Estadística con aplicaciones en R

Manuel Ricardo Contento Rubio

EDITORIAL UTADEO

Manuel Ricardo Contento Rubio

Estadístico y Magíster en Enseñanza de las Ciencias Exactas y Naturales (Universidad Nacional de Colombia), Magíster en Modelado y Simulación (Universidad de Bogotá Jorge Tadeo Lozano). Pertenece al grupo de investigación de Didáctica de las Ciencias en la Línea de evaluación de la educación de la UJTL en donde ha desarrollado investigación en Modelos de la Teoría de respuesta al ítem. Ha sido profesor y director de tesis de estudiantes de la Maestría en Ciencias Ambientales y Maestría en Modelado y Simulación de la UJTL.

Estadística con aplicaciones en R

UTADEO
UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO
FACULTAD DE CIENCIAS NATURALES E INGENIERÍA
DEPARTAMENTO DE CIENCIAS BÁSICAS

Contento Rubio, Manuel Ricardo
Estadística con aplicaciones en R. / Manuel Ricardo Contento Rubio. - Bogotá: Universidad de Bogotá Jorge Tadeo Lozano, 2019.
412 páginas ; 22 cm.

ISBN: 978-958-725-272-9

1. Estadística – Procesamiento de datos. 2. R (Lenguaje de programación de computadores). 3. Estadística descriptiva.
4. Probabilidades. 5. Intervalos de confianza. 6. Prueba de hipótesis estadística. 7. Análisis de varianza. 8. Análisis de regresión. I. Tít.

CDD519.50285

Estadística con aplicaciones en R
ISBN impreso: 978-958-725-272-9
ISBN digital: 978-958-725-273-6
ISBN e-pub: 978-958-725-274-3

Edición:
Diseño de carátula y pauta gráfica: Juanita Giraldo
Adecuación pauta gráfica: Luis Carlos Celis Calderón
Corrección de estilo: Hernando García Bustos
Coordinación editorial: Mary Lidia Molina Bernal
Diagramación: Francisco Jiménez

Fundación Universidad de Bogotá Jorge Tadeo Lozano | Vigilada Mineducación.
Reconocimiento de personería jurídica: Resolución N°. 2613 de 14 de agosto de 1959, Minjusticia.
Acreditación institucional de alta calidad, 6 años: Resolución 4624 del 21 de marzo de 2018, Mineducación.

Estadística con aplicaciones en R

Manuel Ricardo Contento Rubio

Contenido

Presentación

A lo largo de la experiencia como docente de universidad, es frecuente evidenciar que cuando los estudiantes enfrentan el aprendizaje de conceptos de estadística logran la habilidad suficiente para saber utilizar algunos algoritmos y aplicar ciertos modelos de probabilidad. Saben, por ejemplo, cómo proceder cuando se encuentran frente a una variable con distribución normal y cuándo aplicar un determinado intervalo de confianza o algunas pruebas de hipótesis; sin embargo, también se evidencian dificultades al momento de la interpretación en contexto, debido en muchos casos a que no comprenden el sentido de lo que aprenden. ¿Qué caracteriza a los valores que se distribuyen de forma sesgada versus distribuciones simétricas? ¿Cuál es el significado y las implicaciones del teorema central del límite? ¿Cuál es la relación que existe entre el valor de la estadística de prueba y el p-valor?

Tener la posibilidad de explorar las ideas detrás de cada concepto o procedimiento y descubrir relaciones entre ellos puede favorecer la comprensión para así aprender mejor su verdadero significado. La tecnología y el software hacen posible este hecho al permitir la simulación de experimentos aleatorios con rapidez y fiabilidad, así como visualizar datos para revelar patrones que generan detalle y conocimiento a profundidad de ciertos fenómenos y que al final se convierten en información comprensible al usuario. El software elegido es R, el cual tiene dos ventajas: la primera, es de uso libre; la segunda, los estudiantes pueden usarlo en cualquier lugar y momento sin limitarse al entorno del aula.

El software, descargable de manera gratuita en http://cran.r-project.org, es un conjunto integrado de programas que permite, entre otras acciones, manejar bases de datos, hacer cálculos complejos, proveer resultados rápidamente y elaborar gráficos estadísticos de gran calidad. Este software consta de una serie de paquetes básicos y otros que se pueden descargar según las necesidades del usuario, de manera que se potencian las posibilidades de la aplicación.

Este libro presenta el entorno del *software* R como una herramienta para los cálculos y proporciona una justificación para usar este programa sobre cualquier otro, en particular los que requieren de licencia. Inicialmente, se ofrece una introducción general a R en el contexto estadístico (capítulo 1), para luego dar paso al análisis descriptivo (capítulo 2) en el cual se muestran algunas funciones clave de R para visualizar información. Los conceptos de probabilidad y variables aleatorias cobijan el tercer capítulo y son fundamentales para el manejo de los modelos estadísticos más usados para variables discretas y continuas. El capítulo 4 ofrece las funciones de R para el cálculo de probabilidades y cuantiles, sustituyendo el manejo de tablas. El capítulo 5 trata las distribuciones de muestreo y los teoremas asociados más importantes y que constituyen la base de la inferencia mediante intervalos de confianza (capítulo 6) y prueba de hipótesis (capítulo 7). Finalmente, se da paso a dos métodos que tienen un lugar preponderante entre los procedimientos estadísticos: el análisis de varianza (capítulo 8) y la regresión lineal (capítulo 9).

Aprovecho para dar gracias a las sugerencias y comentarios de profesores y estudiantes, quienes usaron las versiones preliminares en la Universidad de Bogotá Jorge Tadeo Lozano y el conjunto de temas que finalmente se plasman en este libro ha evolucionado y se ha enriquecido de las conversaciones mantenidas con ellos.

1
El contexto estadístico

Introducción

Es frecuente que muchos de ustedes se pregunten ¿qué es estadística?, ¿quiénes la usan?, ¿cuándo se debe aplicar?, ¿por qué debo estudiarla? En este capítulo se hace una presentación de la estadística, su contexto y sobre el *software* R que apoya este libro, buscando aportar a responder estos interrogantes y animar al lector a darse una oportunidad de conocer un poco de esta útil herramienta.

¿Por qué estudiar estadística?

En el ámbito académico y profesional es frecuente la lectura de artículos científicos y una diversidad de otras publicaciones en donde se utiliza estadística, con el consabido reto para el lector de comprender aquello que está escrito. A manera de ejemplo, se presenta el aparte de análisis estadístico que fue efectuado en una investigación sobre la contaminación de una fuente de agua (Claret, Urrutia, Ortega, Abarzua, Perez & Palacios, 2005).

Estudio de la contaminación en agua de pozo destinada a consumo humano y su expresión espacial en el secano mediterráneo de Chile

Claret M.*, Urrutia R.**, Ortega R.***, Abarzua M. **, Pérez C*. y Palacios M*.

* Instituto de Investigaciones Agropecuarias (INIA); ** Centro EULA, Universidad de Concepción; *** Pontificia Universidad Católica de Chile.

Análisis Estadísticos

Mediante el uso del sw SAS para análisis de varianza, se obtuvo las estadísticas básicas y Coeficiente de Correlación de Pearson. La Matriz de Resultados muestra una correlación significativa (<0.05 correlación significativa) entre conductividad eléctrica (ce) y concentración de nitratos con un valor de <0.001. los resultados sugieren que ce podría ser un buen indicador del contenido de nitratos.

Al pasar por la lectura de este fragmento de la publicación se puede identificar una serie de términos técnicos, referidos al análisis de los datos obtenidos, que tienen implicaciones estadísticas: sw SAS, análisis de varianza, estadísticas básicas, coeficiente de correlación de Pearson, matriz de resultados, correlación significativa (< 0.05 correlación significativa), valor de < 0.001. Al final aparece una conclusión derivada de los métodos estadísticos usados, anunciando que la conductividad eléctrica está asociada con el contenido de nitratos.

Las preguntas relevantes ahora son: ¿podemos entender lo que allí está escrito?, ¿qué significado tienen los números que aparecen?, ¿qué se debe entender por correlación?, ¿qué implicación tiene el calificativo *significativo* adjuntado a la correlación?, ¿cuáles son los resultados que sugieren la conclusión que se presenta?

La práctica más cómoda para el lector sin mayor preocupación por una postura crítica sobre la publicación es dar por sentado lo que aparece escrito, considerar pertinentes y adecuados los métodos estadísticos usados y validas las conclusiones que se derivan de los datos. En resumen, ser de la opinión "si está publicado debe ser cierto".

Hay que comentar que esta posición implica una situación ideal, donde los editores de las publicaciones científicas "garantizan" el adecuado uso de los métodos estadísticos por parte de los autores, y los lectores, sin preocuparse por este aspecto de la investigación, suponen que los resultados son verídicos.

El afán desmedido por usar métodos estadísticos para el análisis de datos ha llevado a percibir que la estadística es una "herramienta", o que los estadísticos son técnicos a quienes recurrir para solucionar problemas que se aprecian simplemente como "operativos" (aplicar una determinada prueba estadística, sacar el valor-p, calcular la potencia de la prueba, correr el modelo, cuántas observaciones tomar, etcétera), la mayor parte de las veces sin haberlos vinculado al proceso de formulación, diseño o discusión del estudio. Esto ha llevado a un uso de la estadística sin suficiente reconocimiento de su estatus como disciplina científica independiente, lo que ha devenido en una instrumentalización simple y superficial, que impide el aprovechamiento completo de sus potencialidades, pero que además puede afectar el rigor y validez de los estudios (Fernández & Belem, 2016).

Sin embargo, lo cierto es que los consumidores de las publicaciones científicas deben estar en capacidad de valorar los métodos estadísticos aplicados, con el objetivo de evaluar la contundencia de los argumentos a favor o en contra de las conclusiones de la investigación. Esta necesidad puede desaminar a estudiantes, profesionales o investigadores que sin dedicarse a la estadística tienen que hacerle frente. Buena parte de los desaciertos que se encuentran en las publicaciones científicas son errores de diseño fundados, en gran medida, en el desconocimiento de los principios básicos de la estadística. Hay que comentar que la cantidad de errores que se cometen no es despreciable, en parte debido a que cuando se requiere la aplicación de la estadística en las publicaciones, están en manos de los investigadores la sensatez y el cuidado al usar los métodos estadísticos.

La tarea para los consumidores y productores de las publicaciones científicas es adentrarse en el estudio detallado de los principios básicos de la estadística para que estén en capacidad de comprender, criticar y valorar los métodos estadísticos usados en las publicaciones (Guttman, 1979). Afortunadamente las ideas básicas para comprender el contexto estadístico en las publicaciones no son complicadas y generalmente acuden a la lógica y al sentido común.

¿Qué es estadística?

Una idea frecuente es creer que la estadística se reduce a la presentación de cifras, tal como la utilizan los presentadores de las noticias deportivas al referirse, por ejemplo, a los resultados de los partidos de fútbol, posiciones de los equipos, número de partidos jugados, ganados o perdidos, empatados, y así sucesivamente. En otras situaciones, también son presentados en los medios de comunicación datos relativos al número de accidentes automovilísticos que ocurren en un determinado lapso, muertes violentas, opinión de electores, cifras de desempleo, índices de inflación y de precios, etc. Esta es la faceta más notoria de la estadística, pero tiene un espectro más amplio tanto a nivel teórico como en las aplicaciones.

Un aspecto básico para comenzar el estudio y la comprensión de la estadística es acercarse a una delimitación de sus objetivos. Entre las primeras impresiones, se concibe la estadística como un conjunto de métodos para recopilar, analizar, interpretar y presentar grandes volúmenes de datos; el énfasis principal de esta visión es mostrar su aspecto procedimental y revelar la relación estadística-datos, aunque se reconoce que los datos son parte importante de esta disciplina. Otros aspectos y preguntas que tienen mayor relevancia para señalar el verdadero objetivo de la estadística son:

- ¿Para qué se han recolectado los datos? La intención final para disponer de registros (mediciones) debe evidenciar la concordancia con los objetivos de la investigación.
- ¿Fue adecuado el método de recolección de los datos? Son diversas las formas en que puede seleccionarse los datos: encuestas estructuradas, entrevistas, registros administrativos y registros demográficos, entre otros; cada uno de estos tiene ventajas y limitaciones que deben preverse para evitar sesgos en los registros obtenidos.
- ¿Está sustentada la cantidad de datos recolectados? Cuando se recolecta información se cuestiona insistentemente a los investigadores sobre el tamaño de muestra y el diseño muestral, cuyos detalles deben formar parte del protocolo de la investigación.
- ¿Se verificó el cumplimiento de los requisitos de aplicación de los métodos estadísticos utilizados? Todo método estadístico se sustenta en desarrollos teóricos con supuestos distribucionales. Lamentablemente, desde cierto sector despreocupado por la base estadística y matemática, se cree que la verificación de dichos supuestos, como parte del análisis estadístico, es engorrosa y un capricho teórico, sin reconocer que estos constituyen las condiciones bajo las cuales las conclusiones son justificables y válidas.

- ¿Concuerdan las conclusiones de la investigación con la realidad? Esto pone de manifiesto ciertos errores en que se puede incurrir cuando se trabaja con la información proveniente de una muestra, particularmente en las pruebas de hipótesis.

Si el investigador no puede responder adecuadamente estas preguntas de manera técnica y científica, entonces la validez de los resultados queda en entredicho. Lo anterior evidencia que el objetivo de la estadística está lejos de ser la obtención de datos; se debe pensar con anterioridad en puntos como los mencionados y disponer de un detallado plan de investigación antes de proceder a la recolección de los datos, y no esperar enfrentar dudas acerca de la capacidad que tiene la información disponible para responder los objetivos de la investigación. En casos más extremos, darse a la recolección de datos y posteriormente ajustar algunos objetivos con el ánimo de conducir una investigación.

Si bien la estadística se preocupa por la consecución de datos, tiene en principio que ver con el diseño de investigaciones y con la inferencia. Para la primera hay que entender que se preocupa por las etapas que han de seguirse en una investigación y la manera en que se llevan a cabo según el método científico. La inferencia atiende a la manera en que los datos conforman una evidencia sólida para llegar a una conclusión valida.

Uno de los puntos neurálgicos radica en que los datos provienen de un conjunto mayor llamado población, que en la mayoría de los casos no es posible estudiar en su totalidad, es decir, llevar a cabo el CENSO. La estadística proporciona a los investigadores una alternativa al censo, al indagar solo una fracción de la población con el ánimo de emitir conclusiones a nivel general. Lo anterior resulta cómodo y atractivo para los usuarios, pero es necesario planear cuidadosamente ciertas etapas para disfrutar de un conjunto de resultados que sean relevantes y poder extrapolar las conclusiones a la población.

La estadística y el pensamiento estadístico son factores principales en la vida cotidiana y en muchas de las ocupaciones. Casi cada profesión tiene que ver con datos (cuantitativos y cualitativos), y por consiguiente, una necesidad de investigadores y ciudadanos instruidos estadísticamente que puedan contribuir a generar un razonamiento crítico, fundamental para una sociedad bien informada.

Software estadístico R

Otro punto importante que ha hecho posible la incursión de los investigadores en el contexto estadístico es el acceso cada vez más fácil a programas especializados en el procesamiento de información promovido en parte por la llamada "piratería de *software*". Entre los programas estadísticos más conocidos se pueden mencionar SPSS, SAS, STATISTICA, S-Plus, STATA. Existe también una explosión de textos en donde se expone someramente la manera de aplicar un método estadístico (Regresión, Anova, Estadística no paramétrica, por mencionar algunos) con un detrimento de los conceptos involucrados.

Por otra parte, hay que ser precavido pues el solo acceso a un *software* estadístico no es garantía de una adecuada aplicación; el buen uso que se puede dar a la estadística depende en mayor proporción del nivel de dominio conceptual que tenga quien la aplica, más que del *software* usado. No obstante, el uso de *software* puede colaborar para adentrarse en el fascinante mundo de la estadística, y para facilitar aún más dicho proceso este libro usa y promueve el uso del *software* libre, siendo R el caso más sobresaliente (R Core Team, 2016).

El *software* R es uno de los más flexibles, potentes y profesionales que existen actualmente para realizar tareas estadísticas, desde las más elementales hasta las más avanzadas. Está desarrollado y soportado por una comunidad académica a nivel mundial, cuenta además con la ventaja de ser gratuito y su descarga e instalación son sencillas; consulte el sitio http://cran.r-project.org/ para este efecto.

Para dimensionar las posibilidades y características de este *software* se recomienda la lectura del artículo "Por qué comprar un programa estadístico si existe R" (Salas, 2008), donde se compara R con SAS y SPSS, dos de los programas estadísticos licenciados más usados en docencia e investigación. (Disponible en línea: http://www.scielo.org.ar/pdf/ecoaus/v18n2/v18n2a07.pdf)

Con el objetivo de sustentar la versatilidad de R para efectuar análisis estadísticos se mencionan algunas de las ventajas que tiene este *software*, esperando animar a los estudiantes a iniciarse en el manejo de esta potente herramienta:

1. R es gratuito. No necesita hacerse a versiones "piratas" o comprar costosas licencias que deben renovarse periódicamente, claro está, después de pagar.
2. R tiene el sustento de toda una comunidad académica mundial. Además se dispone de una excelente documentación y apoyo en línea.
3. R es empleado por investigadores de múltiples áreas del conocimiento. Esto hace posible conocer diversas facetas de aplicación.

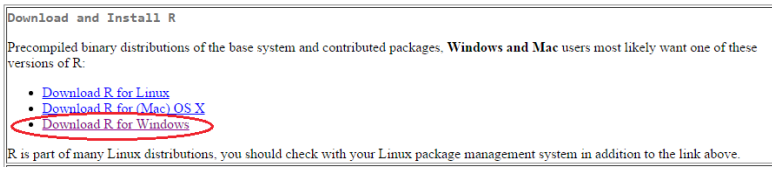
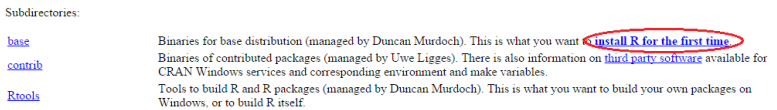
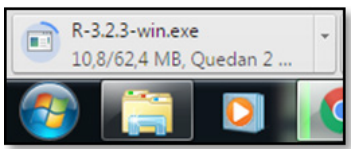
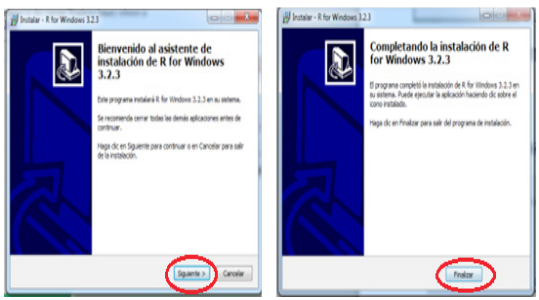
4. R se está mejorando cada día. Continuamente aparecen nuevos paquetes gratuitos que expanden la capacidad de R para solucionar diferentes problemas.
5. Emplea una interfaz de línea de comando (*command-line*) que permite aprender mientras se hacen los cálculos. Los paquetes que se manipulan por ventanas y con clics se asemejan a una caja negra; si el usuario es neófito, nunca sabrá qué hizo el *software*.
6. R es uno de los paquetes estadísticos de mayor crecimiento respecto de su uso en diferentes disciplinas.
7. El lenguaje de programación de R es intuitivo.
8. R crea gráficos de gran calidad y con la posibilidad de adecuarlos a las necesidades de los investigadores.
9. R y LaTex (*software* diseñado para generar documentos científicos) trabajan de manera integrada. Esto hace posible componer documentos técnicos y científicos sin problemas de compatibilidad.
10. R es multiplataforma. Funciona en Mac, Windows o Linux.
11. R hace pensar al usuario en los fundamentos de la estadística.

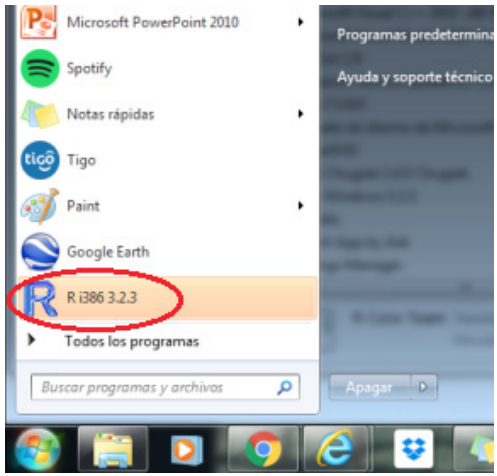
Este libro utiliza R para las diversas aplicaciones; además se provee un número importante de ejemplos que incluyen el código utilizado y cuyas instrucciones pueden modificarse para solucionar los ejercicios propuestos y demás aplicaciones posteriores. Para iniciarse con el *software* se recomienda la lectura de los siguientes documentos (disponibles en línea, ver las referencias bibliográficas), en donde se proporcionan los elementos básicos para trabajar con R en el ámbito de la estadística.

- R para principiantes (Paradis, 2002).
- Introducción a R (R Development Core Team, 2000).

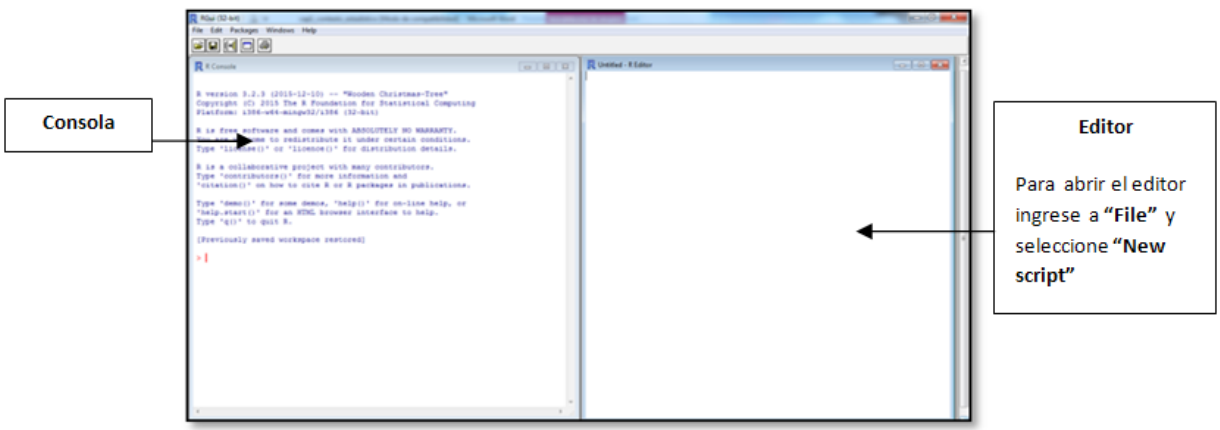
Instalación del *software* R

La descarga e instalación de R es sencilla, a continuación se indican los pasos esenciales para una adecuada instalación de R en Windows (para otros sistemas operativos es similar):

1. Ingrese a la página http://cran.r-project.org/ Seleccione **"Download R for Windows".**	
2. Seleccione **"install R for the first time".**	
3. Espere la descarga, ejecute y seleccione el idioma.	
4. En las ventanas consecutivas, seleccione. **"Siguiente".** Indique **"Finalizar"** para completar la instalación.	

5. Busque R en el menú Inicio para tener acceso al programa.	

Al ingresar a R encontrará una **consola** en la cual serán impresos los resultados de las operaciones y/o instrucciones asignadas desde el **editor.**



Introducción a R

R es un lenguaje interpretado, es decir, que el código no necesita ser preprocesado mediante un compilador; eso significa que el computador es capaz de ejecutar la sucesión de instrucciones dadas por el programador sin necesidad de leer y traducir exhaustivamente todo el código; por otra parte, se tiene que R es un lenguaje que diferencia las letras minúsculas y mayúsculas (*case sensitive*).

R es un lenguaje de formato libre, es decir, que se admiten espacios, tabuladores y comentarios en cualquier parte del código. Se puede entrar un comando a la vez en la línea que identifica el símbolo del sistema o *command prompt* (>) o correr un conjunto de comandos desde un archivo fuente. Las sentencias finalizan en punto y coma o en salto de línea. Cuando se quiere poner más de una sentencia en una línea, es necesario poner un punto y coma (;) para separarlas.

La mayor parte de la funcionalidad en R se proporciona a través de funciones integradas y creadas por el usuario, así como por la manipulación de objetos. Un objeto es básicamente cualquier cosa (datos, variables, cadena de caracteres, funciones, gráficos, resultados analíticos, etc.) a la cual se le puede asignar un valor; cada objeto tiene un atributo de clase que dice a R cómo manejarlo. Es el uso de objetos como entidad básica una diferencia fundamental de la filosofía de R, con el resto del *software* estadístico.

Cualquier expresión evaluada por R se realiza en una serie de pasos, con unos resultados intermedios que se van almacenando en objetos para ser observados o analizados posteriormente, de tal manera que se puede hacer un análisis sin necesidad de mostrar su resultado inmediatamente produciendo unas salidas mínimas. Cada objeto pertenece a una clase, de modo que las funciones pueden tener comportamientos diferentes según sea la clase a la que pertenece su objeto argumento; por ejemplo, no se comporta igual una función cuando su argumento es un vector que cuando es un fichero de datos u otra función.

Todos los objetos de datos se mantienen en la memoria durante una sesión interactiva. Las funciones básicas están disponibles de forma predeterminada. Otras funciones están contenidas en los paquetes que se pueden adjuntar a una sesión actual, según sea necesario; el tema de los paquetes se tratará más adelante en esta sección.

Las declaraciones o sentencias consisten en funciones y asignaciones. R utiliza el símbolo <– para asignaciones, en vez del típico signo de igual (=). Por ejemplo, escriba en el editor las cuatro líneas siguientes:

```
a <- 300
b <- 500
c <- a+b
c
```

Seleccione (sombree) las cuatro líneas de código anteriores, luego presione "**Ctrl**" + "**r**" (o en su defecto, **F5**) para ejecutar los comandos elegidos en el ambiente Windows, mientras que para Mac se ejecuta con "**command**"+"**enter**". Estas instrucciones crean un objeto denominado `a` al cual se asigna el valor `300;` de manera similar se crea un objeto `b` con asignación de `500`, luego se crea un objeto `c` que contiene la suma de `a` y `b`, finalmente se solicita al *software* mostrar el valor consignado en el objeto `c`. Se observará el resultado en la consola.

```
> a <- 300
> b <- 500
> c <- a+b
> c
[1] 800
```

Con la sentencia `x<-rnorm(5);x` se crea un vector `x` que contiene cinco valores provenientes de una distribución normal estándar y se visualiza los datos que se guardan en el vector `x`. Otra forma en que se muestran directamente los resultados guardados en el objeto `x` es escribir la sentencia entre paréntesis, así: `(x<-rnorm(5))`. Para este ejemplo se obtienen los siguientes resultados, pero note que los datos generados son diferentes pues los cinco valores son aleatorios.

```
> x<-rnorm(5);x
[1] 0.1850220 0.6883421 0.3681860 0.2622298 1.2763340
> (x<-rnorm(5))
[1] -0.6414067 -0.9604300  0.2379763 -1.0163363 -0.4056877
```

Aunque el *software* permite que el signo = sea usado para las asignaciones de objetos, no es usual escribir los programas de esa manera, ya que no es la sintaxis estándar; hay algunas situaciones en que no funcionará, y si usted decide usarla puede causar una mala impresión, a tal punto de ser blanco de críticas por parte de los programadores ya diestros en R. También puede invertir la dirección de asignación, `rnorm(5)-> x` es equivalente a la sentencia anterior. Una vez más, hacerlo es poco común y no se recomienda.

Ahora piense que estamos interesados en estudiar la relación entre la edad y el peso en la etapa de la infancia a la adolescencia. Los datos están dados en la Tabla 1.1. Se dispone de la información de nueve personas.

Tabla 1.1 Datos antropométricos en jóvenes.

Edad (años)	Peso (kg)	Edad (años)	Peso (kg)
3	14	3	12
5	17	6	20
6	19	8	26
8	23	14	40
10	32		

Los registros de edad y peso se introducen como vectores usando la función `c()`, la cual combina sus argumentos en vectores o listas; el promedio se obtiene con la función `mean()`, la desviación estándar con `sd()`, la correlación entre la edad y el peso se calcula usando `cor()`. Finalmente se grafica el peso contra la edad mediante la función `plot()`. Las funciones en R, por ejemplo `mean()`, `cor()` y `plot()` se denotan de manera similar a como se hace en matemáticas, con el nombre de la función y el argumento entre paréntesis.

Una práctica muy útil consiste en escribir las instrucciones en el editor y salvar este archivo (*script*) con un nombre corto pero descriptivo y con extensión .R; por ejemplo, los vectores que guardan los datos de Edad y Peso y las funciones que calculan el promedio de la variable Peso, la desviación estándar de la Edad, la correlación entre Edad y Peso y el gráfico de estas variables se puede guardar en un *script* de nombre `relación.R`, y se puede ejecutar en cualquier momento o si lo requiere se pueden copiar algunas instrucciones que se puedan necesitar para ejecutar cálculos similares pero sobre otras variables. A continuación se presentan los resultados que se obtienen al ejecutar las instrucciones que se consignaron en el *script*.

```
> Edad <- c(3,5,6,8,10,3,6,8,14)
> Peso <- c(14,17,19,23,32,12,20,26,40)
> mean(Peso)
[1] 22.55556
> sd(Edad)
[1] 3.5
> cor(Edad, Peso)
[1] 0.9900868
> plot(Edad, Peso)
```

De los resultados anteriores se tiene que el peso promedio es aproximadamente 22.56 kilogramos, la desviación estándar de la edad es 3.5 años y se evidencia una relación directa y fuerte entre la edad y el peso, dado que el coeficiente de correlación es positivo y cercano a uno (correlación = 0.9900868). El diagrama de dispersión se muestra en la Figura 1.1. y dada la relación directa entre las variables, se aprecia gráficamente cómo a medida que aumenta la edad también lo hace el peso.

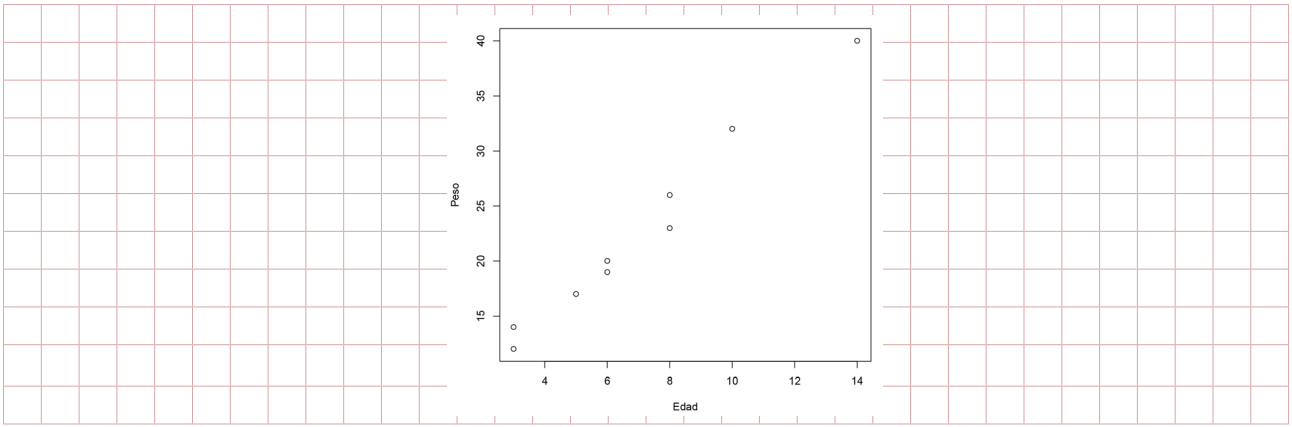


Figura 1.1 Diagrama de dispersión para peso y edad.

Cuando se instala el *software* se incluyen una serie de bases de datos obtenidas por investigadores en distintos contextos y que han sido liberadas con propósitos fundamentalmente académicos y conforman una buena fuente de información para aplicaciones estadísticas. Para obtener una lista de los datos disponibles, ejecute la sentencia `data()`; en el recuadro se presenta una fracción de las mencionadas bases de datos que se visualizan según orden alfabético.

```
Data sets in package 'datasets':

AirPassengers            Monthly Airline Passenger Numbers 1949-1960
BJsales                  Sales Data with Leading Indicator
BJsales.lead (BJsales)
                         Sales Data with Leading Indicator
BOD                      Biochemical Oxygen Demand
CO2                      Carbon Dioxide Uptake in Grass Plants
ChickWeight              Weight versus age of chicks on different diets
DNase                    Elisa assay of DNase
EuStockMarkets           Daily Closing Prices of Major European Stock
                         Indices, 1991-1998
Formaldehyde             Determination of Formaldehyde
```

Para cargar una de estas bases de datos, por ejemplo `BOD`, se debe suministrar dicho nombre como argumento de la función. Si desea visualizar la información contenida en este objeto, simplemente digite el nombre de la base de datos.

```
> data(BOD)
> BOD
  Time demand
1    1    8.3
2    2   10.3
3    3   19.0
4    4   16.0
5    5   15.6
6    7   19.8
```

Se puede acceder a una descripción general de la base de datos usando la función `help(BOD)`. Esta ayuda se ejecuta en la web, por lo cual conviene estar conectado; además `help()` también se puede aplicar para obtener información sobre el manejo de las funciones en R y solo basta poner entre paréntesis el nombre de la respectiva función. La ayuda para la base de datos `BOD` es la siguiente:

```
BOD {datasets}                                          R Documentation

                    Biochemical Oxygen Demand

Description

The BOD data frame has 6 rows and 2 columns giving the biochemical
oxygen demand versus time in an evaluation of water quality.

Usage
BOD

Format

This data frame contains the following columns:
Time
       A numeric vector giving the time of the measurement (days).
demand
       A numeric vector giving the biochemical oxygen demand (mg/l).
```

Source

Bates, D.M., and Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, Appendix A1.4.

Originally from Marske (1967), Biochemical Oxygen Demand Data Interpretation Using Sum of Squares Surface. M.Sc. Thesis, University of Wisconsin – Madison.

Otro aspecto importante que vale la pena mencionar en esta breve introducción es acerca de los denominados paquetes (*package*) en R, los cuales ayudan a obtener mejor provecho del uso del *software*. La instalación básica de R viene equipada con múltiples funciones que permiten efectuar procedimientos como: importar y manejar bases de datos, realizar transformaciones de datos, ajustar y evaluar modelos estadísticos, manejar funciones probabilísticas, diseñar representaciones gráficas, entre otras. Sin embargo, la enorme potencia de R deriva de su capacidad de incorporar en cualquier momento nuevas funciones capaces de realizar procedimientos más sofisticados y completos.

Un paquete es una colección de funciones, datos, código R y documentación que se almacenan en una carpeta conforme a una estructura bien definida y fácilmente accesible para R. En la página web del *software* (indagar la dirección: https://cran.r-project.org/web/packages/) se puede consultar la lista de paquetes disponibles. A mediados de 2018 esta lista incluía algo más de 12.500 paquetes. Al instalar el *software* se incorporan por defecto numerosos paquetes y el usuario puede potenciarlo con algunos otros, según la necesidad; para acceder a la lista de aquellos que actualmente tiene instalados en su computador, use la función `library()`.

La manera más rápida para disponer de un paquete es instalarlo mediante internet. A manera de ejemplo, para instalar el paquete `car` (Companion to Applied Regression) se escribe en la consola o en un *script* la instrucción `install.packages("car")` y al ejecutarla aparece el mensaje: *Please select a CRAN mirror for use in this sesion*, que solicita elegir un servidor (repositorio) de una lista desplegable y de donde se instalará el paquete. Existen muchas universidades y otras instituciones que ofrecen ser repositorios de R pero se puede elegir para este propósito cualquier sitio.

Una vez instalado el paquete (¡proceso que solo se realiza una vez!), el *software* está en capacidad de usar las funciones que contiene, siempre y cuando el usuario solicite tenerlo a disposición en la sesión de trabajo; esta acción se denomina cargar el paquete y se usa la función `library()` con argumento el nombre del paquete, así: `library(car)`.

Es muy importante entonces distinguir entre un paquete instalado en el computador y un paquete cargado en memoria:

- Tener instalado un paquete significa que en algún momento el usuario lo dispuso en su computador a través de internet y fue copiado en algún directorio en donde R lo puede localizar.
- Cargar en memoria significa que durante nuestra sesión de trabajo R se ha leído el contenido del paquete e incorporado las funciones que contiene a su espacio de trabajo mediante la función `library()`; en dicho caso las funciones contenidas en el paquete pueden ya ser invocadas y ejecutadas.

Tener a disposición un recurso como este *software* implica gran responsabilidad en el momento del uso para lograr aplicaciones coherentes o para apoyar el proceso de aprendizaje de la estadística; por tanto, es necesario reconocer algunos elementos que pueden guiar el uso de R para resolver problemas de tipo estadístico; estos son:

1. Se requiere en principio conocer los fundamentos básicos del área estadística que se desea usar. Sin conocimiento previo no se pude pensar en usar R de manera apropiada.
2. El *software* requiere tener algunos conceptos básicos de programación. En los últimos años la programación se ha convertido en un aspecto tan importante en la educación como la lectura, la escritura o el aprendizaje de un nuevo idioma.
3. Es indispensable contar con un manual sobre R, no para leerlo completamente, sino para consultarlo en caso de ser necesario. Hay una extensa bibliografía de libros, documentos y manuales que vinculan el uso del *software* y los métodos estadísticos.

Aunque hay diversas formas de empezar a conocer este programa, bien sea en internet, como a través de manuales, blogs especializados o vídeos, se puede considerar una práctica provechosa examinar códigos de otras personas para comprender su funcionamiento y ajustarlos a la situación de interés. También es una buena manera tratar de comprender cómo actúan las diferentes funciones.

Para finalizar esta breve introducción se explica cómo se puede actualizar R, proceso fácil en particular para Windows; con otros sistemas operativos se debe recurrir al sitio web de CRAN -Comprehensive R Archive Network (https://www.r-project.org/) para instalar la versión más nueva.

Para la actualización se usan los siguientes comandos:

```
install.packages("installr", dependencies = TRUE)
library(installr)
updateR()
```

Se despliega un cuadro de diálogo que lo guía por los siguientes pasos:

- Comprueba una versión más nueva de R.
- Si existe, la función descargará la versión R más actualizada y ejecutará su instalador.
- Una vez hecho esto, la función ofrecerá copiar (o mover) todos los paquetes de la antigua biblioteca R a la nueva biblioteca R.
- A continuación ofrecerá actualizar los paquetes movidos.

Ejercicios

Adelante las actividades propuestas con base en las lecturas recomendadas (Introducción a R: capítulos 1, 2, 5, 6, 7 y R para principiantes: hasta el capítulo 4).

1. Efectúe las asignaciones siguientes en R: $x=5$, $y=0$, $z=0.0005$, $w=0$, $Z=0.00005$. Posteriormente use el *software* para efectuar las siguientes operaciones. Comente sobre el resultado.
   a. w/x
   b. x/z
   c. x/Z
   d. x/y
   e. y/w
   f. $e^{-\frac{x}{y}}$
   g. $\sqrt{\frac{Z}{z}}$

2. Asigne el siguiente vector: $A=(21,35,28,63,9,89,54,19,26,56,54,22,49)$ e indique cuál instrucción debe usar para efectuar las siguientes operaciones sobre A.
   a. Suma de los elementos de A
   b. Mínimo de A
   c. Máximo de A
   d. Número de elementos de A
   e. Elementos ordenados del vector.
   f. Promedio de los elementos de A
   g. Promedio de $6 \cdot A$

3. Usar el código que aparece en el recuadro, para hacer la gráfica para las horas trabajadas en un taller (X), y el número de unidades producidas (Y).

X	80	79	83	84	78	60	82	85	79	84	80	62
Y	300	302	315	330	300	250	300	340	315	330	310	240

```
x <- c(80,79,83,84,78,60,82,85,79,84,80,62)
y <- c(300,302,315,330,300,250,300,340,315,330,310,240)
plot(x,y)
```

a. Para insertar el título apropiado al gráfico, modifique la tercera línea del código y coloque `plot(x, y, main="Diagrama de Dispersión")`. Tenga en cuenta usar una coma para separar las opciones.
b. Coloque el título HORAS TRABAJADAS al eje X usando la opción `xlab="titulo eje x"` y dentro de las comillas el rótulo respectivo.
c. Coloque el título PRODUCCIÓN al eje Y usando la opción `ylab="titulo eje y"` y el rótulo del eje entre comillas.
d. Cambie el color a los puntos que por defecto es negro. Use `col = 1`, o también `col="black"`
e. Cambie el tamaño a los puntos mediante `cex=1`. Si usa número inferior a uno se reduce el tamaño de los puntos.
f. Modifique el símbolo usado para los puntos y ponga un asterisco. La opción por defecto es `pch=1` que usa para los puntos el círculo; a manera de ejemplo, con `pch=2` identifica las coordenadas en la gráfica con un triángulo.
g. Analizando el contexto, ¿qué tipo de relación (directa o inversa) deberían presentar estas variables?
h. Halle el coeficiente de correlación entre las variables. ¿El signo del coeficiente corresponde a lo estipulado en el numeral anterior?
i. Llamando a una relación como fuerte y directa cuando el coeficiente está en el intervalo [0.9 ; 1.0] y si dicho coeficiente cae dentro del intervalo [-0.9 ; -1.0], la relación es fuerte pero inversa. De acuerdo con lo anterior, ¿cómo catalogaría la relación entre las horas trabajadas y la producción?

4. Construya la siguiente base de datos en Excel y luego guarde el archivo como tipo CSV (Comma Separated Values) con el siguiente nombre "indicadores.csv".

ciudad	población	menores5	desempleo2013	desempleo2014
A	125000	31250	8.6	9.1
B	230000	46000	10.6	10.4
C	150000	22500	9.3	9.5
D	80000	14400	8.9	8.2

La definición de las variables incluidas en esta base de datos es:

población: cantidad de habitantes de la ciudad.

menores5: población de personas menores de cinco años en la ciudad.

desempleo2013: tasa de desempleo (%) para el año 2013

desempleo2014: tasa de desempleo (%) para el año 2014

Para leer los datos desde un archivo, debe indicar a R en cuál directorio se encuentran los datos. Para eso acceda al menú *Archivo* y después a la opción *Cambiar dir…* y buscar el directorio en donde guardó sus datos y señalarlo. Luego use la siguiente instrucción para leer esta base de datos,

```
x <- read.table("indicadores.csv", header = T, sep = ";", dec=".")
```

Otro método consiste en leer datos de un archivo que usted pueda elegir, escribiendo:

```
x <- read.table(file.choose(), header = T, sep = ";", dec=".")
```

Sin necesidad de cambiar el directorio, se puede leer el archivo si previamente le indicamos al *software* la ruta de acceso usando la funcion `setwd()`, tal como se ejemplifica a continuación:

```
setwd("C:/Users/manuel.contento/Dropbox/Curso_ESTADISTICA/BasesDatos")
x <- read.table("indicadores.csv", header = T, sep = ";", dec=".")
```

Si lee el archivo de esta última manera, tenga presente que la ruta de acceso depende de la manera en que cada usuario tiene configurado su computador.

Usando cualquiera de las opciones anteriores, el *software* crea un objeto "x" que contiene la base de datos. Si quiere visualizarlo, simplemente digite el nombre del objeto en la siguiente línea y ejecute; el *software* muestra la base de datos.

```
> x
  ciudad población menores5 desempleo2013 desempleo2014
1      A    125000    31250           8.6           9.1
2      B    230000    46000          10.6          10.4
3      C    150000    22500           9.3           9.5
4      D     80000    14400           8.9           8.2
```

Calcule la proporción de menores de cinco años usando

```
tasamenores5 <- x$menores5/x$población
```

Se puede modificar la base de batos adicionando esta columna; tenga la precaución de usar un nuevo nombre para el conjunto de datos con dicha modificación, esto se logra con la instrucción siguiente.

```
xnuevo <- data.frame(x, tasamenores5)
```

La nueva base de datos, con la proporción de menores de cinco años, queda así:

```
> xnuevo
  ciudad población menores5 desempleo2013 desempleo2014 tasamenores5
1      A    125000    31250           8.6           9.1         0.25
2      B    230000    46000          10.6          10.4         0.20
3      C    150000    22500           9.3           9.5         0.15
4      D     80000    14400           8.9           8.2         0.18
```

Genere un diagrama circular (pie) usando el código adjunto:

```
pie(xnuevo$población,labels=xnuevo$ciudad, main="Distribución
porcentual de la población por ciudad", col = c("purple4",
"violetred1", "green3", "cornsilk2"))
```

a. Analice el gráfico que aparece. ¿Qué utilidad tiene la información que se presenta?

Compile el siguiente conjunto de instrucciones y obtenga el diagrama de barras ordenado.

```
#Ordena base datos descendente según tasamenores5 (- indica descendente)
xnuevo1<-xnuevo[order(-xnuevo$tasamenores5),]
barplot(xnuevo1$tasamenores5, main="Proporción de población menor
de 5 años por ciudad", xlab ="Ciudad", names.arg=xnuevo1$ciudad,
ylim=c(0,0.25),col = rainbow(4),cex.names=0.9)
```

b. Describa la proporción de población inferior a cinco años. ¿Qué deduce de la información presentada?

Ahora ejecute las instrucciones siguientes.

```
desempleo<-rbind(xnuevo$desempleo2013,xnuevo$desempleo2014)
barplot(desempleo, main="Tasa de desempleo 2013-2014 por ciudad",
xlab="Ciudad",
names.arg=xnuevo$ciudad, ylim=c(0,15), col=c("lightcyan","lavender"),
cex.names=0.9, beside = T, legend.text = c("2013", "2014"))
```

c. A partir del gráfico obtenido, interprete el comportamiento y evolución de la tasa de desempleo para 2013-2014 en estas cuatro ciudades.

5. La información de la tabla corresponde a los indicadores sobre pobreza y desigualdad de las urbes que hacen parte de la *Red Colombiana de Ciudades Cómo Vamos*, y que ha sido extractados del Informe de Calidad de Vida (http://redcomovamos.org/wp-content/uploads/2015/03/Pobreza_ICV7.pdf). Se incluyen las siguientes variables:

Poblacion2013: cantidad de habitantes para el año 2013 en las áreas metropolitanas así: Barranquilla incluye Soledad, Bucaramanga incluye Floridablanca, Girón y Piedecuesta, Manizales incluye Villamaría, Medellín incluye Valle de Aburrá, Cali incluye Yumbo, Pereira incluye Dosquebradas y La Virginia.

IPM: Incidencia de Pobreza Monetaria. Es el porcentaje de la población que está bajo la línea de pobreza. Se considera pobres aquellos hogares que obtengan ingresos inferiores al valor mensual de una canasta de alimentos y otros bienes básicos como vivienda, educación, salud, transporte y esparcimiento. De acuerdo con las estimaciones del DANE, durante el año 2013 en las trece principales ciudades de Colombia, satisfacer todo el conjunto de necesidades básicas (alimenticias y no alimenticias), costaba $227.118 por persona al mes, lo que corresponde al valor de la línea de pobreza en este conjunto de ciudades. Se provee la incidencia de pobreza monetaria para 2011, 2012 y 2013 denotadas IPM2011, IPM2012 e IPM2013, respectivamente.

TTI: Tasa de Trabajo Infantil. Proporción de menores que participan del mercado laboral, calculado por el DANE para la población entre 5 y 17 años de edad.

TTI_A: Tasa de Trabajo Infantil Ampliada. Proporción de menores que trabajan, agregando aquellos que dedican más de 15 horas a las labores de cuidado (oficios domésticos). Calculado por el DANE para la población entre 5 y 17 años de edad.

PPE: Población en Pobreza Extrema. De acuerdo con las estimaciones del DANE, durante el año 2013 en las trece principales ciudades de Colombia, satisfacer las necesidades alimenticias básicas de una persona costaba $96.422, lo que corresponde al valor de la línea de pobreza monetaria extrema.

Ciudad	Poblacion2013	IPM2011	IPM2012	IPM2013	TTI_A2012	TTI2012	PPE2013
Bogotá	7674366	13.1	11.6	10.2	12	8	122790
Medellín	3685382	19.2	17.7	16.1	11	7	110561
Cali	2431437	25.1	23.1	21.9	10	6	106983
Barranquilla	1822438	34.7	30.4	29.1	7	3	76542
Bucaramanga	1103989	10.7	10.4	10.3	12	9	13248
Cartagena	978600	33.4	32.7	29.2	13	3	56759
Ibagué	542876	22.0	21.3	18.6	18	10	13572
Valledupar	433242	36.0	32.8	31.4	12	6	19063
Manizales	447344	19.2	17.6	16.2	8	2	11631
Pereira	226862	21.6	21.9	24.0	10	6	11570

Construya la base de datos en Excel y guarde el archivo como tipo CSV. Luego cargue en R los datos desde el archivo de tipo CSV que creó en Excel. Visualice el archivo en R para cerciorarse de que la información que contiene corresponda fielmente a los datos que se proporcionan.

a. Calcule la proporción de personas en pobreza extrema como tasaPPE2013=PPE2013 / Poblacion2013 y luego modifique la base de batos adicionando esta columna.
b. Genere un diagrama circular (pie) para la Distribución porcentual de la población por ciudad y analice la información a partir de la gráfica que obtiene.
c. Diseñe un diagrama de barras para mostrar en forma descendente la proporción de personas que están en pobreza extrema para el año 2013. Describa los resultados con base en el diagrama.
d. Elabore un gráfico que muestre el comportamiento y evolución de la incidencia de pobreza monetaria de 2011 a 2013 en estas ciudades. Describa los resultados a partir del gráfico elaborado.

6. El gerente de mercadeo de un banco quiere desarrollar un estudio orientado a conocer mejor a sus clientes, ampliando información relacionada con sus hábitos de consumo de medios con el fin de mejorar la efectividad de sus acciones de marketing. En el cuestionario, además de la información demográfica, se preguntó a los entrevistados en qué medios han visto/obtenido información del producto (cuenta de ahorros, crédito de libre inversión, crédito hipotecario, tarjeta de crédito) que han adquirido con el banco. Se dispone de la información de las siguientes variables:

V0: Identificador
V1: Educación. (1: Primaria incompleta, 2: Primaria completa, 3: Bachillerato incompleto, 4: Bachillerato completo, 5: Técnico, 6: Universitario)
V2: Edad (años cumplidos)
V3: Actividad económica. (1:Ama de Casa, 2:Estudiante, 3:Empleado, 4:Independiente)
V4: Género. (1:Femenino, 2:Masculino)
V5: Estado civil. (1:Soltero, 2:Unión libre, 3:Casado, 4:Divorciado, 5:Viudo)
V6: Número de hijos.
V7: Ingreso (en miles de pesos). Los campos en blanco son de clientes que no informan ingreso
V8: Producto que adquiere el cliente. (1:Cuenta de Ahorros, 2:Crédito Hipotecario, 3:Crédito Libre Inversión, 4: Tarjeta de Crédito)

V9: Medio por el cual se enteró del producto. (1:Cine, 2:Correo Electrónico, 3:Contacto con Asesor, 4:Pagina Web del Banco, 5:Paraderos, 6:Periódico, 7:Publicidad en Buscadores (Google), 8:Publicidad Página Web, 9:Publicidad en redes sociales, 10:Radio, 11:Recomendación Amigo, 12: Recomendación Familiar, 13:TV, 14:Vallas)

Usando la información que se proporciona en la Tabla 1.2., realice las actividades que se describen a continuación.

a. Construya una base de datos en Excel con la información recolectada. Grabe la base como tipo CSV (Comma Separated Values) con un nombre apropiado.
b. Lea en R los datos desde un archivo siguiendo las indicaciones dadas anteriormente. A manera de ejemplo se puede usar el siguiente código para leer el archivo de nombre `DataEjer6Clientes.csv` que en el entorno de R se denominó clientes. La opción `names()` muestra los nombres de las variables. Ejecute esta y las demás instrucciones desde el editor, y guarde todo en un *script* con un nombre apropiado.

```
clientes <- read.table("DataEjer6Clientes.csv", header=TRUE, sep=";",
dec=".")
clientes
names(clientes)
```

Al correr el código anterior, en particular al solicitar los nombre de las variables, notará que están denominadas como V0, V1,…, V9. Se requiere cambiar el nombre de las variables por denominaciones apropiadas y útiles en el momento de hacer futuros procesamientos.

c. Use la siguiente instrucción para denominar ID a la variable V0.

```
colnames(clientes)[ colnames(clientes) == "V0" ] <- "ID"
```

Si vuelve a solicitar los nombres con la instrucción `name(clientes)` se observa el cambio en la designación de esta variable, tal como se evidencia en el resultado que aparece en la consola.

```
[1] "ID" "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8" "V9"
```

Usando como base el código con el que se cambió el nombre a V0 por ID, cambie el nombre de las demás variables a los siguientes:

V1: Educación
V2: Edad
V3: Actividad_Económica
V4: Género
V5: Estado_Civil
V6: Número_Hijos
V7: Ingreso
V8: Producto
V9: Medio

d. En el *software* R, un **factor** representa una variable **cualitativa o categórica**. El factor almacena las categorías en la forma de un vector con valores discretos numéricos (1, 2, 3, 4, etc.) que son los **códigos** de los valores de la variable y otro vector de caracteres interno que contiene las etiquetas de esos códigos. Por ejemplo, la variable educación tiene almacenados números del 1 al 6 que identifican cada nivel de educación. Esta variable es de tipo categórica y además los niveles están ordenados, donde 1 indica el más bajo nivel de educación y 6 el máximo. Para indicar que esta variable es de tal naturaleza se debe convertir a factor, estipular las etiquetas asociadas a cada código y finalmente establecer que los niveles están ordenados.

Puesto que esta base de datos tiene 9 variables, se requiere ser eficiente para hacer mención a ellas; para esto se usa la opción `attach()` que permite hacer referencia a las variables del dataframe `clientes` de manera directa. Si la opción no se usa previamente, es necesario referirse a la variable Educación mediante `clientes$Educación`. El siguiente grupo de instrucciones permite convertir la variable Educación en factor categórico ordinal y definir sus etiquetas.

```
attach(clientes)          #para fijar las variables del data frame
clientes
#Convertir la variable Educación en un factor
Educación <- factor(Educación)
#Asignación de los niveles al factor:
levels(Educación) <- c("Primaria Incompleta", "Primaria Completa",
"Bachillerato Incompleto",
"Bachillerato Completo", "Técnico", "Universitario")
#Indicar que se trata de un factor ordinal:
Educación <- ordered(Educación)
Educación
```

Si la variable es categórica pero los niveles no implican un orden, técnicamente llamada variable nominal, simplemente no se usa la opción `ordered()`. A manera de ejemplo, veamos el caso para la variable Género.

```
#Convertimos la variable Género en un factor
Género <- factor(Género)
#Asignamos los niveles al factor:
levels(Género) <- c("Femenino", "Masculino")
Género
```

Convierta en factor las demás variables categóricas del conjunto de datos, a saber: Actividad_Económica, Estado_Civil, Producto y Medio. Determine si la variable es ordinal o nominal.

e. En el recuadro se proporciona un conjunto de instrucciones que debe ejecutar en R; describa lo que hace la instrucción `table()` y `barplot()`, así como los cambios que percibe en cada una de las gráficas que se obtienen. Indique cuál opción se encarga de la modificación; por ejemplo, la opción `main="Distribución por Género"` asigna un título a la gráfica.

```
table(Género)
#Gráfica 1
barplot(table(Género))
#Gráfica 2
barplot(table(Género), main="Distribución por Género", xlab="Género",
ylab="Frecuencia")
#Gráfica 3
barplot(table(Género), main="Distribución por Género", xlab="Género",
ylab="Frecuencia", names.arg=c("Mujeres","Hombres"))
#Gráfica 4
barplot(table(Género), main="Distribucion por Género", xlab="Género",
ylab="Frecuencia", names.arg=abbreviate(levels(Género)))
```

¿Cuál de las cuatro gráficas considera más conveniente o apropiada para hacer una presentación en público? Explique.

f. Use la opción `barplot()` para obtener la gráfica para la variable Educación, Actividad_Económica, Estado_Civil, Producto y Medio. Decida cuál de las opciones que se presentaron en el anterior numeral conviene usar de manera que la gráfica para cada variable sea simple, clara y funcional. ¿Qué puede deducir de las gráficas para cada variable?

g. Ejecute las instrucciones `summary(Número_Hijos)` y `summary(Ingreso)`. Averigüe qué cálculos proporciona el *software* cuando se usan estas instrucciones.

h. ¿Qué puede deducir de la gráfica que se obtiene mediante `plot(Número_Hijos,Ingreso)`? Explique.

7. Transcriba en un *script* el código que se provee en el recuadro. Ejecute línea a línea las instrucciones y responda las preguntas. Note que con la primera instrucción se instalará el paquete faraway.

```
install.packages("faraway")    #instala el paquete faraway
library(faraway)               #carga el paquete faraway
try(data(package = "faraway")) #lista las BD incluidas en el paquete
"faraway"
help(gala)                     #información de la data gala
data(gala)      #carga la base de datos especificada (gala)
gala            #muestra la base de datos gala
dim(gala)       #indica la dimensión de la base de datos
names(gala)     #Muestra los nombres de las variables
summary(gala)   #Visualiza un resumen de las variables
pairs(gala)     #matriz de diagramas de dispersión
```

a. ¿Cuántas bases de datos contiene el paquete `faraway`?

b. Describa las variables que están en la base de datos de nombre `gala`

c. ¿Qué indica la dimensión de la base de datos?

d. Indague sobre los resultados que provee la función `summary()`

e. ¿Qué tipo de gráfica se obtiene cuando se aplica la función `pairs()` a la base de datos? Averigüe sobre la utilidad de este diagrama y explique

Tabla 1.2 Datos para el ejercicio 6.

V0	V1	V2	V3	V4	V5	V6	V7	V8	V9
1	4	23	3	1	1	1	963	3	5
2	3	49	4	2	3	4	2509	3	8
3	4	50	4	2	3	4	2738	2	10
4	4	27	1	1	2	0	922	1	13
5	3	21	2	1	2	1	894	3	13
6	4	29	4	2	2	1	2092	1	8
7	4	26	3	2	2	1	1350	3	2
8	1	55	4	2	2	0	897	3	13
9	4	18	2	1	1	0	979	3	3
10	5	58	3	1	1	1	1046	3	13
11	5	33	3	2	1	1	2072	3	9
12	4	28	3	2	2	0	1418	1	13
13	5	21	4	2	1	1	915	1	11
14	4	50	3	2	2	1	1737	3	14
15	4	54	4	1	4	4	1073	3	13
16	6	21	3	1	1	0	1328	1	10
17	5	48	3	1	3	3	846	1	10
18	4	46	3	1	1	0	917	3	13
19	3	35	3	1	5	0	1181	1	13
20	6	25	2	2	1	1	1347	3	9
21	6	24	3	2	1	0	1022	3	13
22	4	38	3	2	4	3	988	1	5
23	6	27	3	2	1	1	1827	1	6
24	6	19	2	2	1	1	961	1	13
25	4	20	2	1	1	1	899	1	13
26	6	22	2	1	1	1		3	13
27	5	18	3	1	1	1	1613	3	5
28	6	18	2	1	1	1		1	14
29	5	27	4	1	1	1	1681	3	5
30	6	43	3	1	4	4	2425	4	7
31	5	23	4	2	1	1	1915	1	10
32	6	23	2	2	1	1	885	3	7