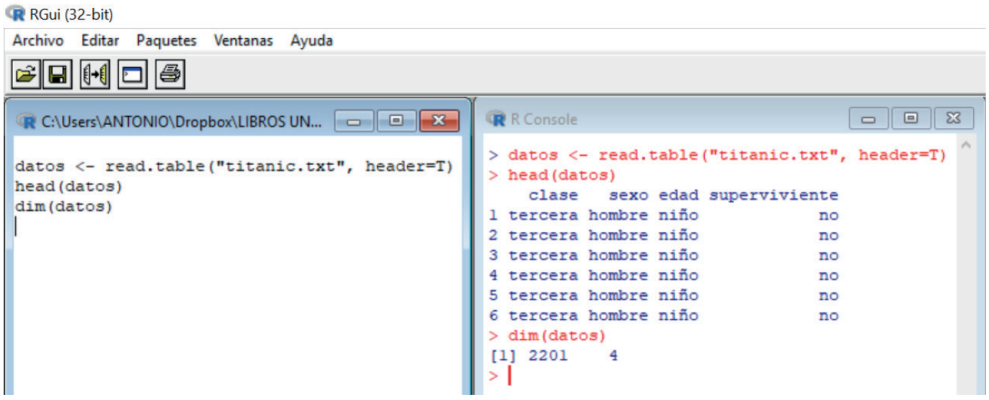


Figura 3.1. Script de datos y salida en la consola. Elaboración propia.

Sea X una variable cualitativa con k posibles valores .

- Por ejemplo la variable *clase* presenta $k = 4$ posibles valores:

, , , y .

Los posibles valores de las variables del ejemplo se obtienen con el siguiente código:

Ejemplo – Titanic

Código R

```
attach(datos)
levels(clase)

                # salida de la consola
[1] "primera"   "segunda"   "tercera"   "tripulación"

levels(sexo)
```

```
                # salida de la consola
[1] "hombre" "mujer"

levels(edad)
                # salida de la consola
[1] "adulto" "niño"

levels(superviviente)

          # salida de la consola
[1] "no" "si"
```

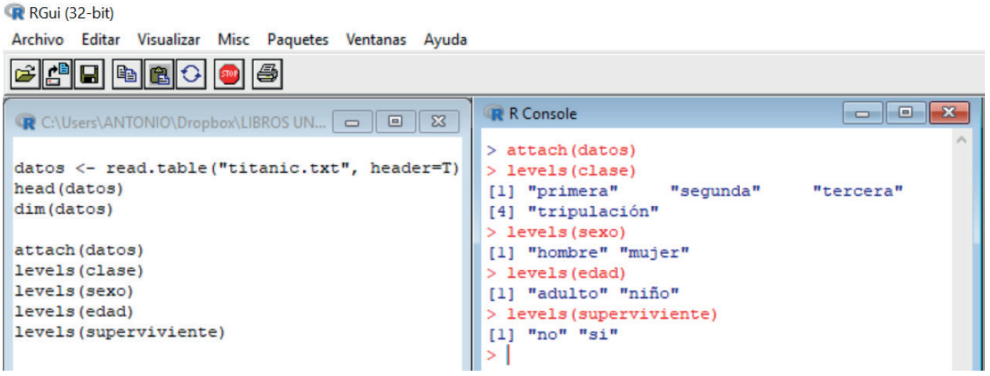


Figura 3.2. Niveles de las variables y salida en la consola. Elaboración propia.

En la Figura 3.2 se observa el código attach(datos) para entrar únicamente bajo el entorno de la base, datos, es decir con este código, se puede utilizar las variables directamente con sus nombres.

3.1 Tablas de frecuencia

Sea $X_1, \ldots, X_n$ una muestra de n observaciones de la variable X. Para cada uno de los posibles valores C_j se define:

- **Frecuencia absoluta de C_j** : número de veces que aparece C_j en la muestra. Se denota por n_j

- **Frecuencia relativa de C_j (se denota como f_j):** proporción de veces que aparece C_j en la muestra. Se denota por f_j y se calcula como el cociente entre la frecuencia absoluta y el total de individuos, es decir $f_j=\frac{n_j}{n}$.
  Si se multiplica la frecuencia relativa por 100 entonces se obtiene un porcentaje.

Ejemplo – Titanic

Código R

```
datos <- read.table("titanic.txt", header=T)

attach(datos)
n = length(clase) ; n     #  longitud de la variable clase
        # salida de la consola
[1] 2201

nj = table(clase) ; nj     #  frecuencia absoluta
        # salida de la consola
clase
   primera     segunda     tercera tripulación
       325         285         706         885

fj = nj/n ; fj   #  frecuencia relativa
        # salida de la consola
clase
   primera     segunda     tercera tripulación
 0.1476602   0.1294866   0.3207633   0.4020900

100*fj     #  porcentaje frecuencia relativa
```

```
        # salida de la consola
clase
   primera    segunda    tercera tripulación
  14.76602   12.94866   32.07633   40.20900
```

En la Figura 3.3 se observa la forma de calcular estas frecuencias absolutas y relativas en R:

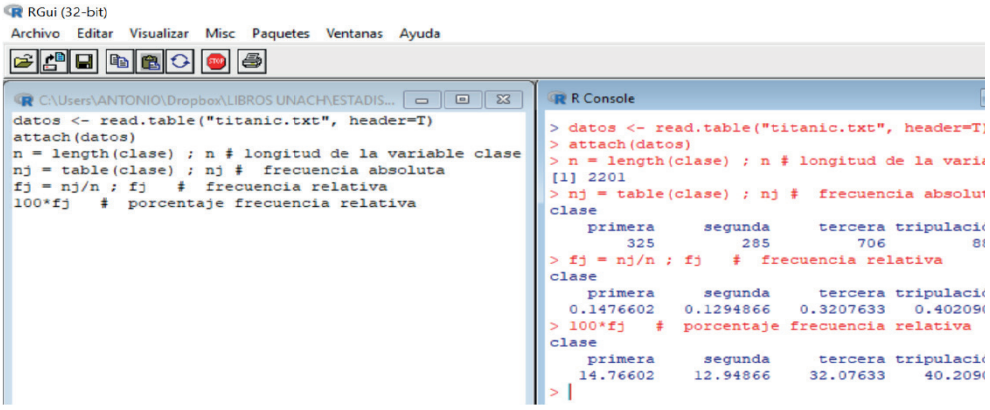


Figura 3.3. Cálculo de la frecuencia absoluta y relativa. Elaboración propia.

Análogamente se pueden calcular las frecuencias absolutas y relativas utilizando menos código en R:

Código R

```
datos <- read.table("titanic.txt", header=T)
tabla <- table(datos$clase)
tabla    #   frecuencias absolutas

        # salida de la consola
   primera    segunda    tercera tripulación
       325        285        706        885
prop.table(tabla)# frecuencias relativas
```

```
# salida de la consola
  primera    segunda    tercera tripulación
0.1476602  0.1294866  0.3207633  0.4020900
```

Tabla 3.1

Presentación de las frecuencias de la variable clase

Valor	frec. absoluta	frec. relativa (%)
Primera	325	14.8 %
Segunda	285	12.9 %
Tercera	706	32.1 %
Tripulación	885	40.2 %
Sum	2201	100 %

Elaboración propia.

En la Tabla 3.1 se observa que el naufragio titanic tiene menor número de pasajeros de segunda clase, en cambio tripulación tiene mayor número.

3.2 Representaciones gráficas de las variables cualitativas

Con el fin de comunicar rápidamente una imagen visual de los datos, se representan las frecuencias mediante distintos tipos de gráficas.

A continuación, se relacionan los tipos de representación más utilizados que conviene conocer para elegir el más adecuado a cada caso.

- Gráfico de barras
- Gráfico de sectores

3.2.1 Diagrama de barras

Para cada , se representa un rectángulo cuya altura coincide con (frecuencia absoluta) o (frecuencia relativa).

Ejemplo, Titanic: El gráfico se obtiene con el siguiente código

Código R

```
datos <- read.table("titanic.txt", header=T)
attach(datos)
n = length(clase) ; n       # longitud de la variable clase
        # salida de la consola
[1] 2201

nj = table(clase) ; nj      #   frecuencia absoluta
        # salida de la consola
clase
   primera     segunda     tercera tripulación
       325         285         706         885

fj = nj/n ; fj              #    frecuencia relativa
        # salida de la consola
clase
   primera     segunda     tercera tripulación
 0.1476602   0.1294866   0.3207633   0.4020900

100*fj                      #  porcentaje frecuencia relativa
        # salida de la consola
clase
   primera     segunda     tercera tripulación
  14.76602    12.94866    32.07633    40.20900

par(mfcol=c(1,2))       #  Gráficos de barras
barplot(nj , main='frecuencias absolutas')
barplot(fj , main='frecuencias relativas')
```

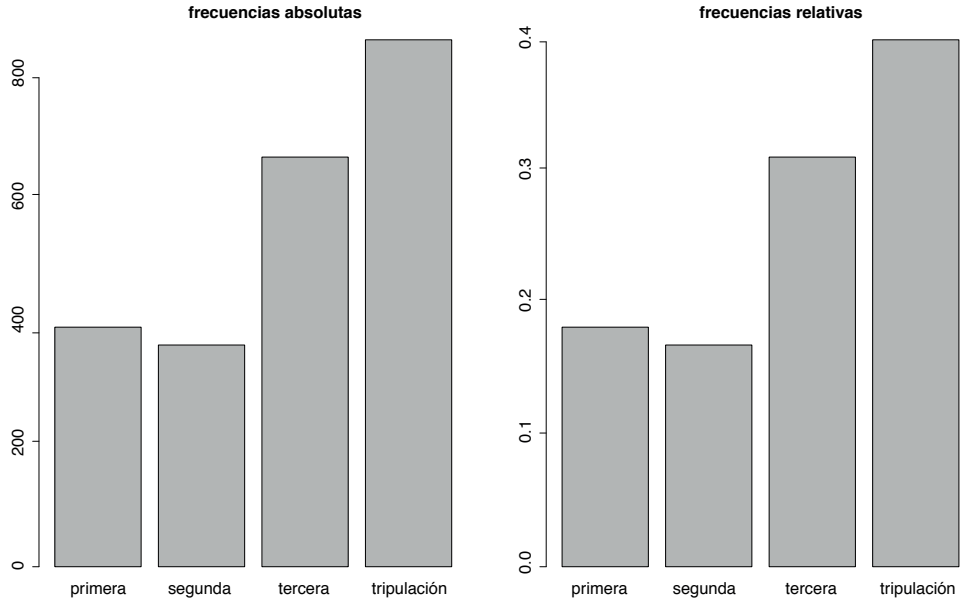


Figura 3.4. Salida gráfica de las frecuencias absolutas y relativas en barras de la variable clase. Elaboración propia.

En la Figura 3.4 se observa que los pasajeros del naufragio titanic de segunda clase tiene la barra de menor altura (menor número de pasajeros), en cambio la clase tripulación tiene la barra más alta (mayor número de pasajeros).

Utilizando la librería *plotrix* de R (ver instalación de librerías en anexos), se puede realizar gráficos de barras en 3 dimensiones de forma cilíndrica mediante la función *barp* de esta librería:

```
datos <- read.table("titanic.txt", header=T)
attach(datos)
library(plotrix)
barp(table(clase), col = "blue", cylindrical = TRUE,
   main = "Frecuencias absolutas", names.arg = names(table(clase)))
```

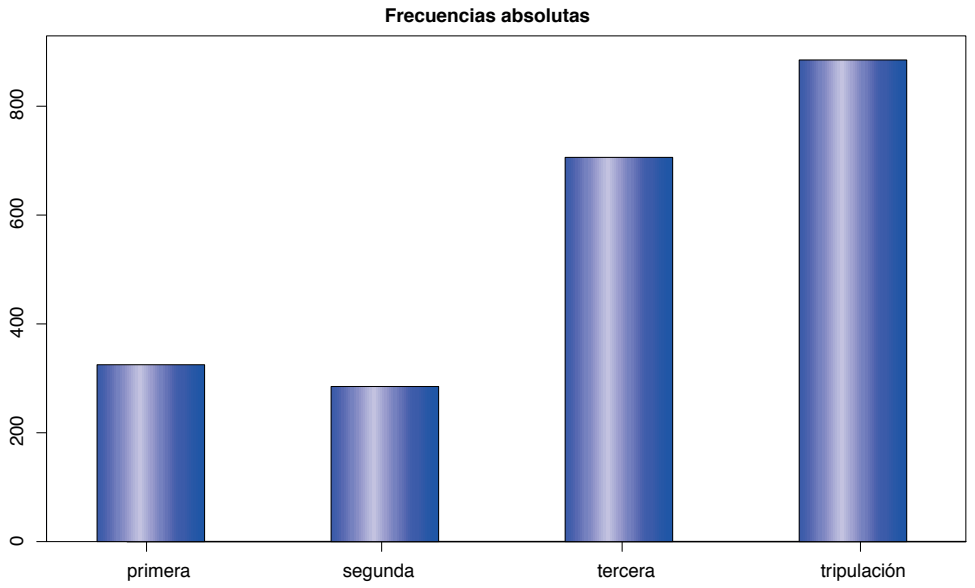


Figura 3.5. Salida gráfica de las frecuencias absolutas y relativas en barras cilíndricas de la variable clase. Elaboración propia.

En la Figura 3.5, se observa que los pasajeros del naufragio titanic de segunda clase tiene la barra cilíndrica de menor altura (menor número de pasajeros), en cambio la clase tripulación tiene la barra cilíndrica más alta (mayor número de pasajeros).

3.2.2 Gráfico de sectores de las variables cualitativas

Se descompone un círculo en sectores de área proporcional a la frecuencia de la modalidad correspondiente.

Ejemplo, Titanic: El gráfico se obtiene con el siguiente código

Código R

```
datos <- read.table("titanic.txt", header = T)
attach(datos)
n = length(clase) ; n          #   longitud de la variable clase
```

```
nj = table(clase) ; nj          #   frecuencia absoluta
pie(nj, col=rainbow(6), main="Diagrama Pastel")      # gráficos de
sectores
```

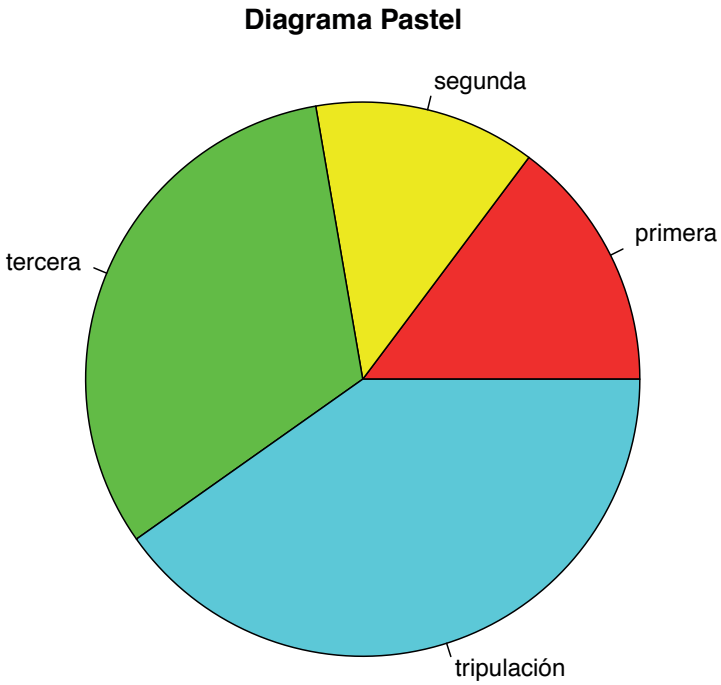


Figura 3.6. Salida gráfica de las frecuencias en sectores de la variable clase. Elaboración Propia.

En la Figura 3.6 se observa que los pasajeros del naufragio titanic de segunda clase tiene menor área (menor número de pasajeros), en cambio la clase tripulación tiene mayor área (mayor número de pasajeros).

Los gráficos de sectores en tres dimensiones también se pueden realizar utilizando la función *pie3D* de la librería *plotrix*:

```
datos <- read.table("titanic.txt", header=T)
attach(datos)
library(plotrix)
pie3D(table(clase), explode=0.1, main="Gráfico de sectores",
      labels = names(table(clase)), labelcex=1)
```

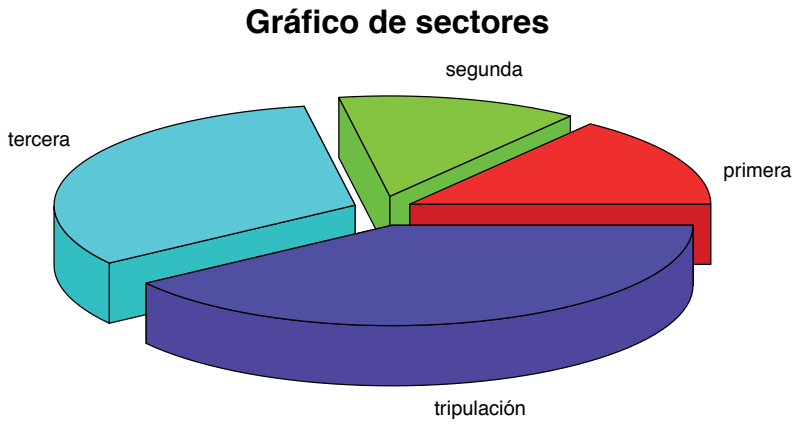


Figura 3.7. Salida gráfica en 3D de las frecuencias en sectores de la variable clase. Elaboración Propia.

En la Figura 3.7 se observa que los pasajeros del naufragio titanic de segunda clase tiene menor volumen (menor número de pasajeros), en cambio la clase tripulación tiene mayor volumen (mayor número de pasajeros).

3.2.3 Tablas multidimensionales

La función *table* puede ser utilizada para hacer tablas de más de una variable de la siguiente manera:

```
Tabla = table(clase, superviviente) ;  Tabla
```

Los resultados se observan en la Tabla 3.2, que son frecuencias absolutas de dos variables, *clase* y *superviviente*:

Tabla 3.2

Frecuencias absolutas de dos variables cualitativas

Clase/superviviente	no	si
Primera	122	203
Segunda	167	118
Tercera	528	178

Tripulación	673	212

Elaboración propia.

En la Tabla 3.2 se observa que los pasajeros de segunda clase son menos supervivientes, mientras que la clase tripulación han logrado la mayor supervivencia.

Ejemplo – Titanic

Código R

```
datos <- read.table("titanic.txt", header=T)
attach(datos)
tabla = table(clase, superviviente) ; tabla
        # Salida de la consola
         superviviente
clase            no    si
 primera        122  203
 segunda        167  118
 tercera        528  178
 tripulación    673  212
```

Con el siguiente código en R,

```
addmargins(tabla)
```

Se añaden a la tabla anterior las frecuencias marginales (sumas por filas y columnas).

Tabla 3.3

Frecuencias marginales y absolutas

	no	**si**	**Sum**
Primera	122	203	325
Segunda	167	118	285

Tercera	528	178	706
Tripulación	673	212	885
Sum	1490.00	711.00	2201

Elaboración propia.

La tabla de frecuencias relativas (por filas) se obtienen con el siguiente código en R:

```
datos <- read.table("titanic.txt", header=T)
attach(datos)
tabla = table(clase, superviviente)
tabla = prop.table(tabla, 1)  # El 1 se refiere al cálculo de frecuencias por filas
tabla
        # Salida de la consola
                    superviviente
clase                no          si
  primera       0.3753846   0.6246154
  segunda       0.5859649   0.4140351
  tercera       0.7478754   0.2521246
  tripulación   0.7604520   0.2395480
```

Nótese como la probabilidad de supervivencia es muy superior en *primera* que en resto de las clases.

También es necesario la representación gráfica en barras de frecuencias de tablas de dos o más variables.

Código R

```
titanic=read.table("titanic.txt", header = T)
attach(titanic)
A = table(clase, superviviente)
barplot(A, legend = rownames(A), beside=T)
```

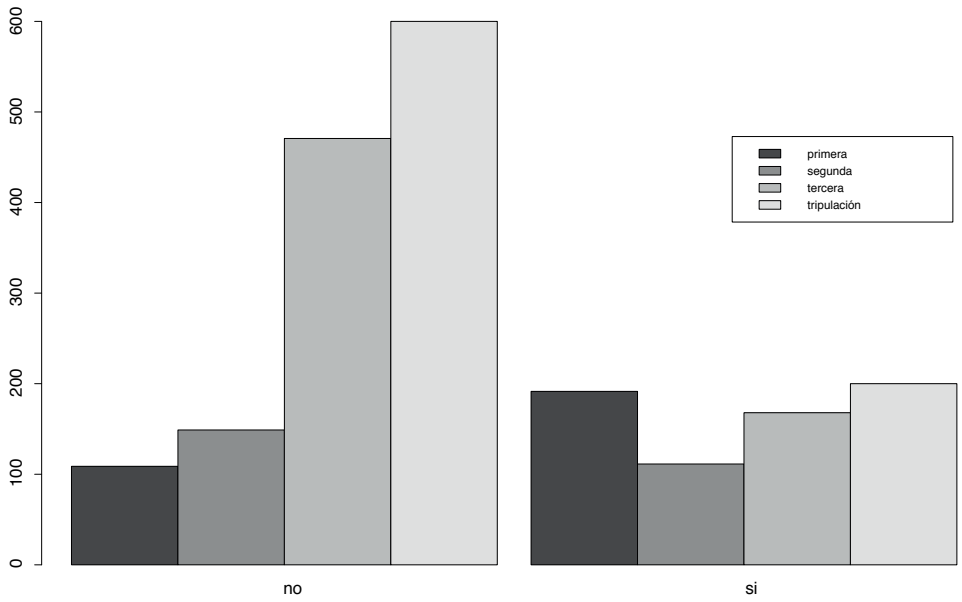


Figura 3.8. Frecuencias absolutas de dos variables cualitativas. Elaboración propia.

En la Figura 3.8 se observa que los pasajeros de segunda clase tiene la altura menor de la barra, mientras que la clase tripulación tiene la altura mayor de la barra con respecto a la variable superviviente.

De forma análoga se utiliza la función *barp* de la librería *plotrix* para realizar el diagrama de barras en 3 dimensiones en forma cilíndrica con el siguiente código en R:

```
titanic=read.table("titanic.txt", header = T)
attach(titanic)
A = table(clase, superviviente)
library(plotrix)
barp(A, names.arg = colnames(A), cylindrical = TRUE, shadow =
TRUE,
        col=rainbow(4), legend.lab = rownames(A), legend.pos =
list(x=0.7,y=600))
```

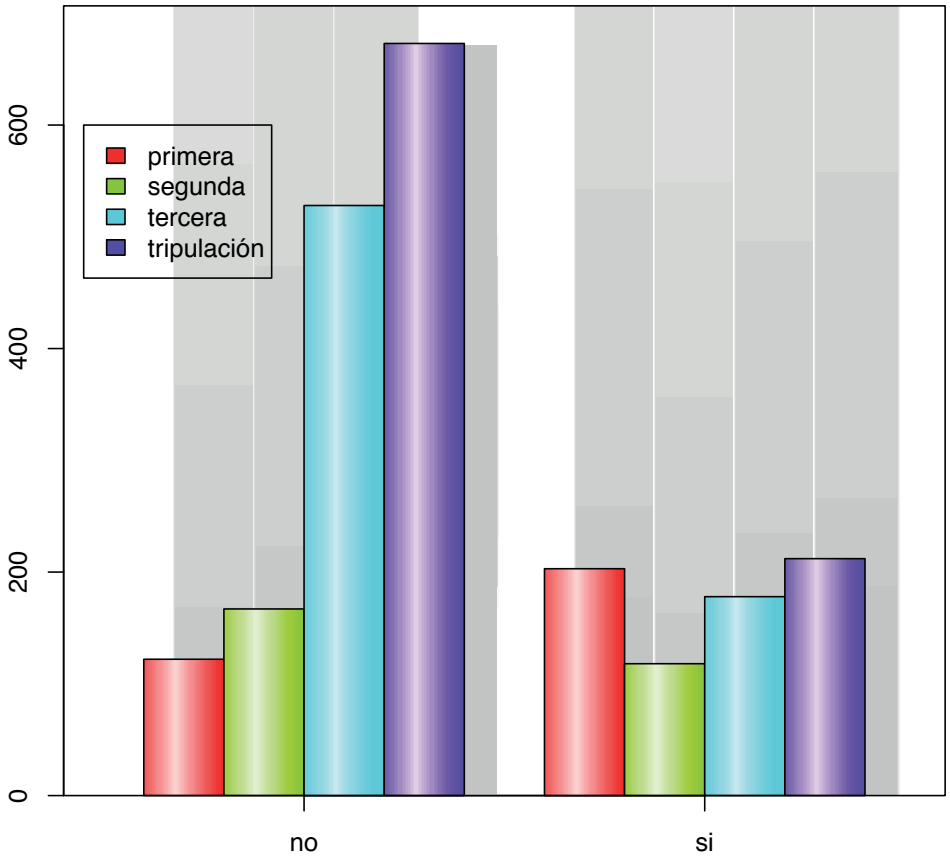


Figura 3.9. Frecuencias absolutas de dos variables cualitativas con apariencia en 3 dimensiones y barras cilíndricas. Elaboación propia.

Aplicación de R en la actualidad

En el capítulo de libro publicado: Gestión de la práctica pre profesional en el currículo de la educación superior - caso Universidad Nacional de Chimborazo (https://redipe.org/editorial/educacion-contemporanea-calidad-educativa-y-buen-vivir/), los autores hacen un estudio descriptivo y de análisis de correspondencias de variables cualitativas, centrando su atención en la práctica pre profesional como eje transversal del currículo de la educación superior, surge como parte de un proyecto de investigación cuyo propósito general fue rediseñar los procesos

académicos de la Universidad Nacional de Chimborazo, institución educativa ubicada en el centro del territorio ecuatoriano. Como parte del estudio indicado, aparece la necesidad de identificar los procedimientos implementados para la gestión de las prácticas pre profesionales. y su impacto en la formación de los futuros profesionales, orientado al diseño de una propuesta enmarcada en la construcción de un modelo de prácticas sistémico e integrador. En esta publicación se observa la importancia de la aplicación de herramientas del *software* R en variables cualitativas.

Problemas propuestos para realizar con el *software* R

1. Las lesiones observadas en edificios construidos con cemento aluminoso, en los años cincuenta en determinada zona geográfica, han sido clasificadas como *leves*, *graves* y *muy graves*. Los siguientes datos reflejan el resultado de la observación de 50 edificios afectados.
http://graduados.unach.edu.ec/estadistica-descriptiva-con-R/Lesiones_edificios.txt

   a. Construya una tabla de frecuencias
   b. Construya un gráfico circular y otro de barras con las funciones *barplot* y *barp* del *software* R.
   c. Comente estos resultados

2. En un estudio con el fin de relacionar el consumo de licor y la hipertensión, se tomaron los siguientes datos correspondientes a una muestra de 280 personas.
http://graduados.unach.edu.ec/estadistica-descriptiva-con-R/Consumo_licor.txt

   a. Haga una representación adecuada tanto gráfica como en tablas de frecuencias a la información anterior.

b. ¿Observa alguna relación entre las dos variables? Explique

3. A un curso de bachillerato de último año, se le pregunto por la carrera por la cual sentían una mayor inclinación, al continuar estudios universitarios. Estos fueron sus respuestas; A-Administración; C-Contabilidad; D-Derecho; E-Economía; I-Ingeniería; M-Medicina; O-Odontología.
http://graduados.unach.edu.ec/estadistica-descriptiva-con-R/Carreras_Universitarias.txt

a. Construya una tabla de frecuencias
b. Construya un gráfico circular y otro de barras con las funciones *barplot* y *barp* del *software* R.
c. Comente estos resultados

Capítulo 4

Variables cuantitativas discretas

La característica principal de las variables cuantitativas discretas es tener un número finito o infinito numerable de valores distintos.

4.1 Tablas de frecuencia

Para variables discretas cuantitativas los posibles valores pueden ser ordenados, de forma que
$C_1 < \cdots < C_k$

Para cada C_j, se definen las frecuencias absolutas n_j y relativas f_j exactamente igual a como ya se ha hecho para las variables cualitativas.

Además, ahora se definen las frecuencias acumuladas:

- **frecuencia absoluta acumulada de C_j**: número de observaciones que presenta dicha modalidad o alguna de las anteriores. Se denota por N_j y viene dada por

  $$N_j = n_1 + \ldots + n_j$$

  Nota: En las variables cualitativas, como son las del ejemplo Titanic, no tenía sentido las frecuencias acumuladas ya que no es posible establecer orden en los valores de la variable.

- **frecuencia relativa acumulada de C_j**: Se denota por F_j y su valor viene dado por

  $$F_j = N_j / n = f_1 + \ldots + f_j$$

De las definiciones anteriores se obtiene la Tabla 4.1.

Tabla 4.1

Frecuencias de la variable cuantitativa discreta

Modalidad	frecuencia absoluta	frecuencia relativa	frec. Absoluta acumulada	frec. Relativa acumulada
C_1	n_1	f_1	N_1	F_1
⋮	⋮	⋮	⋮	⋮
C_j	n_j	f_j	N_j	F_j
⋮	⋮	⋮	⋮	⋮
C_k	n_k	f_k	N_k=n	F_k=1
Total	n	1		

Elaboración propia.

Ejemplo – Tráfico

En un estudio de tráfico se ha recabado información acerca del número de ocupantes en los automóviles. Para ello se contó el número de ocupantes en 40 automóviles.

1 3 2 2 3 1 1 2 2 1 1 4 3 1 3 2 3 2 2 2
1 2 5 1 3 1 2 1 3 1 4 1 1 3 4 2 2 1 1 4

Obteniendo la Tabla 4.2 de frecuencias,

Tabla 4.2

Frecuencias de la variable tráfico

clase	f.abs.	f.rel.	f.abs.acu.	f.rel.acu.
1	15	0.38	15	0.38
2	12	0.30	27	0.68
3	8	0.20	35	0.88
4	4	0.10	39	0.98
5	1	0.03	40	1.00
	40	1		

Elaboración propia.

Las tablas anteriores han sido obtenidas con el siguiente código:

Código R

```
ocupantes <- c(1, 3, 2, 2, 3, 1, 1, 2, 2, 1, 1, 4, 3, 1, 3, 2, 3, 2, 2, 2,
               1, 2, 5, 1, 3, 1, 2, 1, 3, 1, 4, 1, 1, 3, 4, 2, 2, 1, 1, 4)
nj = table(ocupantes) ; nj     #    frecuencias absolutas
        # Salida de la consola
ocupantes
 1  2  3  4  5
15 12  8  4  1


Nj = cumsum(nj) ;  Nj      #     frecuencias absolutas acumuladas

        # Salida de la consola
1   2    3   4   5

15  27  35  39  40

fj = prop.table(nj) ; fj        #  frecuencia relativa

        # Salida de la consola
ocupantes
    1     2     3     4     5
0.375 0.300 0.200 0.100 0.025


Fj = cumsum(fj) ;  Fj     # frecuencia relativa acumulada

        # Salida de la consola
   1       2       3       4       5
0.375   0.675   0.875   0.975   1.000
```

4.2 Representaciones gráficas

Con las frecuencias obtenidas se pueden hacer resúmenes gráficos que se realizan de forma similar al caso de las variables cualitativas.

Ejemplo – Tráfico

Código R

```
ocupantes <- c(1, 3, 2, 2, 3, 1, 1, 2, 2, 1, 1, 4, 3, 1, 3, 2, 3, 2, 2, 2,
               1, 2, 5, 1, 3, 1, 2, 1, 3, 1, 4, 1, 1, 3, 4, 2, 2, 1, 1, 4)
nj = table(ocupantes) ; nj     #     frecuencias absolutas
        # Salida de la consola
ocupantes
 1   2   3   4   5
15  12   8   4   1


Nj = cumsum(nj) ;  Nj      #     frecuencias absolutas acumuladas

        # Salida de la consola
1   2   3   4   5
15  27  35  39  40

fj = prop.table(nj) ; fj        #  frecuencia relativa

        # Salida de la consola
ocupantes
    1     2     3     4     5
0.375 0.300 0.200 0.100 0.025

Fj = cumsum(fj) ;  Fj     # frecuencia relativa acumulada
```

```
        # Salida de la consola
   1       2     3       4     5
0.375 0.675 0.875 0.975 1.000
```

Las representaciones gráficas de las frecuencias obtenidas de la variable *ocupantes* se realizan de forma análoga a las variables cualitativas.

En cuestiones de organizar las gráficas se utiliza la función *layout* que divide la pantalla gráfica en forma matricial en filas y columnas, ubicando en cada posición una gráfica, así por ejemplo se ubica por filas los 5 gráficos siguientes:

```
layout(matrix(c(1, 2, 5, 3, 4, 5), 2, 3, byrow=TRUE), respect=TRUE)
# Ventana gráfica
                    # dividida en 2 filas y 3 columnas donde el quinto
gráfico es el único de
                  # la tercera columna
barplot(nj, main = "frecuencia absolutas", xlab = 'ocupantes')
barplot(fj, main = "frecuencia relativas", xlab = 'ocupantes')
barplot(Nj, main = "frecuencia absolutas acumuladas", xlab =
'ocupantes')
barplot(Fj, main = "frecuencia relativas acumuladas", xlab =
'ocupantes')
pie(nj, col=rainbow(6), main = 'ocupantes')
```

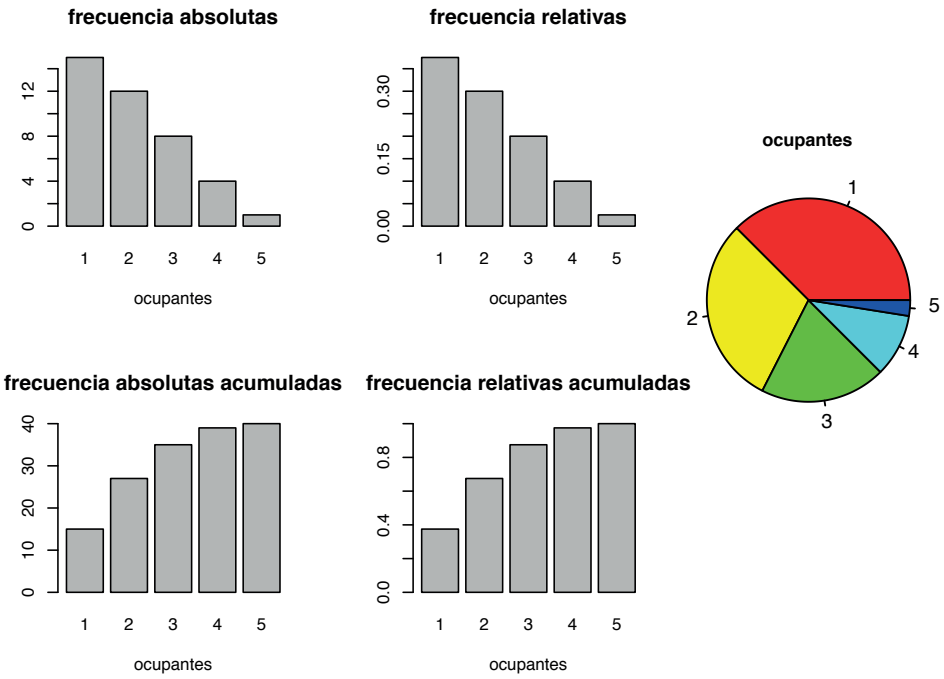


Figura 4.1. Representación gráfica en barras y sectores de las frecuencias de la variable ocupantes. Elaboración propia.

En la Figura 4.1 se observan 5 diagramas, 4 de barras y uno de sectores, en los que el número de automóviles con 1 ocupante es mayor, y también la acumulación de frecuencias empieza con los automóviles de 1 ocupante hasta los automóviles con 5 ocupantes.

Utilizando las funciones de la librería *plotrix* se tiene la presentación gráfica de las frecuencias con visualización en 3 dimensiones, siguiendo el código en R:

```
library(plotrix)
layout(matrix(c(1, 2, 5, 3, 4, 5), 2, 3, byrow=TRUE), respect=TRUE)
# Ventana gráfica
                    # dividida en 2 filas y 3 columnas donde el quinto
gráfico es el único de
```

```
                # la tercera columna
barp(nj, col="blue",cylindrical=TRUE, main="Frecuencias absolutas",
    names.arg = names(nj))
barp(fj, col="blue",cylindrical=TRUE, main = "Frecuencias relativas",
    names.arg = names(fj))
barp(Nj, col = "blue", cylindrical = TRUE, main="Frecuencia absolutas \n acumuladas",
    names.arg=names(Nj))
barp(Fj, col = "blue", cylindrical=TRUE, main = "Frecuencia relativas \n acumuladas",
    names.arg = names(Fj))
pie3D(nj, explode = 0.1, main="Gráfico de sectores", labels = names(nj), labelcex = 1)
```

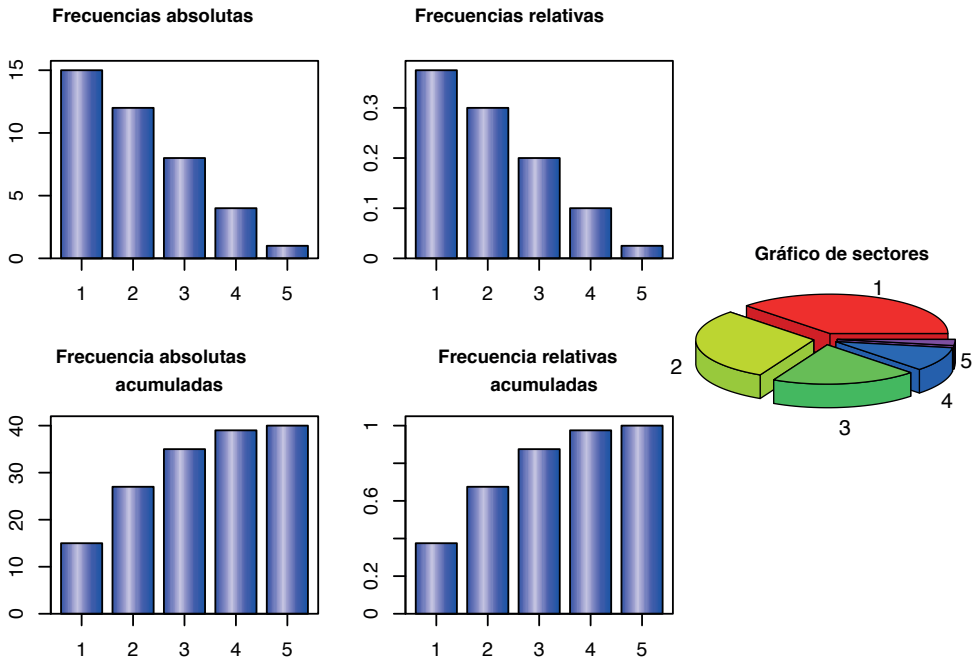


Figura 4.2. Frecuencias con visualización en 3 dimensiones. Elaboración propia.

4.3 Función de distribución empírica

Dada una muestra $X_1, \cdots, X_n$ se define la función de distribución empírica en un punto x como la proporción de puntos en la muestra menores o iguales a x

$$F_n(x) = \frac{\textit{número de valores menores o iguales a } x}{n}$$

Nótese que para variables discretas, se tiene:

- F_n toma valores en el intervalo [0,1],
- Es una función escalonada creciente,
- Los saltos de esta función se dan en cada uno de los valores C_j. Además el salto en cada C_j coincide con la correspondiente frecuencia relativa f_j.

Ejemplo – Tráfico

Código R

```
ocupantes <- c(1, 3, 2, 2, 3, 1, 1, 2, 2, 1, 1, 4, 3, 1, 3, 2, 3, 2, 2, 2,
               1, 2, 5, 1, 3, 1, 2, 1, 3, 1, 4, 1, 1, 3, 4, 2, 2, 1, 1, 4)
plot(ecdf(ocupantes), verticals = T, main = "Distribucion empirica",
xlab = 'ocupantes',
     col = 'red', lwd=2)
```

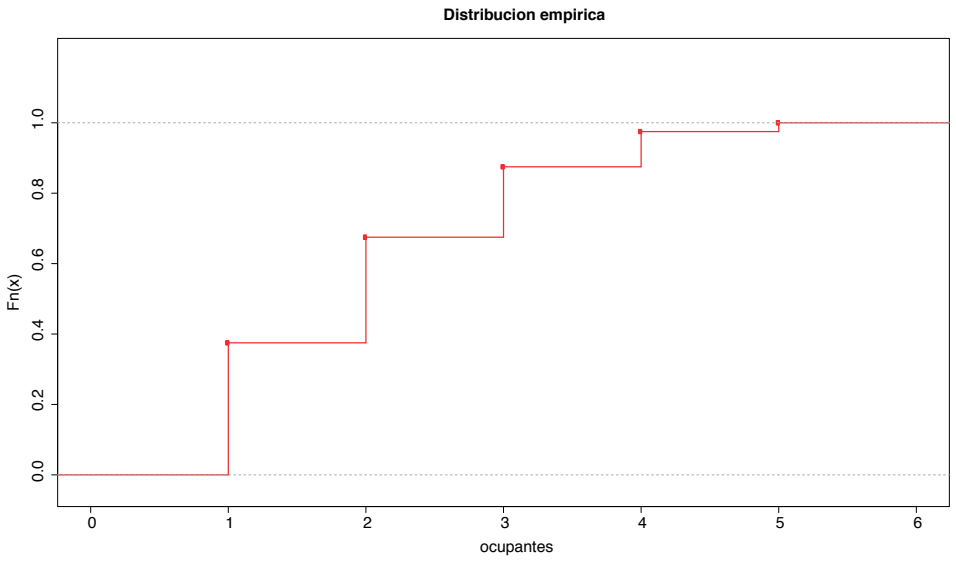


Figura 4.3. Distribución empírica de la variable ocupantes mediante la función ecdf. Elaboración propia.

En la Figura 4.3 se observa la distribución de forma empírica de los datos de número de automóviles por ocupantes.

Ejemplo – Número de hijos por mujer

En la siguiente tabla se presenta el número de hijos por mujer en el año 2008 para aquellas madres que tuvieron hijos hasta este año. Los datos corresponden a una simulación.

Tabla 4.3

Frecuencia del número de hijos por mujer en el año 2008

	1	2	3	4	5	>5
Número de mujeres	13279	8348	1233	228	56	31

Elaboración propia.

Código R

```
hijos = matrix(c(13279, 8348, 1233, 228, 56, 31), 1)
colnames(hijos) = c("1", "2", "3", "4", "5", ">5")
rownames(hijos) = "número de mujeres"
hijos
```

```
# Salida de la consola
                    1     2     3    4  5  >5
número de mujeres 13279 8348 1233 228 56  31
```

Tabla 4.4
Frecuencias de la variable hijos

hijos	f.abs.	f.rel.	f.abs.acu.	f.rel.acu.
1	13279	0.57	13279	0.57
2	8348	0.36	21627	0.93
3	1233	0.05	22860	0.99
4	228	0.01	23088	1.00
5	56	0.00	23144	1.00
>5	31	0.00	23175	1.00

Elaboración propia.

Los resultados de la Tabla 4.4 de frecuencias de la variable hijos se obtienen con el siguiente código:

Código R

```
hijos = matrix(c(13279, 8348, 1233,  228,56, 31), 1)
nj = hijos ; Nj = cumsum(nj) ; Nj
fj = prop.table(nj) ; Fj = cumsum(fj)
tabla = data.frame(nj=as.vector(nj) , fj = as.vector(fj), Nj=as.
vector(Nj),
                   Fj = as.vector(Fj))
rownames(tabla) = colnames(hijos)
tabla
```

```
# Salida de la consola
  nj       fj          Nj      Fj
1 13279 0.572988134 13279 0.5729881
2  8348 0.360215750 21627 0.9332039
3  1233 0.053203883 22860 0.9864078
4   228 0.009838188 23088 0.9962460
5    56 0.002416397 23144 0.9986624
6    31 0.001337648 23175 1.0000000
```

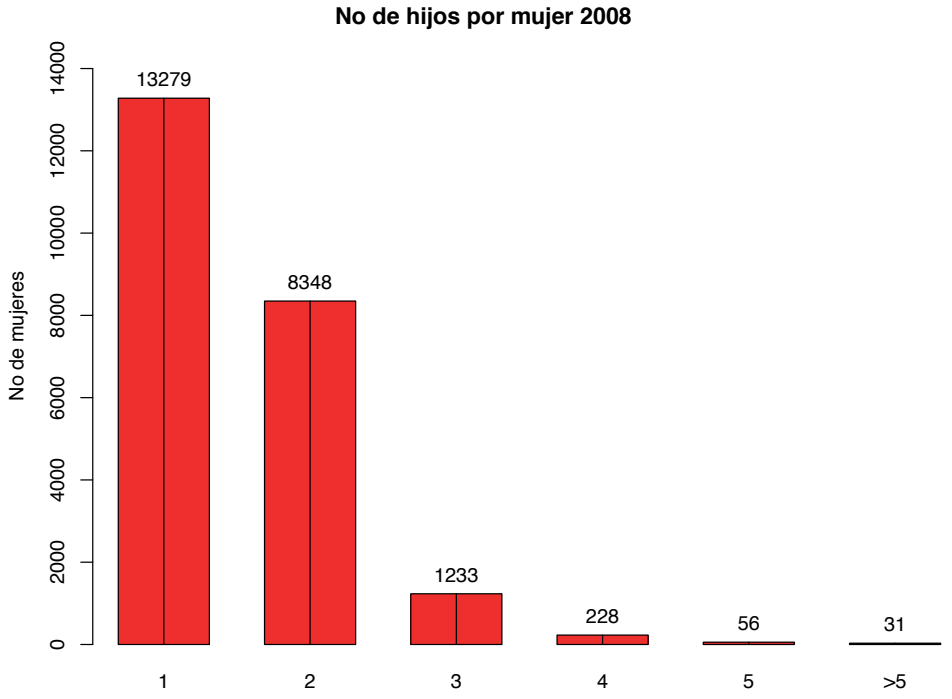


Figura 4.4. Frecuencias absolutas con texto en las barras. Elaboración propia.

En la Figura 4.4 se observa el valor de la frecuencia absoluta en cada barra de la gráfica, y esta se obtiene mediante el siguiente código en R:

Código R

```
hijos = matrix(c(13279, 8348, 1233, 228, 56, 31), 1)
colnames(hijos)=c("1", "2", "3", "4", "5", ">5")
r = barplot(hijos, col = 'red', ylim = c(0, 15000),
                main = "No de hijos por mujer 2008", ylab="No de
mujeres")
lines(r, hijos, type = 'h')
text(r, hijos, hijos, pos=3)
```

Ejemplo – Sector económico

En el fichero económico.txt están los datos de la población (en miles de mujeres) ocupada por sector económico (CNAE 2009) en el año 2009 para Galicia - España (datos obtenidos de http://igualdade.xunta.gal/es/content/estructura-cnae-2009-0).

Código R

```
datos = read.table("económico.txt", header = TRUE)
head(datos)
```

```
· # Salida de la consola
```

	provincia	Agricultura.e.pesca	Industria	Construción	Servizos
1	Galicia	91.6	194.1	115.7	750.1
2	A Coruña	31.0	73.9	53.2	341.0
3	Lugo	24.6	17.9	13.1	82.3
4	Orense	10.2	23.0	10.8	78.6
5	Pontevedra	25.9	79.2	38.6	248.3

La Figura 4.5, se obtiene con el código que sigue:

Código R

```
datos = read.table("económico.txt", header = TRUE)
head(datos)

      · # Salida de la consola
```

	provincia	Agricultura.e.pesca	Industria	Construción	Servizos
1	Galicia	91.6	194.1	115.7	750.1
2	A Coruña	31.0	73.9	53.2	341.0
3	Lugo	24.6	17.9	13.1	82.3
4	Orense	10.2	23.0	10.8	78.6
5	Pontevedra	25.9	79.2	38.6	248.3

```
      # Convertimos en matriz (interesa para hacer los gráficos)
datos2 = as.matrix(datos[,-1])
datos2

      # Salida de la consola
```

	Agricultura.e.pesca	Industria	Construción	Servizos
[1,]	91.6	194.1	115.7	750.1
[2,]	31.0	73.9	53.2	341.0
[3,]	24.6	17.9	13.1	82.3
[4,]	10.2	23.0	10.8	78.6
[5,]	25.9	79.2	38.6	248.3

```
colnames (datos2)[1]="Agri.y.pesca"  #  El nombre original es muy
largo
galicia = datos2[1,]
galicia
      # Salida de la consola
```

```
Agri.y.pesca   Industria  Construción   Servizos
        91.6        194.1        115.7        750.1
r = barplot(galicia, main = "Poblacion ocupada por sector economico",
                ylab = "No de trabajadores")
lines(r, galicia, type = 'h')
text(r, galicia, galicia, pos=1)
```

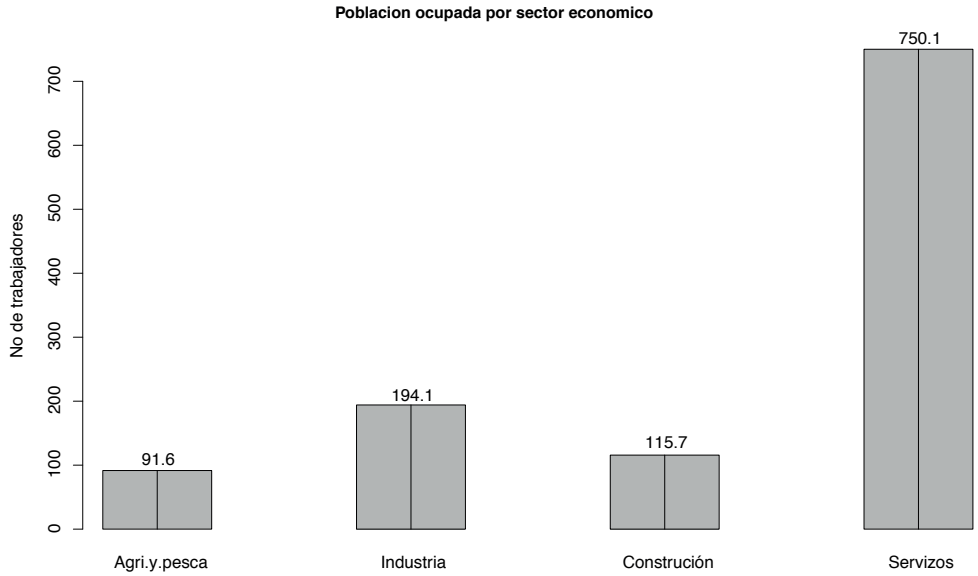


Figura 4.5. Frecuencias absolutas del sector económico en Galicia–España. Elaboración propia.

En la Figura 4.5 se observa el reparto del sector económico en la región de Galicia-España, siendo el menor número de persona dedicadas a la agricultura y pesca.

Estudio gráfico por provincia

Los siguientes gráficos se obtienen con el código que sigue:

```
datos = read.table("económico.txt", header = TRUE)
datos
```

```
# Salida de la consola
```

	Provincia	Agricultura.e.pesca	Industria	Construción	Servizos
1	Galicia	91.6	194.1	115.7	750.1
2	A Coruña	31.0	73.9	53.2	341.0
3	Lugo	24.6	17.9	13.1	82.3
4	Orense	10.2	23.0	10.8	78.6
5	Pontevedra	25.9	79.2	38.6	248.3

```
datos2 = as.matrix(datos[,-1])
datos2
```

```
# Salida de la consola
```

	Agricultura.e.pesca	Industria	Construción	Servizos
[1,]	91.6	194.1	115.7	750.1
[2,]	31.0	73.9	53.2	341.0
[3,]	24.6	17.9	13.1	82.3
[4,]	10.2	23.0	10.8	78.6
[5,]	25.9	79.2	38.6	248.3

```
galicia = datos2[-1,]
galicia
```

```
# Salida de la consola
```

	Agricultura.e.pesca	Industria	Construción	Servizos
[1,]	31.0	73.9	53.2	341.0
[2,]	24.6	17.9	13.1	82.3
[3,]	10.2	23.0	10.8	78.6
[4,]	25.9	79.2	38.6	248.3

```
galicia = t(galicia)    #  interesa trasponer
galicia
```

```
[,1] [,2] [,3]  [,4]
```

Agricultura.e.pesca	31.0	24.6	10.2	25.9
Industria	73.9	17.9	23.0	79.2
Construción	53.2	13.1	10.8	38.6
Servizos	341.0	82.3	78.6	248.3

La Figura 4.6 se obtiene con el siguiente código:

```
colores = c("lightblue", "mistyrose", "lightcyan","lavender")
rownames (galicia)[1] = "Agri.y.pesca"    # El nombre original es
muy largo
colnames(galicia) = c('A Coruña', 'Lugo', 'Orense', 'Pontevedra')
barplot(galicia, col = colores , legend = rownames(galicia),
        main = "Poblacion ocupada por sector economico")
```

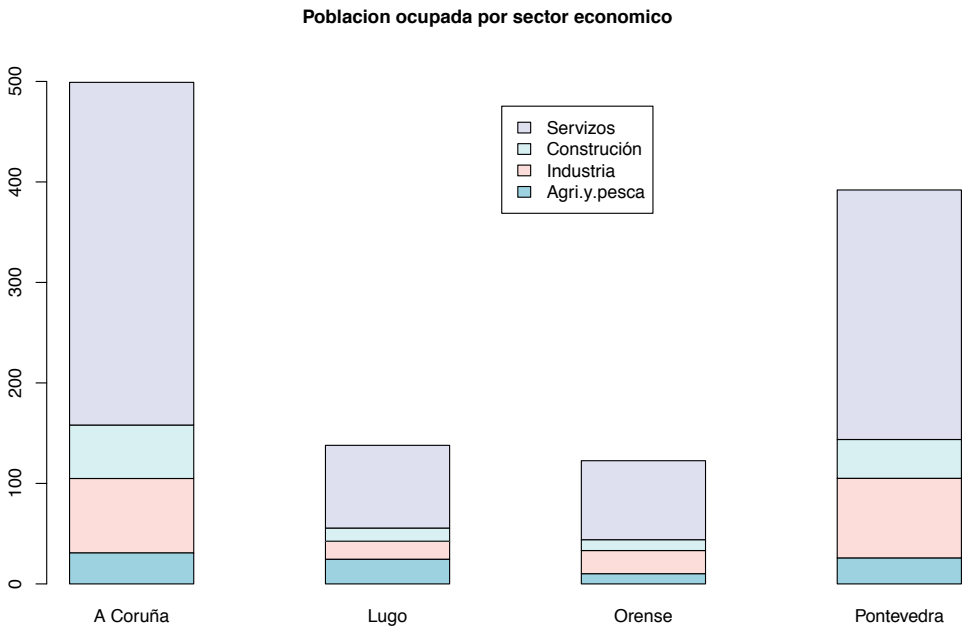


Figura 4.6. Frecuencias absolutas de cada provincia en Galicia-España. Elaboración propia.

La Figura 4.7 se obtiene con el siguiente código:

```
barplot(galicia, col=colores, legend = rownames(galicia),
      beside = T, main = "Población ocupada por sector económico")
```

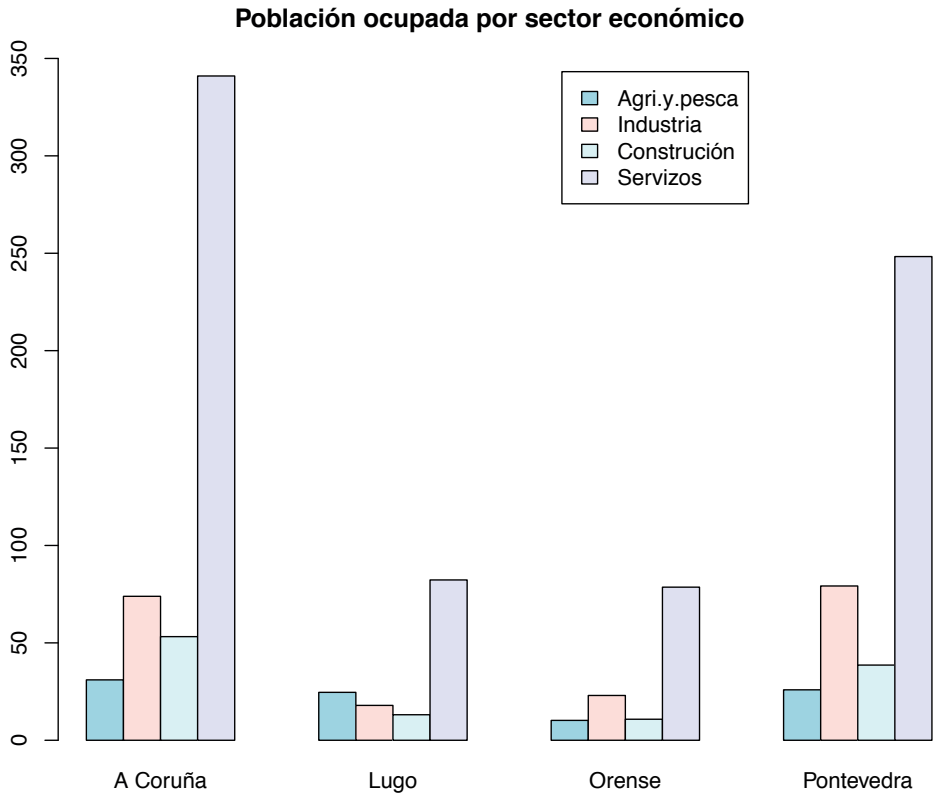


Figura 4.7. Frecuencias absolutas por provincia utilizando el código beside = T en la función barplot. Elaboración propia.

La Figura 4.8 se obtiene con el código siguiente:

```
galicia2 = prop.table(galicia, 2) # probabilidades
galicia2
      # Salida de la consola
```

	[,1]	[,2]	[,3]	[,4]
Agricultura.e.pesca	0.0621118	0.17839014	0.08319739	0.06607143
Industria	0 1480665	0.12980421	0.18760196	0.20204082
Construción	0.1065919	0.09499637	0.08809135	0.09846939
Servizos	0.6832298	0.59680928	0.64110930	0.63341837

```
rownames (galicia)[1] = "Agri.y.pesca"  # El nombre original es muy largo
colnames(galicia2) = c('A Coruña','Lugo', 'Orense', 'Pontevedra')
barplot(galicia2, col=colores,legend = rownames(galicia),
        main="Poblacion ocupada por sector economico")
```

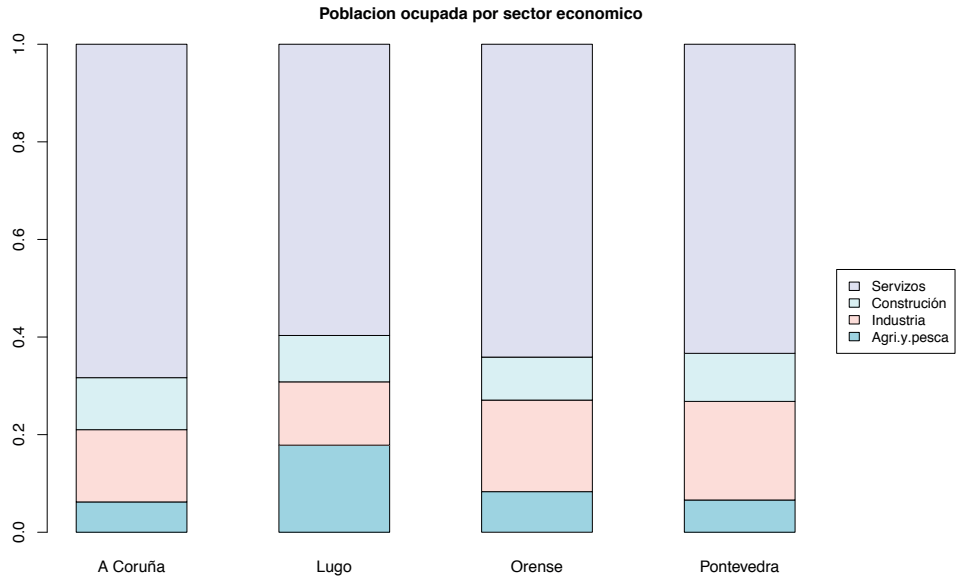


Figura 4.8. Frecuencias en escala de porcentaje de cada provincia de Galicia – España. Elaboración Propia.

En las figuras 4.6, 4.7 y 4.8 se observan que la mayor cantidad de personas se dedican al sector de servizos.

Ejemplo – Sida

El fichero *sida.txt* contiene la serie de casos diagnosticados de sida por año y sexo.

Código R

```
datos = read.table("económico.txt", header = TRUE)
head(datos)
```

```
# Salida de la consola
  año Varones Mujeres Total
1 1981      1       0     1
2 1982      3       1     4
3 1983     13       1    14
4 1984     49       3    52
5 1985    158      19   177
6 1986    407      92   499
```

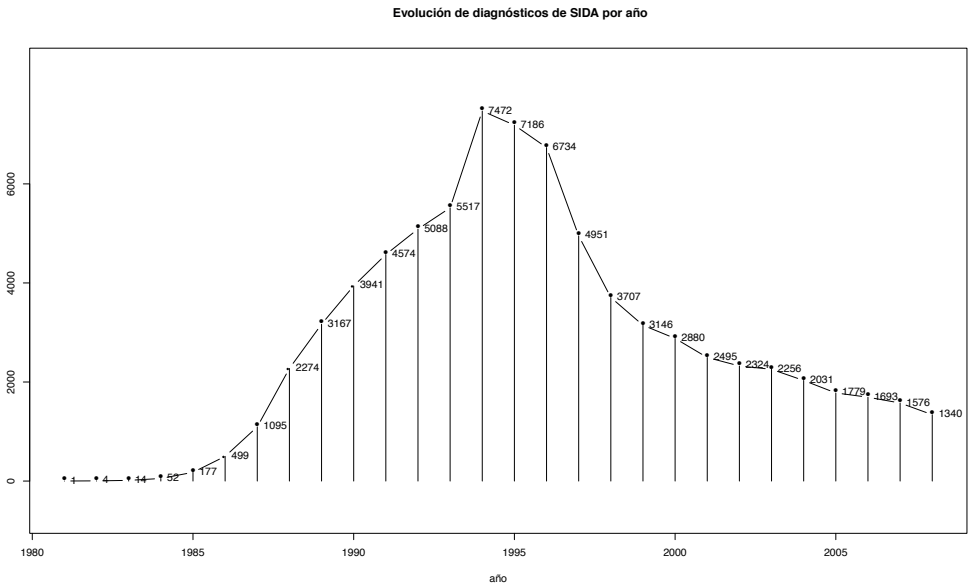


Figura 4.9. Estudio exploratorio del sida desde el año 1981 al 2008. Elaboración propia.

En la Figura 4.9 se observa la evolución del sida en 28 años, obteniendo en 1994 el más alto valor, 7472. Esta figura se obtiene con el siguiente código:

Código R

```
datos = read.table("sida.txt", header = TRUE)
attach(datos)
plot(año, Total, type = 'b', ylab="", main= 'Evolución de
```

```
diagnósticos de SIDA por año')
lines(año, Total, type = 'h')
text(año, Total, Total, pos=4)
```

En la Figura 4.10 se observa la evolución del sida por sexo, obteniendo mayores valores en varones con respecto a las mujeres. Esta figura se obtiene con el siguiente código:

```
datos = read.table("sida.txt", header = TRUE)
attach(datos)
plot(año,  Varones, type = 'b', pch = 0, col = 'red', ylab="",
     main = 'Evolución de diagnósticos de SIDA por año')
lines(año,  Mujeres, type = 'b', pch=1, col = 'blue')
legend("topleft", c("varones", "mujeres"),
       col = c('red', 'blue'), pch = c(0,1), lty = c(1,1), box.lty=0)
```

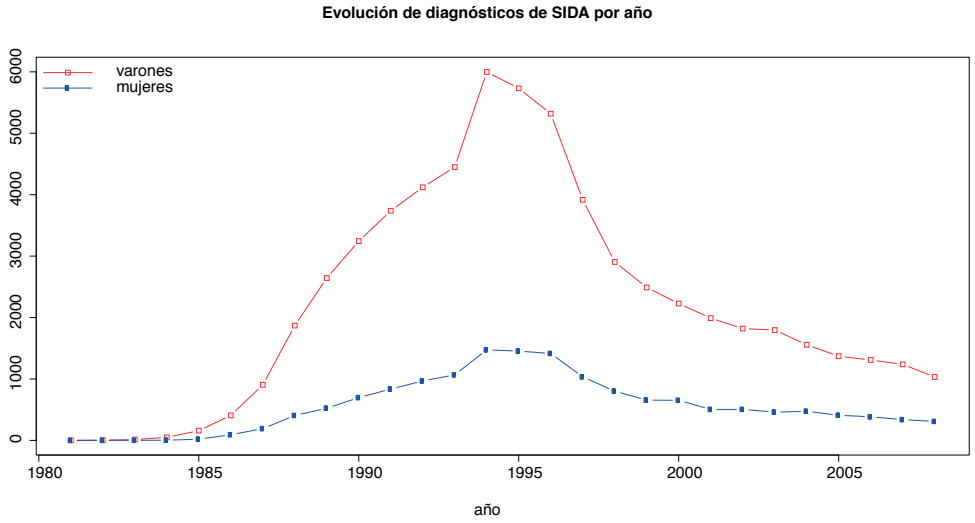


Figura 4.10. Estudio exploratorio del sida desde el año 1981 al 2008 por sexo. Elaboración propia.

Problemas propuestos para realizar con el *software* R

1. La producción de maíz (en toneladas) de unas granjas son las que figuran en la tabla adjunta

Granja	A	B	C	D	E	F	G
Producción	16	12	20	17	23	12	18

Representar gráficamente estos datos en un diagrama de barras

2. En un país en los años que se indican, el número de nacimientos por cada mil habitantes es el que se señala en la siguiente tabla:

Año	1960	1965	1970	1975	1980	1985	1990	1995
Nacimientos	23	20	18	17	14	13	13	15

Represente gráficamente estos datos:

a. En un gráfico cartesiano interpolando linealmente entre cada dos años consecutivos

b. En un diagrama de barras

3. En las elecciones municipales de una cierta localidad concurren tres partidos políticos (PA, PB, PC). Los votos válidos emitidos en las elecciones de los años 1992 y 1996 se distribuyeron entre los partidos como vemos en el siguiente cuadro (en el N y B significan votos nulos y en blanco)
Representar estos datos mediante tres diagramas de barras

	1992	1996
PA	7962	10306
PB	11137	8694
PC	3153	2498
N y B	759	1203

Capítulo 5

Variables continuas

Cuando la variable en estudio es continua (o discreta con un número elevado de valores distintos) toma tantos posibles valores como número de observaciones y, por tanto, no es posible escribirlos todos ellos en una columna, como se hizo anteriormente.

5.1 Tabla de frecuencias

Para tabular estos datos conviene agruparlos en unos cuantos intervalos y determinar el número de individuos que pertenecen a cada uno de ellos.

Tomar el intervalo como unidad de estudio, en lugar de cada valor de la variable, supone: una simplificación del problema, pero a cambio hay una pérdida de información.

Por lo tanto, es importante elegir un número adecuado de intervalos que equilibre estos dos aspectos.

El fichero *cacharros.txt* recoge datos recogidos en una fábrica de cacharros.

Hay 59 datos de 4 variables:

- *artículo*: tipo de cacharro (codificada con números del 1 al 4),
- *diámetro*: diámetro en cm.,
- *tiempo*: tiempo de fabricación en minutos,
- *precio*: precio de venta al público en euros.

Código R

```
Cacharros <- read.table("cacharros.txt", header = T)
```

```
attach(Cacharros)
table(tiempo)

        # Salida de la consola
tiempo
 12.02  13.25  15.61  16.41  16.66  17.46  17.67  17.84  20.21  20.59
     1      1      1      1      1      1      1      1      1      1
 20.83  20.85  21.04  21.34  21.87  22.55  22.82  23.21  23.74  23.77
     1      1      1      1      1      1      1      1      1      1
 23.88  26.09  26.25  26.52  26.53  27.76  28.64  29.48   30.2  31.46
     1      1      1      1      1      1      1      1      1      1
 31.86  32.62   32.9   33.7  33.71  33.89  34.16  34.88  37.11  39.71
     1      1      1      1      1      1      1      1      1      1
 43.14  44.14  44.45  44.95  45.12  45.78  47.65  48.74  49.48  53.18
     1      1      1      1      1      1      1      1      1      1
 54.86  55.53  58.76  63.13   64.3  68.63  74.48  86.42 109.38
     1      1      1      1      1      1      1      1      1
```

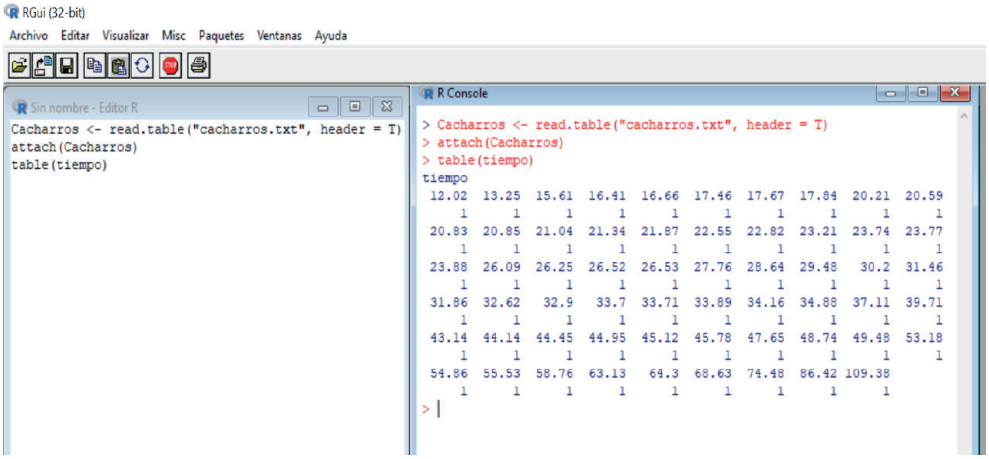


Figura 5.1. Salida de la consola de frecuencias de una variable continua. Elaboración propia.

En la salida de la consola y en Figura 5.1 se observa una tabla con tantas posiciones como datos muestrales y todas las frecuencias

iguales a uno. Está claro que este procedimiento no será válido para variables continuas.

Para hacer una tabla de frecuencias de variables continuas,

- Se discretiza la variable, y
- Se construye la correspondiente tabla de frecuencias.

A continuación, se muestra la Tabla 5.1 de frecuencias para la variable *tiempo*:

Tabla 5.1
Frecuencias: absolutas, absolutas acumuladas, relativas y relativas acumuladas de la variable tiempo

clase	frec. absol.	frec. relat	fr. abs. acum	fr. rel. acum
≤ 35	38	0.64	38	0.64
(35,60]	15	0.25	53	0.90
(60,85]	4	0.07	57	0.97
> 85	2	0.03	59	1.00
Sum	59	1		

Elaboración propia.

En la Tabla 5.1 se observa que la variable tiempo de fabricación de los artículos se divide en 4 clases, obteniendo la mayor cantidad de artículos que se han fabricado en tiempos menores o iguales a 35.

Los resultados de la Tabla 5.1 se obtienen con el siguiente código:

```
Cacharros<-read.table("cacharros.txt", header = T)
attach(Cacharros)
tiempod=cut(tiempo,breaks=c(-Inf,35,60,85,Inf)) # Discretizamos
tiempo
nj = table(tiempod) ; nj    #    Frec. Absolutas
```

```
        # Salida de la consola
tiempod
(-Inf,35]   (35,60]   (60,85] (85, Inf]
      38        15         4         2
```

```
Nj = cumsum(nj) ;   Nj    #    Frec. absolutas acumuladas
```

```
        # Salida de la consola
(-Inf,35]   (35,60]   (60,85] (85, Inf]
      38        53        57        59
```

```
fj=prop.table(nj); fj      #    Frec. Relativa
```

```
        # Salida de la consola
tiempod
 (-Inf,35]     (35,60]       (60,85]       (85, Inf]
0.64406780 0.25423729 0.06779661 0.03389831
```

```
Fj = cumsum(fj) ;  Fj     # Frecuencia relativa acumulada
```

```
        # Salida de la consola
(-Inf,35]    (35,60]      (60,85]      (85, Inf]
0.6440678 0.8983051 0.9661017 1.0000000
```

Las tablas de frecuencias obtenidas dependerán del,

- número de cortes, y
- posición de los mismos.

A continuación, se muestran las frecuencias obtenidas para 5 cortes.

Código R

```
Cacharros<-read.table(cacharros.txt", header = T)
attach(Cacharros)
tiempod=cut(tiempo, breaks = 5)   #    Discretizamos tiempo
nj = table(tiempod) ; nj    #  Frec. Absolutas

        # Salida de la consola
tiempod
(11.9,31.4]    (31.4,50.9]   (50.9,70.5]   (70.5,90]    (90,109]
        29          20             7              2            1
Nj = cumsum(nj) ;  Nj    #  Frec.absolutas acumuladas

        # Salida de la consola
(11.9,31.4]    (31.4,50.9]   (50.9,70.5]   (70.5,90]    (90,109]
        29          49             56             58           59

        # Salida de la consola
fj = prop.table(nj) ; fj     # Frec. Relativa

        # Salida de la consola
tiempod
(11.9,31.4]    (31.4,50.9]    (50.9,70.5]   (70.5,90]      (90,109]
 0.49152542  0.33898305  0.11864407  0.03389831  0.01694915

Fj = cumsum(fj) ;  Fj      # Frecuencia relativa acumulada

        # Salida de la consola
(11.9,31.4]    (31.4,50.9]   (50.9,70.5]   (70.5,90]      (90,109]
  0.4915254   0.8305085   0.9491525      0.9830508   1.0000000
```

5.2 Representaciones gráficas

5.2.1 Histograma

El histograma de un conjunto de datos es un gráfico de barras que representan las frecuencias con que aparecen las mediciones agrupadas en ciertos intervalos y luego contar cuántas observaciones caen en cada intervalo. Sólo se utiliza con variables continuas, y cuando se dispone de una cantidad grande de datos.

Para cada clase, C_j se dibuja un rectángulo apoyado en el eje X cuya base sea el intervalo y cuya área sea proporcional a la frecuencia n_j a representar. Por lo tanto, la altura h_j queda determinada por el cociente n_j/a_j entre la frecuencia n_j y la amplitud a_j del intervalo. En el *software* R se utiliza la función hist.

Ejemplo - Cacharros

```
Cacharros <- read.table("cacharros.txt", header = T)
attach(Cacharros)
hist(tiempo, col = "blue" , main = "Histograma de tiempo",
      xlab = "Tiempo",  ylab = "Frecuencia")
```

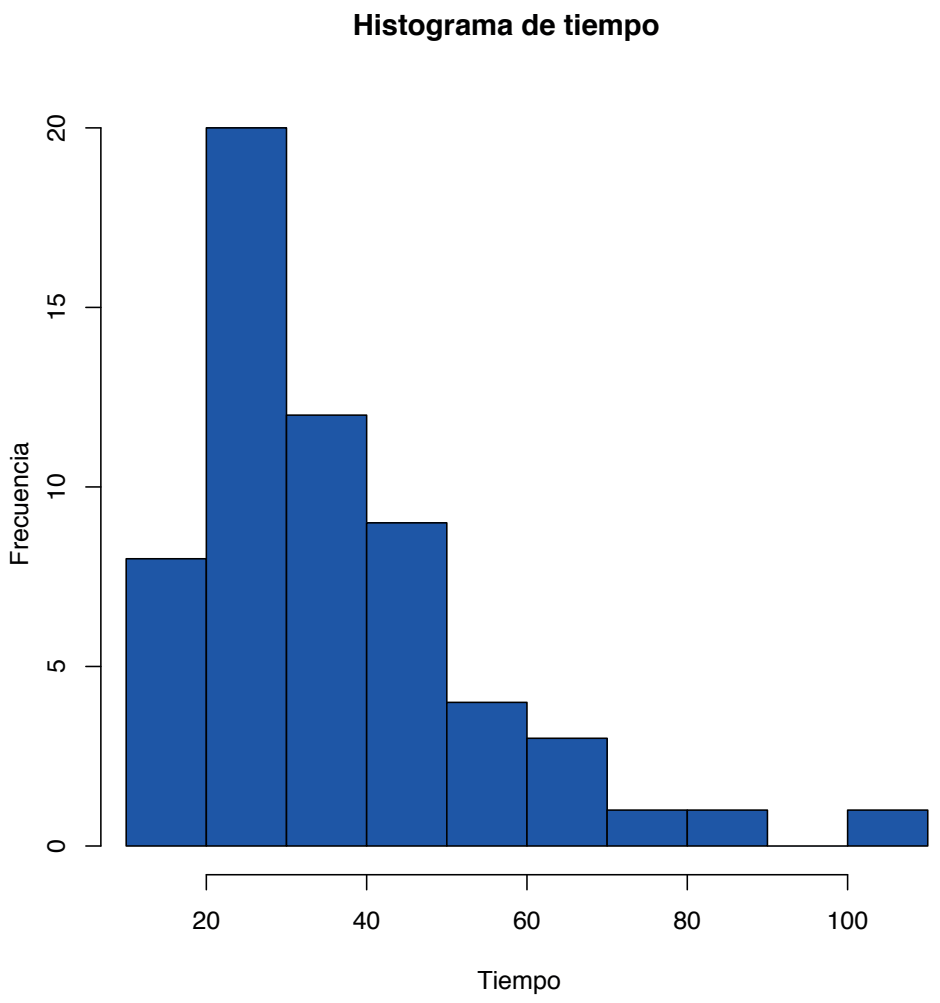


Figura 5.2. Histograma de la variable tiemp. Elaboración propia.

En la Figura 5.2 se observa mediante un histograma la distribución de la variable tiempo de fabricación de los artículos, obteniendo la mayor cantidad entre 20 y 30 minutos.

Los histogramas son muy útiles para apreciar la forma de la distribución de los datos, si se escoge adecuadamente el número de clases y su amplitud.

Sin embargo, la selección del número de clases y su amplitud que adecuadamente representan la distribución puede ser complicado:

Un histograma con muy pocas clases agrupa demasiado las observaciones y un histograma con muchas clases deja muy pocas observaciones en cada una de ellas.

Ninguno de los dos extremos es apropiado.

Existen varias reglas para determinar el número de clases. El *software* R por defecto selecciona el número de clases siguiendo el llamado método de *Sturges* ().

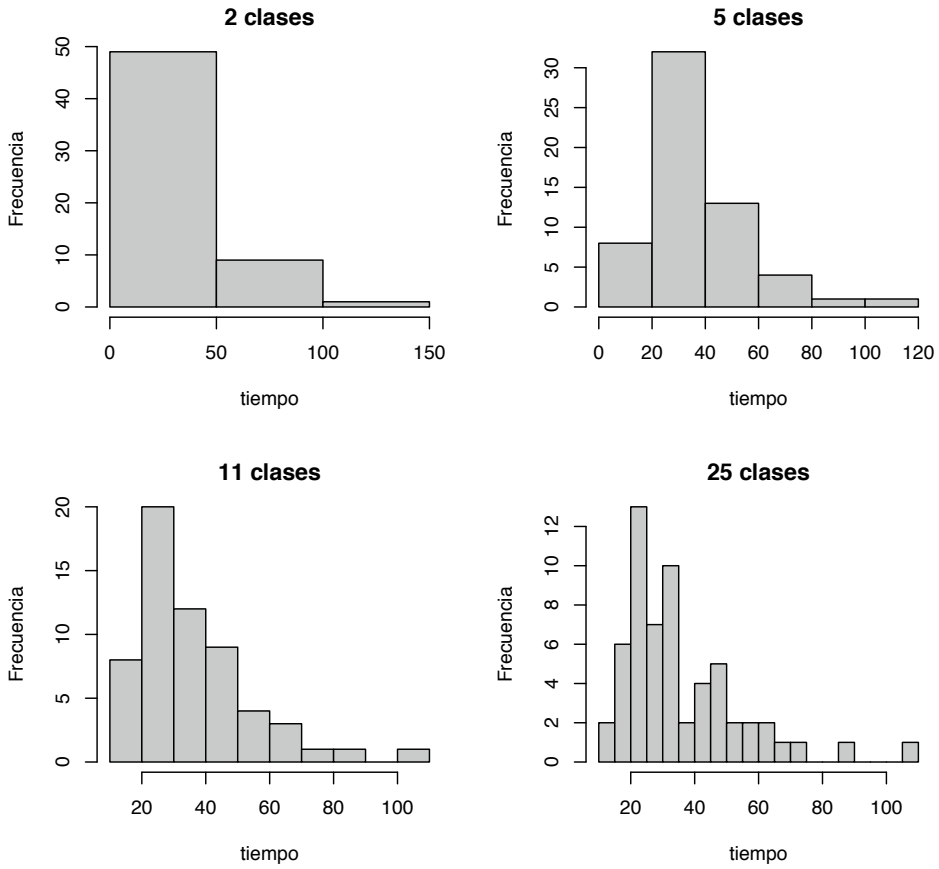


Figura 5.3. Efecto del número de clases en los histogramas de la variable tiempo. Elaboración propia.

La Figura 5.3 muestra el efecto del número de clases de la variable *tiempo*, esta se puede obtener mediante el siguiente código:

Ejemplo – Cachorros

Código R

```
Cacharros <- read.table("cacharros.txt", header = T)
attach(Cacharros)
par(mfrow=c(2,2))   # División de la ventana gráfica en 2 fila y 2 columnas
hist(tiempo,breaks=2,main="2 clases")
hist(tiempo,breaks=5,main="5 clases")
hist(tiempo,breaks=11,main="11 clases")
hist(tiempo,breaks=25,main="25 clases")
```

En la Figura 5.4 se presentan dos histogramas de las variables *tiempo* y *precio*, todas estas son continuas.

```
Cacharros <- read.table("cacharros.txt", header = T)
attach(Cacharros)
par(mfrow=c(1,2))   # División de la ventana gráfica en 1 fila y 2 columnas
hist(tiempo, main = "Histograma de tiempo", ylab = "Frecuencia")
hist(precio, main = "Histograma de precio", ylab = "Frecuencia")
```

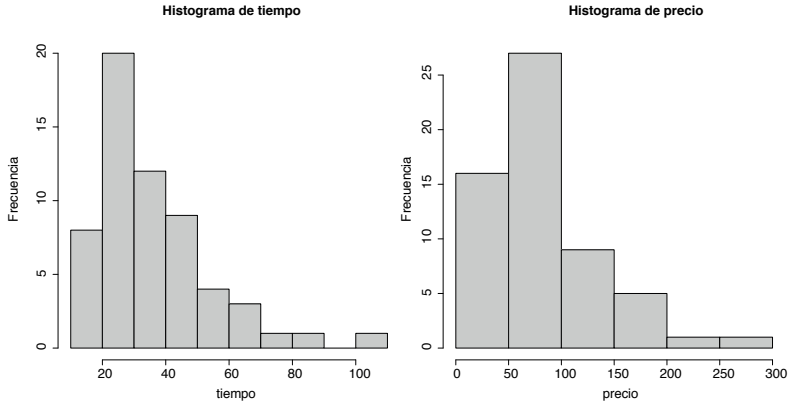


Figura 5.4. Histogramas de las variables tiempo y precio. Elaboración propia.

5.2.2 Árbol de tallo y hojas

Otro gráfico que puede ser utilizado para la representación de variables continuas es el llamado árbol de tallo y hojas. Este tipo de gráfico son fáciles de realizar a mano, y se solían utilizar como una forma rápida (aunque igual no demasiado pulida) de visualizar los datos.

Ejemplo – Cacharros

Código R

```
Cacharros <- read.table("cacharros.txt", header = T)
attach(Cacharros)
stem(tiempo)
```

```
        # Salida de la consola
The decimal point is 1 digit(s) to the right of the |
 1 | 23667788
 2 | 01111123334446677899
 3 | 01233444457
 4 | 0344556899
 5 | 3569
 6 | 349
 7 | 4
 8 | 6
 9 |
10 | 9
```

En esta salida de la consola se observan los tallos (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) y sus hojas que están a su derecha, los valores de los tallos deben leerse con la como desplazada un lugar hacia la derecha, por ejemplo, se escoge el tallo 5 y sus hojas se leen de forma ordenada:

53, 55, 56, 59. También, el aspecto general de la salida es como un histograma de la variable *tiempo*.

5.3 Función de distribución empírica

Dada una muestra $X_1, \cdots, X_n$ se define la función de distribución empírica se define exactamente igual a como se había hecho en el caso discreto.

$$F_n(x) = \frac{\textit{número de valores menores o iguales a } x}{n}$$

Entonces:

- F_n toma valores en el intervalo [0,1],
- Es una función escalonada creciente.

Sin embargo, ahora los valores no se repiten, y

- los saltos de F_n se dan en cada valor muestral X_i, y la amplitud del salto es $1/n$

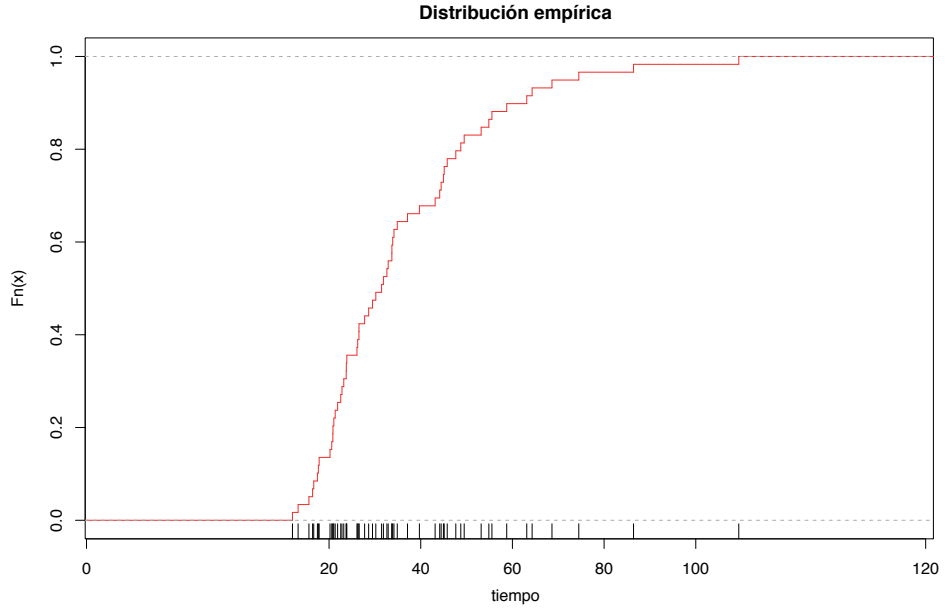


Figura 5.5. Función de distribución empírica de la variable tiempo. Elaboración propia.

En la Figura 5.5 se observa la distribución empírica de la variable tiempo de fabricación de los artículos.

La Figura 5.5 se obtiene con el siguiente código:

```
Cacharros <- read.table("cacharros.txt", header = T)
attach(Cacharros)
plot(ecdf(tiempo), verticals = T, main = "Distribución empírica",
xlab = 'tiempo',
     col = 'red', do.points=F)
rug(tiempo)
```

5.4 Medidas de posición y dispersión

Hasta ahora se han mostrado, para una variable de interés X, distintas formas de presentar en forma de tablas y gráficos una colección de datos de dicha variable

$$X_1, \cdots, X_n$$

A veces conviene reducir toda esta información en una o varias medidas resumen.

Algunas de estas medidas son las que siguen a continuación:

Medidas de Posición

- Media Muestral
- Mediana
- Cuantiles

Medidas de Dispersión

- Varianza y Desviación Típica
- Rango o Rango Intercuartílico
- Coeficiente de Variación

Media Muestral

La media muestral se define como el promedio de los datos:

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

Ejemplo – Consumo de automóviles

La variable "Consumo de combustible (litros/100km a 90km/h)" de seis automóviles de la misma marca ha sido de,

6.7 6.3 6.5 6.5 6.4 6.6

obteniéndose un consumo medio muestral de,

$$\overline{X} = \frac{6.7 + 6.3 + 6.5 + 6.5 + 6.4 + 6.6}{6} = \frac{32.4}{6} = 6.5$$

Código R

```
consumo <- c(6.7, 6.3, 6.5, 6.5, 6.4, 6.6)
mean(consumo)

        # Salida de la consola
[1] 6.5
```

Mediana

La media aritmética puede ser muy sensible a los valores extremos de la variable.

Ejemplo – Diámetro de un cilindro

Diez medidas de la variable X= "diámetro de un cilindro (en cm.)" fueron anotadas por un científico como:

3.88 4.09 3.92 3.97 4.02 3.95 4.03 3.92 3.98 40.6

La media aritmética de los valores anteriores es

$$\overline{X}= \frac{3.88 + \cdots + 40.6}{10} = 7.636$$

Esta medida no representa la posición central de los datos obtenidos ya que está muy influenciada por el valor 40.6 que claramente un valor raro con respecto al resto de los datos obtenidos.

Ante este tipo de situaciones será conveniente utilizar otra medida más robusta como puede ser la mediana.

La mediana es aquel valor *Me* que divide a la población en dos partes de igual tamaño, la mitad son mayores que él y la otra mitad inferior a él.

Supuestos ordenados los datos de menor a mayor $X_1 \leq \cdots \leq X_n$, entonces

- Si n es impar, la mediana coincide con el valor central.
- Si n es par, la mediana se calcula como la media de los dos valores centrales

Ejemplo – Diámetro de un cilindro

Los diámetros ordenados son:

3.88 3.92 3.92 3.95 3.97 3.98 4.02 4.03 4.09 40.6

Como n=10 es un número par, la mediana se calcula como la media de los dos valores centrales situados en las posiciones 5 y 6

$$Me = \frac{3.97 + 3.98}{2} = 3.975$$

Código R

```
diametro=c(3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98, 40.6)
mean(diametro)

        # Salida de la consola
[1] 7.636
        # Salida de la consola
median(diametro)
[1] 3.975
```

Media versus Mediana

Para distribuciones simétricas (sin valores atípicos) de la media y la mediana están muy próximas.

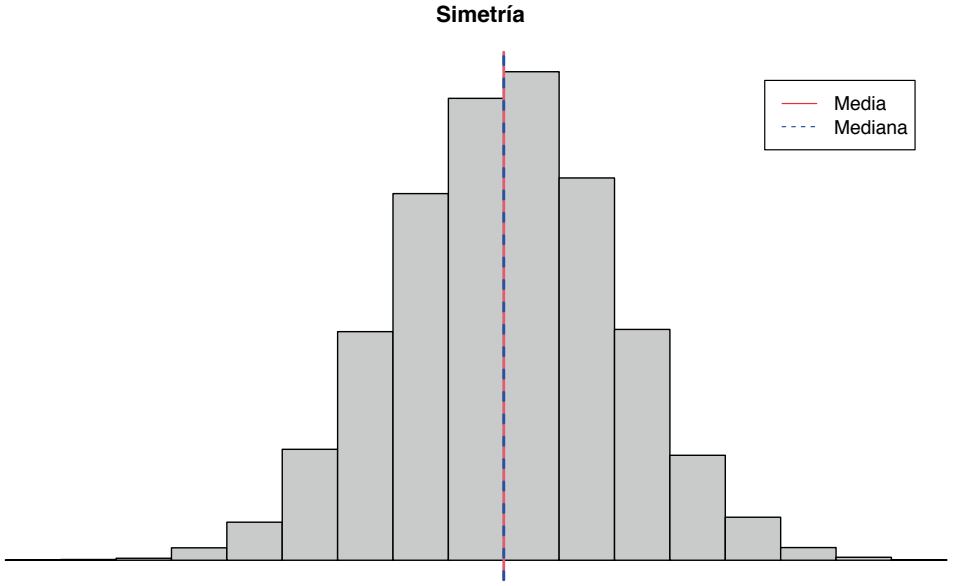


Figura 5.6. Histograma simétrico con respecto a la media. Elaboración propia.

Sin embargo, cuando las distribuciones son asimétricas la media y la mediana no serán coincidentes.
Las asimetrías pueden ser de acuerdo a la distribución de los datos de la variable, esto es:

- Asimetría Derecha, cuando la media de los datos es mayor que la mediana, es decir: $X > Me$
- Asimetría Izquierda, cuando la media de los datos es menor que la mediana, es decir: $X > Me$

.

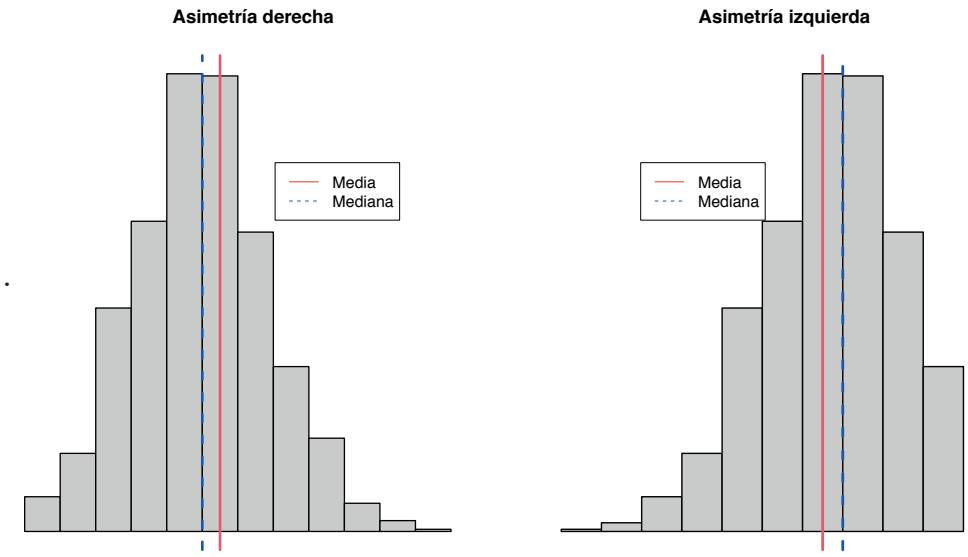


Figura 5.7. Histogramas asimétricos con respecto a la media. Elaboración propia.

Cuantiles

Los cuantiles son una generalización de la mediana.
El cuantil de orden p con $0 < p < 1$ es aquel valor X_p que,

- Una proporción p de la muestra es menor que dicho valor, y
- El resto (es decir una proporción 1 - p mayor).

Nótese que la mediana es el cuantil de orden $p = 0.5$.

Cálculo de los Cuantiles:

A continuación se explica el método utilizado por la función *quantile* con la configuración por defecto de R.

Sea la muestra ya ordenada $X_1 \leq X_2 \cdots X_n$. Denotemos por I a la parte entera de 1+(n-1)p y R E]0,1[el resto, de forma que se establece la relación

$$1+(n-1)p = 1+R$$

El cuantil de orden p viene dado por,

$$x_p = X_I + R(X_{I+1} - X_I)$$

Ejemplo – Diámetro de un cilindro

Para el cálculo de los cuantiles primero se ordenan los valores:

3.88 3.92 3.92 3.95 3.97 3.98 4.02 4.03 4.09 40.6

Para el cálculo del cuantil de orden $p = 0.25$ se realiza la operación,

$1 + (10 - 1)\, 0.25 = 3.25 \Rightarrow I = 3,\ R = 0.25$
obteniéndose que,

$$x_{0.25} = X_3 + R\,(X_4 - X_3) = 3.92 + 0.25\,(3.95 - 3.92) = 3.9275$$

De igual modo, para calcular el cuantil de orden $p = 0.45$ se obtiene

$1 + (10 - 1)\,.\,0.45 = 5.05 \Rightarrow I = 5,\ R = 0.05$

Resultando,

$$x_{0.45} = 3.97 + 0.05\ (3.98 - 3.97) = 3.9705$$

Código R

```
diametro=c(3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98, 40.6)
quantile(diametro, probs = c(0.25, 0.45))

        # Salida de la consola

  25%   45%
3.9275 3.9705

quantile(diametro)

        # Salida de la consola
   0%      25%     50%     75%     100%
3.8800  3.9275  3.9750  4.0275 40.6000
```

Cuartiles

Los cuartiles son los cuantiles de orden 0.25, 0.50 y 0.75 (dividen a la muestra en 4 partes de igual frecuencia).

- Normalmente se denotan por Q1, Q2 y Q3 y se denominan primer, segundo y tercer cuartil muestral, respectivamente.

- El segundo cuartil muestral coincide con la mediana muestral.

Ejemplo – Diámetro de un cilindro

Código R

```
diámetro = c(3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98, 40.6)
quantile(diametro)

# Salida de la consola
    0%     25%     50%     75%    100%
3.8800  3.9275  3.9750  4.0275 40.6000
```

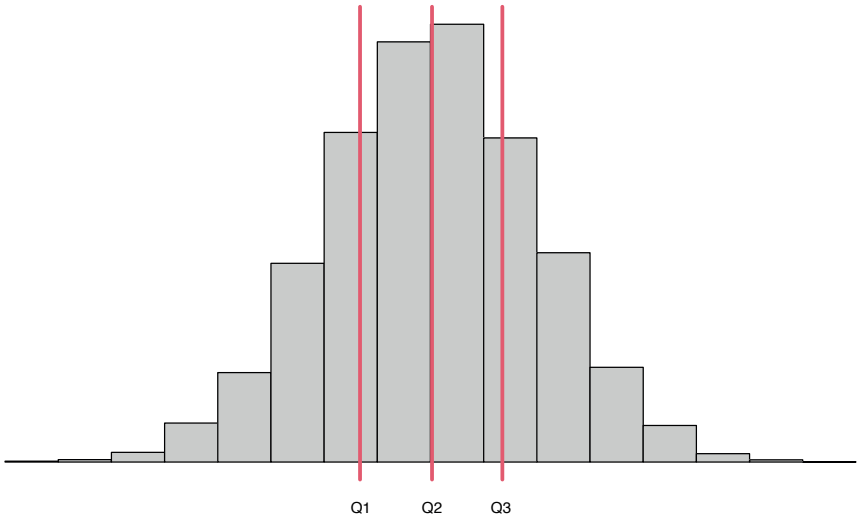


Figura 5.8. Histograma con cuartiles. Elaboración propia.

En la Figura 5.8 se observa la ubicación de cuartiles en un histograma

Deciles y Centiles

- Los deciles: son los cuantiles muestrales de orden 0.1, . . . ,0.9 (dividen a la muestra en 10 partes de igual frecuencia)

Ejemplo – Diámetro de un cilindro

Código R

```
diametro=c(3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98, 40.6)
x=diametro
quantile(x, probs = seq(0.1, 0.9, 0.1))    #    deciles

    # Salida de la consola
 10%  20%  30%  40%  50%  60%  70%  80%  90%
3.916 3.920 3.941 3.962 3.975 3.996 4.023 4.042 7.741
```

- Los centiles: son los cuantiles muestrales de orden 0.01, . . . ,0.99 (dividen a la muestra en 100 partes de igual frecuencia)

Código R

```
diametro=c(3.88,4.09,3.92,3.97,4.02,3.95, 4.03, 3.92, 3.98, 40.6)
x=diametro
quantile(x,probs=seq(0.01,0.99,0.01))      #    centiles

    # Salida de la consola
   1%    2%    3%    4%    5%    6%    7%    8%    9%
3.8836 3.8872 3.8908 3.8944 3.8980 3.9016 3.9052 3.9088 3.9124
  10%   11%   12%   13%   14%   15%   16%   17%   18%
3.9160 3.9196 3.9200 3.9200 3.9200 3.9200 3.9200 3.9200 3.9200
  19%   20%   21%   22%   23%   24%   25%   26%   27%
3.9200 3.9200 3.9200 3.9200 3.9221 3.9248 3.9275 3.9302 3.9329
  28%   29%   30%   31%   32%   33%   34%   35%   36%
3.9356 3.9383 3.9410 3.9437 3.9464 3.9491 3.9512 3.9530 3.9548
  37%   38%   39%   40%   41%   42%   43%   44%   45%
3.9566 3.9584 3.9602 3.9620 3.9638 3.9656 3.9674 3.9692 3.9705
  46%   47%   48%   49%   50%   51%   52%   53%   54%
3.9714 3.9723 3.9732 3.9741 3.9750 3.9759 3.9768 3.9777 3.9786
```

55%	56%	57%	58%	59%	60%	61%	62%	63%
3.9795	3.9816	3.9852	3.9888	3.9924	3.9960	3.9996	4.0032	4.0068
64%	65%	66%	67%	68%	69%	70%	71%	72%
4.0104	4.0140	4.0176	4.0203	4.0212	4.0221	4.0230	4.0239	4.0248
73%	74%	75%	76%	77%	78%	79%	80%	81%
4.0257	4.0266	4.0275	4.0284	4.0293	4.0312	4.0366	4.0420	4.0474
82%	83%	84%	85%	86%	87%	88%	89%	90%
4.0528	4.0582	4.0636	4.0690	4.0744	4.0798	4.0852	4.4551	7.7410
91%	92%	93%	94%	95%	96%	97%	98%	99%
11.0269	14.3128	17.5987	20.8846	24.1705	27.4564	30.7423	34.0282	37.3141

Datos antropométricos

En el fichero *pediatria.txt* están registrados datos antropométricos de 3556 niños cuyas edades están comprendidas entre los 3 años y los 12 años. Las variables disponibles son

- *sexo*: hombre, mujer.
- *edad*: edad en años
- *peso*: peso en kg.
- *talla*: altura en cm.
- *imc*: índice de masa corporal en Kg/m^2

Ejemplo – Pediatria

Código R

```
datos <- read.table("pediatria.txt", header = T)
head(datos)

      # Salida de la consola
  sexo edad peso talla      imc
1 varón    3 14.5  94.4 16.27137
```

```
2 varón   3 13.0  91.5 15.52749
3 varón   3 12.2  90.5 14.89576
4 varón   3 14.4  92.7 16.75726
5 varón   3 13.5  92.5 15.77794
6 varón   3 16.5  96.1 17.86640
```

```
attach(datos)
plot(edad, talla)
```

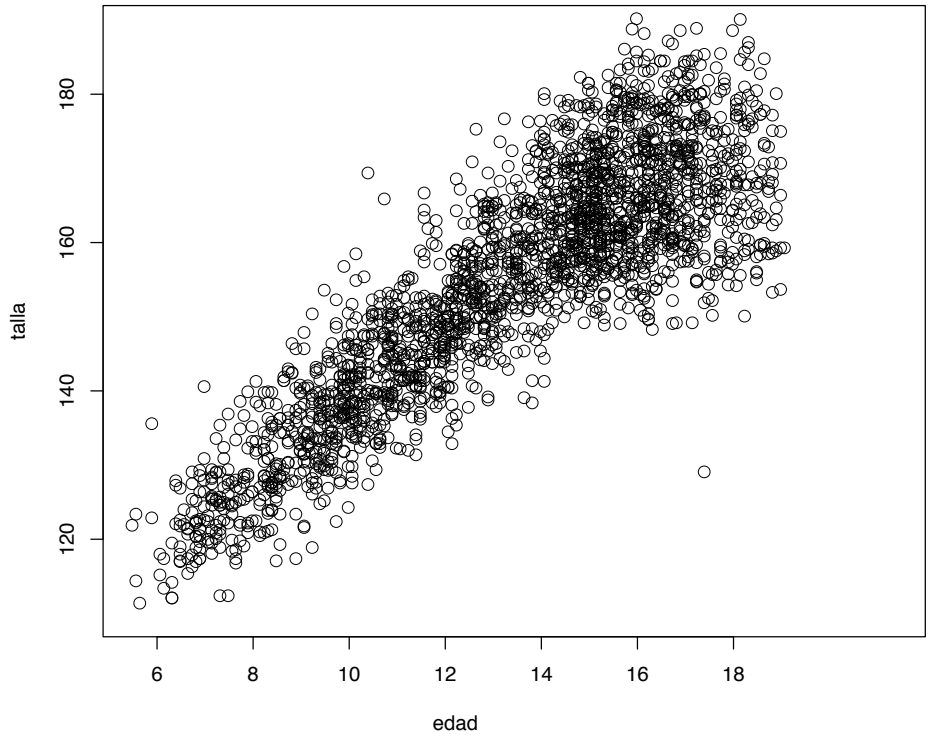


Figura 5.9. Variable talla en función de edad . Elaboración propia.

En la Figura 5.9 se observa un diagrama exploratorio de puntos de la variable talla en función de edad.

Código R

```
datos <- read.table("pediatria.txt", header = T)
attach(datos)
plot(edad, peso)
```

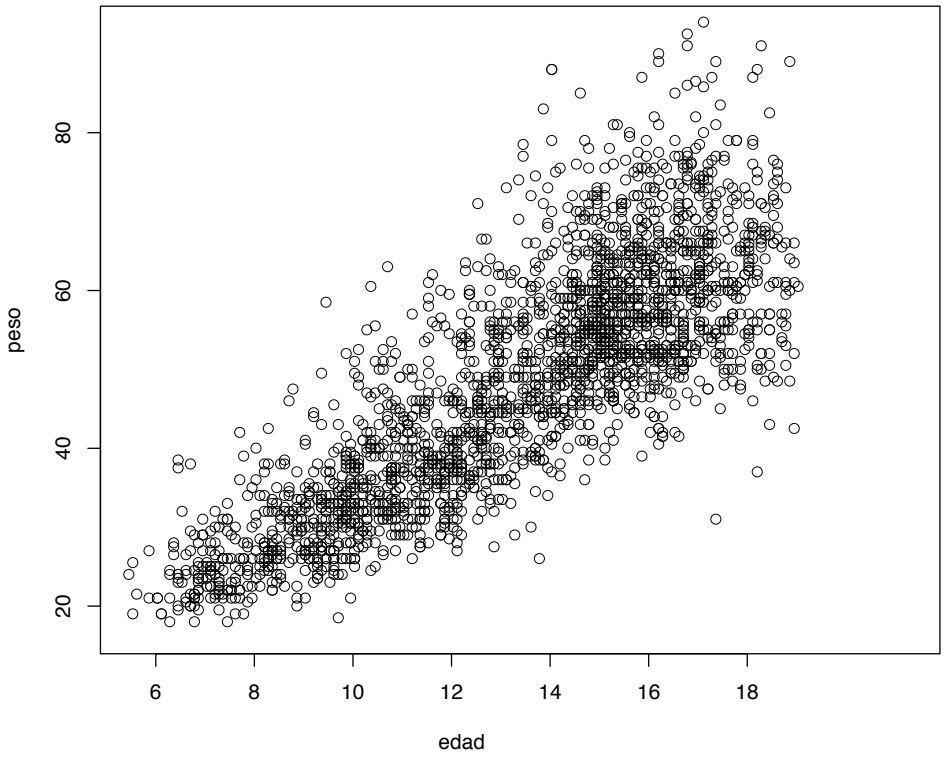


Figura 5.10. Variable peso en función de la edad. Elaboración propia.

En la Figura 5.10 se observa un diagrama exploratorio de puntos de la variable peso en función de edad.

```
plot(edad, imc)
```

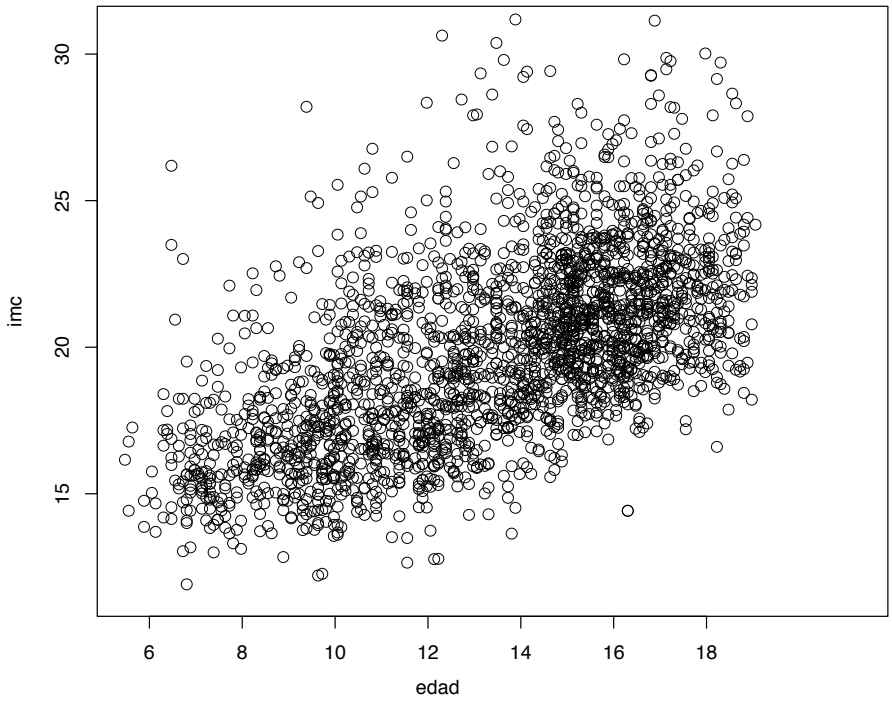


Figura 5.11. Variable imc en función de edad. Elaboración propia.

En la Figura 5.11 se observa un diagrama exploratorio de puntos de la variable imc en función de edad.

Varianza y Desviación típica

La varianza muestral es la medida de dispersión por excelencia

$$s^2 = \frac{(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n}$$

Interpretación:

- Si s^2 es próxima a cero los datos estarán muy concentrados entorno a su media.
- Si s^2 es grande significa que existe que los datos son muy dispares entre sí.

La varianza puede ser calculada de forma más "rápida" utilizando la expresión equivalente

$$s^2 = \frac{X_1^2 + \ldots + X_n^2}{n} - \overline{X}^2$$

Las unidades de s^2 son las mismas que las de X al cuadrado. Para mantener la misma unidad de medida de las observaciones, se define la desviación típica muestral de un conjunto de dados como la raíz cuadrada positiva de la varianza:

$$s = \sqrt{s^2} = \sqrt{\frac{(X_1 - \overline{X})^2 +) \ldots + (X_n - \overline{X})^2}{n}}$$

Ejemplo – Consumo de automóviles

Consideremos de nuevo la variable X = "Consumo de combustible (litros/100km a 90km/h)" de seis automóviles

6.7 6.3 6.5 6.5 6.4 6.6

La media y varianza muestral son dadas por:

$$\overline{X} = \frac{(6.7+6.3+6.5+6.5+6.4+6.6)}{6} = \frac{32.4}{6} = 6.5$$

$$s^2 = \frac{(6.7-6.5)^2 + \cdots + (6.6-6.5)^2}{6} = \frac{0.1}{6} = 0.0167$$

Esta cantidad puede ser calculada de forma equivalente como

$$s^2 = \frac{6.7^2 + \cdots + 6.6^2}{6} = 6.5^2 = \frac{253.6}{6} - 6.5^2 = 0.0167$$

La desviación típica muestral de los datos es

$s=\sqrt{0.0167}=0.129$

Cuasi varianza muestral

Se sabe que la varianza muestral s^2 tiende a dar valores más bajos de los esperados. Por este motivo, en la práctica, se suelen utilizar la cuasi-varianza S^2 y cuasi-desviación típica S muestrales:

$$s^2=\frac{n}{n-1}s^2=\frac{(X_1-X)^2+\ldots+(X_n-X)^2)}{n-1} \text{ y } S=\sqrt{S^2}$$

Ejemplo – Consumo de automóviles

La cuasi-varianza y cuasi-desviación típica muestral de la variable consumo son:

$$s^2=\frac{(6.7-6.5)^2+\cdots+(6.6-6.5)^2}{5}=\frac{0.1}{5}=0.02 \quad \text{y} \quad S=\sqrt{0.02}=0.1414214$$

Código R

```
consumo <- c(6.7, 6.3, 6.5, 6.5, 6.4, 6.6)
var(consumo)

        # Salida de la consola
[1] 0.02

sd(consumo)

        # Salida de la consola

[1] 0.1414214
```

Nota: En lo siguiente *S* diremos simplemente desviación típica muestral y S^2 como varianza muestral.

Rango

El rango o recorrido que corresponde a la diferencia entre el mayor valor observado de la variable y el menor.

Ejemplo – Consumo de automóviles

El rango de consumos es:
rango = 6.7 - 6.3 = 0.4

Código R

```
consumo <- c(6.7, 6.3, 6.5, 6.5, 6.4, 6.6)
(máximo = max(consumo))

        # Salida de la consola
[1] 6.7

(minimo = min(consumo))

        # Salida de la consola
[1] 6.3

(rango = maximo-minimo)

        # Salida de la consola
[1] 0.4
```

De forma análoga se consigue con el siguiente código:

```
consumo <- c(6.7, 6.3, 6.5, 6.5, 6.4, 6.6)
(rango = range(consumo))  #  de forma equivalente
        # Salida de la consola
[1] 6.3 6.7

rango[2] - rango[1]

        # Salida de la consola
[1] 0.4
```

Rango Intercuartílico

Se define el rango intercuartílico como la diferencia entre el tercer y el primer cuartil. Es decir, es la longitud del intervalo donde se encuentran el 50% de los datos centrales.

RI = 3º cuartil – 1º cuartil = $Q3$ - $Q1$

Ejemplo – Consumo de automóviles

El rango de consumos es:
RI = 6.575 - 6.425 = 0.15

Código R

```
consumo <- c(6.7, 6.3, 6.5, 6.5, 6.4, 6.6)
(Q = quantile(consumo, probs = c(0.25, 0.75)))

        # Salida de la consola
  25%   75%
6.425 6.575
```

```
(RI = Q[2] - Q[1])

        # Salida de la consola
 75%
0.15
```

Coeficiente de variación

Otra medida que se suele utilizar es el coeficiente de variación (*CV*). Es una medida de dispersión relativa de los datos y se calcula dividiendo la desviación típica muestral por la media y multiplicando el cociente por 100.

$$CV=100\,\frac{S}{|\overline{X}|}$$

Ejemplo – Consumo de Automóviles

El *CV* de la variable consumo es:

$$CV=100\,\frac{0.1414}{6.50}=2{,}176\%$$

Código R

```
consume <- c(6.7, 6.3, 6.5, 6.5, 6.4, 6.6)
CV<-function(X){100*sd(X)/abs(mean(X))}
CV(consumo)

        # Salida de la consola
[1] 2.175713
```

La utilidad del *CV* radica en que permite comparar la dispersión o variabilidad de dos o más grupos.

Ejemplo – *Peso* versus *tensión*

Se ha registrado el peso *X* (en kg.) y la tensión arterial *Y* (en mmHg.) de 5 pacientes

peso	70	60	56	83	79
tensión	150	170	135	180	195

Obteniéndose

- un peso medio $\overline{X} = 69.6$ kg. con desviación típica $S_X = 11.67$ y
- una tensión media de $\overline{Y} = 166$ mmHg con desviación típica $S_Y = 23,82$.

¿qué distribución es más dispersa, el peso o la tensión arterial?

Si se comparan las desviaciones típicas se observa que la desviación típica de la tensión arterial es mucho mayor. Sin embargo, no se pueden comparar dos variables que tienen escalas de medidas diferentes, por lo que se calculan los coeficientes de variación:

$$CV_{\text{de peso}} = 100\frac{11.67}{69.6} = 16.77\%$$

$$CV_{\text{de tensión}} = 100\frac{23.82}{166} = 14.35\%$$

A la vista de los resultados, se observa que la variable peso tiene una mayor dispersión.

Código R

```
peso <- c(70, 60, 56, 83, 79)
```

```
tension <- c(150, 170, 135, 180, 195)
mean(peso) ; mean(tension)
        # Salida de la consola

[1] 69.6
[1] 166

sd(peso) ; sd(tension)

        # Salida de la consola

[1] 11.67476
[1] 23.82226

CV<-function(X){100*sd(X)/abs(mean(X))}
CV(peso);CV(tension)

        # Salida de la consola

[1] 16.77408
[1] 14.35076
```

Ejemplo – Pediatría

En la Tabla 5.2 se muestra la media y desviación típica de las variables *Talla, Peso* e IMC en función de Sexo.

Tabla 5.2

Media y desviación típica muestrales de pediatría

	PESO		TALLA		IMC	
sexo	**media**	**sd**	**media**	**sd**	**Media**	**Sd**
Hembra	45.43	12.81	150.37	13.96	19.62	3.16
Varón	49.27	16.46	156.31	17.48	19.46	3.23

Elaboración propia.

A continuación, se muestra en la Tabla 5.3 el coeficiente de variación para *talla, peso* e *IMC* en función del *sexo*.

Tabla 5.3
CV de pediatría

SEXO	PESO	TALLA	IMC
	CV	*CV*	*CV*
Hembra	28.20	9.28	16.09
Varón	33.41	11.18	16.60

Elaboración propia.

A la vista de los resultados obtenidos se comprueba que *talla* es la variable con menor dispersión y que *peso* es la variable con mayor dispersión.

Las dos tablas anteriores se obtienen con el siguiente código:

Ejemplo – Pediatría

Código R

```
pediatria = read.table('pediatria.txt', header = T)
attach(pediatria)
library(abind)
library(RcmdrMisc)
tabla = numSummary(pediatria[,3:5], statistics = c('mean', 'sd'),
        groups = pediatria$sexo)
tabla = as.data.frame(tabla$table)
tabla

        # Salida de la consola

          mean.peso  sd.peso   mean.talla  sd.talla   mean.imc
sd.imc
```

hembra	45.43148	12.80945	150.3680	13.95719	19.61894	3.157311
varón	49.27099	16.45915	156.3139	17.48355	19.46337	3.230583

Para calcular el coeficiente de variación se realiza del modo siguiente.

Código R

```
pediatria = read.table('pediatria.txt', header = T)
attach(pediatria)
library(abind)
library(RcmdrMisc)
tabla = numSummary(pediatria[,3:5], statistics = c('mean', 'sd'),
        groups = pediatria$sexo)
tabla = as.data.frame(tabla$table)
tabla$cv.PESO = abs(100*tabla$sd.peso/tabla$mean.peso)
tabla$cv.TALLA = abs(100*tabla$sd.talla/tabla$mean.talla)
tabla$cv.IMC = abs(100*tabla$sd.imc/tabla$mean.imc)
tabla[,7:9]

        # Salida de la consola

          cv.PESO   cv.TALLA   cv.IMC
hembra 28.19510   9.282026    16.09318
varón  33.40536   11.184902   16.59827
```

5.5 Diagrama de cajas. Datos atípicos

Los diagramas de caja son representaciones basadas en los cuartiles, que permiten:

- Mostrar las principales características de la muestra: posición, dispersión, asimetría, etc.

- Identificar la presencia de observaciones atípicas (valores missing)

Ejemplo – Cacharros

Código R

```
cacharros <- read.table("cacharros.txt", header = T)
attach(cacharros)
boxplot(tiempo, horizontal = T, main = 'Tiempo', col= 'blue' )
```

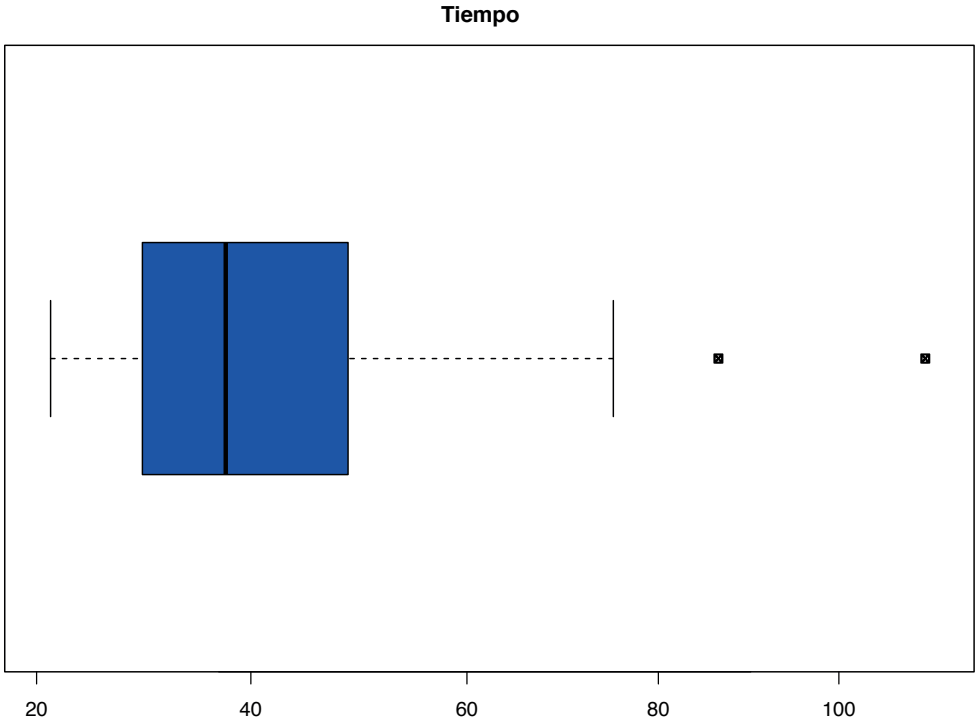


Figura 5.12. Diagrama de caja de la variable tiempo. Elaboración propia.

En la Figura 5.12 se observa la distribución de la variable tiempo con 2 datos atípicos

A continuación, en la Figura 5.13 se muestra el diagrama de caja construido a partir de los siguientes datos,

```
x <- c(-180, -174, 52, 600, 73, -154, 108, -74, 31, -450, 183, -174, -131,
-67, 17, 165,-21, -45, 4, -33, -45, 4, -540)
```

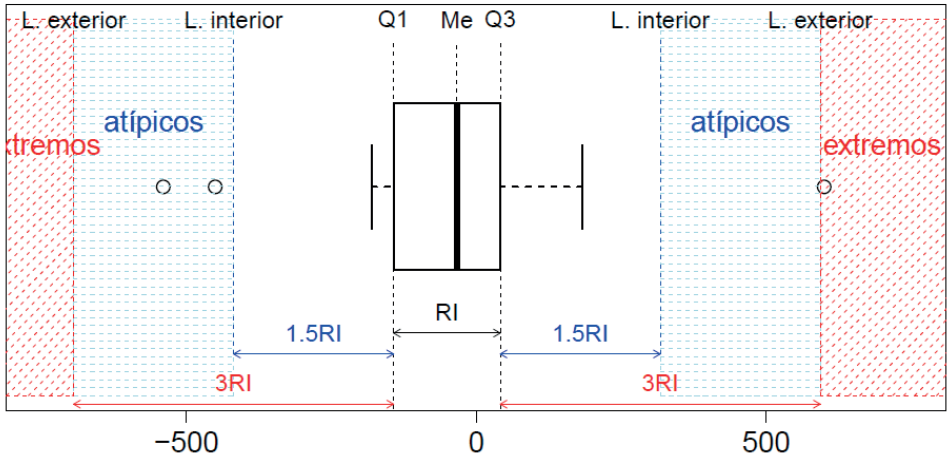


Figura 5.13. Diagrama de caja de la variable x. Elaboración propia.

El procedimiento de construcción del diagrama de caja de la Figura 5.13 es el que sigue:

- Se dibuja una caja horizontal que comienza en el primer cuartil Q1 y termina en el tercer cuartil Q3, con una línea vertical en la mediana Me.

- A continuación, se trazan dos líneas verticales situadas respectivamente a la izquierda de Q1 y derecha Q3 a una distancia de 1.5 RI. Estas constituyen las barreras interiores.

- Después se repite la misma operación a una distancia de 3 RI y estas reciben el nombre de barreras exteriores.

- Finalmente, se traza un segmento desde cada lado de la caja al dato más extremo que aparezca dentro de las barreras interiores.

Datos atípicos

Como ya se ha comentado este tipo de gráficos permiten la detección de datos atípicos:

- La caja del diagrama contiene la mitad central de los datos y cada una de las otras dos cuartas partes queda a uno de los lados de la caja.

- A las observaciones que están fuera de las barreras interiores (área sombreada en azul) se les llama datos atípicos. En particular los que caen fuera de las barreras exteriores (área sombreada en rojo) son los datos atípicos extremos.

Este tipo de datos requieren una atención especial:

- Bien porque corresponden a errores de medida,
-
- O bien porque contienen información relevante de la variable en estudio.

En cualquier caso, será muy importante la detección de dichos valores.

Con los datos anteriores los valores atípicos son -450, -540 y 600, siendo este último un atípico extremo.

Código R

```
x <- c(-180, -174, 52, 600, 73, -154, 108, -74, 31, -450, 183, -174, -131,
-67, 17, 165,
        -21, -45, 4, -33, -45, 4, -540)
boxplot(x, horizontal = T, col = 'green' )
```

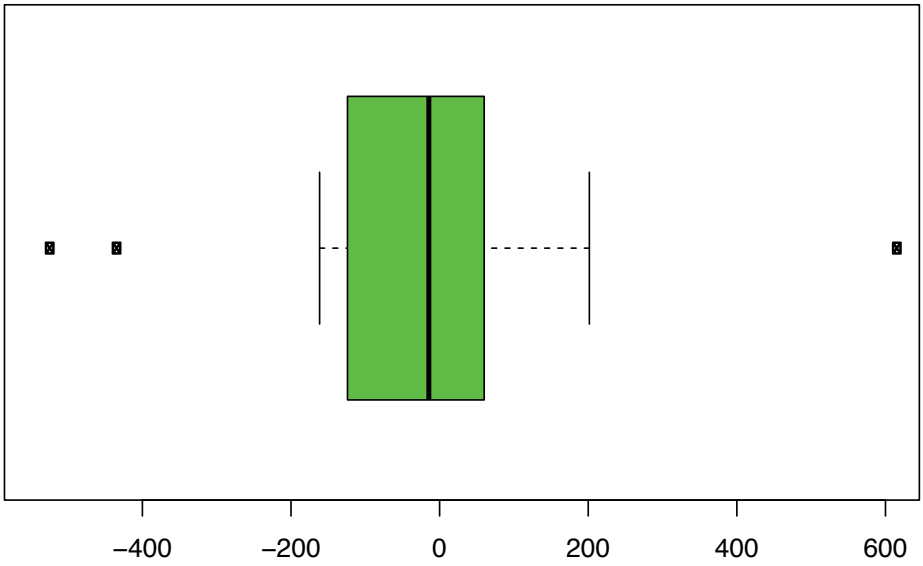


Figura 5.14. Diagrama de caja con la función boxplot de la variable x. Elaboración propia.

En la Figura 5.14 se observa la distribución de la variable x con datos atípicos.

Con el siguiente código se puede calcular los datos atípicos de la variable *x*,

Código R

```
x <- c(-180, -174, 52, 600, 73, -154, 108, -74, 31, -450, 183, -174, -131,
-67, 17, 165,
        -21, -45, 4, -33, -45, 4, -540)
Q = quantile(x, probs = c(0.25, 0.5, 0.75))
```

```
Q1 = as.numeric(Q[1]) ; Q2 = as.numeric(Q[2])
Q3 = as.numeric(Q[3]) ; RI = Q3 - Q1

# Calcular atípicos
x[x < Q1 - 1.5*RI]    #  inferiores

        # Salida de la consola
[1] -450 -540

x[x > Q3 + 1.5*RI]   #  superiores

# Salida de la consola
[1] 600

# Calcular atípicos extremos
x[x < Q1 - 3*RI]   #  inferiores

        # Salida de la consola
numeric(0)

x[x > Q3 + 3*RI]   #  superiores

        # Salida de la consola
[1] 600
```

Histograma versus Boxplot

Código R

```
cacharros <- read.table("cacharros.txt", header = T)
attach(cacharros)
par(mfcol=c(2,3), mar=c(3, 4, 2, 2))
```

```
hist(tiempo,main='tiempo',xlab="",ylab="")
boxplot(tiempo,horizontal=T)
hist(diametro,main='diametro',xlab="",ylab="")
boxplot(diametro,horizontal=T)
hist(precio,main='precio',xlab="",ylab="")
boxplot(precio,horizontal=T)
```

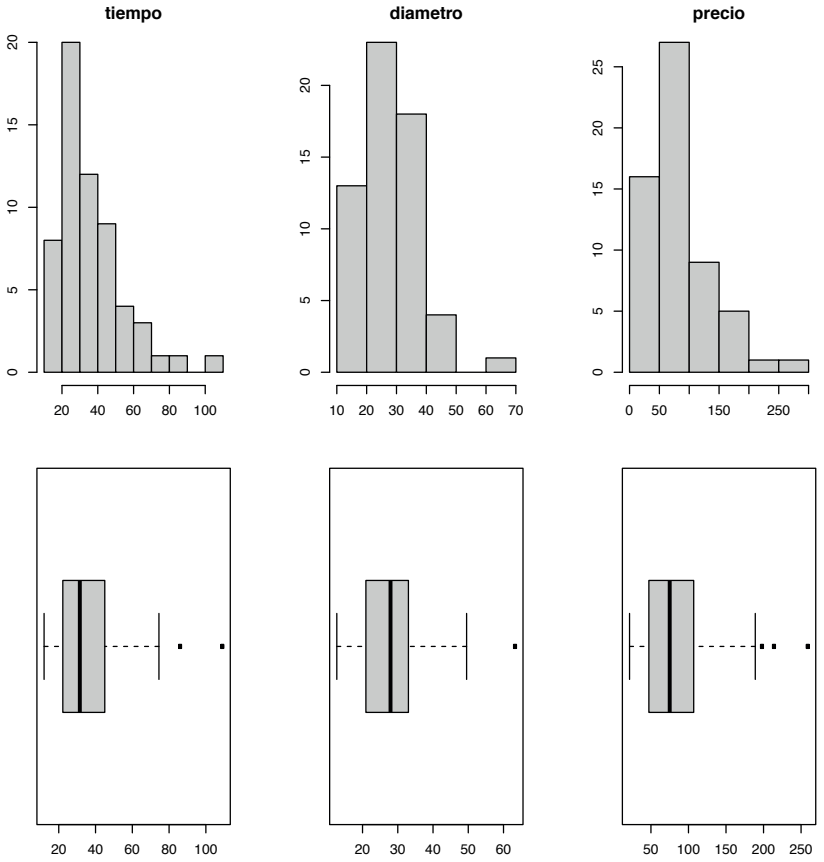


Figura 5.15. Histogramas versus boxplots. Elaboración propia.

En la Figura 5.15 se observan la distribución de las variables tiempo, diámetro y precio de la fabricación de artículos, tanto en histogramas como en diagramas de cajas.

Aplicación de R en la actualidad

En el artículo publicado: Método bootstrap para hipótesis concernientes a la diferencia de medias en muestras pareadas (http://dspace.espoch.edu.ec/handle/123456789/9394), los autores hacen un estudio descriptivo e inferencial con dos variables: producción mensual de leche y producción promedio mensual de leche, en vacas alimentadas solo con forraje y con forraje más ensilaje de maíz en los meses de julio y agosto respectivamente. Estas muestras son tomadas en la Estación Experimental Tunshi, Facultad de Ciencias Pecuarias de la Escuela Superior Politécnica de Chimborazo, Riobamba-Ecuador. En esta publicación se observa la importancia de la aplicación de herramientas del *software* R en variables continuas.

Problemas propuestos para realizar con el *software* R

La base de datos *crabs.txt* se puede obtener del paquete *MASS* del *software* R. Esta base se refiere a mediciones morfológicas en cangrejos leptograpsus dispuestas en 200 filas y 8 columnas, que describen 5 medidas morfológicas en 50 cangrejos, cada uno de dos formas de color y ambos sexos, de la especie *Leptograpsus variegatus* recolectada en Fremantle, W. Australia.
Las variables de la base son:

sp: especies, "B" o "O" para azul o naranja.
sex: sexo, "M" o "F"
index: variable con números del 1 al 50 para cada uno de los cuatro grupos.

FL: tamaño del lóbulo frontal (mm).
RW: ancho posterior (mm).
CL: longitud del caparazón (mm).

CW: ancho del caparazón (mm).
BD: profundidad del cuerpo (mm).

1. Representar gráficamente las variables FL, RW, CL, CW y BD en histogramas con la función *hist*, además en los argumentos de esta función añadir labels=T para visualizar los valores de las frecuencias en cada barra.

2. Realizar tablas de frecuencias absolutas, relativas, frecuencias absolutas acumuladas, frecuencias relativas acumuladas y porcentaje de frecuencias de las variables FL, RW, CL, CW y BD.

3. Calcular las medidas de posición y de dispersión de las variables FL, RW, CL, CW y BD.

4. Representar gráficamente las variables FL, RW, CL, CW y BD en diagramas de caja mediante la función *boxplot* para analizar la existencia de datos atípicos.

5. Realizar diagramas de dispersión utilizando la función *plot* de los siguientes pares de variables y establecer si existe alguna relación de dependencia gráfica (lineal o polinómica):

- FL versus RW
- FL versus CL
- FL versus CW
- FL versus BD
- RW versus CL
- RW versus CW

- RW versus BD
- CL versus CW
- CL versus BD
- CW versus BD

6. Utilizando la función *boxplot* realizar los diagramas de caja de cada una de las variables FL, RW, CL, CW y BD en dependencia de la variable sp. Por ejemplo, *boxplot*(FL~sp), comparar las distribuciones de los grupos de variables.

7. Utilizando la función *boxplot* realizar los diagramas de caja de cada una de las variables FL, RW, CL, CW y BD en dependencia de la variable sp. Por ejemplo, *boxplot*(FL~sex), comparar las distribuciones de los grupos de variables.

Capítulo 6

Gráficos avanzados para variables estadísticas en R

Los gráficos que se desarrollan en este capítulo se clasifican como especiales porque en su contexto tanto gráfico como en código del *software* R necesita más detalle con respecto a los que se han visto en los capítulos anteriores, así como la utilización de funciones de otras librerías o paquetes.

6.1 Gráfico de telaraña

Estos gráficos utilizan la función *radial.plot* del paquete *plotrix*, su aspecto general tiene la forma de una telaraña, que es muy utilizada para describir gráficamente variables cualitativas o cuantitativas.

Veamos el ejemplo de dos variables cualitativas que describen la influencia de las tutorías de los Docentes hacia los Estudiantes de la Universidad Nacional de Chimborazo.

Los datos del ejemplo fueron obtenidos de una encuesta con 2 preguntas, realizada a 180 estudiantes, donde estos tienen la posibilidad de elegir una de las cuatro opciones: Siempre, Frecuentemente, A veces, Nunca.

Las preguntas son:

P1: ¿Las actividades planificadas para tutorías despiertan su interés y compromiso para
participar en ellas?

P2: ¿Las actividades que realiza en las tutorías, constituyen un proceso de apoyo que
beneficia su formación como estudiante?

La información de la encuesta a los estudiantes de la Universidad Nacional de Chimborazo (Unach) está tabulada en la Tabla 6.1.

Tabla 6.1

Información de la encuesta a 180 estudiantes

Opciones	P1	P2
Siempre	43	60
Frecuentemente	66	61
A veces	55	43
Nunca	16	16

Elaboración propia.

La Figura 6.1, muestra la parte exploratoria de los valores de las frecuencias absolutas en cuatro ejes designados por las 4 opciones.

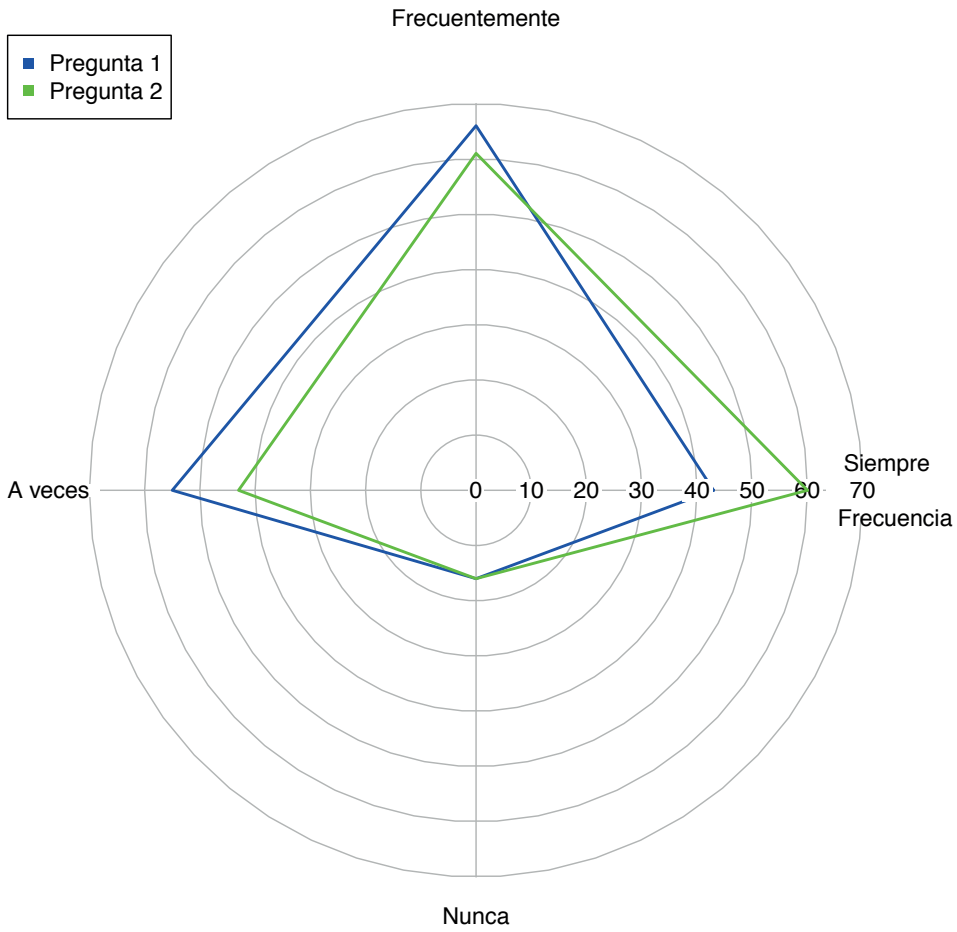


Figura 6.1. Frecuencias absolutas descritas en forma de telaraña. Elaboración propia.

El código para realizar la Figura 6.1 se detalla a continuación:

```
library(plotrix)
datos1 = c(43 ,66, 55, 16)
datos2 = c(60, 61, 43, 16)
opcion = c("Siempre\n \n Frecuencia ", "Frecuentemente", "A
veces", "Nunca")
radial.plot(datos1, labels = opcion, rp.type = "p", line.col = "blue",
            radial.lim = c(0, 70), lwd = 2)
radial.plot(datos2, labels = opcion, rp.type = "p", line.col = "green",
            radial.lim = c(0, 70), lwd = 2, add=T)
legend("topleft", c("Pregunta 1", "Pregunta 2"), pch=c(15,15),
        col = c("blue", "green"), bty = "y")
```

6.2 Gráfico de escalera

Este tipo de gráficos se utilizan para realizar diagramas en barras de una variable cualitativa o cuantitativa discreta con la función *staircase.plot* del paquete *plotrix*.

Para realizar el diagrama tomaremos los datos de la variable P1 del ejemplo anterior:

Tabla 6.2
Información de la encuesta a 180 estudiantes

Opciones	P1
Siempre	43
Frecuentemente	66
A veces	55
Nunca	16

Elaboración propia.

El código para realizar el diagrama es el siguiente:

```
library(plotrix)
datos1 = c(43 ,66, 55, 16)
totals = c("TRUE", "FALSE", "FALSE", "TRUE")
labels = c("Siempre", "Frecuentemente", "A veces", "Nunca")
staircase.plot(datos1, totals, labels, direction = "s")
```

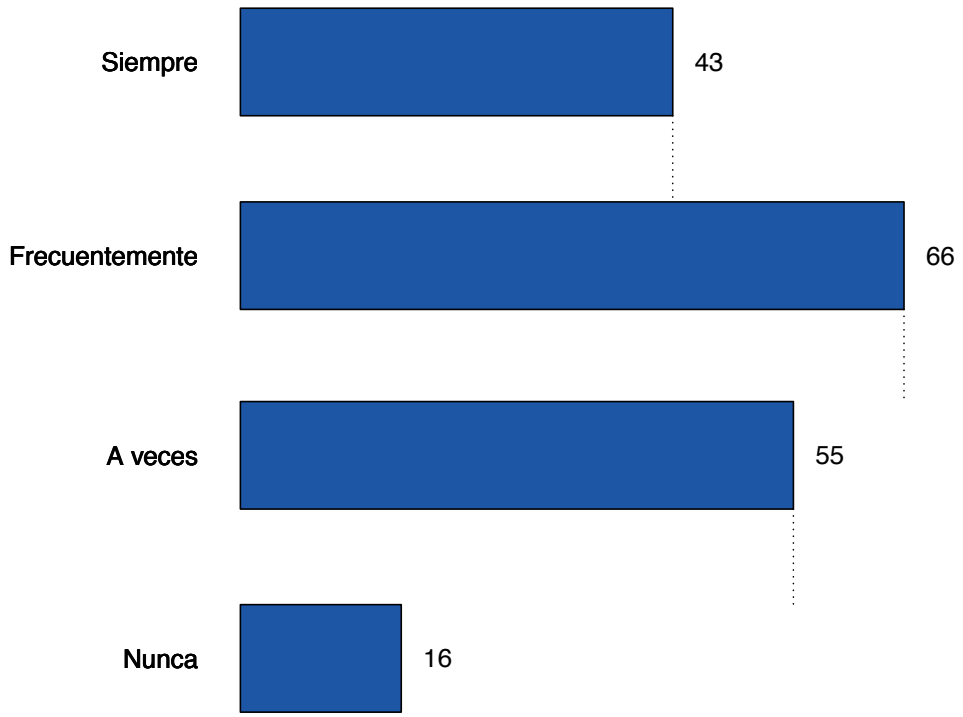


Figura 6.2. Información de la encuesta a 180 estudiantes de la variable P1. Elaboración propia.

En la Figura 6.2 se observa que cada barra tiene el valor de la frecuencia lo que hace a este diagrama que la descripción de la variable sea más concreta.

6.3 Tabla de gráficos

Los gráficos dispuestos en una tabla permiten representar diagramas de dispersión de todas las variables de una base de datos.

La base de datos *iris.txt* se refiere al estudio de la morfología de tres especies de flores. Las mediciones se realizan de la longitud de sépalo y pétalo, además del ancho de sépalo y pétalo de cada una de estas especies.

Las variables de estudio son:

Cuatro variables continuas, *Sepal.Length, Sepal.Width, Petal.Length y Petal.Width*
Una variable cualitativa: *Species*

Para la tabla de gráficos de dispersión se utiliza la función *pairs* del paquete *vcd* con el código siguiente:

```
library(vcd)
Iris = read.table("iris.txt", header = T)
head(Iris)
pairs(Iris[, 2:5], pch = 21, bg = c("brown", "blue", "cyan")
[unclass(Iris$Species)],
      cex=1.8)
grid_legend(0.2, 0.98, pch = 19, col = "brown", "setosa", frame =
FALSE)
grid_legend(0.5, 0.98, pch = 19, col = "blue", "versicolor", frame =
FALSE)
grid_legend(0.8, 0.98, pch = 19, col = "cyan", "virginica", frame =
FALSE)
```

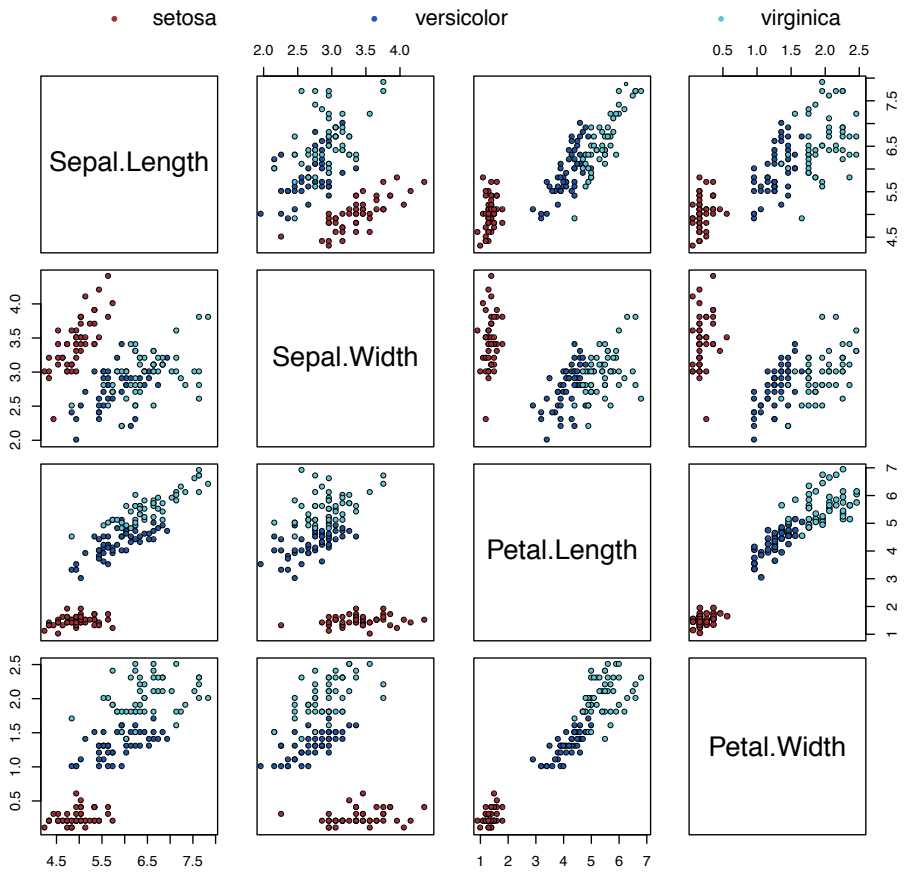


Figura 6.3. Gráficos de dispersión de una base de datos. Elaboración propia.

En la Figura 6.3 se observa una tabla de gráficas exploratorias de puntos de variables de la base de datos iris.

También el paquete *vcd* contiene las funciones *spine* y *cd_plot* para la realización de gráficos de comparación de grupos de una variable continua con respecto a una variable cualitativa.

Para la aplicación de estas funciones se utilizan las variables *sexo* y *talla* de la base de datos *pediatría.txt*, referente al estudio de mediciones de dichas variables junto con las de *edad* e *imc*, de una muestra de 2345 personas entre 5 y 19 años.

El código para las gráficas se desarrolla a continuación:

```
library(vcd)
Pediatria = read.table("pediatria.txt", header = T)
attach(Pediatria)
spine(sexo~talla, main = "Diagrama de comparación \n Talla de hombres y mujeres")
```

Figura 6.4. Comparación de talla con respecto a sexo. Elaboración propia.

En la Figura 6.4 se observa el discernimiento de la variable talla en función de la variable *sexo* (hembra, varón) y además se puede observar en el eje vertical de la parte derecha del gráfico una escala de 0 a 1 que se interpreta la diferencia en porcentaje. Por ejemplo, para las personas de talla de 1.75 cm el porcentaje de hembras es menor al 5% y el de varones en mayor al 95%.

Existe otro gráfico muy similar que se realiza con la función *cd_plot* que se puede interpretar de forma análoga al anterior.

```
library(vcd)
Pediatria = read.table("pediatria.txt", header = T)
attach(Pediatria)
cd_plot(sexo ~ talla, main = "Comparación \n Talla de hombres y mujeres")
```

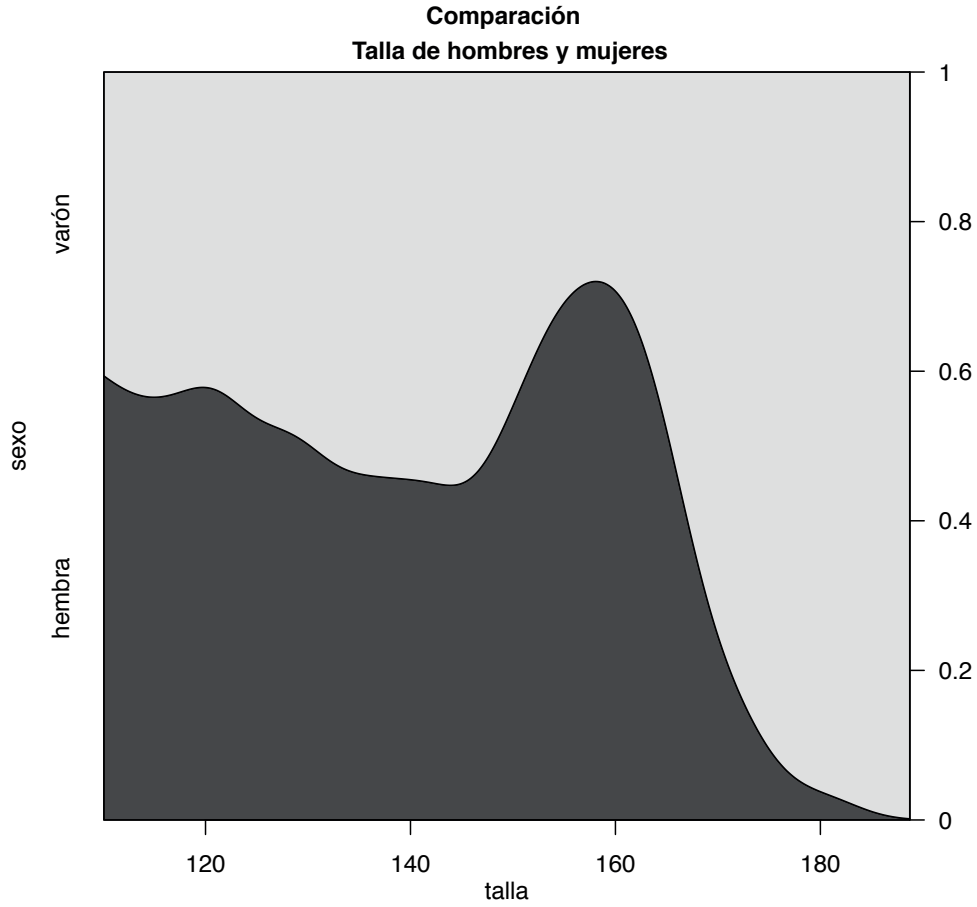


Figura 6.5. Comparación de talla con respecto a sexo con cd_plot. Elaboración propia.

6.4 Gráfico de comparación entre distribuciones de variables continuas

El paquete *beanplot* tiene una función del mismo nombre que realiza gráficos de comparación entre distribuciones de variables continuas.

Para la aplicación de esta función se utiliza la variable *talla* en dependencia de la variable *sexo*. El código para desarrollar esta gráfica es el siguiente:

```
library(beanplot)
Pediatria = read.table("pediatria.txt", header = T)
attach(Pediatria)
beanplot(talla ~ sexo, side = "both", ll = 0, col = list("pink", "blue"),
        overallline = "median", border = NA, ylab = "Talla", show.
names = F,
      main = "Comparación de distribuciones \n Talla de hombres
y mujeres")
legend("topleft", fill = c("pink", "blue"), legend = c("Mujer",
"Hombre"), bty = "n")
```

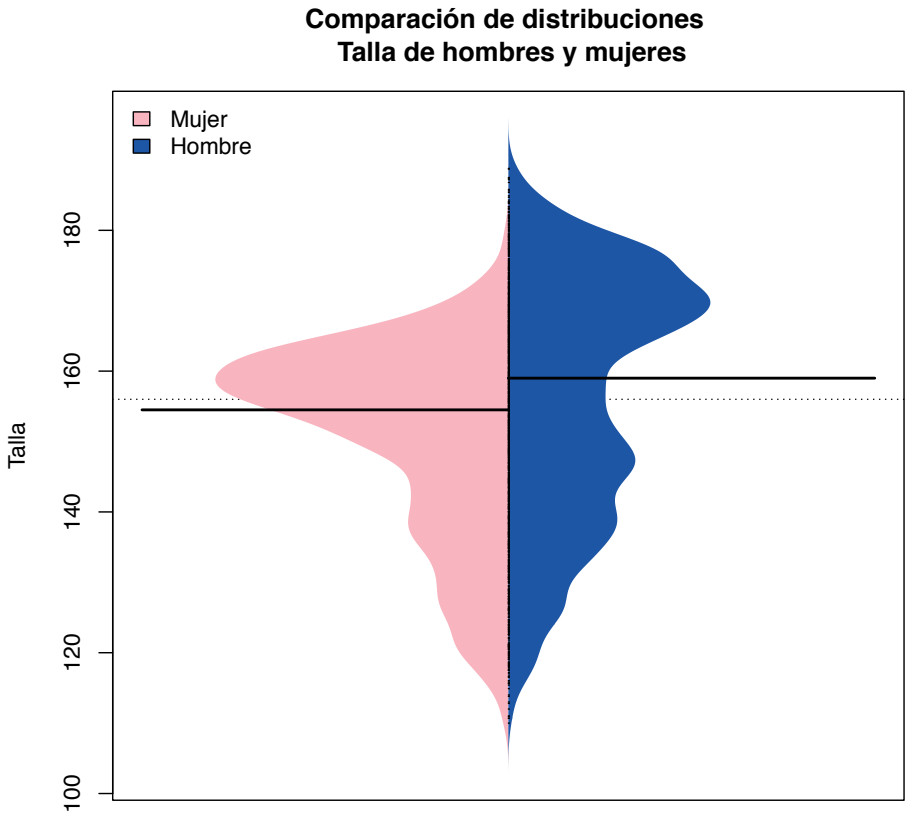


Figura 6.6. Comparación de distribuciones de talla con respecto a sexo. Elaboración propia.

En la Figura 6.6, se observan las distribuciones de la variable talla referente a hombres en la parte derecha y referente a mujeres en la parte izquierda del gráfico central, también se dibujan las rectas horizontales a la altura de sus medianas en estas distribuciones, cuya mediana de la talla de las mujeres es menor que la mediana de la talla de hombres.

6.5 Gráfico de dispersión

El gráfico más utilizado para el estudio de variables continuas es el de dispersión, por este motivo se realiza el código con la función básica *plot* con varios argumentos que hacen posible una mejor visualización de las características de las variables.

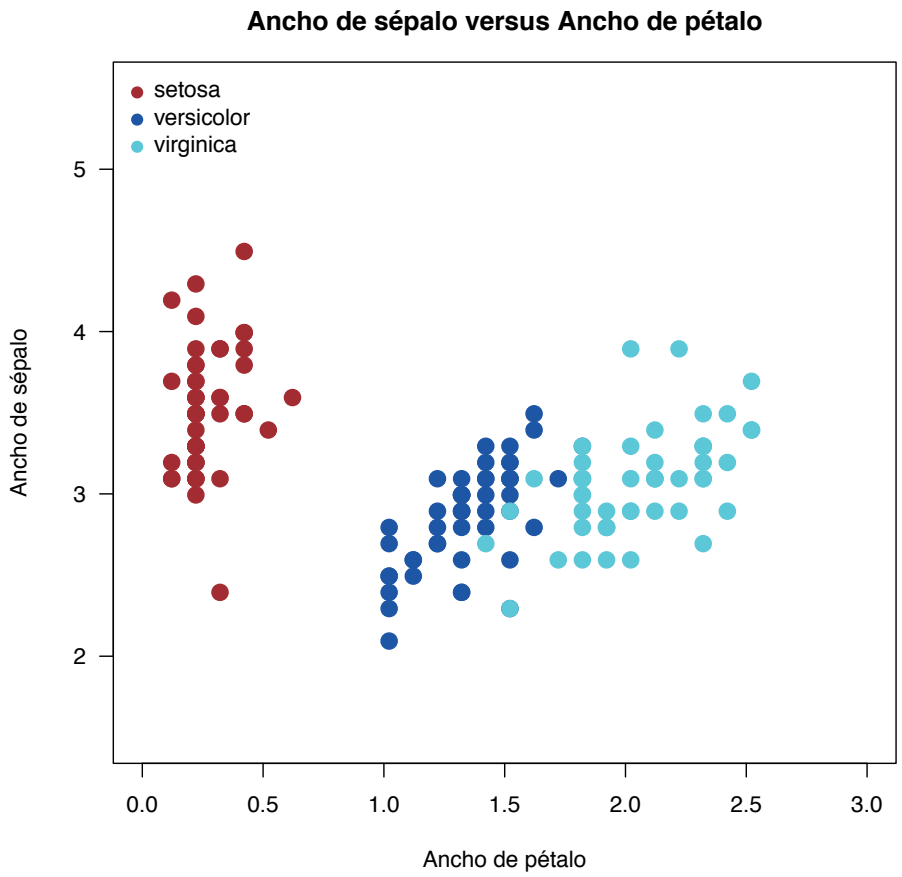


Figura 6.7. Dispersión de la variable ancho de sépalo en función del ancho de pétalo. Elaboración propia.

En la Figura 6.7 se observa la dispersión de puntos de la variable ancho de sépalo en función de la variable ancho de pétalo, identificados por colores con respecto a las especies.

El código para la realización de la Figura 6.7 se desarrolla a continuación:

```
Iris = read.table("iris.txt", header = T)
attach(Iris)
plot(Iris[Species == "setosa", "Petal.Width"],
     Iris[Species == "setosa", "Sepal.Width"],
```

```
      xlim = c(0,3), ylim = c(1.5, 5.5), col = "brown",
      xlab = "", ylab = "", las = 1, pch = 20, cex=1.5)
points(Iris[Species == "versicolor", "Petal.Width"],
         Iris[Species == "versicolor", "Sepal.Width"],
         col = "blue", pch = 20, cex=1.5)
points(Iris[Species == "virginica", "Petal.Width"],
         Iris[Species == "virginica", "Sepal.Width"],
         col = "cyan", pch = 20, cex=1.5)
title(main = "Ancho de sépalo versus Ancho de pétalo",
     ylab = "Ancho de sépalo", xlab="Ancho de pétalo")
legend("topleft", c("setosa", "versicolor", "virginica"),
       pch = c(20, 20, 20), col = c("brown", "blue", "cyan"), bty="n")
```

Problemas propuestos para realizar con el *software* R

La base de datos *crabs.txt* se puede obtener del paquete *MASS* del *software* R. Esta base se refiere a mediciones morfológicas en cangrejos leptograpsus dispuestas en 200 filas y 8 columnas, que describen 5 medidas morfológicas en 50 cangrejos, cada uno de dos formas de color y ambos sexos, de la especie Leptograpsus variegatus recolectada en Fremantle, W. Australia.

Las variables de la base son:
sp: especies, "B" o "O" para azul o naranja.
sex: sexo, "M" o "F"
index: variable con números del 1 al 50 para cada uno de los cuatro grupos.
FL: tamaño del lóbulo frontal (mm).
RW: ancho posterior (mm).
CL: longitud del caparazón (mm).
CW: ancho del caparazón (mm).
BD: profundidad del cuerpo (mm).

1. Representar gráficamente las variables FL, RW, CL, CW y BD en dependencia de la variable sp mediante la función *pairs*.

2. Representar gráficamente las variables FL, RW, CL, CW y BD en dependencia de la variable sex mediante la función *pairs*.

3. Utilizando la función *cd_plot* realizar los gráficos de comparación con cada una de las variables FL, RW, CL, CW y BD en dependencia de la variable sp. Por ejemplo, *cd_plot*(sp ~ FL), comparar las distribuciones de los grupos de variables.

4. Utilizando la función *cd_plot* realizar los gráficos de comparación con cada una de las variables FL, RW, CL, CW y BD en dependencia de la variable sex. Por ejemplo, *cd_plot*(sex ~ FL), comparar las distribuciones de los grupos de variables.

5. Utilizando la función *beanplot* realizar los gráficos de comparación con cada una de las variables FL, RW, CL, CW y BD en dependencia de la variable sp. Por ejemplo, *beanplot* (FL~ sp), comparar las distribuciones de los grupos de variables.

6. Utilizando la función *beanplot* realizar los gráficos de comparación con cada una de las variables FL, RW, CL, CW y BD en dependencia de la variable sex. Por ejemplo, *beanplot* (FL~ sex), comparar las distribuciones de los grupos de variables.

7. Realizar diagramas de dispersión utilizando la función *plot* de los siguientes pares de variables e identificando con colores diferentes los puntos que corresponden a los niveles de la variable sex.

- FL versus RW
- FL versus CL
- FL versus CW
- FL versus BD
- RW versus CL
- RW versus CW
- RW versus BD
- CL versus CW
- CL versus BD
- CW versus BD

Anexos

Ayudas en el *software* R

El *software* R tiene varias ayudas dependiendo de la necesidad, por ejemplo:

A. Componentes de una librería o paquete

Para saber sobre las componentes de una librería o paquete, se necesita la función, *help* para el sistema on-line de ayuda. La demostración de esta función se realiza con uno de los paquetes básicos instalado por defecto en el *software*, denominado *car*.

El código en la consola es como sigue:

```
help(package = "car")
```

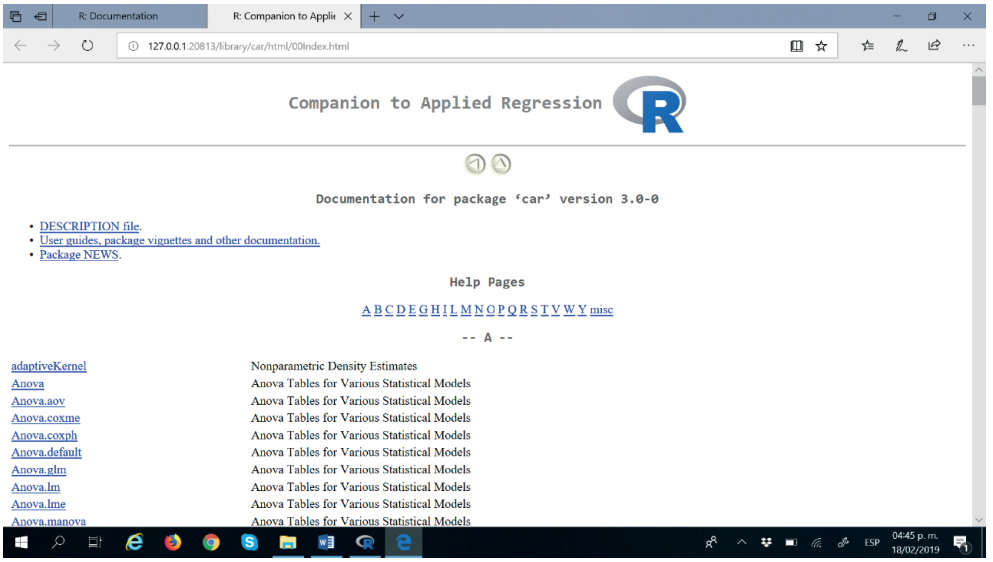


Figura A.1. Ayuda del paquete car. Elaboración propia.

En la Figura A.1 se observan todos los componentes que tiene el paquete car, desde la A hasta la Y, estos son las funciones

principales con las que cuenta. Al hacer click en una de ellas se despliega también su respectiva ayuda.

Otra manera de obtener la ayuda de una función principal (por ejemplo, *anova*) de un paquete es,

```
help(anova)
```

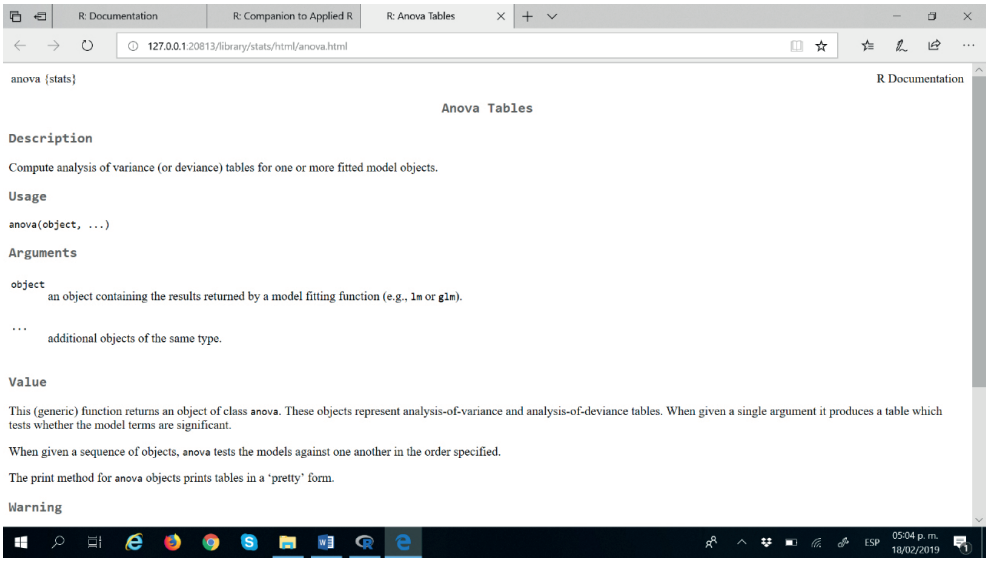


Figura A.2. Ayuda dela función anova del paquete car. Elaboración propia.

B. Instalación de nuevas librerías o paquetes

El *software* R contiene paquetes básicos que por defecto vienen listos para ser utilizados una vez que se instala el *software*.

Por la necesidad de nuevos paquetes se procede a instalarlos desde los servidores ubicados en varios países del mundo, se aconseja ubicar el más cercano.

A continuación se realiza paso a paso la instalación de un paquete, por ejemplo el que se ha venido utilizando, *plotrix*.

Paso 1. Se selecciona el espejo CRAN

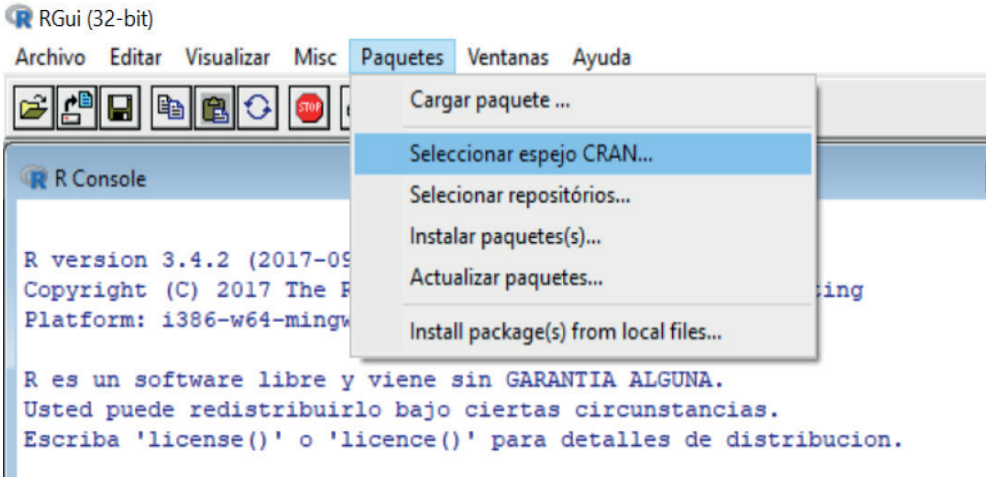


Figura B.1. Espejo CRAN. Elaboración propia.

Paso 2. Se elige el país

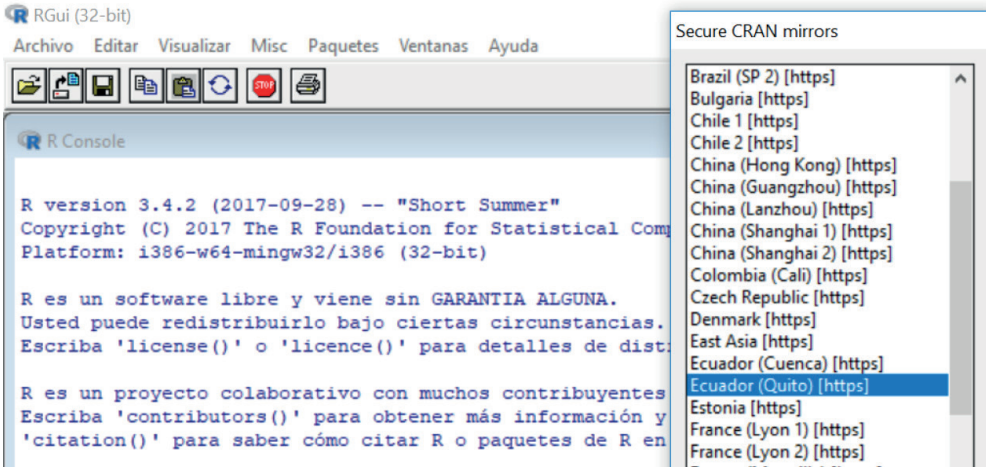


Figura B.2. Ubicación del país donde está un servidor. Elaboración propia.

En este paso, Figura B.2, es necesario elegir el país donde se encuentra un servidor, desde el que se va a descargar el paquete, en este caso será conveniente Ecuador (Quito) o también se puede desde Ecuador (Cuenca).

Paso 3. Instalación del paquete

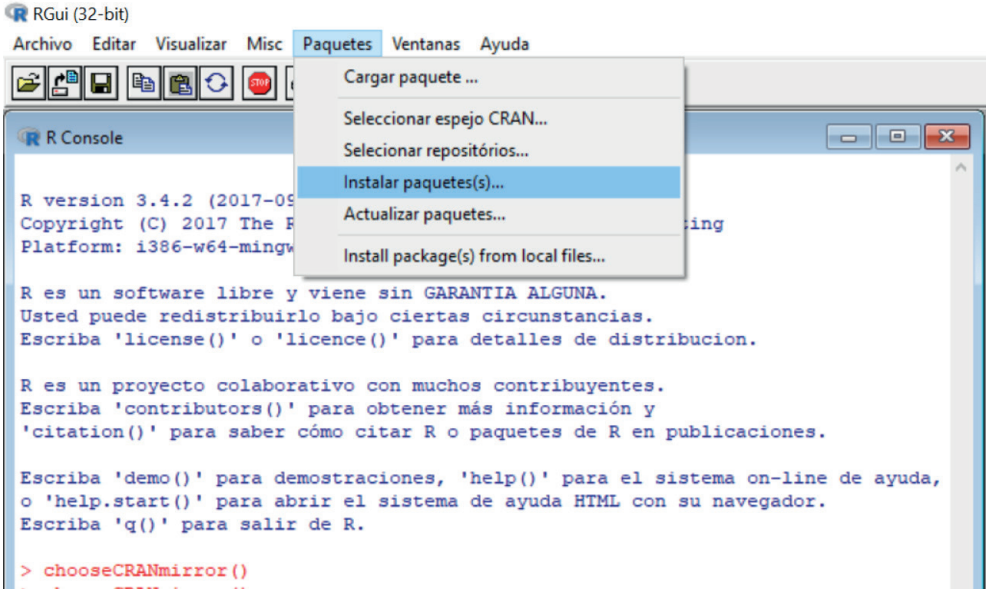


Figura B.3. De paquetes elegir la instalación. Elaboración propia.

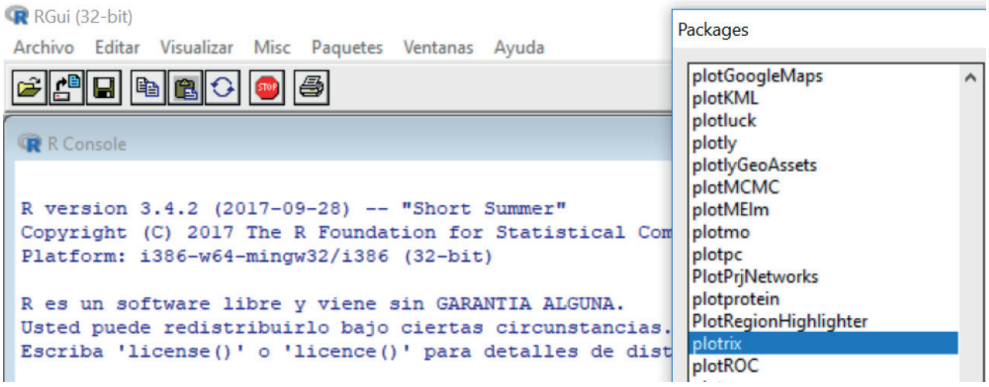


Figura B.4. Elección de un paquete (plotrix). Elaboración propia.

En la Figura B.4 se observa un listado de paquetes, del que se puede elegir uno de ellos de acuerdo a la necesidad, por ejemplo el paquete plotrix, que hemos utilizado para realizar varios gráficos en el libro. Una vez instalado el paquete se puede utilizar cada vez que se requiera.

C. Uso de la consola

El capítulo 1 ya se realizó algunos ejemplos sobre el uso de la consola, pero cabe señalar también aspectos importantes sobre la limpieza y el borrado de la memoria de la consola.

Limpieza de la consola

La limpieza se realiza de tres maneras:

Primera, la más simple es pulsar la teclas *Control+L*

Segunda, es haciendo click derecho del mouse y aparece una ventana de la que se elige *Limpiar pantalla*, Figura C.1:

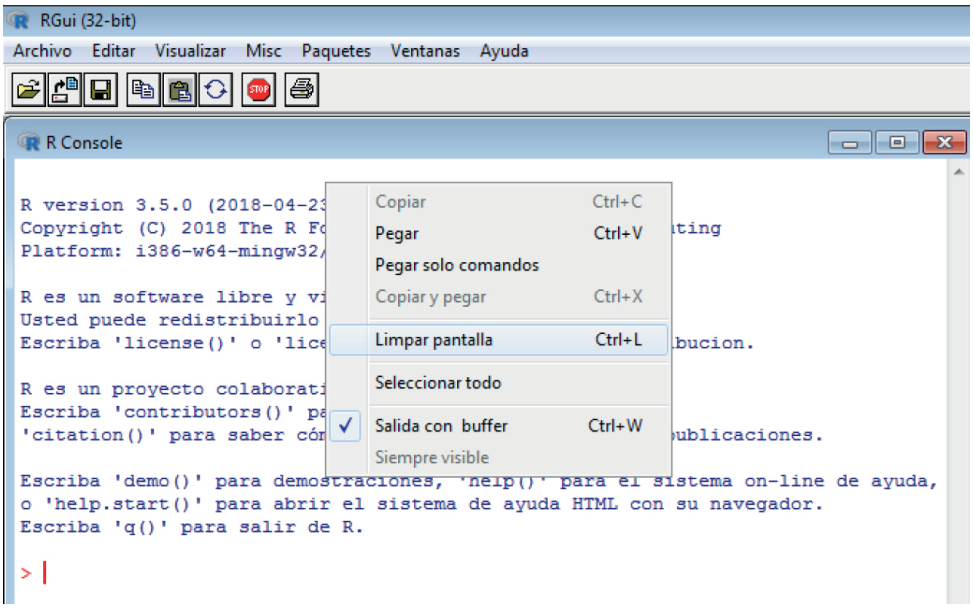


Figura C.1. Limpieza de la consola. Elaboración propia.

Tercera, la limpieza es desde *Editar*, se elige *Limpiar consola,* y la pantalla queda en blanco.

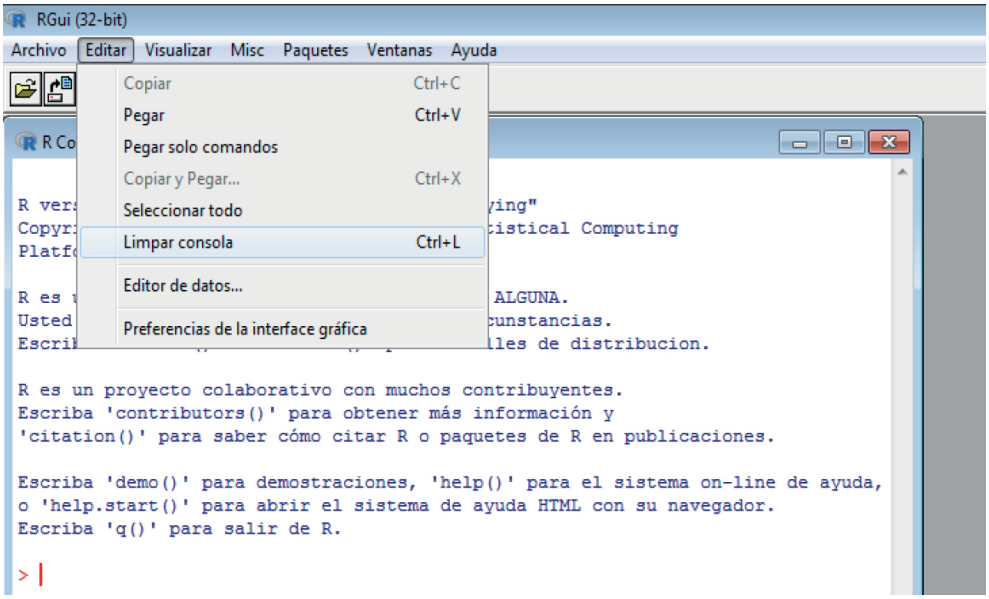


Figura C.2. Editar y limpieza de la consola. Elaboración propia.

Remover todos los objetos de la consola

Al remover los objetos, se está limpiando todas las variables, gráficos, y otras operaciones realizadas en la memoria de la consola. Por lo que se debe tener mucho cuidado, ya que nunca más se puede usar estos objetos borrados.

En la Figura C.3 se observa que al hacer click en la pestaña Misc, se despliega una ventana con varios íconos, de los que tomamos, Remover todos los objetos, y aparece en la consola el código:

```
rm(list = ls(all = TRUE))
```

Y la consola queda vacía en su totalidad. Para confirmar, nuevamente hacemos click en Misc y se pulsa Listar objetos, y se tiene el código:

```
ls()
        # Salida de la consola
character(0)
```

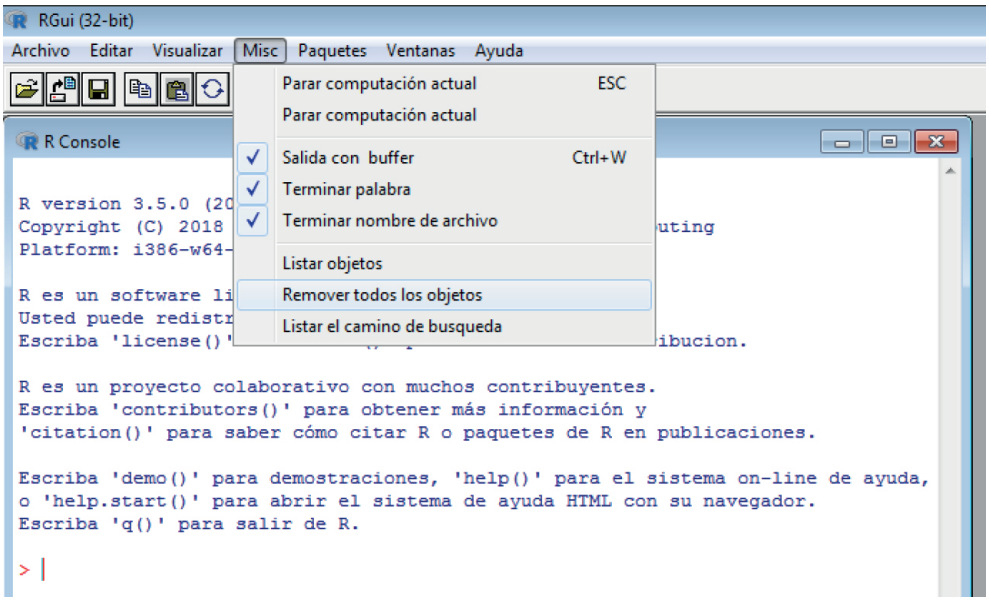


Figura C.3. Remover todos los objetos de la consola. Elaboración propia.

D. lectura y manejo de archivos

En general existen dos clases de archivos que se pueden manipular en el *software* R.

Iniciamos con los que se encuentran en las librerías o paquetes del *software*. Primero ubicamos el paquete del que se va a obtener la base de datos o archivo, por ejemplo:

```
library(datasets)
data(iris)          # Ingreso a la base de datos iris
head(iris)          # Para observar las 6 primeras filas de iris
tail(iris)          # Para observar las 6 últimas filas de iris
```

Se note que los códigos *head* y *tail* son necesarios únicamente cuando la base de datos es demasiada extensa.

En la Figura D.1 se observan dos ventanas dispuestas en forma vertical, la superior es la del script, donde están los códigos anteriores y en la inferior es la consola en la que se corre dichos códigos mediante la pestaña Correr línea o seleccionar.

Entonces en este momento la base de datos *iris* está dispuesta para realizar el respectivo análisis.

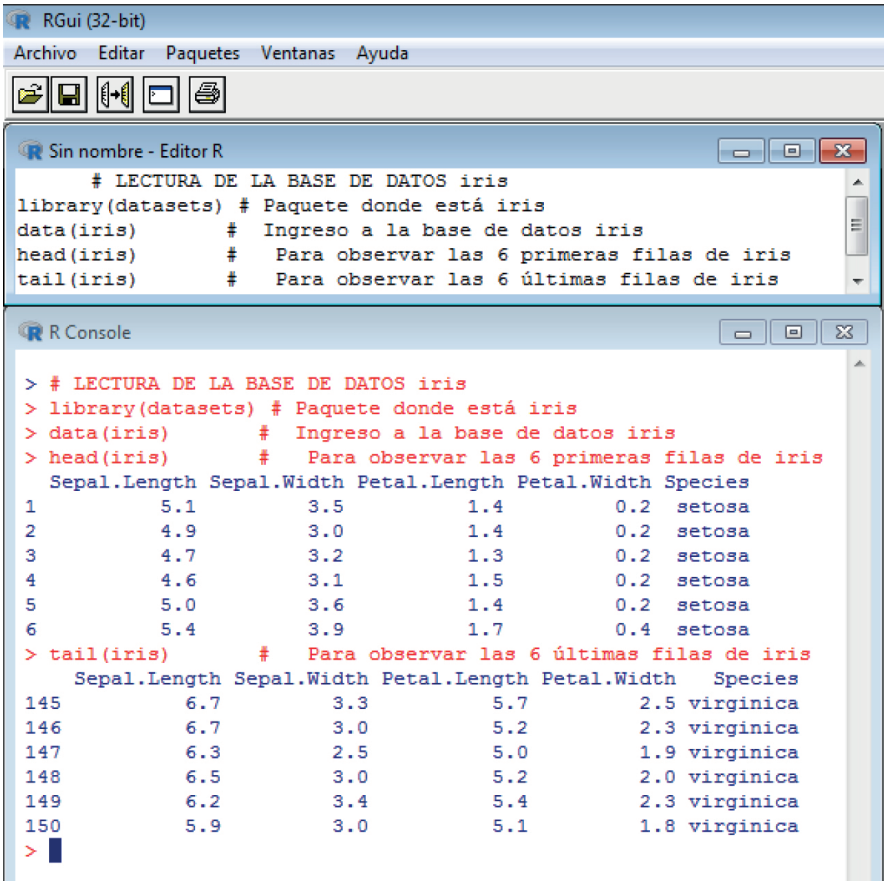


Figura D.1. Lectura de la base de datos iris. Elaboración propia

Para conocer de manera detallada esta base datos iris, se accede con la ayuda en la consola, *help(iris)* y a continuación se despliega una ventana on-line con la descripción, uso, formato, lugar donde se tomó los datos, referencias y ejemplo en el R.

También se puede acceder a todas las bases de datos de un paquete por ejemplo del *datasets*, insertando en la consola *help(package=datasets)*.

Todas o la mayoría de las librerías del *software* R contienen bases de datos que han sido utilizadas para desarrollar y aplicar sus funciones principales.

Existen otros archivos externos con datos de investigaciones personales o grupales, los más simples son los realizados en formato texto, con extensión .txt, los que se dan lectura en el R con la función *read.table*.

Se debe seguir dos pasos para la lectura de un archivo externo:

Primer paso. Direccionar la consola hacia el lugar donde se encuentra el archivo

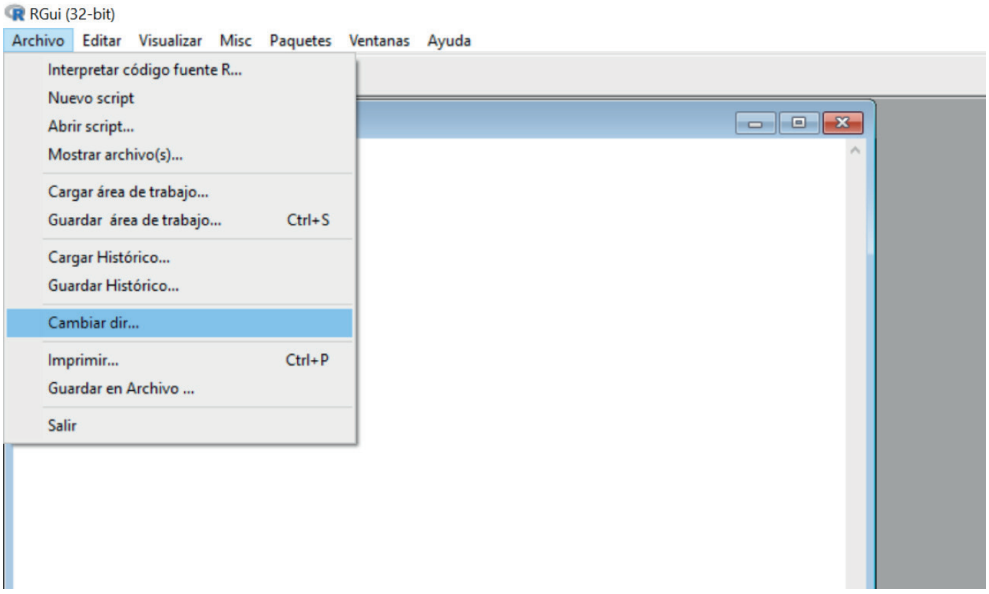


Figura D.2. Dirección de la consola. Elaboración propia.

En la Figura D.2 se observa la pestaña, Cambiar dir…, la que direcciona la consola al lugar de trabajo donde se encuentran los archivos que van a ser leídos.

A continuación, se identifica la ubicación del archivo en el computador, por ejemplo, si un archivo se encuentra en el escritorio la dirección será: C:\Users\ANTONIO\Desktop y esta se copia en el recuadro a lado derecho de, Carpeta, y luego se pulsa aceptar (Figura D.3).

Este procedimiento nos ayuda a leer directamente los archivos desde el escritorio en este caso, así, análogamente se puede realizar la lectura de un archivo desde cualquiera que sea la ubicación o dirección.

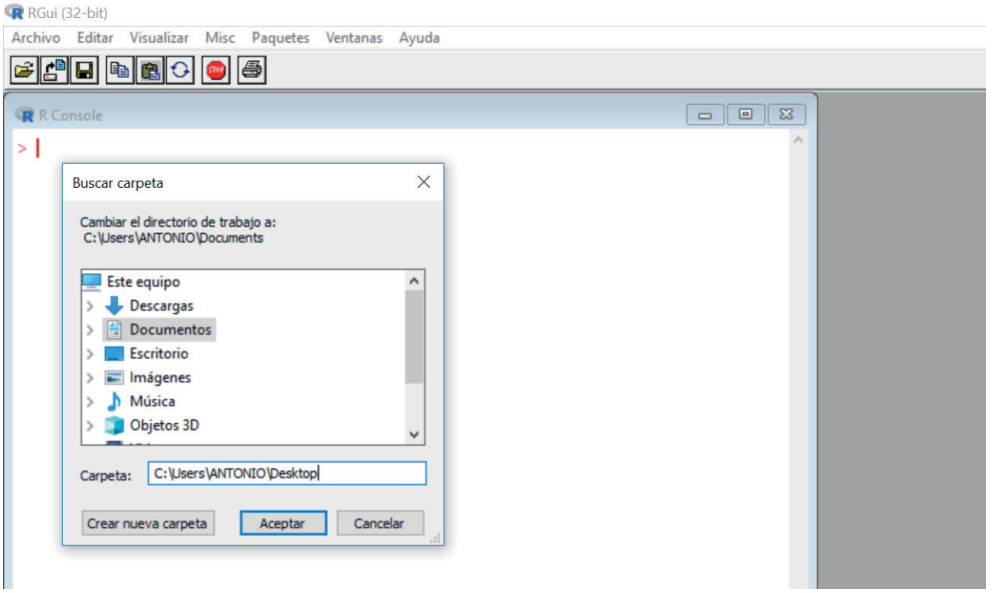


Figura D.3. Dirección del archivo. Elaboración propia.

Segundo paso. En esta parte se crea un nuevo script, y se divide la ventana principal de forma horizontal o vertical. Si, se elige vertical se ubica el script a la izquierda y la consola a la derecha (es opcional, puede ser viceversa).

En el script se escribe el nombre que se desea dar a la base de datos, seguido del sigo "=" o "<-" y de la función *read.table*.

El código de lectura es el siguiente:

```
Rendimiento = read.table("pruebas.txt", header=TRUE)
Rendimiento
```

```
      # Salida de la consola
     prueba1  Prueba2
1          5         9
2          6        10
3          7         9
4          8         8
5          9         7
6         10         6
7          4         3
```

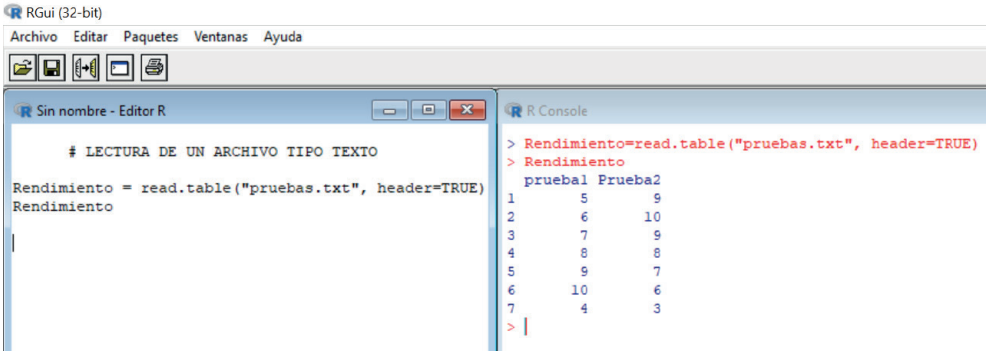


Figura D.3. Lectura del archivo de texto. Elaboración propia.

Referencias Bibliográficas

Fox, J. (2018). *RcmdrMisc: R Commander Miscellaneous Functions*. R package version 1.0-10. Obtenido de: https://CRAN.R-project.org/package=RcmdrMisc

Fox, J., & Weisberg, S. (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. Obtenido de: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gutierrez, E. (2014). Probabilidad y estadística. Aplicaciones a la ingeniería y ciencias. México: Ed. Patria.

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software*, Code Snippets 28(1). 1-9. Obtenido de: http://www.jstatsoft.org/v28/c01/

Lemon, J. (2006). Plotrix: a package in the red light district of R. R-News, 6(4): 8-12.

Mendenhall, W. (2010). *Introducción a la probabilidad y estadística*. México: Ed. Cengage Learning

Miller, J. (1999). Estadística matemática con aplicaciones. México: Ed. Pearson.

R Core Team. (2017). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Obtenido de: https://www.R-project.org/

Ross, S. (2005). Introducción a la Estadística. México: Ed. Reverté.

Spiegel, M (1976). Teoría y problemas de probabilidad y estadística. México: Ed. McGraw-Hill, Serie Schaum

Walpole, R. (2012). Probabilidad y estadística para ingenieros y ciencias. México: Ed. Pearson-Prentice Hall.

UNIVERSIDAD NACIONAL DE CHIMBORAZO

Gestión del Conocimiento y Propiedad Intelectual

ESTADÍSTICA DESCRIPTIVA CON R. GRÁFICOS AVANZADOS Y APLICACIONES*;* se publicó en el mes de julio de 2021 en la Universidad Nacional de Chimborazo.

Estadística Descriptiva con R. Gráficos avanzados y aplicaciones

En la actualidad, la demanda de bases de datos es muy elevada, debido a que existe bastante información de investigaciones en áreas de Ingeniería, Salud, Educación, Ciencias Políticas y en otras en general. Estas bases necesitan un análisis detallado para determinar sus características y cualidades relevantes las que se deben representar de forma resumida y clara.

El presente libro titulado, Estadística Descriptiva con R. Gráficos avanzados y aplicaciones, es una alternativa para realizar el análisis de datos utilizando el software estadístico R de libre acceso en internet. El desarrollo de este libro está dividido en seis capítulos, los anexos y la bibliografía; a continuación se realiza su descripción.

En el primer capítulo, se realiza los primeros pasos que hay que hacer para entrar en confianza con los códigos, ventanas e íconos del software R. Este software es de libre acceso, que se puede obtener de la página de internet https://www.r-project.org/ y elegir el cran mirror más próximo para descargarlo de la forma más rápida. También se desarrolla varios ejemplos y gráficos sencillos, así como la manera de utilizar un script y la consola.

VICERRECTORADO DE **Investigación, Vinculación y Posgrado**
DIRECCIÓN DE **Investigación**
GESTIÓN DEL **Conocimiento y Propiedad Intelectual**

Campus "La Dolorosa" Avda. Eloy Alfaro y 10 de Agosto | Teléfonos: (593-3) 37 30 880 • ext. 2007- 2008
http://investigacion.unach.edu.ec/