

# RStudio para Estadística Descriptiva en Ciencias Sociales

Manual de apoyo docente para la asignatura Estadística Descriptiva. Carrera de Sociología,  
Universidad de Chile (segunda edición)

*Giorgio Boccardo Bosoni y Felipe Ruiz Bruzzone*

*2019-06-10*



# Índice general

<b>Nota técnica</b>	<b>5</b>
<b>1 Presentación: del sentido de este material</b>	<b>7</b>
<b>2 ¿Cómo definir qué y cuantos software de análisis estadístico manejar?</b>	<b>9</b>
2.1 SPSS: uno de los más usado en Ciencias Sociales . . . . .	10
2.2 EXCEL: uno de los olvidados . . . . .	10
2.3 Stata: un programa de nicho . . . . .	11
2.4 Python: un lenguaje de programación de mayor potencia . . . . .	12
2.5 R: el temido . . . . .	13
2.6 Ventajas del uso de R . . . . .	15
2.7 Una mirada comparada: limitantes y potencialidades . . . . .	16
<b>3 Instalación de los softwares a utilizar en este manual</b>	<b>19</b>
3.1 Instalación de R y RStudio en Microsoft Windows . . . . .	20
3.2 Instalación de R y RStudio en Mac OS . . . . .	23
3.3 Instalación de R y RStudio en Linux . . . . .	24
<b>4 Uso básico de RStudio</b>	<b>33</b>
4.1 ¿Qué es RStudio?: una interfaz para usar R . . . . .	33
4.2 Carpeta de trabajo y memoria temporal del programa . . . . .	33
4.3 Elementos fundamentales del uso de sintaxis en RStudio . . . . .	35
4.4 Manejo básico de la sintaxis de R: creación de objetos, inclusión de anotaciones y definición de secciones . . . . .	36
4.5 Tipos de objetos en R (vectores) . . . . .	38
4.6 Construcción de una base de datos . . . . .	39
<b>5 Manejo de la biblioteca y gestión de paquetes</b>	<b>43</b>
5.1 Descargar paquetes . . . . .	43
5.2 Cargar paquetes . . . . .	43
5.3 Actualizar la versión básica de R y los paquetes instalados . . . . .	44

<b>6</b>	<b>Gestión de bases de datos</b>	<b>47</b>
6.1	Descarga de una base de datos de interés . . . . .	47
6.2	¿Cómo abrir bases de datos desde formato SPSS? . . . . .	48
6.3	¿Cómo abrir bases de datos desde diferentes formatos de Microsoft Excel? . . . . .	51
6.4	Construir una base de datos sólo con variables de interés: exploración de bases de datos y recodificación de variables . . . . .	56
<b>7</b>	<b>Estadística descriptiva con RStudio</b>	<b>67</b>
7.1	Estimación puntual de estadísticos descriptivos usando R . . . . .	67
7.2	Inferencia estadística univariada: de la estimación puntual al parámetro . . . . .	79
<b>8</b>	<b>Construcción de gráficos usando RStudio: funcionalidades básicas y uso del paquete <i>ggplot2</i></b>	<b>85</b>
8.1	Funciones básicas para la construcción de gráficos . . . . .	85
8.2	Introducción al uso del paquete <i>ggplot2</i> . . . . .	92
<b>9</b>	<b>Introducción al uso de RMarkdown para la compilación de resultados de RStudio en diferentes formatos</b>	<b>101</b>
9.1	¿Qué es RMarkdown y para qué sirve? . . . . .	101
9.2	Los diferentes elementos de una sintaxis de RMarkdown . . . . .	103
9.3	Aspectos generales (formato de texto) . . . . .	106
9.4	Aspectos generales (configuración de trozos de código) . . . . .	113
9.5	Presentación de resultados básicos en RMarkdown . . . . .	114
<b>10</b>	<b>Materiales de apoyo para el aprendizaje</b>	<b>121</b>
10.1	Instalación de R y RStudio . . . . .	121
10.2	Uso de RMarkdown . . . . .	121
10.3	Uso de Zotero como gestor de referencias bibliográficas . . . . .	121
10.4	LaTeX y RMarkdown . . . . .	121
	<b>Bibliografía</b>	<b>123</b>

# Nota técnica

En su segunda edición este libro fue editado en RStudio mediante RMarkdown y compilado usando el paquete Bookdown. Su ejecución se efectuó en una distribución del sistema operativo Linux de tipo Mint, específicamente en su actualización 19.1 “*Tessa*” y variante **Cinnamon Edition**. Para evitar algunos problemas derivados de la actualización del kernel de Linux se utilizó el software R en su versión 3.4.4, aunque a la fecha de su publicación, R ya había lanzado a su versión 3.6. Para evitar incongruencias entre algunas dependencias de sistema operativo y paquetes adicionales, se decidió utilizar algunos paquetes en versiones anteriores a las actualmente disponibles en el CRAN mediante el paquete *devtools*, lo que se detallan a continuación. Ante dudas u observaciones técnicas, no dude en contactar a los autores.

Paquete	Versión instalada	Última versión
modeest	2.1	2.3.3
summarytools	0.8.7	0.9.3
multcomp	1.4-8	1.4-10



# Capítulo 1

## Presentación: del sentido de este material

Luego de varias ediciones de la asignatura Estadística Descriptiva de la Carrera de Sociología en la Universidad de Chile, hemos considerado necesario ofrecer a los y las estudiantes de la Carrera una sistematización del conocimiento acumulado en el uso de lenguaje de programación R y en estadística descriptiva como herramientas para la investigación sociológica.

En tal perspectiva, el presente documento de apoyo docente tiene un doble propósito.

En primer lugar, convertirse en un material de apoyo para los y las estudiantes del ramo *Estadística Descriptiva*, para facilitar el aprendizaje de herramientas que en un inicio pueden resultar un poco intimidantes y apoyar el desarrollo de sus trabajos y tareas en la asignatura.

En segundo lugar, proporcionar un documento en español a investigadores e investigadoras en ciencias sociales que se interesen en conocer las herramientas básicas de R y RStudio.

En tanto apoyo docente, el presente documento se organiza en una lógica de aprendizaje incremental. Es decir, se enseñarán herramientas de R y RStudio para el análisis de datos sociales mediante diversos ejemplos, comenzando desde los usos más simple hasta cuestiones de mayor complejidad. Con el objetivo de facilitar el aprendizaje, en la siguiente carpeta en línea se encuentra toda la documentación para poder replicar los ejemplos propuestos a lo largo de este documento (sintaxis y bases de datos), así como una plantilla básica para el uso de RMarkdown.

La segunda edición de este manual revisó errores de la primera edición y profundizó contenidos, en particular incorporó contenidos más avanzados sobre el uso de *ggplot2* para la construcción de gráficos e incluyó el uso del paquete *survey* para la ponderación de resultados. Esta versión se encuentra disponible de forma gratuita en formato de libro interactivo digital (documento HTML) en un repositorio oficial del paquete Bookdown; por otra parte, puede encontrarse para su descarga en formato de libro para imprimir (documento PDF) en el repositorio Researchgate de los autores.<sup>1</sup>

Este documento de apoyo docente no es completamente exhaustivo sobre el uso de R y RStudio, pues está delimitado por las especificidades de la asignatura y por el campo disciplinar en que se inscribe. Para usuarios interesados en profundizar sobre Lenguaje R se pueden consultar algunas de las referencias bibliográficas incluidas en este documento.

Agradecemos a Cristóbal Moya por su incansable y desinteresado apoyo para la introducción de R y RStudio a la Carrera de Sociología y al Magíster en Ciencias Sociales de la Universidad de Chile, como también a los valiosos comentarios realizados por Rocio Salas y Francisca Torres a diferentes borradores de este documento.

---

<sup>1</sup>Sugerimos usar el siguiente estilo para citar este documento como referencia bibliográfica: Boccardo, Giorgio y Ruiz, Felipe (2019). *RStudio para Estadística Descriptiva en Ciencias Sociales. Manual de apoyo docente para la asignatura Estadística Descriptiva*. Santiago: Departamento de Sociología, Facultad de Ciencias Sociales, Universidad de Chile. Segunda Edición.

**Los autores.**<sup>2</sup>

*Junio de 2019*

---

<sup>2</sup>**Giorgio Boccardo Bosoni.** Sociólogo y académico del Departamento de Sociología de la Universidad de Chile. Doctor(c) en Ciencias Sociales y Magíster en Estudios Latinoamericanos de la misma casa de estudios. Email de contacto: gboccardo@u.uchile.cl - **Felipe Ruiz Bruzzone.** Sociólogo y docente de apoyo del Departamento de Sociología de la Universidad de Chile. Magíster en Ciencias Sociales de la misma casa de estudios. Email de contacto: felipe.ruiz@ug.uchile.cl.



## Capítulo 2

# ¿Cómo definir qué y cuantos software de análisis estadístico manejar?

Existen múltiples lenguajes de producción y análisis de datos. Para quienes leen este documento nombres como SPSS, Microsoft Excel, Stata o Python quizá no sean desconocidos: varias de estas herramientas computacionales son ampliamente utilizadas en el campo de las Ciencias Sociales.

Si bien presentan características disimiles en cuanto a atributos como en su facilidad de uso, generalidad o especificidad de las herramientas de análisis que incorporan y el costo asociado a su utilización, es posible afirmar que su incorporación en los procesos de investigación - profesional o académica - ha contribuido de manera positiva al facilitar el procesamiento computacional de conjuntos extensos de datos y la ejecución de análisis estadísticos que en general resultan de una elevada complejidad ante volúmenes elevados de información (Elousa, 2009).

La decisión de qué software de análisis estadístico utilizar no tiene una respuesta predeterminada: la elección dependerá de las necesidades de la investigación. Esto pues los lenguajes de programación son *herramientas* y el principal criterio para decidir el uso de uno u otro debe efectuarse en función de la particularidad de los objetivos y alcances de la investigación que se busque desarrollar.

Es por eso que aunque puedan existir diferencias sustanciales entre los diferentes software mencionados cada programa puede tener una utilidad específica. En tal sentido, su aplicación debe efectuarse sin perder de vista su cualidad de herramientas con atributos y potencialidades particulares. En todo momento es la investigadora o investigador quien indica instrucciones de análisis a estos programas computacionales a partir del problema que se está investigando, el tipo de información con que se trabaja, los objetivos del estudio, los recursos disponibles y conocimientos que tiene el equipo de investigación. Así, el uso de un programa de análisis estadístico no reemplaza los procesos de decisiones metodológicas como tampoco evidencia de manera automática los errores ni las incoherencias de los análisis: por eso, su uso debe hacerse en relación a las decisiones teóricas y metodológicas realizadas previamente en el contexto de un diseño de investigación particular.

En tal medida la decisión de usar uno u otro software - o una combinación de varios - no es arbitraria. Debe basarse en una observación razonada e informada de las limitantes y potencialidades que cada herramienta ofrece, siempre pensando en los requerimientos específicos que demandan los procesos de investigación. A continuación se ofrece una breve introducción a diferentes software de análisis de datos que pueden ser considerados para el análisis estadístico de datos sociales.

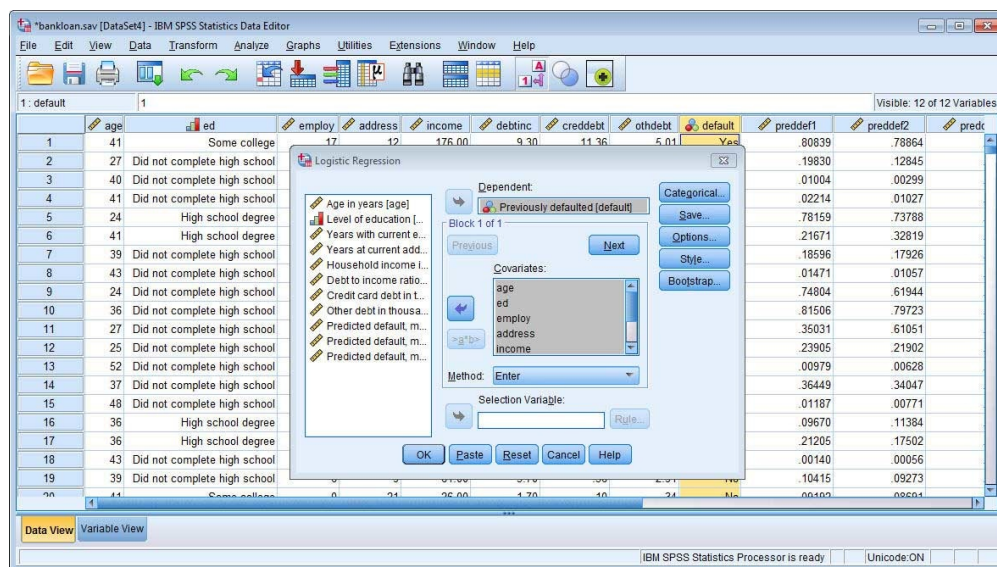


Figura 2.1: Interfaz del programa Statistical Package for Social Sciences, de IBM

## 2.1 SPSS: uno de los más usado en Ciencias Sociales

SPSS es una de las herramientas de análisis estadístico más extendidas en el campo de las Ciencias Sociales. De hecho, muchos de los manuales para realizar distintos tipos de análisis estadísticos traen ejemplos e instrucciones obtenidas con este programa. Este software se basa en una estructura del tipo planilla de datos pero incluye una interfaz basada en botones y ventanas que hace muy amigable su uso.

SPSS permite crear bases de datos que contengan la información de un estudio cuantitativo y realizar distintas operaciones utilizando aquellos datos: análisis uni o multivariado, análisis descriptivo o explicativo, gestión y visualización de datos, entre otros.

Este paquete estadístico requiere de una licencia pagada, aunque puede ser utilizado temporalmente mediante una versión de prueba. De esta forma puede adquirirse el programa básico e incorporar otros paquetes de análisis de datos (pagados), según las diferentes herramientas de análisis requeridas. También existen actualizaciones periódicas del programa, las cuales incluyen tanto mejoras en la interfaz de usuario como nuevas aplicaciones para el análisis de datos.

## 2.2 EXCEL: uno de los olvidados

Microsoft Excel es un software que permite crear bases de datos, analizar información y calcular diferentes estadísticos. Este tipo de software se denomina *software de hoja de cálculo* debido a que su interfaz se presenta como una planilla ordenada a partir de filas y columnas donde la unidad básica de almacenamiento de información es la *celda*. Además permite realizar operaciones de estadística descriptiva y multivariada y cuenta con una interfaz más cómoda para la digitación de datos.

En términos generales Excel permite crear tablas que calculan de forma automática los valores de ciertos análisis que el investigador o investigadora específica, imprimir tablas con diseños cuidados y crear gráficos de manera muy simple. Este programa forma parte de “Office”, un conjunto de herramientas de Microsoft que combina varios tipos de software para crear documentos de texto, hojas de cálculo y presentaciones y para administrar el correo electrónico. También corresponde a un software pagado (requiere comprar una licencia de Microsoft Office), sin embargo es usual contar ya con una licencia básica al comprar un computador que

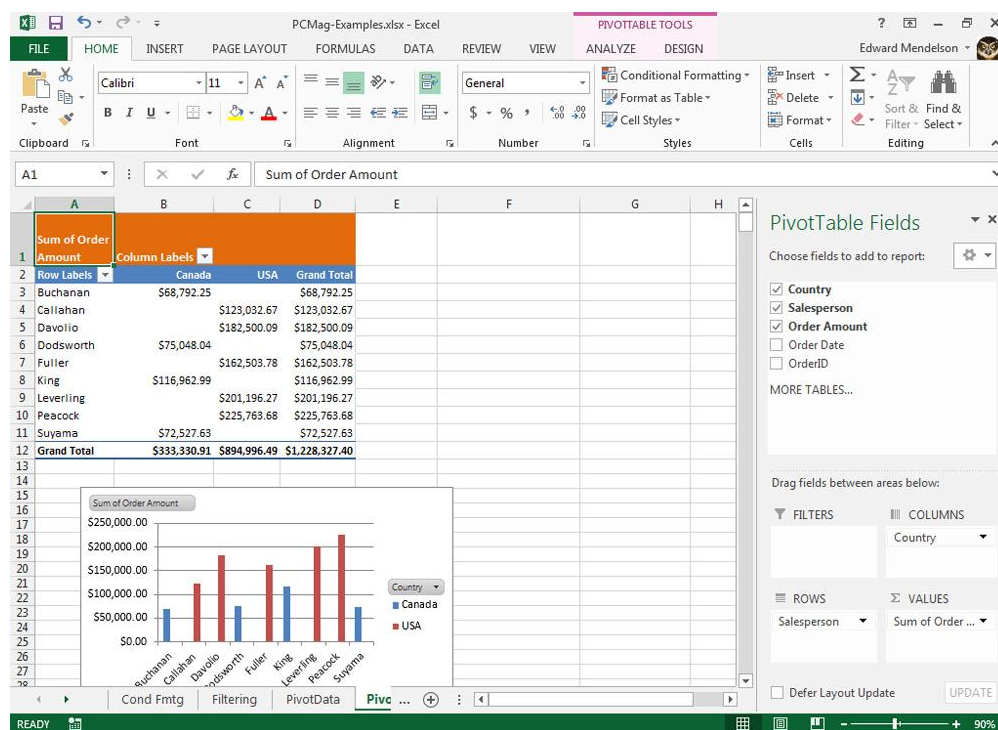


Figura 2.2: Interfaz del software Excel, de Microsoft Office

incluye Microsoft Windows como sistema operativo o conseguir una clave de licencia para uso individual o institucional.<sup>1</sup>

Muchas de las bases de datos disponibles, ya sea que provengan de instituciones de gobierno o de otro tipo se encuentran en formato .xls (planilla Excel), lo que implica que como investigadoras e investigadores se vuelve necesario manejar herramientas de este tipo, pues en muchos casos será necesario trabajar con ellas de manera articulada con softwares de análisis estadístico más especializados (nociones claves en este sentido son **extensión de archivo** y **exportar/importar**, cuestiones que serán revisadas más adelante). La experiencia en investigación social académica y profesional indica que por lo general los datos se digitan y trabajan en bruto en hojas de cálculo para posteriormente efectuar los procesamiento estadísticos en otros programas de mayor especialización.

## 2.3 Stata: un programa de nicho

Stata es un software de uso pagado - aunque también dispone de una versión de prueba - desarrollado por la compañía StataCorp. De manera similar a SPSS este programa combina un formato de entrada de datos basado en la estructura de planilla de cálculo, a la vez que presenta una amigable interfaz de botones junto con la posibilidad de manejarlo directamente a través de sintaxis.

Incorpora una gran cantidad de herramientas para la gestión de bases de datos y análisis estadísticos de diferente nivel de complejidad. Se evidencia un software que pretende incorporar en una sola plataforma

<sup>1</sup>Para quienes deseen acceder a una versión actualizada de este paquete de Microsoft, la Universidad de Chile ha puesto a disposición de sus alumnos una licencia gratuita para hasta 3 computadores. Por otro lado, en el mundo del *software libre* existen plataformas alternativas que ofrecen de manera gratuita el mismo conjunto de herramientas que Microsoft Office, de manera altamente compatible con ellas: una de las más conocidas es el *paquete de oficina libre* desarrollado por The Software Foundation, conocido como LibreOffice (disponible para distintos sistemas operativos). Una opción similar es Open Office, desarrollado por The Apache Software Foundation. La primera alternativa (Libre Office) presenta mayor compatibilidad con las aplicaciones de Microsoft Office, pero sus actualizaciones son menos estables que la segunda (Open Office).

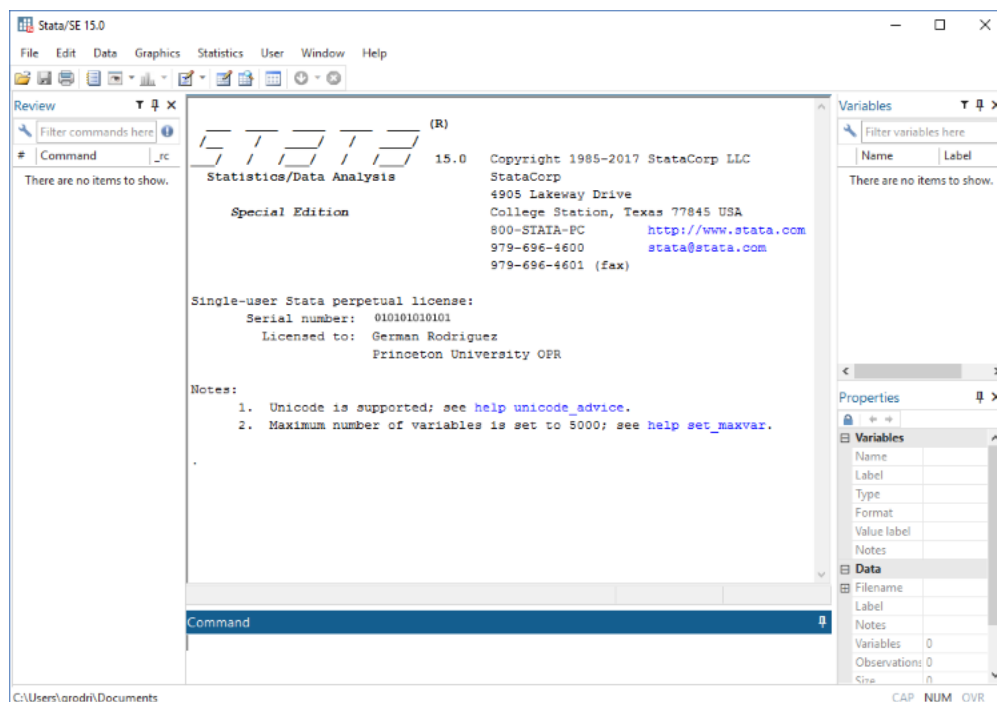


Figura 2.3: Interfaz del software Stata

todos los procesos de una investigación cuantitativa: construcción, validación y mantención de bases de datos, análisis estadísticos simples y multivariados, construcción de gráficos y resultados para su publicación. Es ampliamente utilizado en el campo de las ciencias económicas.

Para su uso oficial o institucional requiere la compra de una licencia pagada. Así se adquiere el programa básico que incluye todas las posibilidades de manejo y análisis de datos mencionadas, sin hacer necesaria la compra de paquetes adicionales. También existen actualizaciones periódicas del programa las cuales incluyen tanto mejoras en la interfaz de usuario como nuevas aplicaciones para el análisis de datos, que aseguran la estabilidad del lenguaje de programación entre diferentes versiones del software.

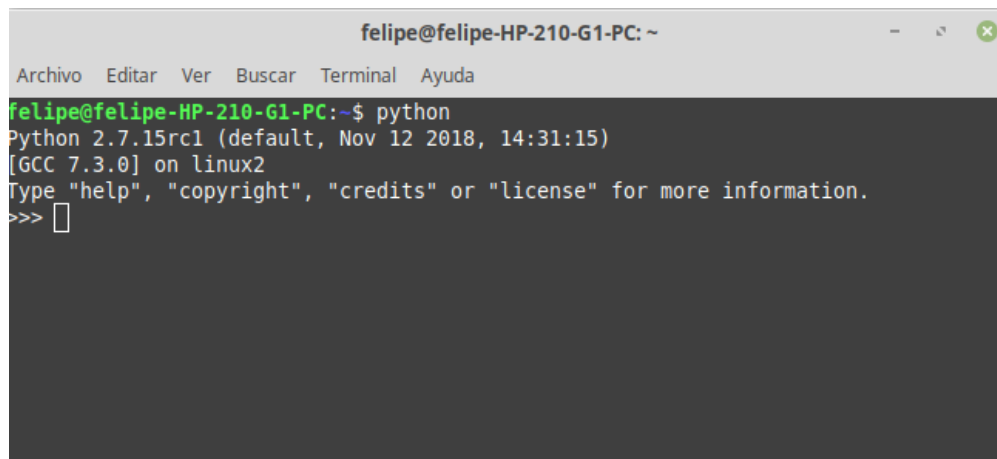
## 2.4 Python: un lenguaje de programación de mayor potencia

Python es un lenguaje de programación multiparadigma que soporta programación orientada a objetos, programación imperativa y programación funcional. Su principal premisa es desarrollar un lenguaje de programación que sea *simple* y *entendible* por los usuarios.

Al ser un *lenguaje de programación* sus usos son más amplios que los anteriores softwares señalados. Sus características se pueden aplicar al desarrollo de páginas web, sistemas de almacenamiento de información, programación de otros softwares, operaciones matemáticas y estadísticas. En términos generales brinda un lenguaje de sintaxis con una estructura que permite una programación clara en escalas pequeñas y grandes.

Como se verá, tiene características similares a R: es un programa gratuito de código abierto, que cuenta con una comunidad científica activa que colabora permanentemente en su desarrollo, elementos que permiten que integre múltiples aplicaciones adicionales a su versión básica, para diferentes campos de aplicación.

Contar con un lenguaje de programación intuitivo hace que una de sus principales ventajas frente a R sea la de una curva de aprendizaje más rápida. Por otro lado, presenta una mayor capacidad para trabajar con el procesamiento de datos textuales (*text mining*, *text analysis*), es más eficiente en el manejo de grandes bases de datos y también presenta mejores recursos para manejar grandes volúmenes de información - incluyendo



```
felipe@felipe-HP-210-G1-PC: ~
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
felipe@felipe-HP-210-G1-PC:~$ python
Python 2.7.15rc1 (default, Nov 12 2018, 14:31:15)
[GCC 7.3.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> 
```

Figura 2.4: Interfaz del software Python

*metadatos* - como aquellos que pueden ser extraídos desde Internet (técnicas conocidas como *webscrapping*). Sin embargo, presenta menos desarrollo que R en relación al desarrollo de herramientas de visualización de datos y en paquetes específicos para análisis estadístico multivariado.

## 2.5 R: el temido

Se trata de un software de distribución gratuita (un *freeware*) desarrollado por *The R Foundation for Statistical Computing*. R es un lenguaje de programación utilizado para el análisis de datos cuyo énfasis de uso está en la configuración directa de los análisis de parte del usuario antes que en una interfaz amigable.

Se trata de un proyecto colaborativo en la medida que los mismos usuarios van desarrollando nuevas aplicaciones que son compartidas gratuitamente en la página oficial del software; así, R está en permanente ampliación: es un proyecto abierto y gratuito.

*“R, en tanto en cuanto software libre, se inscribe dentro del proyecto GNU General Public Licence (Licencia Pública General, GNU). Se trata de una licencia creada por The Free Software Foundation (Fundación para el software libre), organización fundada por Richard Matthew Stallman en el año 1985. El principal propósito de la licencia GNU es declarar la libertad del uso, modificación y distribución del software y protegerlo de intentos de privatización que puedan de algún modo restringir su uso (...). Parte de la vasta información disponible sobre R es accesible a través de la página CRAN (Comprehensive R Archive Network; <http://cran.r-project.org/>), sitio oficial de R. Es la página base del proyecto R, desde la cual se puede descargar la última versión del programa (un equipo formado por unas doce personas, R Development Core Team; asumió en 1997 las labores de actualización semestral del código de R), consultar manuales sobre R, obtener ayuda sobre su funcionamiento a través de un sistema de ayuda on line, y, en definitiva, estar al corriente de los movimientos en este entorno de trabajo.”* (Elousa, 2009, 653)

La modalidad básica de este programa de análisis estadístico presenta una interfaz poco amigable para el usuario. Como se observa en la imagen anterior la interfaz no dista mucho de la sofisticación de un bloc de notas pues sólo presenta un espacio en blanco donde escribir los comandos, mientras que ninguno de los botones sirve para realizar análisis estadísticos.

Básicamente es un editor de sintaxis, lo que en principio requeriría la habilidad de manejar al dedillo todos y cada uno de los comandos necesarios para ejecutar algún tipo de análisis mediante el particular lenguaje de programación de este *freeware*. Es fundamentalmente esta característica la que inhibe al usuario inicial en su

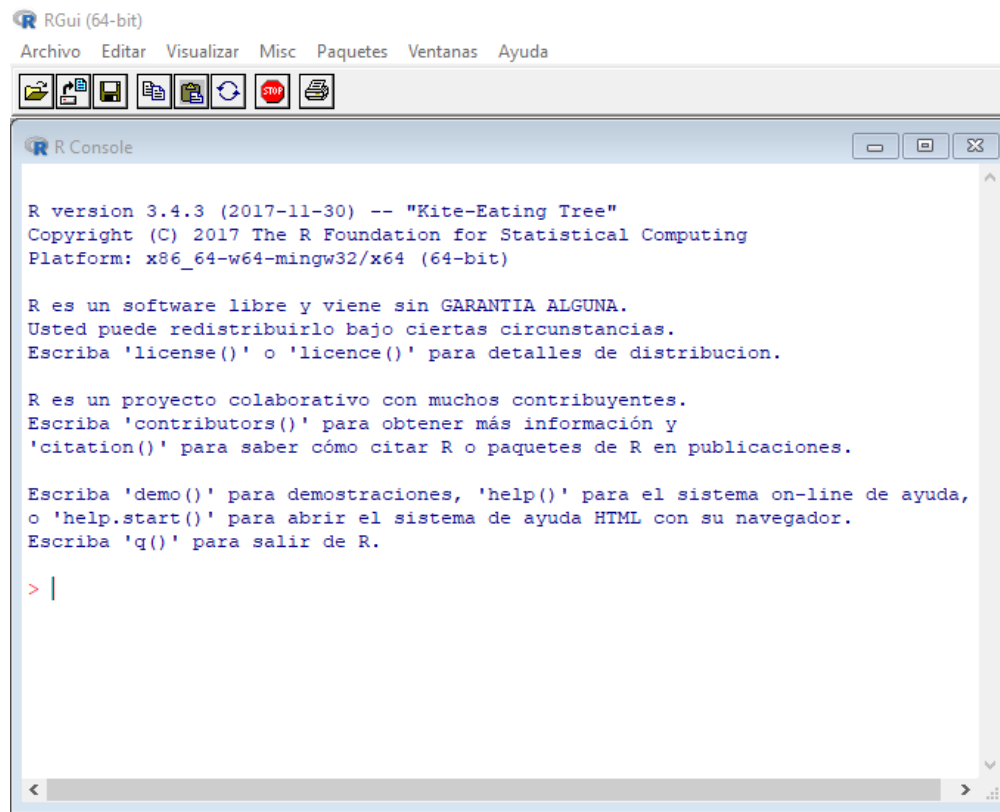


Figura 2.5: Interfaz del software R, en su versión básica

uso aunque presenta atributos que significan ventajas relativas en relación a los otros softwares mencionados en este capítulo.

Es uno de los softwares con una mayor variedad de herramientas de análisis estadístico univariado y multivariado, no sólo para las ciencias sociales sino también para otras disciplinas.<sup>2</sup> Al ser un programa de contribución libre (código abierto) no requiere la adquisición de una licencia pagada para su uso; existe además una amplia variedad de paquetes descargables de modo gratuito que son desarrollados por usuarios a lo largo de todo el mundo, lo que permite estar al día en cuanto a la exigencia de incorporar nuevas y más sofisticadas herramientas de un modo gratuito y libre. Adicionalmente se trata de un programa “liviano” que gasta poca memoria computacional para ser ejecutado. Todos estos atributos lo configuran como una alternativa bastante interesante para el análisis estadístico de datos sociales.

Si bien la versión básica del programa dista mucho de ser amigable no hay que desesperar pues existen soluciones (gratuitas) para este asunto. Es posible descargar e instalar softwares adicionales que, a partir de la instalación del software base, permitan contar con una máscara que actúe como una interfaz amigable entre el usuario y el entorno computacional que opera con códigos; todos detalles que se verán a lo largo de este documento.

## 2.6 Ventajas del uso de R

A diferencia de otros programas, para ocupar R nos adentraremos en el uso pleno de la modalidad de sintaxis; esto es: no emplearemos botones para realizar los análisis, sino que nos comunicaremos con el software de manera directa, a partir de lenguaje (código) computacional. El uso de la sintaxis (en cualquier software) y el conocimiento de los comandos de R nos permite las siguientes ventajas para nuestros análisis (Elousa, 2009):

1. *Replicabilidad*: Elemento fundamental en la investigación científica y cada vez más en las revistas académicas donde se exigen los archivos de sintaxis para la publicación de resultados. Permite que cualquier persona a quien enviemos nuestros análisis podrá entender cómo fueron construidos y replicarlos de manera exacta.
2. *Eficiencia*: En condiciones “reales” de trabajo continuado, el uso de sintaxis representa un incremento exponencial de la eficacia; por ejemplo, para hacer un solo cálculo (como un calcular una media aritmética), en una modalidad de definición de procedimientos de análisis estadístico mediante botones debemos presionar (por ejemplo) al menos cinco botones para llegar al resultado. Esto es tiempo acumulado, y en instancias de manejo estadístico complejo de datos, consume tiempo y esfuerzo. Como contracara, el uso de sintaxis tiende a aminorar la realización de tales tareas, pues puede llegar a tratarse de una sola línea de comandos.
3. *Control*: Permite un control casi total en el trabajo de análisis, pues permite a quienes investigamos ir definiendo detalles que los programas con botones configuran por defecto; esto además permite detectar errores y potencia el trabajo colaborativo, ya que el lenguaje que diferentes investigadores(as) emplearán, es el mismo.

Si bien este tipo de atributos también pueden alcanzarse utilizando las sintaxis de otros softwares mencionados como Stata o SPSS, la exclusividad del uso de sintaxis en R como lenguaje de programación, acentúa tales características en comparación a las otras herramientas computacionales indicadas.

---

<sup>2</sup>Como lo plantea Paula Elousa, R “está constituido por más de 1.400 paquetes integrados con los que es posible ejecutar simples análisis descriptivos o aplicar los más complejos y novedosos modelos formales. Además, la incorporación a R de interfaces gráficas (...) que crean entornos de trabajo amigables muy similares al entorno del SPSS permiten saltar la barrera de la accesibilidad, y utilizarlo sin ningún tipo de reparo en la docencia. ¿Existe algo mejor? Libre, gratuito, asequible, accesible y siempre a la vanguardia.” (Elousa, 2009, 652)

## 2.7 Una mirada comparada: limitantes y potencialidades

Ya se han revisado las principales herramientas computacionales que existen para el análisis estadístico de datos. ¿Es posible sintetizar en una sola evaluación las potencialidades y límites de cada una de estas alternativas? Creemos que sí, y para ello usaremos seis criterios: 1) generalidad, 2) costo, 3) facilidad de uso, 4) popularidad, 5) el valor del software libre y 6) desarrollo y actualización. Las tres primeras son señaladas siguiendo lo planteado por Elousa (2009) mientras que las últimas tres se proponen como criterios propios para considerar:

1. **Generalidad.** Bajo este criterio, sin lugar a dudas R es una de las alternativas más convenientes. Su estructura como plataforma de código abierto permite incorporar de manera gratuita paquetes adicionales a su versión básica que posibilitan el desarrollo de una gran variedad de técnicas de procesamiento de datos y análisis estadístico. Si bien para los análisis que se enseñarán en este manual programas como Microsoft Excel o SPSS cuentan con las herramientas suficientes, para análisis más sofisticados resultan limitados por las escasas herramientas que disponen, así como en la posibilidad de realizar análisis personalizados y no vía una configuración de fábrica que como usuario no es posible modificar. Además, su configuración como lenguaje de programación hace que R sea una herramienta de mayor flexibilidad que las otras alternativas presentadas: permite trabajar con datos cuantitativos y cualitativos así como usar diferentes fuentes de información (por ejemplo, datos existentes en la web en un sentido amplio). Este tipo de características lo distinguen radicalmente de las alternativas de software construidas en torno a la lógica de la planilla de cálculo (como SPSS y Stata).
2. **Costo.** Herramientas como R y Python son las únicas que presentan una modalidad gratuita de distribución y uso, lo que las pone por delante de las otras alternativas presentadas. Como ya ha sido señalado Microsoft Excel, SPSS y Stata requieren la adquisición de una licencia pagada para su uso en computadores de uso personal o institucional, lo que obliga a incurrir en costos de instalación bastante elevados. Así y todo en el caso de SPSS, por ejemplo, la versión básica no viene con todos los paquetes de análisis (el paquete AMOS, para ecuaciones estructurales por ejemplo, se vende por separado) lo que entorpece aún más la utilidad de estos programas en la medida que los requerimientos de análisis sofisticados aumentan.
3. **Facilidad de uso.** Tanto SPSS como Excel y Stata cumplen con el criterio de facilidad de uso. Son softwares amigables, que ponen a disposición del usuario una amplia variedad de herramientas de uso intuitivo lo que ayuda mucho en la introducción al análisis de datos empleando softwares computacionales. Por otra parte, Python y R presentan una interfaz de mayor complejidad al basarse en un uso estrictamente a partir de instrucciones computacionales o sintaxis. Sin embargo, tal dificultad sólo implica una curva de aprendizaje más lenta que se compensa con las ventajas ya señaladas en relación a los criterios de *costo* y *generalidad*.
4. **Popularidad.** Este criterio se vincula con la “extensión del uso” de una herramienta computacional en el campo científico o profesional. Por mucho que se pueda argumentar a favor de la utilización de softwares como el que enseñaremos en este manual, también se debe considerar que hay herramientas computacionales que gozan de mayor popularidad y uso (por ejemplo SPSS, Stata o Microsoft Excel). En tal sentido, una formación integral debe propiciar un uso combinado de estas herramientas, que permita una flexibilidad y adaptación a diferentes contextos profesionales o académicos, que por lo general no son flexibles en relación a cambiar de manera rápida el software de “cabecera” utilizado para sus procesos de investigación social.
5. **El valor del software libre (gratuito y de código abierto).** Que un software sea *libre* quiere decir que sus usuarios son libres de usarlo, copiarlo, distribuirlo, editarlo y modificarlo según sus propias inquietudes (FSF, 2019). En tal medida, el valor del software libre no sólo refiere a su distribución libre de pagos, sino también a su transparencia en cuanto al diseño de sus diferentes funcionalidades. Esto permite que lo realizado por el software libre no sea una “caja negra” para las y los usuarios, lo que es de alta relevancia para su uso en investigación científica, proceso en el cual se busca tener un control razonado de los límites y potencialidades de cada decisión de análisis tomada.
6. **Desarrollo y actualización.** En el caso de R y Python se trata de plataformas que están en continuo desarrollo y actualización, que cuentan con una comunidad científica activa e involucrada en la producción de nuevas herramientas y soluciones para problemas y desafíos de programación. Ello



implica que este tipo de herramientas tienen un potencial de desarrollo ilimitado, que no se encuentra sujeto a un uso de “moda” o según criterios netamente económicos, pues un software libre se usa siempre que la comunidad científica decida hacerlo. Si bien pueden existir otras alternativas de softwares altamente especializados en técnicas específicas de análisis de datos (sea de análisis cuantitativo o cualitativo) éstas tienden a quedar desactualizadas una vez pierden popularidad y su negocio deja de ser rentable.

La siguiente tabla resume de forma comparada los atributos y características de los distintos software analizados.

Tabla 2.1: Resumen comparativo de diferentes herramientas computacionales para análisis estadístico

Dimensión / Lenguaje	R	Python	SPSS	Excel	Stata
Alcance	General, orientación multidisciplinaria	General, orientación multidisciplinaria	Limitado, orientado a Ciencias Sociales	Limitado, orientado a administración	Limitado, orientado a Economía
Licencia	Libre (freeware)	Libre (freeware)	Pagada (versión de prueba limitada)	Pagada (versión de prueba limitada)	Pagada (versión de prueba limitada)
Aprendizaje	Sintaxis, poco intuitivo	Sintaxis, poco intuitivo	Botones y sintaxis, intuitivo	Botones y sintaxis, intuitivo	Botones y sintaxis, intuitivo
Visualización	Avanzada	Intermedia	Básica	Intermedia	Intermedia
Análisis de texto	Intermedio, poca eficiencia	Avanzado, amplia eficiencia	No	No	No
Minería de Datos	Intermedio, poca eficiencia	Avanzado, amplia eficiencia	No	No	No
Sistema operativo	Windows, Mac OS, Linux	Windows, Mac OS, Linux	Windows, Mac OS	Windows, Mac OS	Windows, Mac OS

Teniendo en cuenta todos los elementos que ya han sido expuestos es que en la formación estadística de la carrera de sociología de la Universidad de Chile se ha decidido adoptar progresivamente un uso extendido de R. Es por ello que resulta de central importancia poder acercar de manera amigable esta herramienta a todos los estudiantes desde el inicio de su formación estadística. Tal es el sentido del presente manual.



## Capítulo 3

# Instalación de los softwares a utilizar en este manual

Para efectos de este manual las y los lectores deberán saber como instalar dos softwares: R y RStudio. Si bien **siempre se usará el segundo**, es importante entender que éste es solo una forma de visualizar el primero. Por tanto, siempre se instalará primero R y luego RStudio.

Para instalar R se descarga un **installer** - software que permite instalar el programa deseado en nuestros computadores. Para poder descargarlo se debe ingresar a la página oficial de R, donde se observa un enlace con el texto CRAN bajo el título **Download** (imagen 3.1).

Al hacer clic sobre tal enlace, la página redirige a una nueva en la cual debemos escoger un “mirror” o servidor específico para descargar los archivos. Como se ve en la imagen 3.2, existen muchos servidores con diferentes ubicaciones alrededor del mundo. En la práctica, se puede seleccionar cualquiera, pero idealmente elegiremos el de Chile, pues por su cercanía debería ser más rápida la transferencia de información.

Al seleccionar un “mirror” específico, se abrirá la página oficial del CRAN - The Comprehensive R Archive Network - que permite ingresar a las páginas de descarga de los **installers** para distintos sistemas operativos (imagen 3.3).

Como se ve, desde tal página se accede a la descarga de los instaladores del programa para diferentes sistemas operativos. A continuación se explicará como conducir el proceso de instalación, primero para el sistema operativo Windows de Microsoft, luego para el sistema operativo de Apple, Mac OS, y finalmente para diferentes distribuciones de Linux.

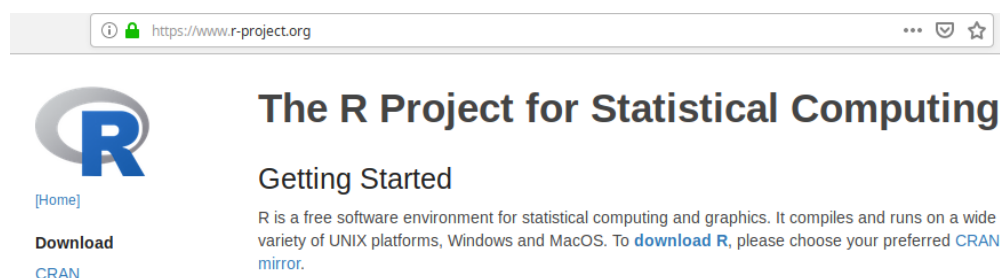


Figura 3.1: Página Oficial del Proyecto R

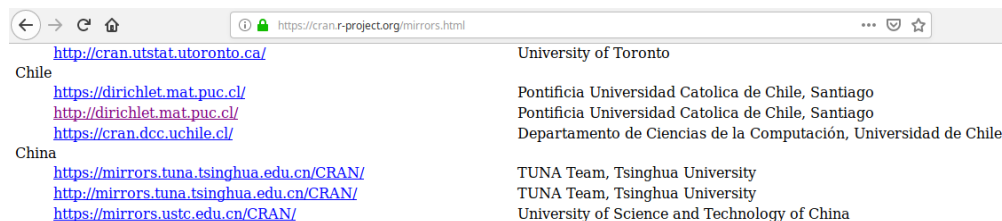


Figura 3.2: Distintos servidores (mirrors) para descargar R

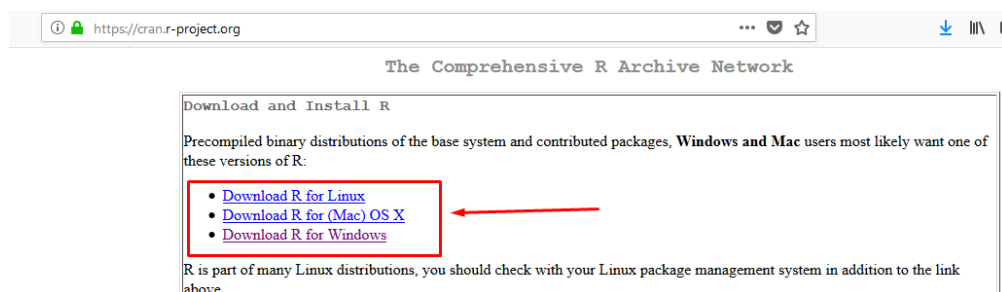


Figura 3.3: Página de descarga de installers de R para distintos sistemas operativos

## 3.1 Instalación de R y RStudio en Microsoft Windows

### 3.1.1 Instalar R en Windows

Al seleccionar la opción Download R for Windows se accederá a la siguiente página en la cual se debe hacer clic en el enlace *install R for the first time*, para descargar el instalador.

La siguiente página (imagen 3.4) ofrecerá la opción de descarga de R para Windows. Al presionar el enlace se descargará un “installer”, es decir, un archivo con extensión “.exe”.<sup>1</sup>

Luego de descargarlo es preciso ubicar tal archivo en la carpeta donde haya sido descargado, para hacer doble clic sobre el mismo y ejecutarlo.

Posterior a eso, se deben seguir las instrucciones de instalación que va ofreciendo el software, hasta completar la instalación. Una vez concluido el proceso, el software habrá quedado instalado en el computador.

<sup>1</sup>Al 10 de junio de 2019 la versión disponible para descarga era la 3.6, publicada el 26 de abril de 2019 cuyo nombre fue *Planting of a Tree*. Con el paso de los meses, es posible que la versión de descarga se haya actualizado

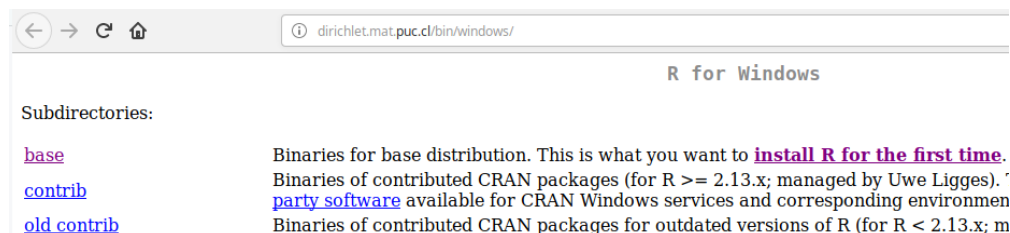


Figura 3.4: R base para Windows

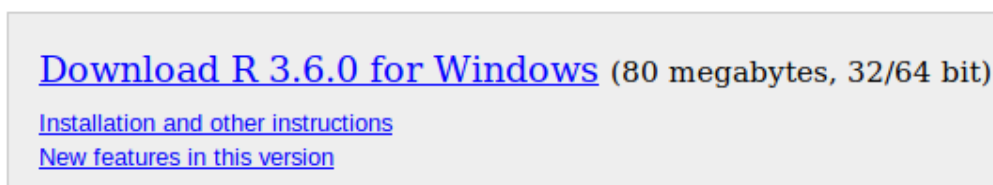


Figura 3.5: Descarga del installer de R para Windows

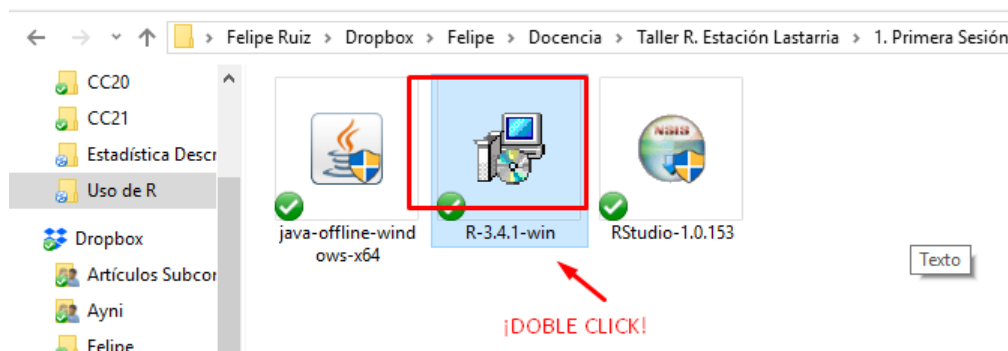


Figura 3.6: Ejecución del installer de R para Windows

### 3.1.2 Instalar RStudio en Windows

La lógica de instalación de RStudio es similar a lo ya presentado para R. En la página oficial de RStudio se debe acceder a la pestaña **Products** y seleccionar la opción *RStudio*. Dentro de esa página hay que acceder al apartado de descarga del installer presionando el enlace *RStudio Desktop* (imagen 3.7).

Luego se debe seleccionar la opción *Download RStudio Desktop* para acceder a la página de descarga de las diferentes versiones de los instaladores.

Dentro de la nueva página que se abrirá, se encuentra la siguiente sección donde se puede descargar el **installer** de RStudio para Windows (también se encuentran para otros sistemas operativos).

Del mismo modo que para la instalación R en su versión básica es preciso buscar y ejecutar el instalador para RStudio desde la carpeta donde fue descargado. Luego de seguir todas las instrucciones del proceso de instalación el programa habrá quedado instalado correctamente. Es necesario enfatizar que en lo posterior *solamente se ejecutará RStudio, y nunca el software R*. Este último queda instalado y RStudio se encarga de ejecutarlo. En breve, siempre se usará el icono de RStudio para abrir una nueva sesión del programa.

RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).



Desktop

Run RStudio on your desktop

[RStudio Desktop >](#)



Server

Centralize access and computation

[RStudio Server >](#)

Figura 3.7: Página de descarga de RStudio

**Support**

Community forums only

- Priority Email Support
- 8 hour response during business hours (ET)

**License**

AGPL v3

[RStudio License Agreement](#)

**Pricing**

Free

\$995/year

DOWNLOAD RSTUDIO DESKTOP

BUY NOW

Figura 3.8: Acceso a instaladores RStudio Desktop

**RStudio Desktop 1.2.1335 — Release Notes**

RStudio requires R 3.0.1+. If you don't already have R, download it [here](#).

Linux users may need to [import RStudio's public code-signing key](#) prior to installation, depending on the operating system's security policy.

**Installers for Supported Platforms**

Installers	Size	Date	MD5
RStudio 1.2.1335 - Windows 7+	126.9 MB	2019-04-08	d0e2470f1f8ef4cd35a669aa323a2136
RStudio 1.2.1335 - Mac OS X 10.12+ (64-bit)	121.1 MB	2019-04-08	6c570b0e2144583f7c48c284ce299eef
RStudio 1.2.1335 - Ubuntu 14/Debian 8 (64-bit)	92.2 MB	2019-04-08	c1b07d0511469abfe582919b183eee83
RStudio 1.2.1335 - Ubuntu 16 (64-bit)	99.3 MB	2019-04-08	c142d69c210257fb10d18c045fff13c7
RStudio 1.2.1335 - Ubuntu 18 (64-bit)	100.4 MB	2019-04-08	71a8d1990c0d97939804b46cfb0aea75
RStudio 1.2.1335 - Fedora 19+/RedHat 7+ (64-bit)	114.1 MB	2019-04-08	296b6ef88969a91297fab6545f256a7a
RStudio 1.2.1335 - Debian 9+ (64-bit)	100.6 MB	2019-04-08	1e32d4d6f6e216f086a81ca82ef65a91
RStudio 1.2.1335 - OpenSUSE 15+ (64-bit)	101.6 MB	2019-04-08	2795a63c7efd8e2aa2dae86ba09a81e5
RStudio 1.2.1335 - SLES/OpenSUSE 12+ (64-bit)	94.4 MB	2019-04-08	c65424b06ef6737279d982db9eefcae1

Figura 3.9: Descarga del instalador de RStudio para Windows Vista, 7, 8 y 10

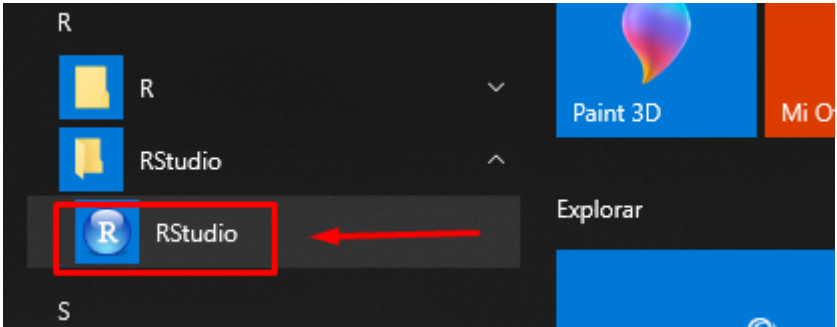


Figura 3.10: Ícono de RStudio en el listado de programas existentes

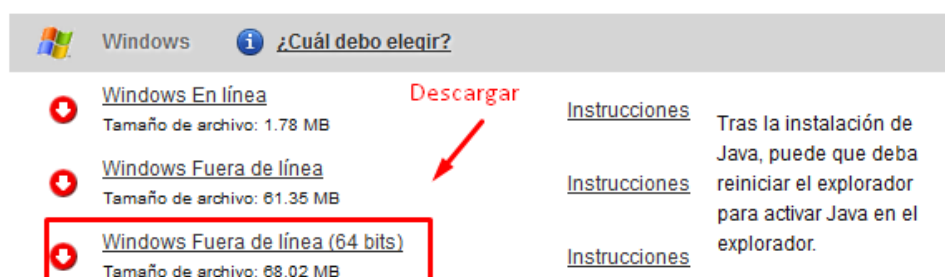


Figura 3.11: Descarga del installer para aplicación Java 64 Bits/Offline

### 3.1.3 Instalación adicional: Java 64 Bits Offline

Al usar Windows como sistema operativo existe un problema que se produce entre la configuración del sistema operativo del computador y la configuración de los exploradores de Internet. Por lo general, el sistema operativo Windows de cualquier computador (de escritorio o portátil tipo notebook) se desenvuelve en una configuración denominada de *64 bits*; por otro lado, los exploradores de Internet (Microsoft Edge, Mozilla Firefox, Google Chrome, etc.) funcionan en *32 bits* de manera predeterminada. En general, para acceder a contenido vía Internet se usa una aplicación externa, llamada Java, que nos permite reproducir vídeos y otras aplicaciones web. Ésta se asocia al explorador de Internet y por tanto *por defecto se instala, funciona y actualiza en formato 32 bits*.

Distintos paquetes de R usan esta aplicación, por lo que suele producirse incompatibilidades entre un software R instalado de acuerdo al sistema operativo (y por tanto, en modo 64 bits) y una aplicación Java instalada de acuerdo a los requerimientos de los exploradores de Internet (32 bits). Esto produce un error en diversos análisis cotidianos. Para prevenirlo, es preciso descargar e instalar una versión de Java denominada *Offline*, en su modalidad de 64 bits. Es una versión de Java que viene en 64 bits y que no se actualizará automáticamente de manera constante, previniendo los errores de ejecución mencionados. Tal versión puede descargarse en el siguiente enlace.

Se deben seguir las mismas indicaciones de instalación que señaladas en los subapartados anteriores de este capítulo: ubicar el instalador en la carpeta de descarga, ejecutarlo y seguir las instrucciones de instalación.

Habiendo realizado estos procedimientos ya es posible introducirse de manera específica en el funcionamiento del programa.

## 3.2 Instalación de R y RStudio en Mac OS

### 3.2.1 Instalación de R en Mac OS

Para instalar R en un sistema operativo Mac OS, propio de dispositivos Apple, se debe acceder a la opción Download R for (Mac) OS X del CRAN (imagen 3.3).

En la siguiente página (imagen 3.12) se debe seleccionar la versión de R para descargar su archivo de instalación.

Desde allí se puede acceder a la última versión de R si es que es compatible con la correspondiente versión del sistema operativo Mac OS (en este caso *El Capitan* o superior). A la fecha de publicación de este material, se podía descargar el archivo *R-3.6.0.pkg*. Si no se tiene compatibilidad por sistema operativo, se debe seleccionar un instalador para una versión anterior de R o bien actualizar el sistema operativo desde la *App Store*. Haciendo clic sobre el enlace, comienza la descarga.

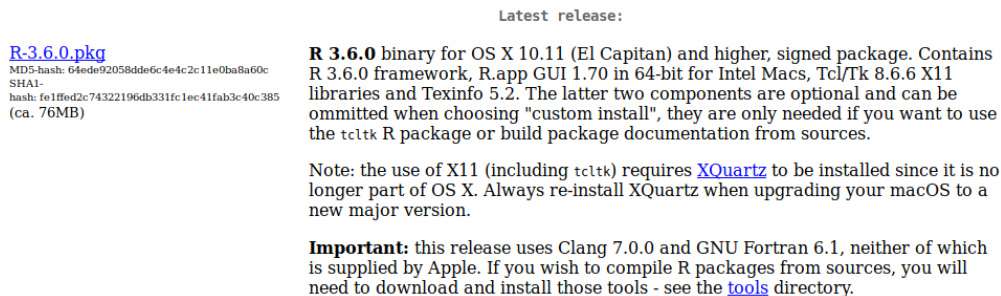


Figura 3.12: Installer para sistema operativo Mac OS

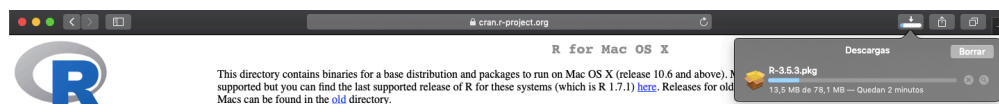


Figura 3.13: Descarga del installer para sistema operativo Mac OS

Una vez descargado, el instalador puede buscarse en las descargas del explorador de Internet *Safari* o en la ventana de descargas del explorador de archivos *Finder*. Haciendo doble clic sobre tal archivo se abrirá la ventana de instalación. luego de ejecutarlo se deben seguir las instrucciones indicadas por el sistema.

Una vez completados todos los pasos la instalación es correcta. Esto puede corroborarse observando que el programa se encuentra disponible en el *Launchpad*, en donde aparecen todas las aplicaciones instaladas en el computador.

### 3.2.2 Instalación de RStudio en Mac OS

Para la instalación de RStudio en Mac OS hay que seguir el mismo tipo de procedimiento ya indicado para Microsoft Windows. En la página de descargas de RStudio Desktop (imagen 3.9) se debe seleccionar la opción denominada *Mac OS X 10.12+ (64-bit)*. Una vez hecho clic el instalador se empieza a descargar automáticamente.

Una vez descargado, el programa puede buscarse en las descargas del explorador de Internet *Safari* o en la ventana de descargas del explorador de archivos *Finder*. Haciendo doble clic sobre el archivo se abre la ventana de instalación. En ella debe arrastrarse el icono de *RStudio* al de aplicaciones (*Applications*) para concretar la instalación del software (imagen 3.1 6).

Para corroborar la correcta instalación de RStudio, puede observarse que el programa se encuentre disponible en el launchpad, en donde aparecen todas las aplicaciones instaladas en el computador.

## 3.3 Instalación de R y RStudio en Linux

### 3.3.1 Instalación de R en Linux

La instalación y puesta en marcha de ambos softwares presenta diferencias de mayor importancia respecto a la instalación en sistemas operativos de Microsoft o Apple, de forma independiente a la distribución de Linux que se esté utilizando. Como fue señalado en la "Nota técnica" al inicio de este manual, todas las operaciones aquí realizadas se efectuaron desde un sistema operativo Linux Mint 19.1 "Tessa", **Cinnamon Edition**.





Figura 3.14: Instalación de R para sistema operativo Mac OS

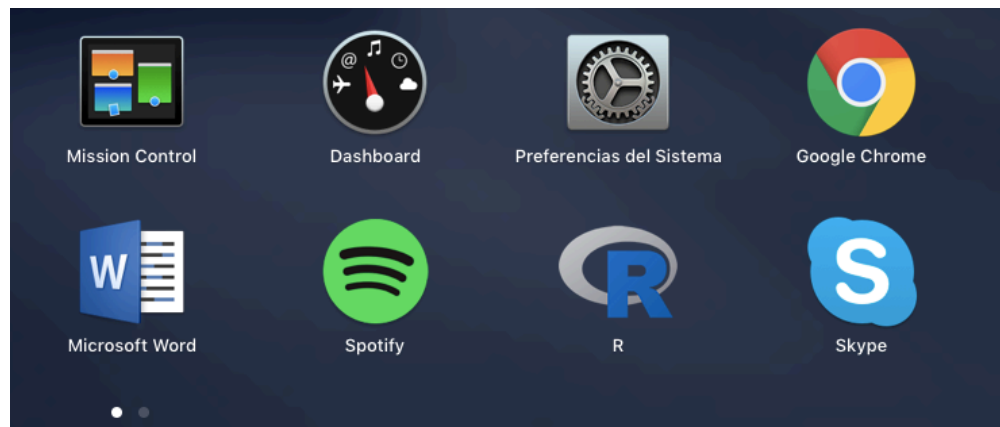


Figura 3.15: R disponible en Launchpad de Mac OS

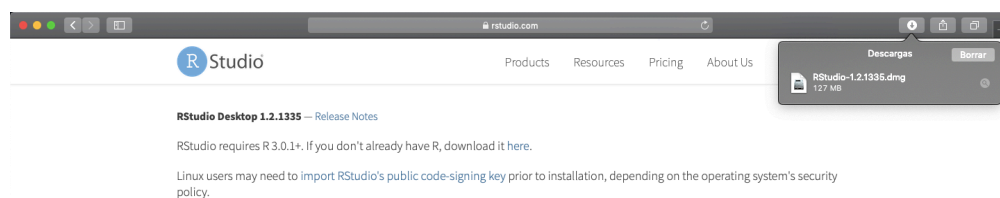


Figura 3.16: Descarga del installer de RStudio para Mac OS

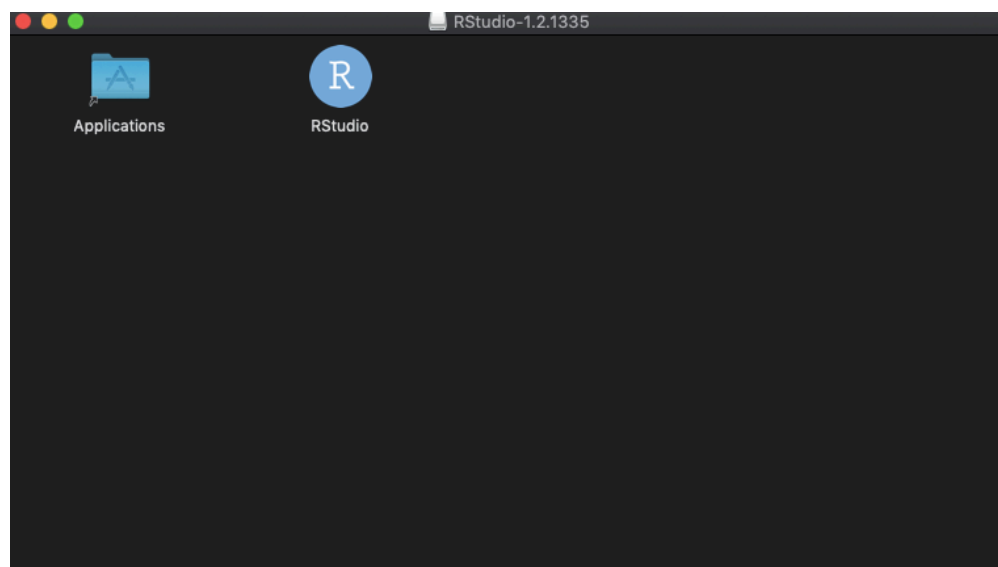


Figura 3.17: Instalación de RStudio en Mac OS

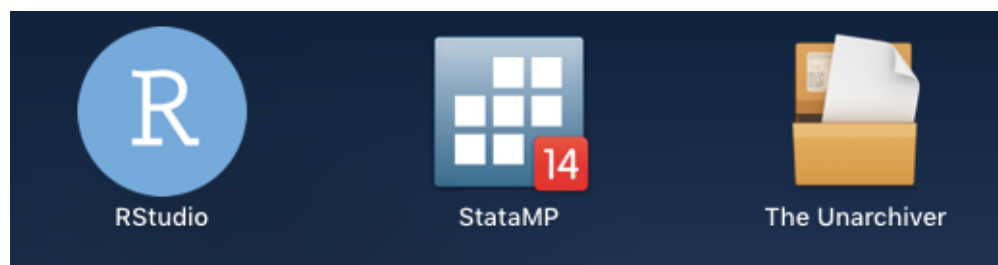


Figura 3.18: Instalación de RStudio en Mac OS

```

felipe@felipe-HP-210-G1-PC: ~
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
felipe@felipe-HP-210-G1-PC:~$ R

R version 3.4.4 (2018-03-15) -- "Someone to Lean On"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

>

```

Figura 3.19: Visualización de R desde un terminal de Linux

Al menos en distribuciones de Linux como Mint o Ubuntu - entre las más populares - R viene como un software previamente instalado. En efecto, esto puede corroborarse abriendo un terminal de comandos en Linux y ejecutando la instrucción R.

Como se observa en la imagen 3.19, la versión de R ya instalada en el sistema corresponde a la 3.4.4. Sin embargo, esto no siempre es así: debido a que el repositorio oficial de paquetes de Linux presenta algunos baches de actualización, es posible que la única versión disponible para actualizar R sea una más antigua (por ejemplo, una versión 3.1.2). Esto es problemático pues genera múltiples incompatibilidades con los paquetes actualmente existentes en el CRAN, que dependen de versiones más recientes del software.

Para instalar una versión más reciente del software, se sugiere ejecutar las siguientes líneas de comando desde un terminal de Linux (una línea por vez):

```

#Esta comando permite agregar claves de acceso para editar el repositorio de Linux
sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys E298A3A825C0D65DFD57CBB651716619E084DAB9

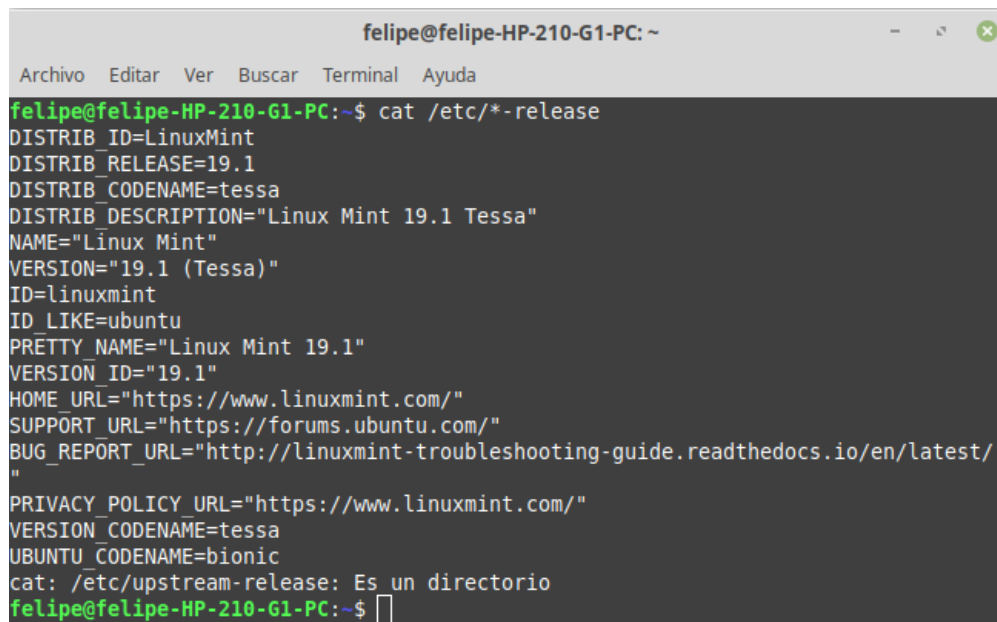
#Este comando permite incorporar un enlace específico desde donde Linux debe instalar/actualizar R
sudo add-apt-repository 'deb [arch=amd64,i386] https://cloud.r-project.org/bin/linux/ubuntu xenial/'

#Este comando indica instalar la última versión de R disponible para Linux
sudo apt-get install r-base

```

Con ello se obtendrá una versión más reciente de R. En muchos casos no será la última, también debido a actualizaciones algo tardías del repositorio de R.

No obstante haber realizado esto, es altamente probable que sigan existiendo algunos problemas para administrar la instalación de paquetes: de forma específica, hay veces que no funciona bien el comando `install.packages()`. Esto porque el sistema operativo Linux no incorpora algunas dependencias específicas de sistema operativo que son requeridas para usar todas las funcionalidades de R.



```
felipe@felipe-HP-210-G1-PC: ~  
Archivo  Editar  Ver  Buscar  Terminal  Ayuda  
felipe@felipe-HP-210-G1-PC:~$ cat /etc/*-release  
DISTRIB_ID=LinuxMint  
DISTRIB_RELEASE=19.1  
DISTRIB_CODENAME=tessa  
DISTRIB_DESCRIPTION="Linux Mint 19.1 Tessa"  
NAME="Linux Mint"  
VERSION="19.1 (Tessa)"  
ID=linuxmint  
ID_LIKE=ubuntu  
PRETTY_NAME="Linux Mint 19.1"  
VERSION_ID="19.1"  
HOME_URL="https://www.linuxmint.com/"  
SUPPORT_URL="https://forums.ubuntu.com/"  
BUG_REPORT_URL="http://linuxmint-troubleshooting-guide.readthedocs.io/en/latest/"  
PRIVACY_POLICY_URL="https://www.linuxmint.com/"  
VERSION_CODENAME=tessa  
UBUNTU_CODENAME=bionic  
cat: /etc/upstream-release: Es un directorio  
felipe@felipe-HP-210-G1-PC:~$
```

Figura 3.20: Información del sistema operativo, desde la consola de Linux

Para solucionar estos problemas recomendamos lo siguiente. Una vez actualizada la versión de R, pero antes de cualquier operación dentro del software, se deben ejecutar los siguientes comandos - una línea por vez - desde la terminal de Linux.

```
#Instala dependencia g++  
sudo apt install g++  
  
#Actualiza el sistema con la nueva dependencia  
sudo apt-get update  
  
#Instala tres dependencias adicionales requeridas para el funcionamiento de R  
sudo apt-get install libcurl4-openssl-dev libssl-dev libxml2-dev  
  
#Instala/actualiza la versión "developer" de R  
sudo apt-get install r-base-dev
```

Una vez hechas estas operaciones, el software debería tener un funcionamiento óptimo.

### 3.3.2 Instalación de RStudio en Linux

La instalación de RStudio en Linux es similar a la de los otros sistemas operativos. Desde la página de descarga de los installers para RStudio se puede acceder a diferentes versiones de instalación indicadas para la distribución de Linux *Ubuntu* (imagen 3.9).

Para el caso de este manual se descargará la última actualización disponible para Ubuntu. es decir el archivo *RStudio 1.2.1335 - Ubuntu 18 (64-bit)* A pesar de que se esté operando una distribución de Linux tipo *Mint*, ésta se basa en una arquitectura tipo *Ubuntu*, por lo que el instalador antes mencionado funcionará correctamente. Esto último puede corroborarse solicitando la información del sistema operativo desde la consola de Linux mediante el comando `cat /etc/*-release`.

Una vez descargado el instalador de RStudio, éste se muestra como un archivo de extensión *.deb* en la carpeta donde se haya guardado el archivo.

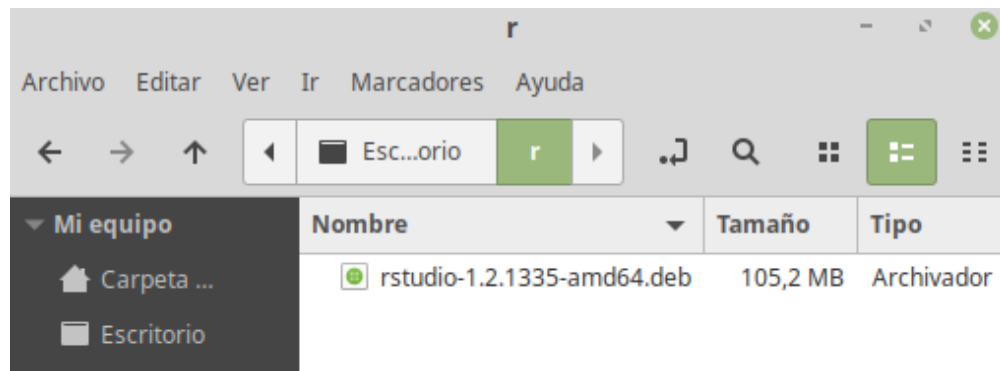


Figura 3.21: Instalador de RStudio para Linux

Tal archivo puede abrirse con el gestor de softwares de *Linux Mint* haciendo doble clic, e instalarlo manualmente mediante una interfaz de ventanas. Como se observa en la imagen 3.22, el instalador indica que se requieren paquetes adicionales para completar la instalación, los que serán instalados de forma automática por el sistema.

Una vez completada la instalación y habiendo aceptado todos los pasos, el software estará disponible en el listado de archivos del sistema, listo para ser utilizado.

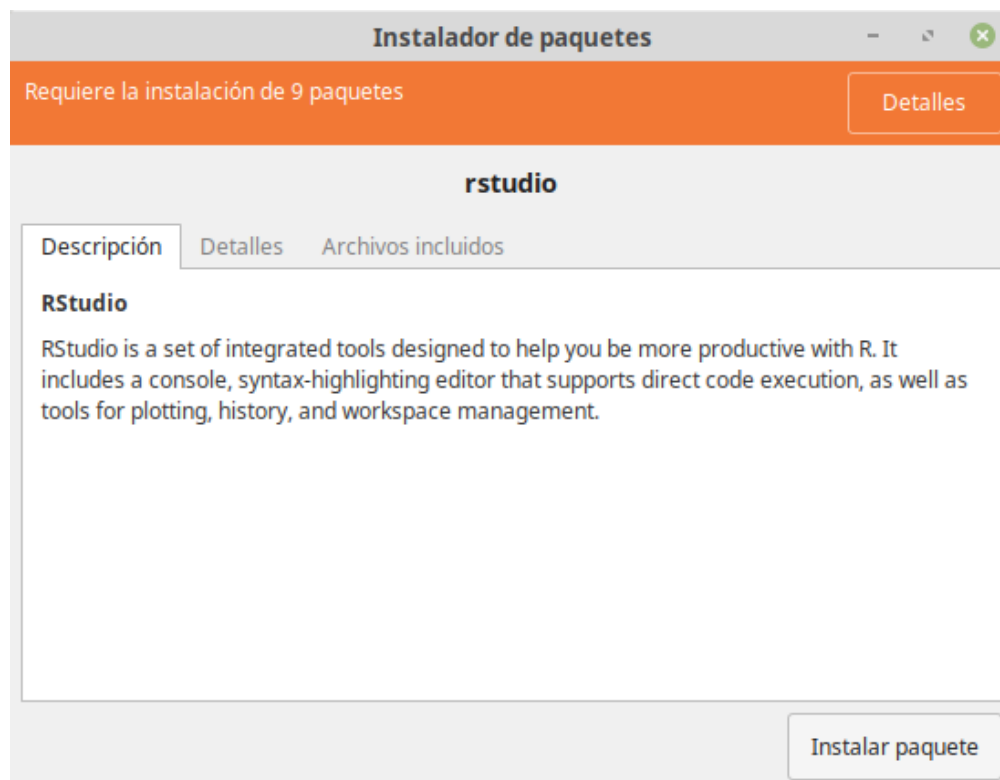


Figura 3.22: Instalación de RStudio para Linux

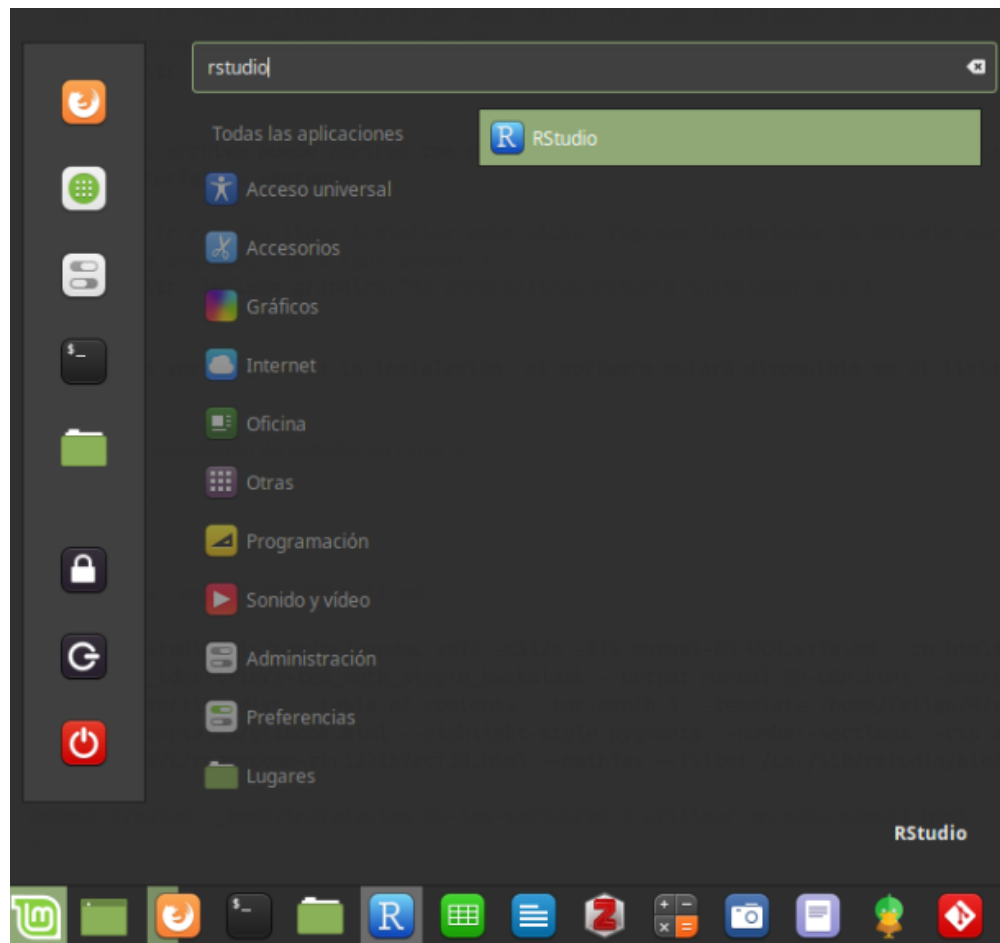


Figura 3.23: RStudio instalado en Linux





## Capítulo 4

# Uso básico de RStudio

### 4.1 ¿Qué es RStudio?: una interfaz para usar R

Para contar con más herramientas de apoyo en el uso de R emplearemos el software RStudio. Este software es una interfase - entre otras existentes como RCommander - que permite contar con una interacción más fluida con el programa R. Básicamente se trata de una máscara para visualizar el software que tiene como principales ventajas (1) el orden y (2) la visualización de los procesos que son llevados a cabo con R, todo de manera simultánea.

Se pueden ver 4 ventanas, además de la barra de opciones en la parte superior.

1. *Ventana (1)*: es el editor de sintaxis: se trata del lugar donde editamos la sintaxis para posteriormente ejecutarla. Al escribir allí no sucederá nada, a no ser que se apriete algún botón para ejecutar los comandos o la tecla `ctrl+enter`.
2. *Ventana (2)*: es el “entorno de trabajo” del programa: en este lugar se muestra el conjunto de datos y los “objetos” (resultados, variables, gráficos, etc.) que se almacenan al ejecutar diferentes análisis.
3. *Ventana (3)* tiene varias sub pestañas: (i) la pestaña **files** permite ver el historial de archivos trabajados con el programa; (ii) la pestaña **plots** permite visualizar los gráficos que se generen; (iii) la pestaña **packages** permite ver los paquetes descargados y guardados en el disco duro así como gestionar su instalación o actualización; (iv) la ventana **help** permite acceder al CRAN - Comprehensive R Archive Network (siempre que se cuente con conexión a Internet), página oficial del software que ofrece diferentes recursos para el programa: manuales para el usuario, cursos on line, información general, descarga de paquetes, información de los paquetes instalados, etc. Esta última pestaña es bastante útil: empleando el motor de búsqueda se accede de manera rápida a manuales de uso de los diferentes paquetes (y sus funciones) instalados en el computador (esto no requiere conexión a Internet).<sup>1</sup>; (v) la ventana **viewer** muestra los resultados al construir reportes mediante funcionalidades tipo *rmarkdown*.
4. *Ventana (4)*: es la consola. Corresponde a lo que sería el software R en su versión básica. Allí el software ejecuta las operaciones realizadas desde el editor de sintaxis.

### 4.2 Carpeta de trabajo y memoria temporal del programa

El software R funciona como un **entorno temporal de trabajo**, esto quiere decir que el usuario va agregando datos y objetos (conjuntos de datos con diferentes atributos) a una “hoja en blanco”. Hay que tener

---

<sup>1</sup>Al trabajar desde el editor de sintaxis, si se escribe el comando `help()` poniendo entre los paréntesis algún comando o palabra clave (por ejemplo `help(setwd)`) se visualizará la ayuda que brinda el software en relación a la búsqueda especificada.

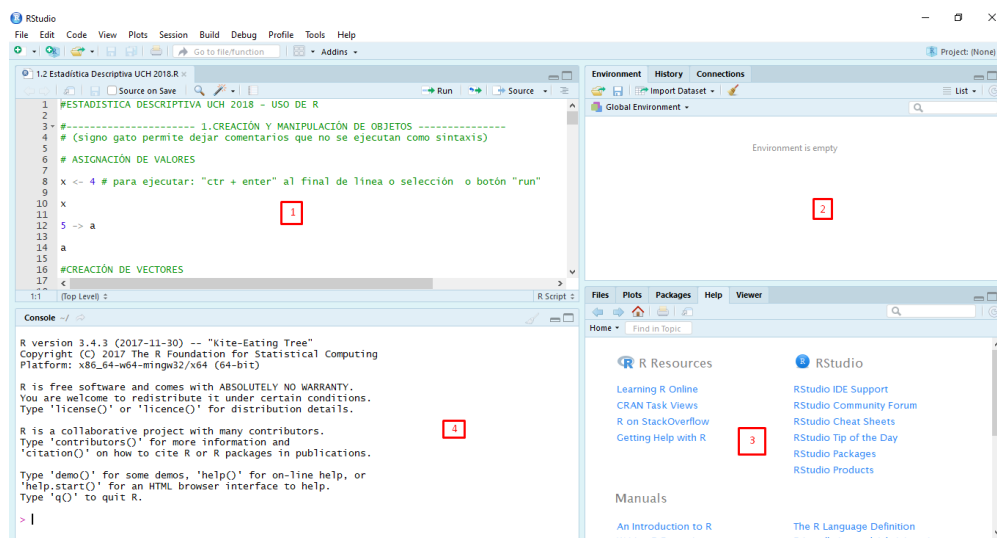


Figura 4.1: Interfaz del software RStudio

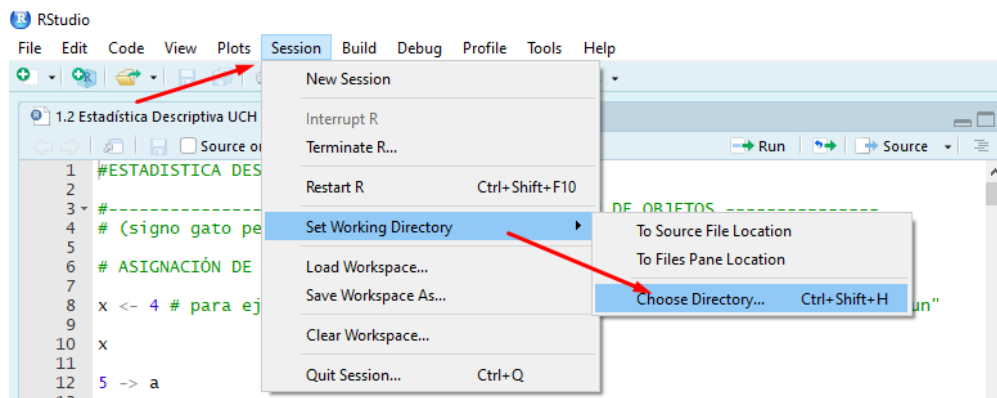


Figura 4.2: Secuencia de botones para establecer carpeta de trabajo

en cuenta que R trabaja con la memoria activa (RAM) del computador, por lo tanto cualquier análisis sólo mostrará la información resultante pero no permanecerá como archivo posible de utilizar de modo posterior. Es decir, si los análisis no son guardados como objetos (vectores, matrices, listas u otros tipos de objetos) se deberán repetir las instrucciones para obtener otra vez el resultado.

Dado que R opera como un espacio temporal y autónomo de trabajo, también es preciso indicarle en qué parte del disco duro del computador están los archivos a utilizar. Para evitar el engorroso procedimiento de indicar todo el tiempo las rutas de acceso a los archivos (elementos del tipo: `C:\escritorio\Curso R\base_datos.xlsx`) es posible establecer una *carpeta de trabajo*: esto es, una carpeta predeterminada donde el programa buscará los archivos a ejecutar y guardará los archivos a conservar con cambios.

Existen dos alternativas para definir la carpeta de trabajo. La primera es emplear el siguiente comando:<sup>2</sup>

```
setwd("ruta de acceso a la carpeta especificada")
```

Otra forma de hacerlo es mediante la botonera superior; presionando el botón **Session**, luego **Set Working Directory**, **Choose Directory** y en la pestaña **examinar** se selecciona la carpeta a utilizar.

<sup>2</sup>Para efectos de este manual, quienes busquen replicar estos ejemplos deben crear una carpeta en su computador definiéndola como carpeta de trabajo para RStudio. Allí deben descargar y guardar el material a utilizar. Tales elementos se pueden encontrar en la siguiente carpeta en línea almacenada en un repositorio Github



Figura 4.3: Consola en ejecución de instrucciones

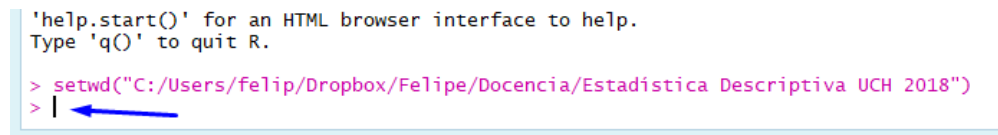


Figura 4.4: Consola de R lista para ejecutar instrucciones

Todas las operaciones de R - sean indicadas vía sintaxis o botones - son ejecutadas según comando computacional que es visualizado en la consola. La ejecución de comandos entrega diferentes señales respecto a su funcionamiento. Por ejemplo, mientras se está ejecutando un comando, el programa muestra un signo “Stop!” en la esquina superior derecha de la consola (como se ve en la imagen). Eso indica que el programa está ocupado ejecutando una acción. Si se presiona tal símbolo, se cancelará la operación en curso.

Una vez ejecutado el comando, debiera observarse el siguiente resultado en la consola. Cuando esté todo en azul (o con otro color diferente al básico de la sintaxis dependiendo de la configuración de colores de RStudio) indica que el comando fue correctamente ejecutado. Asimismo, el signo `>|` con la barra parpadeando, indica que R está listo para procesar nuevas instrucciones. En el siguiente ejemplo (imagen 4.4) se muestra el establecimiento de una carpeta de trabajo cuyo desarrollo ha sido exitosamente ejecutado.

### 4.3 Elementos fundamentales del uso de sintaxis en RStudio

R es un lenguaje de programación. En el caso de la versión básica del software, así como de la interfaz RStudio, el usuario interactuará con el programa mediante **códigos**. La *sintaxis* es un conjunto de **códigos**. Su uso en R es bastante intuitivo y sigue un patrón lógico. De modo general el **lenguaje de programación de R** (o sintaxis) sigue la siguiente estructura básica:

#### Ejercicio 4.1

**comando** (datos a utilizar)

El ejemplo indicado es una operación simple. Al aumentar la complejidad de los análisis la especificación de los comandos irá requiriendo una mayor cantidad de información. Por ejemplo, para construir una tabla de frecuencias de una variable ubicada en una base de datos se utiliza el siguiente comando, que le indica al software el conjunto de datos y variable específica a analizar (ejercicio 4.2):

#### Ejercicio 4.2

**table** (base de datos, variable)

En un sentido global, como se observa en la imagen 4.5, la estructura general de una sintaxis puede resumirse como sigue: a un objeto dado se le asigna el resultado de una función, que a su vez se ejecuta sobre un conjunto de datos especificado, con una serie de configuraciones particulares. En este caso, la sintaxis de la imagen es ficticia pero nos servirá con un propósito pedagógico. Si se lee de izquierda a derecha, la línea de comando puede explicarse como sigue:

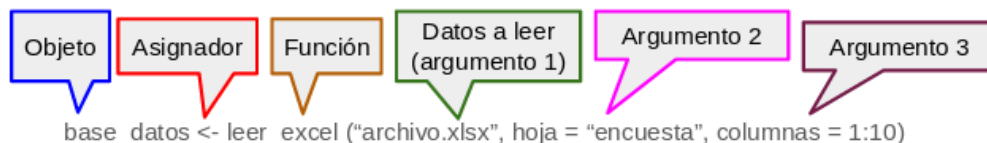


Figura 4.5: Estructura básica de una sintaxis de R

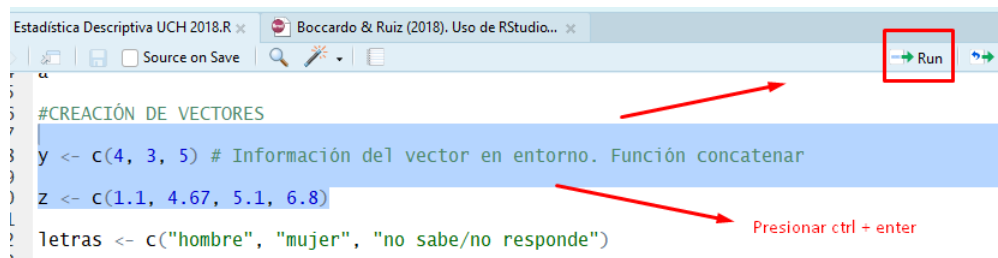


Figura 4.6: Formas de ejecutar una sintaxis en RStudio

- Primero se indica un **objeto** a crear, con un nombre arbitrario, definido por nosotros.
- Luego se indica el **asignador**, que expresa que todo lo que esté a la derecha de la flecha se guardará en el **objeto** creado.
- Luego viene la **función** que en este caso permite leer archivos tipo Microsoft Excel. Luego de la función se abre un paréntesis que contiene los argumentos, es decir, instrucciones que especifican ciertos detalles de lo que queramos la función realice.
  1. El primer **argumento** por lo general indica la información a leer, en este caso indica un archivo de tipo Excel (extensión *xlsx*).
  2. El segundo **argumento** indica la hoja del archivo a leer.
  3. El tercer **argumento** indica qué columnas se leerán de forma específica (en este caso, las primeras diez).

Ahora bien, escribir tales instrucciones en el editor de sintaxis permite contar con una sintaxis editable que no se ejecutará de manera automática como operaciones computacionales (cuestión que sí sucede al escribir sintaxis directamente sobre la consola). Para ejecutar una sintaxis se debe seleccionar con el cursor - como cuando se busca *copiar* un texto - el trozo de código (*code chunk* en inglés) que interesa utilizar para luego:

1. Usando la botonera superior, apretar con el cursor del mouse el botón “Run”.
2. Apretar la combinación de teclas *ctrl + enter*.

Otra opción es ejecutar una sola línea de código. Para esto se posiciona el cursor *sobre* la línea que interesa ejecutar (en cualquier parte de ella) y se aplica algunas de las dos operaciones ya indicadas.

## 4.4 Manejo básico de la sintaxis de R: creación de objetos, inclusión de anotaciones y definición de secciones

Dado que el programa funciona en la memoria temporal del computador - memoria RAM - una vez que se ejecuta un proceso y se cierra el programa la operación y su resultado desaparecerán si no han sido almacenadas de alguna forma en el disco duro.

Para entender esto de manera práctica se pondrá un ejemplo con el comando más básico que se puede utilizar. Este comando consiste en crear un objeto y asignarle un dato numérico. El objetivo del ejercicio 4.1 es darle

#### 4.4. MANEJO BÁSICO DE LA SINTAXIS DE R: CREACIÓN DE OBJETOS, INCLUSIÓN DE ANOTACIONES Y DEFINICIONES

el valor 10 a X, luego pedirle a R que entregue el valor de X, para después cambiar el valor de X a 5. Para cualquiera de tales operaciones, se escribe una letra (en este caso es una X, pero podría ser cualquier otra letra) y con el asignador <- se le asigna el valor numérico deseado.

##### Ejercicio 4.3

```
#Asignación del valor "10" al objeto "X"
```

```
X <- 10
```

```
X
```

```
## [1] 10
```

```
#Cambiar el valor de X: cambiar el "10" por un "5"
```

```
X <- 5
```

```
X
```

```
## [1] 5
```

Como se observa en el ejemplo, si luego de efectuar una asignación se ejecuta el objeto creado ("X"), se obtendrá como resultado el último valor almacenado. De tal modo, se entiende que un objeto es un elemento computacional que puede almacenarse en el programa para ser usado en análisis posteriores. Además, debe destacarse que al ejecutar una operación esta no es permanente: *puede ser transformada si se sobrescribe la información*.<sup>3</sup> La estructura de la sintaxis computacional es bastante flexible. Sólo a modo de ejemplo, puede ejecutarse - ejercicio 4.4 - una asignación de un valor a un objeto alterando el orden de los elementos. Vale la pena destacar que el orden de los elementos no altera el resultado, mientras el asignador apunte desde los valores que nos interesa guardar hacia el nombre de un objeto a crear (o previamente existente). No obstante ello, por una cuestión de orden sugerimos la estructura de sintaxis `objeto <- valores`, con el asignador apuntando hacia la izquierda de la pantalla.

##### Ejercicio 4.4

```
#Asignación del valor "5" al objeto "Y"
```

```
5 -> Y
```

```
Y
```

```
## [1] 5
```

Una cuestión importante es que el editor de sintaxis acepta la inclusión de comentarios en el cuerpo de los comandos. Como se ve en la imagen, al anteponer un signo gato (#) antes de un comando, este se pone de un color diferente a los códigos que representan instrucciones válidas (en la visualización por defecto de RStudio, el color será automáticamente *verde*).

¿Para qué sirve esto? Toda línea que esté con un signo gato antepuesto el programa la ignorará como comando de R. Esto permite incluir notas que indiquen lo que se está haciendo, mensajes o ayudas de memoria. Esto es relevante sobre todo en el contexto de trabajos que toman varios días - en los que será difícil recordar todo lo efectuado o probado - y en contextos de investigaciones colectivas donde se intercambian propuestas de análisis con otros investigadores.

Así, un aspecto bastante útil de la *funcionalidad de comentarios* es que permite separar mediante encabezados cualquier sintaxis. Como se ve en la siguiente imagen, si se indica el signo # y el título se encierra entre *al menos cuatro guiones* a cada lado, el programa considera tal formato como si se tratara de un título; esto permite que usando el explorador de la sintaxis quien investiga pueda **navegar** de manera más rápida por las diferentes secciones de una sintaxis.

<sup>3</sup>Se puede ver que R funciona mediante la creación de objetos que representan un dato o un conjunto de datos. En este caso el valor es simple (un valor numérico) pero también puede tratarse de un objeto que almacene datos de una variable que refiera a categorías sociales tales como Grupo Socio-económico o Género, de un conjunto amplio de casos.

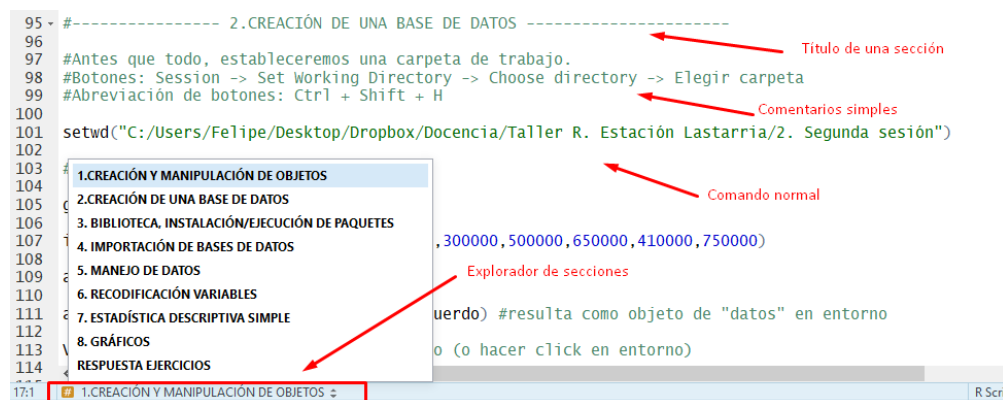


Figura 4.7: Explorador de secciones de una sintaxis de RStudio

## 4.5 Tipos de objetos en R (vectores)

Si los objetos creados contienen un *conjunto de datos del mismo tipo*, para el lenguaje del software (lenguaje de la informática) esto es un vector.<sup>4</sup> Aplicado al ámbito de las ciencias sociales se trata de una variable. Así, en R las variables (o vectores) pueden ser de 4 tipos; el tipo de variable depende del tipo de valores que se le asigna a cada objeto:

1. *numeric*: valores numéricos, incluye decimales.
2. *integer*: números enteros, no incluye decimales.
3. *character* valores alfanuméricos, es decir, letras, números y signos mezclados.
4. *logical*: valores lógicos, TRUE o FALSE.

De esta forma si una variable es *numérica*, tendrá asignado números; en caso de ser una variable de tipo *cadena* (**character**) los valores se deben ingresar entre comillas (“ejemplo”), y en caso de ser *lógica* sus valores serán alguna de las opciones TRUE o FALSE.

Hasta ahora sólo se ha explicado cómo asignar un valor singular a un objeto de R. Para ingresar más de un valor en un vector se deben indicar los elementos a almacenar entre paréntesis y separados por comas, anteceditos de la función *concatenar*, que se ejecuta anteponiendo una *c* al conjunto de objetos a agrupar.<sup>5</sup> Utilizando la función *concatenar* *c()* se puede crear un objeto que agrupe un conjunto de datos (un vector). En este caso se construirá una variable *Edad* con el siguiente conjunto de datos (deben separarse por comas): 15, 12, 27, 55, 63, 63, 24, 21, 70. En este caso será una variable numérica:

### Ejercicio 4.5

```
#Concatenar valores para crear un vector de más de un sólo valor
edad <- c(15, 12, 27, 55, 63, 63, 24, 21, 70)
```

Al ejecutar el comando se observa automáticamente el nuevo objeto en el entorno de R (junto con los objetos *x* e *y* ya creados).

<sup>4</sup>En el ámbito de la informática, un vector (también conocido como matriz) es una zona de almacenamiento contiguo que alberga elementos de un mismo tipo. Puede entenderse como una serie de elementos ordenados en filas o columnas. El vector ofrece una estructura que facilita el acceso a los datos para su manejo computacional. Mayores detalles de esta definición en Wikipedia.

<sup>5</sup>Para nombrar a los objetos es importante saber que estos deben empezar con una letra, aunque puede contener dígitos y puntos (.). Además, se debe tener en cuenta que R diferencia entre mayúsculas y minúsculas.

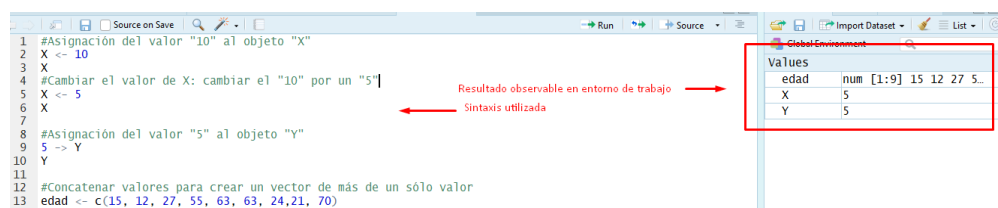


Figura 4.8: Entorno de trabajo de RStudio con objetos

Como se observa, el programa indica que el objeto creado `edad` es un variable o vector de tipo numérico (**numeric**) de una sola dimensión y que tiene nueve casos ( [1:9] ).<sup>6</sup>

En el ejercicio 4.3 se construyó un objeto que consideraremos como una variable resultante de, por ejemplo, haber aplicado una breve encuesta a los estudiantes de una clase. Ahora bien, usando el comando `table` (`edad`) - ejercicio 4.5 (continuación) podemos construir una tabla de frecuencias de los valores; el número debajo del caso indica la frecuencia absoluta de ocurrencia de cada valor.

#### Ejercicio 4.5 (continuación)

```
table(edad)
```

```
## edad
## 12 15 21 24 27 55 63 70
##  1  1  1  1  1  1  2  1
```

En el lenguaje computacional de R las variables se denominan “vectores”. A continuación se observa la creación de tres vectores, uno con números enteros, otro con números reales (incluye decimales y negativos) y un tercero con caracteres alfabéticos.

#### Ejercicio 4.6

```

#Vector numérico (función concatenar)
a <- c(4, 3, 5)

#Vector numérico (función concatenar)
b <- c(1.1, 4.67, 5.1, 6.8)

#Vector alfanumérico (función concatenar)
letras <- c("hombre", "mujer", "no sabe/no responde")

```

En el entorno de trabajo se observa la presencia de tres nuevas variables, con su respectiva cantidad de casos y el tipo de vector asignado por el programa.

## 4.6 Construcción de una base de datos

A continuación se construirá la primera base de datos a partir de tres variables. Para esto, como se observa en los siguientes comandos, se parte por la construcción de tres variables de 9 casos cada una:

<sup>6</sup>También se puede ejecutar el comando `attributes(edad)` para conocer tal información. En el mismo sentido, el comando `class(edad)` sirve para inspeccionar qué tipo de variable estamos considerando.

Variable	Values
a	num [1:3] 4 3 5
b	num [1:4] 1.1 4.67 5.1 6.8
edad	num [1:9] 15 12 27 55 63 63 ...
letras	chr [1:3] "hombre" "mujer" "..."
X	5
Y	5

Figura 4.9: Información de objetos en entorno de trabajo de RStudio

1. **Género.** Variable nominal con valores 1 y 2, que representan las categorías de respuesta “hombre” y “mujer”.
2. **Ingreso.** Variable de razón con valores arbitrarios de ingreso monetario.
3. **Acuerdo en torno al aborto libre.** Variable ordinal tipo *escala Likert* con valores 1, 2, 3, 4 o 5, que representan las categorías “nada de acuerdo”, “un poco de acuerdo”, “ni de acuerdo ni en desacuerdo”, “bastante de acuerdo”, “muy de acuerdo”.

#### Ejercicio 4.7

```
#Creación de las variables: todas tienen la misma cantidad de casos

genero <- c(1,1,2,1,2,2,2,1,2)

ingreso <- c(100000,300000,500000,340000,300000,500000,650000,410000,750000)

acuerdo <- c(1,1,3,2,4,1,5,3,2)
```

A partir de las variables ya creadas se puede construir una base de datos. Para esto se utiliza el comando `data.frame` asignando su resultado al objeto `aborto` que contendrá la base de datos construida.

Si la ejecución del comando es exitosa se verá un nuevo objeto de tipo `data` en el entorno de trabajo, donde además se indicará la dimensión de la base de datos (cantidad de casos y variables).

#### Ejercicio 4.7 (continuación)

```
#Creación de base de datos a partir de variables previamente creadas
#Se asigna el resultado al objeto "aborto"
aborto <- data.frame(genero, ingreso, acuerdo)
```

Posteriormente se puede visualizar la base de datos de dos maneras: utilizando el comando `View` o presionando un botón que se encuentra al lado del objeto base de datos en el entorno de trabajo del software. En el siguiente código se explicita la primera forma y en la imagen posterior la segunda.



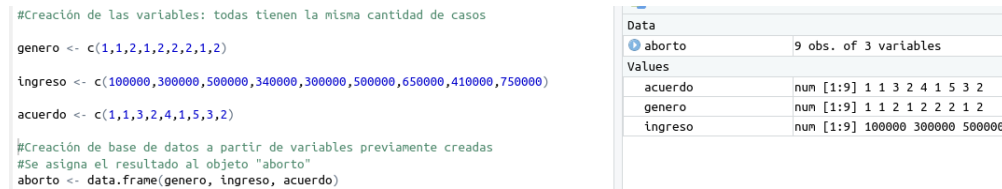


Figura 4.10: Entorno de trabajo con variables y base de datos creadas

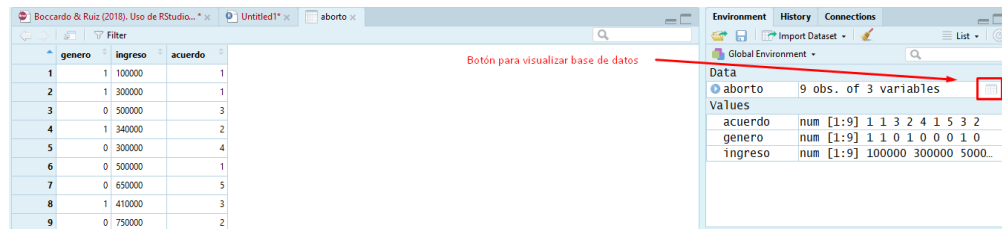


Figura 4.11: Uso de botón para visualizar base de datos

### Ejercicio 4.8

`View(aborto)` *#Comando para visualizar base vía sintaxis (o hacer click en entorno)*

Luego de construida esta base de datos interesa guardarla como un archivo reutilizable para posteriores análisis. Para ello es útil el comando `save`: indicando el nombre del objeto a guardar como archivo y definiendo también el nombre del archivo. Como se ve en la siguiente línea de comando, el nombre del archivo resultante se indica con el argumento `file = "nombre_archivo.extensión"` (el nombre del archivo va entre comillas).

### Ejercicio 4.9

`save(aborto, file = "aborto.RData")` *#Se indica primero el objeto a guardar  
#y luego el nombre del archivo, entre comillas.*

Ejecutando tal comando, se creará un nuevo archivo en la carpeta que hayamos indicado como *carpeta de trabajo*. Como se ve en la imagen a continuación, este archivo debería ser reconocido como un archivo de formato RStudio; se observa que es un archivo recién creado gracias a la información proporcionada por la columna *Fecha de modificación*.

Este archivo es la base de datos recién construida pero almacenada como archivo de formato R en el disco duro, específicamente en la carpeta de trabajo fijada en la sesión de R. Este archivo puede usarse para posteriores análisis, puede ser enviado a otras personas, etc.

Finalmente, se muestran dos maneras para “limpiar” el entorno de trabajo. Esto resulta útil pues luego de hacer múltiples cálculos exploratorios, mientras se depura un esquema de análisis, el entorno de trabajo se irá llenando paulatinamente con objetos que no sirven para continuar trabajando y sólo pueden confundir en los pasos posteriores del análisis.

Nuevamente, para limpiar el entorno de trabajo existen dos maneras:



Figura 4.12: Carpeta de trabajo con base de datos guardada

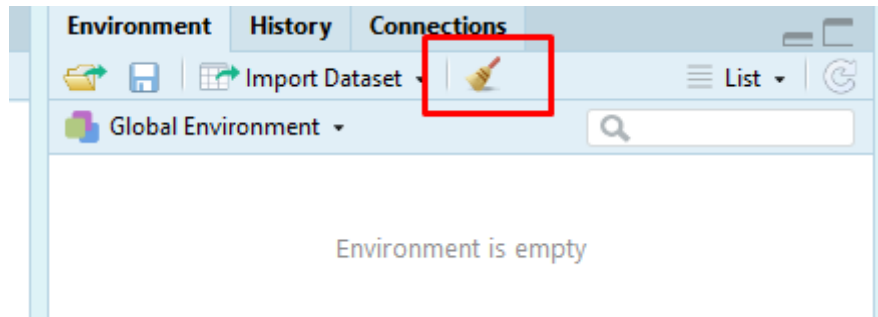


Figura 4.13: Uso de botón para limpiar entorno de trabajo

1. La primera es vía sintaxis, con dos comandos específicos que se detallan a continuación:
  - a. El primero permite eliminar elementos específicos y,
  - b. El segundo permite vaciar totalmente el entorno de trabajo.
2. La segunda forma es utilizando un botón existente en la botonera superior del entorno de trabajo (con forma de escoba) que permite borrar todos los elementos del entorno de trabajo, que se detalla en la siguiente imagen.

#### Ejercicio 4.10

```
#Comando para eliminar elementos específicos, aquí se elimina  
#la base creada.  
remove(aborto)  
  
#Comando para limpiar todos los objetos del entorno de trabajo.  
rm(list = ls())
```

Como se ve en la última imagen, ya sea ejecutando el comando `rm(list = ls())` o apretando el botón indicado, se puede limpiar completamente el entorno de trabajo.

## Capítulo 5

# Manejo de la biblioteca y gestión de paquetes

La versión básica del software R trae una cantidad limitada de herramientas para análisis estadístico. Como generalmente buscaremos usar otras funcionalidades, se vuelve necesario descargar nuevos paquetes y saber cómo cargarlos en la sesión de trabajo.

### 5.1 Descargar paquetes

La descarga e instalación de paquetes adicionales a la versión básica de R se realiza mediante el comando `install.packages()` indicando entre comillas el nombre del paquete a descargar. Este comando conecta la sesión de R directamente con el CRAN y descarga al computador - en la carpeta de instalación de R - los paquetes requeridos. Por ello, se requiere tener conexión a Internet para efectuar tal operación. En el siguiente ejemplo se descarga el paquete *readxl* que permite abrir bases de datos desde formato Excel en la sesión de R.

#### Ejercicio 5.1

```
install.packages("readxl") #Se descarga e instala el paquete readxl.
```

Una vez ejecutado tal comando saldrán distintos mensajes en la consola de R, generalmente con una apariencia como la que se muestra en la imagen a continuación. Vale destacar que muchas veces salen mensajes en rojo e incluso mensajes de alerta (*warnings*); por lo general se trata de mensajes que el programa muestra al usuario, pero que no implican que algo haya salido mal en la ejecución del comando. Como se aprecia en la siguiente imagen, el software indica cuándo el paquete se descargó e instaló de manera adecuada en el computador, a la vez que la consola queda lista para seguir ejecutando análisis.

### 5.2 Cargar paquetes

Ya se ha indicado que R funciona en la memoria temporal del programa (memoria RAM). Esto hace que cada vez que se abre el programa (vía RStudio) éste se despliega en su versión básica. Es por ello que no basta con descargar al disco duro los paquetes para poder utilizarlos. Para usar una función correspondiente a un paquete adicional a la versión básica de R tal paquete se debe **cargar** en la sesión de R en que se está

```

Console | R Markdown x
~/
> install.packages("readxl")
Installing package into 'C:/Users/felip/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/readxl_1.0.0.zip'
Content type 'application/zip' length 1458126 bytes (1.4 MB)
downloaded 1.4 MB

package 'readxl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\felip\AppData\Local\Temp\RtmpwRLXfn\downloaded_packages
> |

```

Este tipo de mensajes no constituyen errores.

Mensaje que indica que el paquete se revisó e instaló de manera exitosa

Software listo para recibir más instrucciones.

Figura 5.1: Consola indicando instalación exitosa del paquete readxl

```

Console | R Markdown x
~/
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/readxl_1.0.0.zip'
Content type 'application/zip' length 1458126 bytes (1.4 MB)
downloaded 1.4 MB

package 'readxl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\felip\AppData\Local\Temp\RtmpwRLXfn\downloaded_packages
> library(readxl)
Warning message:
package 'readxl' was built under R version 3.4.4
> |

```

Mensaje informativo  
No es un error

Figura 5.2: Consola indicando mensajes de alerta al cargar un paquete

trabajando.<sup>1</sup> Esto se efectúa con el comando `library()` indicando entre paréntesis el nombre del paquete a cargar. A diferencia de la función `install.packages` acá el nombre del paquete no se indica entre comillas.

### Ejercicio 5.2

```
library(readxl) #Carga el paquete descargado a la sesión de trabajo de R.
```

Como se observa en la imagen a continuación, en esta operación el software también puede arrojar mensajes en rojo que no significan que haya ocurrido un error. En este caso es un mensaje de alerta para el usuario (*Warning message*) que informa que el paquete “fue construido para una versión del software menor a la 3.4.4”.

## 5.3 Actualizar la versión básica de R y los paquetes instalados

R está en permanente actualización por lo que luego de algunos meses la versión que haya sido instalada quedará desactualizada. Cada versión busca introducir mejoras, robustecer funciones ya existentes, etc.

Si se precisa actualizar la versión instalada del software una primera opción es repetir las operaciones indicadas en el capítulo 3. No obstante, es posible efectuar una actualización del software mediante la función

<sup>1</sup>No es necesario ejecutar el comando `install.packages` cada vez que se requiera usar un paquete adicional a la versión básica de R. Basta con hacerlo una sola vez pues los paquetes se descargan al disco duro del computador, quedando almacenados allí. Lo que se precisa es **cargarlos** en la sesión temporal de R cada vez que se abra una sesión nueva donde no hayan sido cargados previamente.

`updateR()`. Esta función es parte del paquete `installr` y funciona solamente para el sistema operativo Windows. Como se observa en las siguientes líneas de comando, se trata de un procedimiento simple.

### Ejercicio 5.3

```
install.packages("installr")  
  
library(installr)  
  
updateR()
```

Los diversos paquetes existentes para R son desarrollados a lo largo del mundo por una extensa red de colaboradores. Cuando estos paquetes superan su período de prueba son enviados al *R Core Team* (equipo a cargo de mantener y mejorar el software en sus diferentes versiones) donde son testeados y luego subidos de manera oficial al *CRAN*. Así como el software R, estos paquetes están sujetos a permanentes actualizaciones y mejoras. Por ello, también es preciso conocer alguna forma de actualizar de manera rápida y simultánea todos los paquetes que estén instalados en el disco duro. Para ello, se sugiere utilizar el siguiente comando.

### Ejercicio 5.4

```
update.packages()
```



## Capítulo 6

# Gestión de bases de datos

Un aspecto importante en el uso de RStudio enfocado en el análisis de datos sociales es el manejo de base de datos. Esto puede referir tanto a bases que quien efectúa el análisis haya construido como a bases de datos de estudio sociales realizados por otros.

En este manual de apoyo docente se utilizará una base de datos secundaria. Tanto para enseñar los procedimientos de importación, validación y modificación de datos, como para los posteriores aspectos de análisis estadístico descriptivo.

En los siguientes ejemplos se trabajará con la base de datos de la **Encuesta Nacional de Opinión Pública** del Centro de Estudios Públicos (**Encuesta CEP**, de aquí en más) correspondiente a su versión 81 cuyo trabajo en terreno se efectuó entre Septiembre y Octubre de 2017. Esta encuesta busca caracterizar las actitudes y opiniones, políticas, sociales y económicas de la población chilena, destacando las necesidades, principales preocupaciones y preferencias de todos los habitantes del territorio nacional. Es una de las fuentes de información más importantes para estudiar la opinión pública en Chile, respecto a temas coyunturales (CEP, 2017).

Toda la información de esta encuesta se puede encontrar en el apartado Encuesta CEP de la página institucional del centro de estudios mencionado.<sup>1</sup> Desde este sitio en línea el usuario podrá descargar las bases de datos históricas de esta encuesta, junto con sus manuales de uso metodológico.

### 6.1 Descarga de una base de datos de interés

Una primera acción a realizar es la descarga de la base de datos indicada al computador. Esto puede realizarse accediendo a la página web indicada del CEP. En la botonera superior de esta página se encuentra un botón de acceso a la *Encuesta CEP* (fecha de consulta: 10 de abril 2019).

---

<sup>1</sup>Última fecha de consulta: 15 de abril de 2019

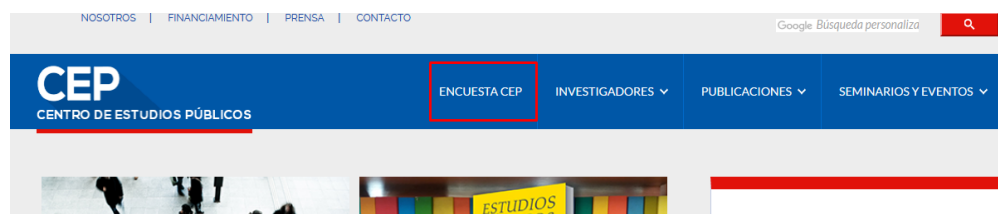


Figura 6.1: Página del CEP

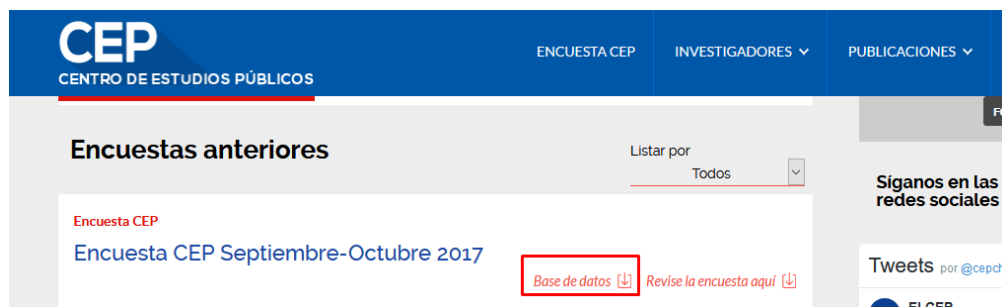


Figura 6.2: Repositorio de bases de datos históricas de la encuesta CEP

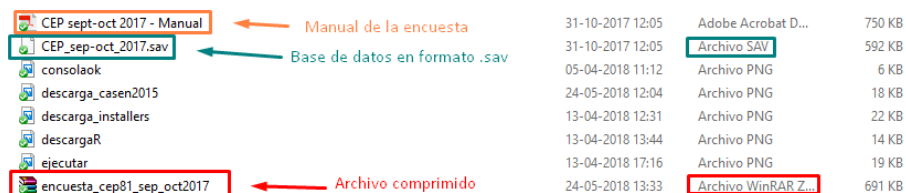


Figura 6.3: Archivos resultantes luego de descomprimir el archivo descargado

Dentro de esa página se debe buscar el apartado *Encuestas anteriores* para encontrar en el repositorio la *Encuesta CEP Septiembre-Octubre 2017*.

Allí, oprimiendo el botón *Base de datos* se podrán descargar los archivos de interés.<sup>2</sup>

Luego de clickear el enlace de descarga se debe guardar el archivo comprimido como el que se ve en la siguiente imagen (con la extensión “.rar”). Para poder *descomprimir* el archivo original basta con tener instalado el programa **WinRAR** (puede descargarse desde esta página). Haciendo clic derecho sobre el archivo descargado, y seleccionando la opción *Extraer aquí* se obtendrá el archivo original (de extensión “.sav”)

Como se ve en la imagen 6.3 el archivo resultante es un archivo de extensión .sav, lo que indica que se trata de un archivo para abrir y trabajar en SPSS. De aquí en más se usará este archivo para desarrollar ejemplos de análisis estadístico. Además de la base de datos, se descarga el manual de uso de la encuesta en formato PDF.<sup>3</sup>

## 6.2 ¿Cómo abrir bases de datos desde formato SPSS?

¿Por qué puede ser importante manejar herramientas que permitan que desde RStudio interactuemos con diferentes formatos de bases de datos? Imagine lo siguiente: durante la formación universitaria usted ha aprendido a usar RStudio como herramienta de análisis estadístico. Pero sucede que al incorporarse a un centro de estudios sociales y observa que el resto de los analistas trabaja fundamentalmente con otro software de análisis estadístico: SPSS. Si usted quiere lograr dialogar con tales especialistas - que muy probablemente no estarán dispuestos a aprender a usar RStudio - deberá lograr al menos abrir bases de datos en formato SPSS, para así poder hacer los análisis que se le encomiende.

Para ello se utiliza un paquete específico llamado **haven** que cuenta con funciones para abrir bases de datos desde diferentes formatos.<sup>4</sup> Para eso primero se descarga e instala el paquete en el computador:

<sup>2</sup>Para los siguientes apartados se trabajará con esta base de datos. El CEP sólo la dispone en formato SPSS (extensión .sav).

<sup>3</sup>Es importante respetar el nombre del archivo para leerlo desde RStudio. En este caso se ha decidido nombrar al archivo de la siguiente forma: *CEP\_sep-oct\_2017.sav*.

<sup>4</sup>Para efectos de este documento tutorial solo se explicará cómo importar bases de datos desde el formato SPSS, que es uno de los más utilizados en ciencias sociales, como también desde archivos generalmente usados vía Microsoft Excel. No obstante, el paquete **haven** también incorpora funciones para importar datos desde el formato para el software Stata, programa más



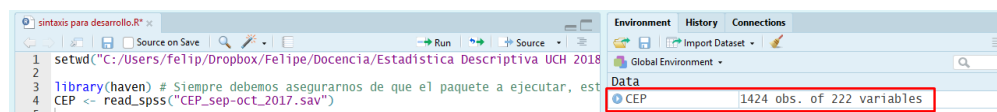


Figura 6.4: Base de datos CEP cargada como objeto en entorno de trabajo

### Ejercicio 6.1

```
install.packages("haven")
```

La función específica a utilizar es `read_spss` cuya ejecución es muy simple. Luego de la función se indica el nombre del archivo a leer entre paréntesis y entre comillas. Como se observa a continuación, esta función sirve para leer la base de datos de la Encuesta CEP descargada en el anterior apartado. Es importante recordar que para usar una fuente de datos en análisis futuros se debe guardar como un “objeto” en el entorno de trabajo. Para eso, la ejecución de la función debe *asignarse* a un objeto nuevo o preexistente mediante la función de asignación. En este caso se asigna a un nuevo objeto llamado “CEP” (CEP <- lectura base de datos):<sup>5</sup>

### Ejercicio 6.2

```
library(haven) #Debemos asegurarnos de que el paquete a ejecutar,
#está cargado en la sesión de Rstudio.

CEP <- read_spss("CEP_sep-oct_2017.sav") # Nombre del archivo entre comillas.
```

Como se observa en la imagen a continuación, luego de tal operación ya se cuenta con un nuevo objeto (*CEP*) en el entorno de trabajo, que almacena la base de datos de la Encuesta CEP.<sup>6</sup>

Con la base de datos cargada en el entorno de trabajo podemos explorar sus principales características. Como se observa en la imagen 6.5 existe un botón (flecha azul indicando hacia abajo) que permite desplegar la estructura interna de la base de datos existente en el entorno de trabajo.

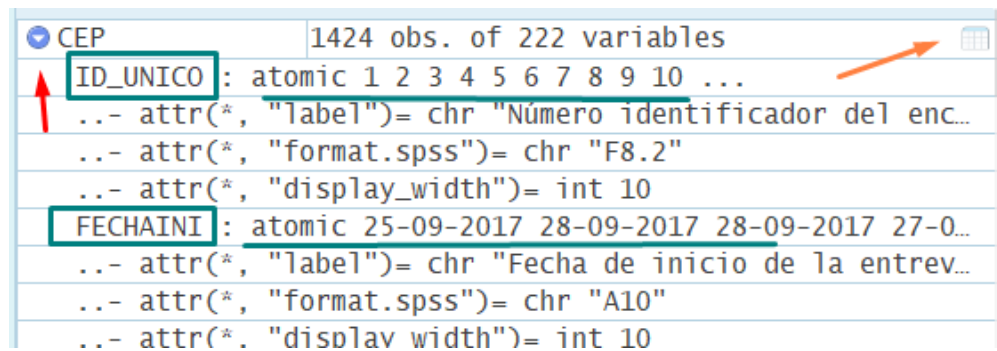
Así se logra explorar rápidamente las principales características de la estructura interna de la base de datos cargada, es decir, los atributos de las diferentes variables en su interior. Entre estas características - a las que se puede acceder con el comando `attributes` - se encuentran las etiquetas, entre otros elementos de formato, que generalmente se observan en la *vista de variables* en SPSS. Si bien en R no se cuenta con una visualización directa de tales elementos, estos se almacenan como atributos en las variables importadas.

Como ya fue indicado en el apartado 4.5 *Construcción de una base de datos* aquí también es posible usar el botón con forma de planilla ubicado a la derecha del objeto en el entorno de trabajo para visualizar la base de datos como una planilla y así poder inspeccionarla visualmente. Esto permitirá explorar rápidamente su estructura y contenidos: observar los nombres de las variables, el tipo de valores que almacenan, etc.

utilizado en el campo de la economía. Estas funciones son similares a los códigos utilizados para trabajar con bases de datos en formato en SPSS. Para más información, leer el siguiente enlace.


<sup>5</sup>Es importante recordar que para que el programa sepa donde buscar los archivos indicados, hay que especificar una “carpeta de trabajo” donde se encuentren los archivos a usar. En este caso, se debe definir una carpeta donde se guarde la base de datos de la Encuesta CEP y la sintaxis donde se irán diseñando los análisis. Para más detalles, repasar el apartado 4.2 *Carpeta de trabajo y memoria temporal del programa*.

<sup>6</sup>Un elemento que se torna relevante al trabajar con más de una base de datos es cómo se nombran los objetos a manipular con R. Como regla general se sugiere trabajar con nombres simples que en el caso de reflejar diferentes momentos en el tiempo, expresen tal orden en los nombres asignados. Por ejemplo, si tuvieramos más de una base de datos correspondiente a diferentes versiones de la Encuesta CEP registradas en 2017, valdría la pena nombrarlas al cargarlas en R con la forma iterativa CEP2017\_1, CEP2017\_2 y así de forma sucesiva, donde el último número representa la posición temporal del estudio dentro del año indicado. Tal formato de nombrar los objetos facilitará enormemente el manejo de los datos al programar análisis con R.



Variable	Class	Label	Format	Display Width
ID_UNICO	atomic	Número identificador del enc...	F8.2	10
FECHAINI	atomic	Fecha de inicio de la entrev...	A10	10

Figura 6.5: Información de los vectores que componen la base de datos



ID_UNICO	FECHAINI	VOTACION_1	VOTACION_2	VOTACION_3	VOTACION_4
1	25-09-2017	10	4	4	4
2	28-09-2017	2	3	3	3
3	28-09-2017	8	5	1	1
4	27-09-2017	7	2	2	2
5	02-10-2017	7	2	2	2
6	29-09-2017	7	2	2	2
7	01-10-2017	4	1	2	2
8	01-10-2017	8	1	1	1
9	03-10-2017	4	1	1	1
10	15-10-2017	7	2	2	2
11	22-09-2017	4	1	1	1
12	29-09-2017	7	2	2	2
13	01-10-2017	8	1	1	1
14	07-10-2017	11	5	1	1

Figura 6.6: Visualización de base CEP como planilla

	A	B	C	D	E	F	G	H
1	ID_UNICO	FECHAINI						
2	1,00	25-09-2017						
3	2,00	28-09-2017						
4	3,00	28-09-2017						
5	4,00	27-09-2017						
6	5,00	02-10-2017						
7	6,00	29-09-2017						
8	7,00	01-10-2017						
9	8,00	01-10-2017						

Figura 6.7: Hoja con identificación de respondientes

### 6.3 ¿Cómo abrir bases de datos desde diferentes formatos de Microsoft Excel?

Otro formato relevante para el trabajo con base de datos es el ecosistema integrado por los diferentes tipos de planillas de cálculo vinculadas al software Microsoft Excel. Muchas veces la digitación de encuestas se efectúa en softwares de estas características, por lo que un formato primario para almacenar bases de datos es, sin duda, las planillas de cálculo. Es más, muchas veces diversos fenómenos sociales son registrados por la actividad humana en este tipo de archivos.<sup>7</sup>

Si bien el formato *planilla de cálculo* generalmente se asocia a Microsoft Excel, debido a la masividad de sus productos, refiere a un formato más general. Hoy en día, también se asocia a aplicaciones en línea como Hojas de Cálculo de Google u otros software de funcionalidad de oficina como *Calc* en sus versiones de Libre Office y Open Office.

A continuación se indican dos modalidades para importar bases de datos en formato de hoja de cálculo a R.

La primera opción es trabajar directamente con un archivo con formato para Microsoft Excel 2007 o superior. Se trata de archivos con una extensión *.xlsx* que tienen un formato de libro, es decir, pueden soportar en su interior a más de una hoja de trabajo. Para efectos caso de este manual se ha convertido la base de datos de la Encuesta CEP trabajada en el apartado anterior a tal formato. Este procedimiento es sencillo si se tiene instalado SPSS.<sup>8</sup> Para efectos de este tutorial el archivo se puede descargar desde el siguiente enlace (indicado en la presentación de este documento), y que lleva por nombre *CEP\_sep-oct\_2017.xlsx*.

Al abrir el archivo desde el explorador de archivos se observa que tiene dos hojas. Una primera almacena el registro de las respuestas con un número de identificación por caso y la fecha de respuesta de la encuesta. La segunda hoja es la base de datos propiamente tal y es la que interesa cargar en el entorno de trabajo de R.

Entonces, ¿cómo cargar una base de datos desde este formato al entorno de trabajo en R? En esta instancia se usará el paquete *readxl* que previamente se debe descargar e instalar en el computador. Específicamente se usará la función *read\_excel* de este paquete en su versión más simple - sin argumentos adicionales - indicando solamente el nombre del archivo a leer. Su resultado se guardará en un nuevo objeto llamado *CEP\_excel*.

#### Ejercicio 6.3

<sup>7</sup>Evidentemente con fines prácticos antes que de investigación. Por ejemplo, procesos contables o de salud, por ejemplo. Estos registros primarios que muchas instituciones realizan - por ejemplo una empresa o una clínica - muchas veces constituyen una interesante fuente de información que puede ser usada para el análisis social.

<sup>8</sup>Basta con ir a *Guardar como* y seleccionar el formato compatible con Excel (*.xlsx*) para obtener una planilla de Excel con la base de datos originalmente guardada en formato *.sav*.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	VOTACION_1	VOTACION_2	VOTACION_3	VOTACION_4	SV_1	SV_2	MB_P1_1	MB_P1_2	MB_P1_3	MB_P2	MB_P3	MB_P4	
2	10	4	4	4	8	3	3	5	7	1	3	4	
3	2	3	3	3	10	5	10	1	2	2	2	4	
4	8	5	1	5	10	10	11	10	7	4	1	4	
5	7	2	2	2	8	5	7	12	6	3	1	3	
6	7	2	2	2	5	5	13	1	11	2	1	2	
7	7	2	2	2	9	5	11	1	5	2	1	3	
8	4	1	2	2	9	5	16	3	1	4	2	3	
9	8	1	1	1	6	8	2	8	1	2	2	2	
10	4	1	1	1	6	7	10	1	7	3	1	3	

Figura 6.8: Hoja con respuestas de encuesta

	ID_UNICO	FECHAINI
1	1	25-09-2017
2	2	28-09-2017
3	3	28-09-2017
4	4	27-09-2017
5	5	02-10-2017
6	6	29-09-2017

Figura 6.9: Hoja de identificación de respondientes cargada como base de datos

```
install.packages("readxl") #Descarga e instalación del paquete
library(readxl) #Cargar paquete en sesión de trabajo de R

CEP_excel <- read_excel("CEP_sep-oct_2017.xlsx") #Leer libro excel
```

Al observar la planilla cargada en el entorno de trabajo se verá que no es la que interesa para desarrollar análisis estadísticos. Por defecto la función lee la primera hoja de del libro de trabajo Excel, por lo que en este caso cargó la planilla con los datos de identificación de cada respondiente, siendo que interesa la lectura de la segunda hoja del libro de trabajo que contiene la base de datos propiamente tal.

Para solucionar este problema se repetirá la operación pero utilizando un argumento extra en la función. Mediante el argumento `sheet` = se le indica al programa la posición o nombre de la hoja que interesa leer en el interior del libro de trabajo. En este caso se indicará que interesa leer la hoja ubicada la posición “2” del libro, o la hoja de nombre “DATOS” (que es lo mismo para efectos de este manual). A continuación se sobrescribe el objeto creado en el entorno de trabajo, actualizando la función con esta información.

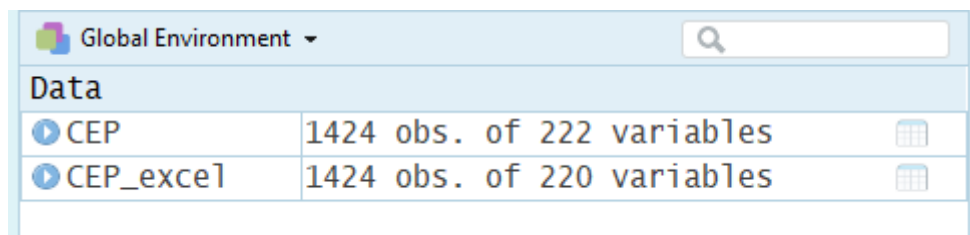
### Ejercicio 6.3 (continuación)

```
CEP_excel <- read_excel("CEP_sep-oct_2017.xlsx", sheet = 2)
#indica posición de hoja en el libro de trabajo.

CEP_excel <- read_excel("CEP_sep-oct_2017.xlsx", sheet = "DATOS")
#indica nombre de hoja en libro de trabajo.
```

Observando el entorno de trabajo se verá que el objeto `CEP_excel` se ha actualizado y ahora presenta 220 variables.<sup>9</sup>

<sup>9</sup>Dos variables menos que las 222 presentes en la base original. Tal diferencia se explica pues manualmente se pasaron a otra



Global Environment		
Data		
CEP	1424 obs. of 222 variables	
CEP_excel	1424 obs. of 220 variables	

Figura 6.10: Hoja con respuestas a encuesta cargada como base de datos

Ahora bien, la función `read_excel` considera a la primera línea de datos como los nombres de las variables de forma automática. Hay veces en que en la primera fila no se encuentra el nombre de las variables habiendo primero otro tipo de información. Por eso resulta importante indicarle a la función desde qué fila comenzar a leer los datos. En el caso de que la información relevante comience en la fila 2, siendo tal fila la que contiene el nombre de las variables se agrega a estos argumentos la opción `skip = 1` para que el software salte u omita la primera fila y comience la lectura de los datos en la fila 2.

### Ejercicio 6.3 (continuación)

```
CEP_excel <- read_excel("CEP_sep-oct_2017.xlsx", sheet = 2, skip = 1)
#indica posición de la hoja en el libro de trabajo.
```

Una segunda forma de importar a R archivos del tipo *hoja de cálculo* es trabajar con el formato CSV (*comma separated values*) o *archivo de valores delimitados por comas*. Se trata de un tipo de archivo más simple que un libro de Microsoft Excel, en la medida que puede contener sólo una - y no varias - hoja de trabajo. Si bien este tipo de archivos pueden encontrarse al descargar una base de datos secundaria, para efectos de este manual de apoyo docente se creará un archivo de este tipo a partir del archivo originalmente almacenado en formato Excel (*.xlsx*).

Para esto, como se observa en las imágenes 6.11 y 6.12, basta con que desde el archivo excel en cuestión, se utilice a la opción *Guardar como* y seleccionando el formato señalado. El programa debería advertir al usuario que el formato seleccionado no soporta múltiples hojas de trabajo (imagen 6.12), por lo que guardará sólo la hoja activa. Asegurando que la planilla que interesa guardar es la hoja que está seleccionada se hace clic sobre la opción *Aceptar*.

El archivo resultante puede ser leído como planilla de Microsoft Excel. Sin embargo, para entender su estructura interna primero se abrirá con la aplicación *Bloc de Notas* mediante la opción *Abrir con....*

Se observa en la imagen 6.14 que el archivo presenta una tabulación del tipo *matriz de datos* - en este caso, la base de datos que contiene las respuestas a una encuesta social - cuyos valores ( cada variable) están separados por el signo punto y coma (;). Este signo delimita cada columna de casos.

Resulta importante considerar esta estructura pues en la notación anglosajona que subyace al lenguaje original de R (el inglés) se usa la coma para separar valores y el punto para denotar valores decimales. En el caso de la notación de habla hispana se emplea la coma para denotar decimales y el punto y coma para separar las observaciones.

Este “detalle” importa pues las funciones básicas para leer archivos CSV viene configuradas por defecto para entender a las comas como separador de los casos. Esto se muestra en el siguiente ejemplo.

### Ejercicio 6.4

---

hoja las dos primeras variables de identificación de las variables.

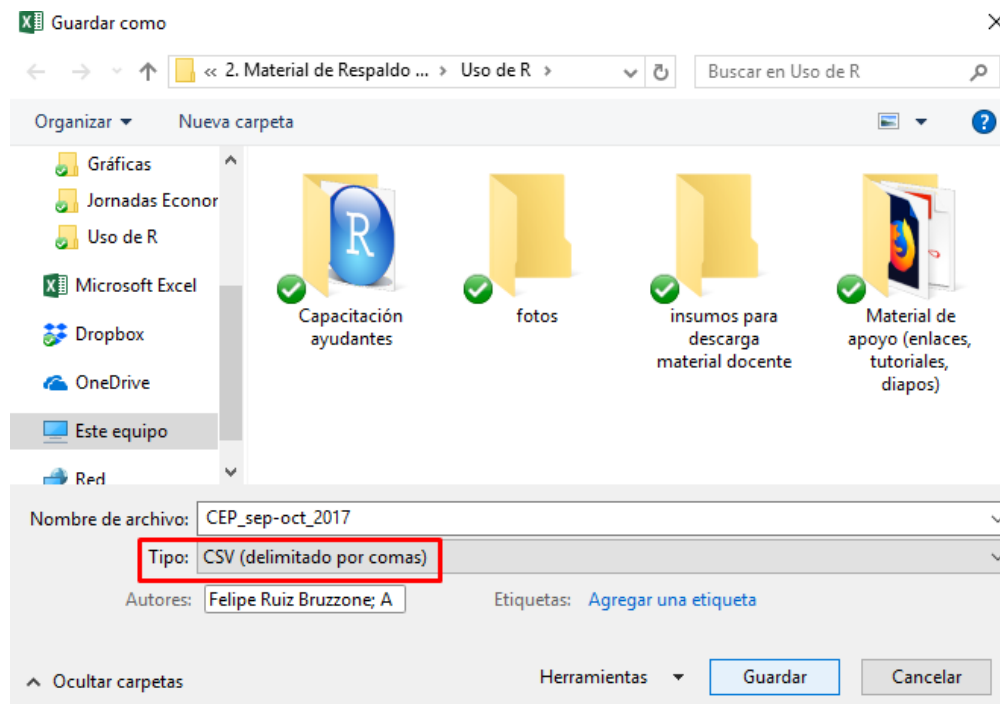


Figura 6.11: Guardar planilla de libro de Excel en formato CSV

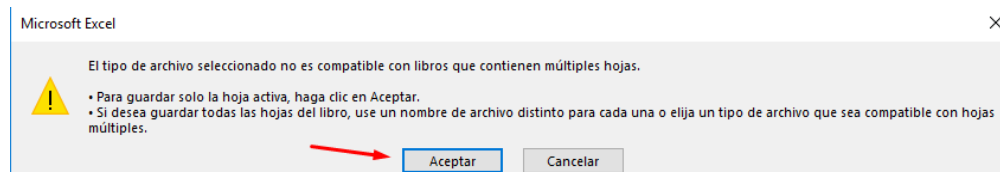


Figura 6.12: Advertencia al guardar en formato CSV

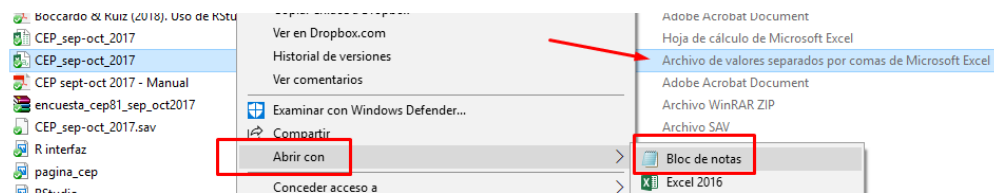


Figura 6.13: Abrir archivo CSV con Bloc de Notas

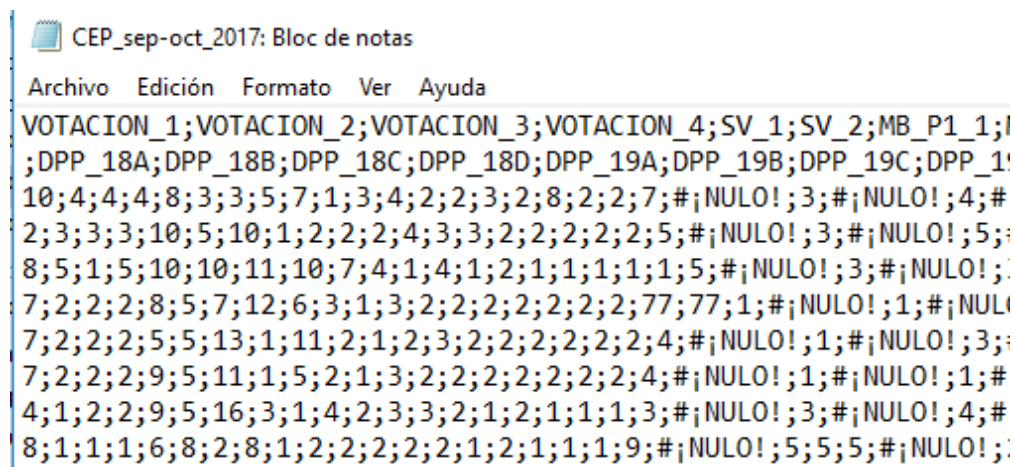


Figura 6.14: Estructura interna del archivo CSV

```

aldo Clases/Uso de R")
> CEP_csv <- read.csv("CEP_sep-oct_2017.csv")
Error in read.table(file = file, header = header, sep = sep, quote = quote, :
  more columns than column names
> CEP_csv <- read.csv("CEP_sep-oct_2017.csv", sep = ";")
> View(CEP_csv)

```

Figura 6.15: Mensaje de error al leer archivo CSV

```
CEP_csv <- read.csv("CEP_sep-oct_2017.csv")
```

En este caso, la ejecución del comando implica un error de ejecución.

El programa avisa que la cantidad de columnas que resultan de la lectura de la planilla es mayor a la cantidad de nombres de variables, por lo que no logra leer el archivo. Esto sucede debido a que las comas presentes en los casos, que denotan valores decimales, el software las ha entendido como separador de casos.

Para solucionar tal problema se indican dos opciones.

- i) La primera es ocupar la misma función, pero agregando un argumento `sep =` mediante el cual se indica qué signo debe considerar para separar los valores.
- ii) La segunda opción es usar una variación de la función original (`read.csv2`), configurada para que considere las comas como notación de decimales y el punto y coma como separador de valores.

#### Ejercicio 6.4 (continuación)

```
CEP_csv <- read.csv("CEP_sep-oct_2017.csv", sep = ";")
CEP_csv2 <- read.csv2("CEP_sep-oct_2017.csv")
```

El resultado muestra que se han leído tres bases de datos coincidentes en términos de estructura, por lo que se ha logrado llegar al mismo resultado usando funciones diferentes.

Para el desarrollo de todos los ejemplos posteriores de este manual se considerará una de las bases de datos leídas. Para ello se “limpiará” por primera vez el entorno de trabajo dejando solamente la base de datos nombrada como `CEP_csv`.

#### Ejercicio 6.5

Global Environment	
Data	
CEP_csv	1424 obs. of 220 variables
CEP_csv2	1424 obs. of 220 variables
CEP_excel	1424 obs. of 220 variables

Figura 6.16: Bases de datos Excel y CSV con la misma estructura de casos y variables

```
remove(CEP_csv2, CEP_excel)
```

## 6.4 Construir una base de datos sólo con variables de interés: exploración de bases de datos y recodificación de variables

Lograr *cargar* una base de datos a la sesión de R es un paso inicial que permite disponer de un conjunto de datos “en bruto” que, muy probablemente se deberán configurar para lograr efectuar los análisis de interés. Se sugiere trabajar solamente con aquellas variables a analizar y crear una nueva base de datos sólo con la información de interés, sin editar la fuente original de datos. Para los siguientes ejercicios se seleccionarán siete variables desde la base de datos de la Encuesta CEP ya mencionada. En la siguiente tabla se indica una descripción de la variable, su nombre en la base original y el nuevo nombre a asignar en la nueva base de datos.<sup>10</sup>

Tabla 6.1: Detalle de las variables específicas a considerar para los análisis

Descripción de la variable	Nombre en base original	Nuevo nombre	Valores	Nivel de medición
Ponderador	POND	<i>pond</i>	Números simples para ponderación.	No aplica.
Género informado	SEXO	<i>sexo</i>	1 = hombre, 2 = mujer.	Nominal.
Región de residencia	REGION	<i>region</i>	Número de región, del 1 al 15.	Nominal.
Fecha de nacimiento (edad)	DS_P2_EXACTA	<i>edad</i>	Edad cumplida en número entero.	Razón.
Satisfacción con la propia vida	SV_1	<i>satisfaccion</i>	Escala del 1 al 10.	Intervalar.
Percepción de satisfacción de chilenos con su vida	SV_2	<i>satisfaccion_chilenos</i>	Escala de 1 al 10	Intervalar.
Evaluación de la situación económica nacional	MB_P2	<i>eval_econ</i>	Escala del 1 al 5	Ordinal.

Para construir una nueva base de datos solo con las variables especificadas será preciso efectuar varias operaciones, que se detallan en este apartado. Para este tipo de manipulación de bases de datos y variables existen diferentes herramientas: algunas forman parte de las funcionalidades básicas de R, pero otras provienen del paquete `dplyr`.<sup>11</sup> En la siguiente tabla se resumen las principales características de las funciones a utilizar.

<sup>10</sup>Para evitar problemas de redacción en los códigos de sintaxis, así como para facilitar la redacción de instrucciones computacionales, se sugiere: i) usar nombres cortos pero descriptivos de las variables, ii) usar sólo mayúsculas o minúsculas, sin combinarlas, iii) evitar usar espacios, privilegiando guiones normales o guiones bajos, sin combinarlos y iv) asignar nombres a los archivos y objetos de forma tal que resulten coherentes para la lectura del computador; por ejemplo, combinar letras y números de forma alfabética o numérica ascendente, de forma tal que se establezca un orden lógico entre los diferentes elementos.

<sup>11</sup>El paquete `dplyr` forma parte de un conjunto más amplio de herramientas integradas en una colección de paquetes de R denominada *tidyverse* (Wickham, 2017), donde se incluyen paquetes especializados en la manipulación y visualización



Tabla 6.2: Funciones de utilidad para la manipulación de bases de datos

Función	Paquete	Utilidad
<i>View</i>	<b>utils</b>	Visualizar un objeto tipo matriz de datos en formato planilla.
<i>names</i>	<b>base</b>	Muestra los nombres de cada elemento incluido en un objeto determinado, por ejemplo, una base de datos.
<i>dim</i>	<b>base</b>	Entrega la dimensionalidad del objeto. En el caso de bases de datos (formato matriz) indica el número de filas (casos) y columnas (variables), en ese orden.
<i>select</i>	<b>dplyr</b>	Seleccionar variables (columnas) específicas de un objeto del tipo <b>data.frame</b> .
<i>rename</i>	<b>dplyr</b>	Renombrar variables dentro de una misma base de datos. Usada como la función <b>select</b> permite seleccionar y al mismo tiempo renombrar variables.
<i>mutate</i>	<b>dplyr</b>	Transformar variables en una nueva, sin alterar la original.
<i>recode</i>	<b>dplyr</b>	Recodificar categorías de una variable, estableciendo una a una las equivalencias entre las categorías originales y las categorías a crear.
<i>recode</i>	<b>car</b>	Recodificar categorías de una variable, permitiendo la recodificación por tramos. De especial utilidad cuando se precisa reducir las categorías de variables de nivel de medición de intervalo o razón, según tramos específicos de respuesta.
<i>save</i>	<b>base</b>	Guardar objetos desde el entorno de trabajo de R al disco duro del computador. Especialmente útil para guardar nuestras bases de datos en formato <i>.RData</i>
<i>load</i>	<b>base</b>	Cargar objetos desde el disco duro a nuestra sesión de trabajo en R. Especialmente útil para cargar bases de datos archivadas en formato <i>.RData</i>
<i>class</i>	<b>base</b>	Informar el tipo de objeto. Permite determinar cómo R ha configurado un conjunto específico de información (una base de datos, una variable en específico, etc.)

Así, de forma específica para la selección de las variables ya señaladas se utilizará la función **select** del paquete **dplyr**, como se observa a continuación:

### Ejercicio 6.6

```
library(dplyr) #Cargar paquete, si no está cargado desde antes.

CEP <- select(CEP_csv, pond = POND, sexo = SEXO, region = REGION, edad = DS_P2_EXACTA,
             satisfaccion_vida = SV_1, satisfaccion_chilenos = SV_2, eval_econ = MB_P2)
#Se indica base de datos, el nombre de variable a crear y los datos que la compondrán.
View(CEP) #Visualización de la base
```

Luego de ejecutar ese comando se habrá creado un nuevo objeto llamado CEP (del tipo *base de datos*) en el entorno de trabajo. Esta base de datos tiene la misma cantidad de casos (1.424) que la base de datos original (CEP\_csv) pero presenta solamente las siete variables seleccionadas y renombradas mediante el comando **select**. Para corroborar el resultado de la construcción de esta nueva base de datos se la visualizará en formato planilla. Como se observa a continuación, esta base de datos contiene solamente las siete variables de interés.

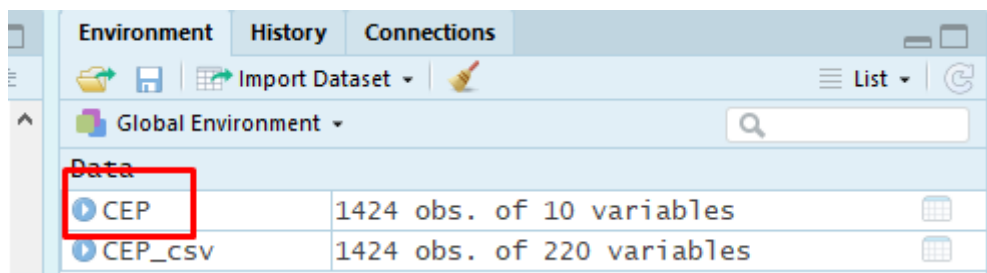


Figura 6.17: Base de datos con selección de variables

	pond	sexo	region	edad	voto	satisfaccion_vida	satisfaccion_chilenos	identificacion_partidos	prob_voto	matri_igualitario
1	1	2	13	18	10	8	3	7	10	7
2	1	2	1	57	2	10	5	5	10	1
3	1	2	14	25	8	10	10	5	1	1
4	1	2	13	37	7	8	5	77	10	10
5	1	2	14	50	7	5	5	4	10	1
6	0	2	8	60	7	9	5	4	10	1
7	1	2	9	66	4	9	5	3	10	1
8	1	2	13	19	8	6	8	9	9	10
9	1	2	7	34	4	6	7	6	10	10
10	1	2	13	39	7	10	10	77	10	10
11	0	2	7	76	4	5	6	2	10	8
12	1	2	6	49	7	10	5	1	10	1
13	1	2	13	32	8	8	5	2	0	4
14	1	2	13	44	11	9	6	77	0	7
15	1	2	14	61	7	8	6	77	10	1
16	1	2	7	40	9	6	7	5	5	3

Figura 6.18: Visualización base de datos construida

Una vez hecha esta operación de selección de variables se habrá creado una nueva base de datos que contiene solo aquellas variables que resultan de interés para los análisis. Antes de proseguir conviene guardar tal objeto como una base de datos de formato R. Para esto se usa el comando `save` para guardar un objeto de formato (o extensión) `.RData`.

### Ejercicio 6.7

```
save(CEP, file = "seleccion_CEP.RData")
```

Este comando creará un nuevo archivo de formato `RData` en la carpeta de trabajo. Este archivo puede usarse para enviar esta base de datos específica a un equipo de trabajo en el contexto de una investigación colectiva, o sencillamente contar con un archivo que ya contenga la base de datos solamente con las variables de interés.

El siguiente paso es explorar las características de tales variables y configurarlas para poder ejecutar los análisis estadísticos que precisemos para nuestro proceso de investigación. Se sugiere dejar en cero el espacio de trabajo utilizando algunas de las formas ya indicadas para hacerlo.

### Ejercicio 6.8

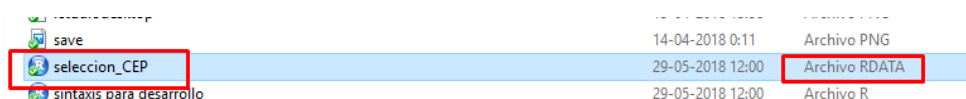


Figura 6.19: Nueva base de datos guardada en carpeta de trabajo

```
rm(list = ls())
```

Para continuar con este manual se sugiere cargar a la sesión de R (desde la carpeta de trabajo) la base construida solo con las variables de interés y guardada en formato RData (`seleccion_CEP.RData`).

### Ejercicio 6.9

```
load("seleccion_CEP.RData")
```

Una primera opción para conocer las características de la base de datos es explorar los nombres de sus variables. Mediante el comando `names` aplicado sobre el objeto `CEP` se obtiene una lista con los nombres de cada columna de la base de datos. Esto resulta de utilidad para los procedimientos de manejo de datos pues no siempre se puede determinar con certeza el nombre de un objeto leyendo el encabezado de la columna. Al visualizarlos como se observa en el siguiente resultado se puede estar seguro del nombre exacto determinando si hay espacios en blanco antes o después de las letras que impliquen un nombre diferente al leído de manera directa.<sup>12</sup>

### Ejercicio 6.10

```
names(CEP)
```

```
## [1] "pond"           "sexo"           "region"
## [4] "edad"           "satisfaccion_vida" "satisfaccion_chilenos"
## [7] "eval_econ"
```

Otro comando de utilidad para conocer características generales de la base de datos es la función `dim`. Este comando, como se observa en el ejercicio 6.11, permite conocer la dimensión de la base de datos que se está explorando. El resultado de esta función arroja dos números: el primer número indica la cantidad de filas de la base de datos, mientras que el segundo número indica la cantidad de columnas. Las bases de datos de estudios sociales están construidas de manera que cada fila representa un caso y cada columna una variable, por lo que la dimensión de una base de datos indicará la cantidad de casos y variables (en ese orden).

### Ejercicio 6.11

```
dim(CEP)
```

```
## [1] 1424    7
```

Finalmente, para conocer las características generales de las variables en la base de datos, y comenzar a evaluar que tipo de recodificaciones se deben realizar, se utiliza la función `summary` para obtener estadísticos descriptivos de cada una de las variables.

### Ejercicio 6.12

---

<sup>12</sup>Al momento de codificar una encuesta, debido a que se trata de un proceso manual, pueden producirse errores humanos. Por ejemplo, la persona que codifica puede haber introducido un espacio en blanco antes o después del nombre de la variable; tanto las planillas de cálculo como la visualización de bases de datos de R no muestran tales espacios en blanco pues sólo visualizan las letras. Es por eso que a simple vista resulta difícil ver si el nombre de una variable es “variable” o “\_variable” (el guión bajo representa la existencia de un espacio en blanco, no se observa así cuando R entrega la información del nombre de la variable). Por otra parte, para prevenir errores de tipeo en la redacción de las sintaxis de análisis, es preferible copiar el nombre de variables u otros objetos directamente desde el listado que resulta al aplicar la función `names`

```
> summary(CEP)
      pond      sexo      region      edad      satisfaccion_vida
Min.   :0.0000  Min.   :1.000  Min.   : 1.00  Min.   :18.00  Min.   : 1.000
1st Qu.:1.0000  1st Qu.:1.000  1st Qu.: 6.00  1st Qu.:36.00  1st Qu.: 6.000
Median :1.0000  Median :2.000  Median : 9.00  Median :50.00  Median : 7.000
Mean   :0.9993  Mean   :1.612  Mean   : 9.16  Mean   :49.87  Mean   : 7.924
3rd Qu.:1.0000  3rd Qu.:2.000  3rd Qu.:13.00  3rd Qu.:64.00  3rd Qu.: 9.000
Max.   :9.0000  Max.   :2.000  Max.   :15.00  Max.   :97.00  Max.   :99.000

satisfaccion_chilenos  eval_econ
Min.   : 1.000  Min.   :1.000
1st Qu.: 4.000  1st Qu.:2.000
Median : 5.000  Median :3.000
Mean   : 8.879  Mean   :2.821
3rd Qu.: 6.000  3rd Qu.:3.000
Max.   :99.000  Max.   :9.000
> |
```

Figura 6.20: Observación general de las variables

```
summary(CEP)
```

¿Por qué resultan de interés estos resultados? Considerando los estadísticos descriptivos construidos se observa que algunas variables contienen códigos que refieren a casos perdidos. Es el caso de las variables `satisfaccion_vida`, `satisfaccion_chilenos` y `eval_econ`. Como se observa en la imagen 6.20, estas variables presentan valores 9 o 99 como máximos, siendo que se trata de variables que presentan un rango menor de valores posibles: la variable `satisfaccion_vida` presenta valores 99 como máximos cuando es una escala de 1 a 10; lo mismo ocurre con la variable `satisfaccion_chilenos`; finalmente, algo similar ocurre con la variable `eval_econ`, que presenta valores 9 como máximos, cuando es una escala del 1 al 5.

Tal información constituye una primera alerta sobre los ajustes a hacer sobre los datos y por tanto son un valioso insumo para el proceso de recodificación de variables.

Para avanzar en los análisis se efectuarán algunas recodificaciones. Primero que todo se observan las características de cada variable a recodificar. Para eso se utiliza las funciones `table` para crear una tabla de frecuencias absoluta y observar los valores de la variable, y `class` para conocer de qué tipo de objeto se trata (en este caso, la variable `sexo`).

### Ejercicio 6.13

```
table(CEP$sexo)
```

```
##
##    1    2
## 553 871
```

```
class(CEP$sexo)
```

```
## [1] "integer"
```

El primer resultado muestra una variable que presenta sólo valores 1 y 2, a la vez que el segundo resultado informa que se trata de un vector de tipo *integer*.

Luego de conocer estas características se cuenta con información suficiente para modificar la variable e incorporarla al análisis. Para recodificar se usará el comando `mutate` del paquete `dplyr`. Este comando permite transformar una variable guardando el resultado de tal operación en una **nueva variable** dentro de la misma base de datos. Resulta útil toda vez que permite editar variables sin perder la información (la variable) original.

Como se observa en las siguientes líneas de comando, el resultado de `mutate` se asigna a la base de datos. Luego de la función `mutate`, los argumentos son: i) el conjunto de datos del que provienen las variables (CEP, nuestra base de datos), ii) el nombre de la nueva variable a crear (`sexo_chr`) seguido de un igual y la operación de transformación de la variable original; en este caso, se emplea el comando `recode` del paquete `dplyr` cuyos argumentos son: la variable que se quiere transformar (`CEP$sexo`) y luego las equivalencias de cada categoría de respuesta. En este caso se indica que “1” se asigna a “hombre” y que “2” se asigna a “mujer”. Vale la pena enfatizar que cada categoría se expresa entre comillas y que cada equivalencia se separa de la siguiente mediante una coma.

#### Ejercicio 6.14

```
CEP <- mutate(CEP, sexo_chr = recode(CEP$sexo, "1" = "hombre", "2" = "mujer"))
table(CEP$sexo_chr)
```

```
##
## hombre  mujer
##      553    871
```

```
class(CEP$sexo_chr)
```

```
## [1] "character"
```

Al ejecutar tal recodificación y solicitar una tabla de frecuencias simple así como la información sobre el tipo de variable creada, ahora la tabla contiene las categorías “hombre” y “mujer”, haciendo más fácil su lectura, mientras que la variable recodificada es del tipo “character”, al contener ahora una codificación basada en letras. Si bien es un ejemplo muy simple, permite entender una primera utilidad de la recodificación de variables, que tiene que ver con facilitar nuestro acercamiento a los datos.

Adicionalmente, se aplicará una segunda recodificación sobre esta variable para convertirla en un vector de tipo factor, aplicando etiquetas para las dos categorías de respuesta (“Hombre” y “Mujer”). ¿Para qué resulta de utilidad esto si ya contamos con resultados entendibles?: pues bien, en el ejercicio anterior al recodificar a valores de tipo *character*, la información numérica almacenada en la variable `CEP$sexo` se perdió al ser reemplazada por letras. Para determinados análisis, precisaremos contar con ambos niveles de información: los valores numéricos asociados a las respuestas, pero también las *etiquetas* de tales valores, que indican lo que representan los números almacenados en la base de datos en términos analíticos.

#### Ejercicio 6.14 (continuación)

```
CEP <- mutate(CEP, sexo_factor = factor(CEP$sexo,
                                       labels = c("Hombre", "Mujer")))
```

Como se observa en el ejercicio 6.14, el comando `factor` permite incorporar las *etiquetas* a cada código de respuesta. En este caso, al ingresar los valores *Hombre* y *Mujer*, en el argumento `labels`, lo que estamos haciendo es asociar las etiquetas indicadas a los valores 1 y 2 en que está codificada la variable.<sup>13</sup>

Ahora se recodificará la variable `región` indicando que sólo habrán dos categorías: “otras regiones” y “región Metropolitana”. Como en el caso anterior, primero conviene observar las características generales de la variable: interesa saber cuantos casos están en la categoría Región Metropolitana para así asegurar que luego de la transformación de la variable, la estructura de casos siga el mismo patrón. Primero exploraremos las características generales de la variable de interés.

<sup>13</sup>Resulta importante indicar que R por lo general funciona con una lógica ascendente, y de izquierda a derecha, en relación a la lectura de información. Por tanto, como los valores de la variable `CEP$sexo` son 1 y 2, al indicar las etiquetas `labels = c("Hombre", "Mujer")` estamos asignando tales etiquetas a los valores, sabiendo que R respetará el orden numérico para efectuar la asignación. Ello se debe tener presente en cualquier operación de recodificación de variables para no errar en lo que buscamos realizar.

## Ejercicio 6.15

```
table(CEP$region) #Observar características de la variables
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##  24   57   24   52  150   82   94  192   98   69    5   17  501   39   20
```

```
class(CEP$region)
```

```
## [1] "integer"
```

En base a estos primeros resultados se observa que se trata de una variable cuyo nivel de medición es nominal. La categoría “13”, que equivale a la región Metropolitana, presenta 501 casos. También sabemos que la variable está configurada como un vector de tipo *integer*. Así, usando el comando `mutate` se recodificará la variable en una nueva denominada `region_factor`; sin embargo, al interior del comando que se utilizaría por defecto, introduciremos una variante.

Si introducimos una función como argumento dentro una función determinada, R siempre aplica las funciones que corresponden al paquete de la función principal inicialmente indicada (en este caso, es el paquete `dplyr`). Para el caso del comando `recode`, el provisto por este paquete no resulta de utilidad para recodificar variables con gran cantidad de categorías, pues no permite realizar una recodificación por tramos de información. Por ello, antes de la función `recode` - que implicaría utilizar aquella incorporada en `dplyr` - incorporamos el argumento `car::`, lo que fuerza a que se ejecute el comando `recode` incluido en un paquete distinto llamado `car`, para así poder recodificar indicando tramos de datos.<sup>14</sup>

Para el caso de la siguiente recodificación se indica el mismo valor para los códigos del 1 al 12, y del 13 al 14, mientras que se asigna un valor diferente para el valor 13 (correspondiente a la región Metropolitana). Primero se recodifica en una nueva variable indicando los tramos de recodificación ya explicados. Posteriormente se sobrescribe la variable asignándole el resultado de la operación de convertirla en una variable de tipo `factor`, aplicando etiquetas para las dos categorías de respuesta ahora existentes.

## Ejercicio 6.15 (continuación)

```
#Transformar a una variable distinta, categorías "Otras regiones" y "Región Metropolitana".
CEP <- mutate(CEP, region_factor = car::recode(CEP$region, "1:12 = 1; 13 = 2; 14:15 = 1"))
```

```
#Sobreescribir variable con resultado de convertir a factor incorporando etiquetas
CEP$region_factor <- factor(CEP$region_factor,
                             labels = c("Otras regiones", "Región Metropolitana"))
```

```
table(CEP$region_factor) #Se sigue manteniendo la estructura de casos.
```

```
##
##      Otras regiones  Región Metropolitana
##             923                501
```

<sup>14</sup>Esto resulta de utilidad en variables de nivel de medición nominal u ordinal con muchas categorías, pero especialmente en variables de nivel de medición intervalar o de razón como ingreso o edad. Dado que presentan un rango muy amplio de valores resulta mucho más sencillo recodificar indicando tramos de valores. En el caso de que los últimos tramos sean valores perdidos, la función construida puede incorporar un último tramo indicando el argumento `else` y asignando el valor lógico usado por R para denotar casos perdidos (NA): `CEP <- mutate(CEP, region_factor = car::recode(CEP$region, "1:12 = 1; 13 = 2; 14:15 = 1; else = NA"))`. Veremos una aplicación concreta de esto en la siguiente recodificación.

```
class(CEP$region_factor) #Cambia el formato del objeto.
```

```
## [1] "factor"
```

El resultado muestra que, por un lado, se mantuvo la estructura de casos, con 501 casos en la región Metropolitana y 923 casos en otras regiones. Eso indica que la recodificación se hizo de manera adecuada. Además, la tabla muestra etiquetas y al solicitar el formato de la variable indica que es un vector de tipo *factor*. En síntesis, la recodificación se efectuó de manera óptima.

Ahora se recodificarán las variables *satisfacción con la vida propia* y percepción de *satisfacción que los chilenos en general sienten con su propia vida*. Primero se exploran ambas variables.

### Ejercicio 6.16

```
#Explorar variable "satisfacción con la propia vida"
class(CEP$satisfaccion_vida)
```

```
## [1] "integer"
```

```
table(CEP$satisfaccion_vida)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 88 99
## 20 10 30 63 176 169 256 244 142 304  4  6
```

```
summary(CEP$satisfaccion_vida)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.000   6.000   7.000   7.924   9.000  99.000
```

El análisis exploratorio de esta primera variable <sup>15</sup> permite identificar que la variable está compuesta por valores discretos entre 1 y 10, aunque se observa también valores 88 y 99 en la distribución que - debido a la lectura del cuestionario y la ficha técnica de la encuesta - se sabe que denotan las categorías “no sabe” y “no responde”.<sup>16</sup> Es posible afirmar entonces que el nivel de medición de esta variables de tipo intervalar, pues incluye más de 7 categorías de respuesta.

```
#Explorar variable "percepción de la satisfacción que los chilenos tienen con su vida"
class(CEP$satisfaccion_chilenos)
```

```
## [1] "integer"
```

<sup>15</sup>En este caso, se indican dos maneras de explorar los valores de la variable. Una opción es el comando `table` que ya ha sido utilizado con anterioridad. Sin embargo, otra opción que sirve es el comando `summary`, mediante el cual se construye un resumen de estadísticos descriptivos. No obstante, esta última opción puede resultar poco clara, ya que impide observar el valor 88 presente en la distribución, pues solamente muestra el valor máximo de la distribución de datos de la variable (99).

<sup>16</sup>En este manual no se profundizará de forma específica en los dilemas que implica el tratamiento de casos perdidos. No obstante, se debe considerar que la decisión de excluir los valores perdidos de los análisis no será siempre la primera opción de trabajo por dos razones: la primera tiene que ver con que la eliminación de tales datos puede provocar sesgos en nuestros resultados si es que las respuestas no se distribuían de forma aleatoria, mientras que la segunda tiene que ver con la potencial de reducción de casos alterando los diseños muestrales que permiten la inferencia estadística. En la actualidad existen diversas herramientas para imputar casos que deben considerarse sobre todo cuando la distribución de las no respuestas no es aleatoria, sino que responde a un patrón o sesgo presente en nuestros datos. Para profundizar en este asunto sugerimos una reciente publicación sobre el tema (van Buuren, 2018), disponible en línea.

```
table(CEP$satisfaccion_chilenos)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 88 99
## 33 21 97 216 469 241 160 56 24 47 52 8
```

```
summary(CEP$satisfaccion_chilenos)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  1.000   4.000   5.000   8.879   6.000   99.000
```

Algo similar ocurre con esta segunda variable. Al igual que la anterior, presenta valores 88 y 99 como casos válidos, lo que altera la distribución de casos ceñida a la escala de 1 a 10. Tal como en la variable anterior, también es posible afirmar - dado que presenta más de 7 categorías - que el nivel de medición de esta variable es de intervalo.

Por tanto, se efectuará una operación de transformación de estas variables, asignando el valor lógico NA a aquellos valores que representan casos perdidos (sin respuesta). Este valor lógico servirá pues R lo reconoce como un valor que no puede utilizarse para operaciones matemáticas, a la vez que no representa un valor alfanumérico contable, por lo que no alterará el tipo de vector (en este caso, numérico de tipo *integer*). Entonces se ejecutan las siguientes funciones, para cada variable:<sup>17</sup>

### Ejercicio 6.16 (continuación)

```
CEP$satisfaccion_vida[CEP$satisfaccion_vida==88]<- NA #Asignar NA a casos con valor 88
CEP$satisfaccion_vida[CEP$satisfaccion_vida==99]<- NA #Asignar NA a casos con valor 99

CEP$satisfaccion_chilenos[CEP$satisfaccion_chilenos==88]<- NA
CEP$satisfaccion_chilenos[CEP$satisfaccion_chilenos==99]<- NA
```

Habiendo ejecutado tales instrucciones, R habrá cambiado los valores 88 y 99 de cada variable por el valor lógico NA. Para evaluar si la transformación de datos resultó en el sentido esperado, a continuación se solicita una tabla de frecuencias y estadísticos de resumen para cada variable.

```
table(CEP$satisfaccion_vida) #Ver resultado de codificación
```

```
##
##  1  2  3  4  5  6  7  8  9 10
## 20 10 30 63 176 169 256 244 142 304
```

```
summary(CEP$satisfaccion_vida)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##  1.000   6.000   7.000   7.311   9.000  10.000     10
```

<sup>17</sup>Como se observa en el siguiente código de R, el comando para codificar valores como NA en la misma variable tiene la siguiente estructura: se indica la variable (base/variable). luego entre corchetes se indica un valor específico dentro de la misma, eso se hace escribiendo nuevamente la variable seguida de dos signos igual y el valor de interés ([base\$variable==88]). Con eso se indica qué valor se busca; luego, con el asignador se indica que a tal selección se asigna el valor NA (<- NA).



```
table(CEP$satisfaccion_chilenos)
```

```
##
##    1    2    3    4    5    6    7    8    9   10
##  33   21   97  216  469  241  160   56   24   47
```

```
summary(CEP$satisfaccion_chilenos)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.000   4.000   5.000   5.334   6.000   10.000     60
```

Como se observa en los resultados, si luego de la recodificación se solicitan tablas de frecuencias para ambas variables, estas ya no mostrarán los casos *perdidos* (88 y 99). Y si se solicitan estadísticos descriptivos, se observará que ahora los valores mínimo y máximo de la variable en la base de datos, coinciden con los valores mínimos y máximos del rango de la variable discernible desde el cuestionario y ficha técnica del estudio, indicándose además la cantidad de casos NA que han resultado excluidos de los análisis.

Ahora se explorará la variable que mide la percepción de la persona entrevistada respecto a la situación económica del país. Como se observa en el ejercicio 6.17, se utiliza el comando `class`, `table` y `summary` para efectuar una aproximación exploratoria a la variable.

### Ejercicio 6.17

```
#Explorar variable "evaluación de la economía"
class(CEP$eval_econ)
```

```
## [1] "integer"
```

```
table(CEP$eval_econ)
```

```
##
##    1    2    3    4    5    8    9
##   74  397  730  204    5    8    6
```

```
summary(CEP$eval_econ)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.000   2.000   3.000   2.821   3.000   9.000
```

Como se observa en los resultados, la variable es una *escala de acuerdo* compuesta por valores del 1 al 5,<sup>18</sup> aunque también se observan valores 88 y 99 en la distribución que gracias a la información disponible en el cuestionario y ficha técnica de este estudio, sabemos que representan a las categorías “no sabe” y “no contesta” (*casos perdidos*). Así, es posible afirmar que se trata de una variable de nivel de medición ordinal.

Para efectos de simplificar las respuestas para su análisis, esta variable se recodificará en tres categorías de evaluación (negativa, neutra y positiva). Asimismo, en el ejercicio 6.17 también se indicará una segunda forma de asignar casos perdidos (valores NA) esta vez desde una recodificación vía comando `mutate`.

la variable está compuesta por valores discretos entre 1 y 10, aunque se observa también valores 88 y 99 en la distribución que - debido a la lectura del cuestionario y la ficha técnica de la encuesta - se sabe que denotan las categorías “no sabe” y “no responde”. Es posible afirmar entonces que el nivel de medición de esta variables de tipo intervalar, pues incluye más de 7 categorías de respuesta.

<sup>18</sup>Los valores originales son “1 = Muy mala”, “2 = Mala” “3 = Ni buena ni mala”, “4 = buena” y “5 = Muy buena”.

**Ejercicio 6.17 (continuación)**

```
#Recodificar variable a 3 tramos: positiva, neutra, negativa

#Valores perdidos se asignan en la misma codificación,
#con argumento else ("todos los demás valores")
CEP <- mutate(CEP, eval_econ_factor =
               car::recode(CEP$eval_econ,
                           "1:2 = 1; 3 = 2; 4:5 = 3; else = NA"))

CEP$eval_econ_factor <- factor(CEP$eval_econ_factor,
                              labels = c("Positiva", "Neutra", "Negativa"))

table(CEP$eval_econ_factor)
```

```
##
## Positiva   Neutra Negativa
##      471      730      209
```

```
summary(CEP$eval_econ_factor)
```

```
## Positiva   Neutra Negativa   NA's
##      471      730      209      14
```

Habiendo efectuado estas transformaciones sobre las variables originales, ya se cuenta con variables de los diferentes tipos para efectuar análisis de estadística univariada: variables cuyo nivel de medición es nominal, ordinal, intervalar y de razón. En el siguiente capítulo se verá como ejecutar análisis estadísticos univariados, en sus variantes de alcance muestral y poblacional.

## Capítulo 7

# Estadística descriptiva con RStudio

Habiendo construido y configurado una base de datos *ad hoc* para nuestros análisis, en el presente capítulo se presenta un modo de calcular frecuencias, medidas de tendencia central, de posición y de dispersión a nivel de la estimación puntual y del parámetro poblacional.<sup>1</sup>

Para efectos de este manual distinguiremos aquellos estadísticos que se calculan desde una muestra (probabilística o no probabilística) de aquellos que se obtienen vía registros administrativos o de censos. Específicamente, en el caso de estadísticos que se estiman a partir de una muestra probabilística avanzaremos desde la estimación puntual al cálculo del parámetro poblacional (con su respectivo intervalo de confianza, nivel de confianza y error de estimación). Esto último nos permitirá estimar los valores que la estimación puntual alcanza en la población (que representa la muestra) y si existen diferencias estadísticamente significativas a nivel del parámetro entre dos grupos (prueba de hipótesis).

El presente capítulo se ordena como sigue: en el primer apartado se indicará cómo realizar la estimación puntual de diferentes estadísticos descriptivos usando R (medidas de tendencia central, distribuciones de frecuencia y medidas de posición, medidas de dispersión y forma de distribuciones) y su exportación a planillas de cálculo para su presentación en reportes de investigación académica o profesional; en el segundo apartado se presentará la forma de realizar inferencia estadística, construyendo intervalos de confianza de parámetros, a partir de estimaciones puntuales, y calculando indicadores insesgados a partir de factores de expansión.

## 7.1 Estimación puntual de estadísticos descriptivos usando R

### 7.1.1 Medidas de tendencia central: cálculo de la media, mediana y moda

El cálculo de la media es sencillo en cuanto a los códigos y aritmética necesarios. En las dos siguientes líneas de comandos se observa el cálculo de una media simple y luego de una media recortada. En ambos casos se utiliza el comando `mean`, indicando la variable sobre la cual se desea hacer el cálculo, acompañado del argumento `na.rm = TRUE` para excluir del cálculo a los casos dados como perdidos (valores NA que ya fueron codificados al preparar las variables). Para el caso de la media recortada, se agrega el argumento `trim`, que permite indicar la proporción de casos que se eliminan en cada extremo de la distribución.<sup>2</sup> Así, se obtiene la estimación puntual de ambos estadísticos.

<sup>1</sup>En este apartado no se profundizará en métodos avanzados para optimizar la presentación de resultados: es importante distinguir entre la construcción de resultados estadísticos para su uso en el proceso de investigación, respecto a la construcción de informes y resultados con un formato adecuado para su divulgación. Éste último aspecto lo revisaremos en los dos últimos capítulos de este documento, enfocados en el uso de los paquetes *ggplot2* y las funcionalidades para la construcción de informes vía *RMarkdown*.

<sup>2</sup>Es por eso que si se busca contar con una media recortada al 5%, se indica una proporción de 0.025, que se recortará a ambos extremos de la distribución, alcanzando un recorte total de una proporción de casos del 0.05 o 5%.

### Ejercicio 7.1

```
mean(CEP$satisfaccion_vida, na.rm = TRUE)
```

```
## [1] 7.311174
```

```
mean(CEP$satisfaccion_vida, na.rm = TRUE, trim = 0.025)
```

```
## [1] 7.390625
```

Como se observa en el ejercicio 7.1 la media aritmética arroja un valor de 7,31 en una una escala de 1 a 10, mientras la media recortada al 5% presenta un valor de 7,39 (ambos valores se redondean al segundo decimal). Como fue anticipado, la interpretación de estos resultados (y sus diferencias) se explorarán en el siguiente apartado.

Para el caso de la mediana, se trata de un comando similar. Se utiliza el comando `median`, indicando la variable sobre la cual se desea hacer el cálculo, acompañado del argumento `na.rm = TRUE` para excluir del cálculo a los casos dados como perdidos (valores NA que ya fueron codificados al preparar las variables). De esa forma se obtiene la estimación puntual de este estadístico.

### Ejercicio 7.2

```
median(CEP$satisfaccion_vida, na.rm = TRUE)
```

```
## [1] 7
```

Como se ve en el ejercicio 7.2, el valor de la mediana es 7, lo que indica que la mitad de los casos indica tener una satisfacción con la propia vida menor a 7, considerando una escala de 1 a 10.

Para el cálculo de la moda, debe haberse instalado previamente el paquete `modeest` y haberlo cargado en la sesión de trabajo; esto permitirá contar con una función que calculará de forma automática el o los valores más frecuentes de la distribución, herramienta que no existe en los paquetes básicos de R. Así, como se ve en el ejercicio 7.3, se utiliza el comando `mfv`<sup>3</sup> para calcular la moda de la variable `edad`. El resultado de este comando indica el valor, o los valores, con más frecuencia dentro de la distribución de casos de la variable señalada.

### Ejercicio 7.3

```
#Ejecutar previamente "install.packages("modeest")"
library(modeest)
mfv(CEP$edad) #Indica el o los valores con más frecuencia
```

```
## [1] 50
```

En este caso, la moda de la variable `edad` es 50 años, lo que indica que el valor que más se repite en la distribución de casos de tal variable corresponde a tal respuesta.

<sup>3</sup>Sigla en inglés correspondiente a la expresión *most frequent values* o *valores más frecuentes* en castellano.

### 7.1.2 Medidas de posición: cálculo de frecuencias absolutas y relativas, cuantiles

El cálculo de tablas de frecuencias absolutas para una variable se efectúa mediante el comando `table`, indicando como argumento de la función la variable sobre la cual se ejecuta el cálculo. En este sub apartado, trabajaremos con la variable ordinal de nuestra base de datos que refiere a una *evaluación de la economía* (CEP\$eval\_econ\_factor) por parte de la persona encuestada.

#### Ejercicio 7.4

```
tabla <- table(CEP$eval_econ_factor)
tabla
```

```
##
## Positiva   Neutra Negativa
##      471      730      209
```

De forma muy sucinta, los resultados del ejercicio 7.4 muestran un importante predominio de opiniones “neutras” con 730 casos, seguida por una concentración de 471 casos en respuestas “positivas” y 209 casos en respuestas “negativas”.

Para facilitar la comparación con otros datos, estos resultados pueden expresarse como frecuencias relativas. Para el cálculo de frecuencias relativas, se usa el objeto tabla de frecuencias simples ya construido (`tabla`): como se ve en el ejercicio 7.5, sobre tal objeto, se ejecuta la función `prop.table` que construye una nueva tabla en la que cada celda es la división simple entre la cantidad de casos de la categoría y el total de casos.

#### Ejercicio 7.5

```
#Frecuencias relativas (proporciones)
prop.table(tabla)
```

```
##
## Positiva   Neutra Negativa
## 0.3340426 0.5177305 0.1482270
```

De tal forma, los resultados del ejercicio 7.5 confirman lo ya observado con frecuencias absolutas. En este caso, los valores se presentan como proporciones de una unidad (1). Así, se observa que la categoría “neutra” concentra la mayor cantidad de casos con aproximadamente la mitad de los mismos (0,52), mientras que la categoría “positiva” concentra un tercio de las respuestas (0,33).

Ahora bien, para facilitar aún más la lectura y divulgación de estos resultados resultará de utilidad convertir estas frecuencias relativas en porcentajes. Como se observa en el ejercicio 7.6, para construir una tabla de porcentajes se multiplica el resultado de calcular la tabla de proporciones por 100; si a ello además se le agrega la función `round`, es posible configurar un resultado redondeado a dos decimales.

#### Ejercicio 7.6

```
#Frecuencias relativas (porcentajes)
prop.table(tabla)*100
```

```
##
## Positiva   Neutra Negativa
## 33.40426 51.77305 14.82270
```

```
round((prop.table(tabla)*100),2)
```

```
##
## Positiva   Neutra Negativa
##    33.40    51.77    14.82
```

De tal modo, los resultados del ejercicio 7.6 permiten afirmar que la mayor cantidad de respuestas se concentran en una evaluación “neutra” con el 51,8% del total, seguido por evaluaciones positivas que representan un 33,4% del total de respuestas, y finalmente las respuestas que reflejan evaluaciones negativas, que alcanzan un 14,8% de los casos.

Una última configuración que resulta de utilidad para el examen de distribuciones de frecuencias es la modalidad de cantidades acumuladas. Este tipo de tablas se construye a partir de tablas de frecuencias absolutas y relativas (sea en su modalidad proporcional o porcentual). En el ejercicio 7.7 se aplica la función `cumsum` a las diferentes modalidades de tablas de frecuencias construidas a partir del objeto `tabla`, que almacena una tabla de frecuencias absolutas. Esta función (`cumsum`) permite construir tablas de frecuencias acumuladas, a partir del cálculo de cada distribución de frecuencias ya construidas en los ejercicios anteriores.

### Ejercicio 7.7

```
#Frecuencias absolutas acumuladas
cumsum(tabla)
```

```
## Positiva   Neutra Negativa
##    471     1201    1410
```

```
#Frecuencias relativas acumuladas
cumsum(prop.table(tabla))
```

```
## Positiva   Neutra Negativa
## 0.3340426 0.8517730 1.0000000
```

```
#Porcentaje acumulado redondado en dos decimales
round(cumsum(prop.table(tabla)*100),2)
```

```
## Positiva   Neutra Negativa
##    33.40    85.18   100.00
```

Así los resultados del ejercicio 7.7, permiten observar que una amplia mayoría de los casos presenta una evaluación neutra o positiva de la economía nacional con el 85,2% de los casos (en términos relativos), lo que equivale a 1.201 casos en términos absolutos. Sólo un 14,8% de las respuestas se concentran en una evaluación negativa de la economía.

Una última forma de describir la concentración de casos de una distribución es mediante los cuantiles. Tales estadísticos permiten describir la concentración de casos según cualquier posición relativa de los datos en la distribución, que resulte de interés. Para el cálculo de cuantiles la función `quantile` permite calcular los casos equivalentes a diferentes proporciones de la distribución (ejercicio 7.8). El primer argumento de la función es la variable a considerar, en este caso `CEP$satisfaccion_chilenos`, que refiere a la evaluación que la persona cree los chilenos tiene de su vida, en una escala de 1 a 10; luego, mediante el argumento `prob` se indica en formato vector los cuantiles a calcular (expresados como proporción). Si existen casos perdidos (codificados como NA) se debe indicar el argumento `na.rm = TRUE`.

## Ejercicio 7.8

```
quantile(CEP$satisfaccion_chilenos, prob = c(0.25, 0.5, 0.75), na.rm = TRUE)
```

```
## 25% 50% 75%
##   4   5   6
```

De tal forma, en el ejercicio 7.8 se solicita al software el cálculo del cuartil 1 (percentil 25 o caso que corta el 25% de la distribución), el cuartil 2 (percentil 50, mediana o caso que corta el 50% de la distribución) y el cuartil 3 (percentil 75, o caso que corta el 75% de la distribución). Los resultados obtenidos en el ejercicio 7.6 permiten concluir que solamente el 25% superior de los casos cree que los chilenos tienen una evaluación de su vida mayor o igual a 6, mientras que la mitad de los casos cree que los chilenos tienen una evaluación de su vida igual o menor a 5, en una escala del 1 al 10.

### 7.1.3 Medidas de dispersión: rango, varianza, desviación estándar y coeficiente de variación

En relación a las medidas de dispersión se partirá por cálculo del rango. Los valores mínimo y máximo de la distribución pueden calcularse de manera simultánea con la función `range`, indicando como argumentos la variable de interés y adicionando también el argumento `na.rm = TRUE` en el caso de que hubieran sido codificados como NA los valores perdidos. Como se observa en el ejercicio 7.9, los valores mínimo y máximo pueden calcularse de forma independiente, con las funciones `min` y `max` respectivamente, que siguen la misma lógica que la función `range`. En este caso se trabajará sobre la variable `CEP$edad`.

## Ejercicio 7.9

```
range(CEP$edad, na.rm = TRUE)
```

```
## [1] 18 97
```

```
min(CEP$edad, na.rm = TRUE)
```

```
## [1] 18
```

```
max(CEP$edad, na.rm = TRUE)
```

```
## [1] 97
```

```
max(CEP$edad, na.rm = TRUE) - min(CEP$edad, na.rm = TRUE)
```

```
## [1] 79
```

Luego, para conocer el rango de los valores se debe restar el valor máximo al mínimo. Los resultados del ejercicio 7.9 indican que, para el caso de la variable `edad` el rango es de 79 años, siendo su valor mínimo 18 años y su valor máximo 97 años.

Otras dos medidas de uso generalizado para describir la dispersión de una variable son la varianza y la desviación estándar. El cálculo de ambas medidas, sigue la misma lógica en lo que refiere a la estructura de la función utilizada. Respectivamente, las funciones utilizadas son `var` para varianza y `sd`, para desviación estándar. Como se observa en el ejercicio 7.10 también debe incluirse el argumento `na.rm = TRUE` en el caso de que hubieran sido codificados como NA los valores perdidos. En este ejercicio se trabaja sobre la variable `CEP$satisfaccion_chilenos`, que refiere a la evaluación que la persona cree los chilenos tienen de su vida, en una escala de 1 a 10.

**Ejercicio 7.10**

```
var(CEP$satisfaccion_chilenos, na.rm = TRUE)
```

```
## [1] 3.017771
```

```
sd(CEP$satisfaccion_chilenos, na.rm = TRUE)
```

```
## [1] 1.737173
```

Atendiendo a los resultados del ejercicio 7.10 puede observarse que la dispersión de la variable que representa la *evaluación que cada entrevistado/a cree que los chilenos tienen sobre su propia vida* es de 3,02 unidades atendiendo a la varianza y de 1,73 unidades atendiendo a la desviación estándar.

Otra medida que debemos aprender a calcular para nuestros análisis es el coeficiente de variación. Como se observa en el siguiente ejercicio (7.11), el coeficiente de variación puede calcularse de manera sencilla, ejecutando una división simple entre la desviación estándar y la media de la variable de interés. Sin embargo, también existe la función `coefficient.variation` (del paquete `FinCal`) que permite calcular este estadístico indicando como argumento la media y desviación estándar de la variable de interés. En este caso tales valores se incluyen indicando la función para calcularlos: de tal modo se evitan errores humanos al momento de especificar los valores, ya sea por tipeo del valor o por su aproximación decimal. En este ejercicio trabajamos con la variable `edad`.

**Ejercicio 7.11**

```
sd(CEP$edad)/mean(CEP$edad)
```

```
## [1] 0.3566407
```

```
#Asegurarse de ejecutar previamente el comando "install.packages("FinCal")"
library(FinCal)
coefficient.variation(sd=sd(CEP$edad), avg = mean(CEP$edad))
```

```
## [1] 0.3566407
```

De este modo, los resultados del ejercicio 7.11 muestran que la variable `edad` presenta una dispersión del 35,67% según el valor obtenido en el coeficiente de variación.<sup>4</sup>

**7.1.4 Forma de una distribución: simetría, curtosis y normalidad**

Un último grupo de estadísticos que resulta de importancia para el análisis estadístico de variables refiere a aquellos que dan cuenta de la forma general que asume la distribución de una variable. En específico, mostraremos en este apartado el cálculo e interpretación de estadísticos que dan cuenta de la simetría, curtosis y normalidad de una variable.

En el siguiente ejercicio (7.12) se muestra cómo calcular los coeficientes de simetría y curtosis mediante las funciones `skew` y `kurtosi` presentes en el paquete `psych`.

**Ejercicio 7.12**


---

<sup>4</sup>En el resultado del ejercicio se obtiene un valor proporcional, es decir una fracción de 1. Cuando expresamos el resultado final en formato de porcentaje, simplemente hemos multiplicado tal resultado por 100.



```
library(psych)
skew(CEP$satisfaccion_vida)
```

```
## [1] -0.5335721
```

```
kurtosi(CEP$satisfaccion_vida)
```

```
## [1] -0.1461258
```

Como se observa en los resultados del ejercicio 7.12, los coeficientes de simetría y curtosis para la variable `satisfaccion_vida` son de -0,53 y -0.15 unidades de forma respectiva. Dado que tales valores están expresados en la unidad de medida de la variable en cuestión, no son útiles para comparar variables de diferentes unidades de medición. Es por eso que en el ejercicio 7.13 se calculan los coeficientes de simetría y curtosis estandarizados: para ello, se divide la simetría calculada, por la raíz cuadrada del valor 6 (este número es una constante en la fórmula) dividido en la cantidad de casos de la variable (en este caso, 1.401).

### Ejercicio 7.13

```
skew(CEP$satisfaccion_vida)/sqrt(6/1401)
```

```
## [1] -8.153359
```

```
kurtosi(CEP$satisfaccion_vida)/sqrt(6/1401)
```

```
## [1] -2.232906
```

Un criterio general para determinar si los coeficientes de simetría y curtosis reflejan una variable semejante a una distribución normal es que ambos valores se encuentren entre -2 y 2. Como puede observarse en los resultados, ambos coeficientes escapan a tal rango por lo que se observa una distribución poco similar a una normal.

Finalmente, la forma más certera para evaluar si la distribución de datos de una variable se comporta según los parámetros que asume una distribución normal es aplicando un test estadístico específicamente orientado a tal evaluación. Para ello, se diferencia el test de Shapiro Wilk y el de Kolmogorov Smirnov. El primero se adecua a muestras pequeñas (menores a 50 casos), mientras que el segundo sirve para muestras de entre 50 y 1.000 casos.

El ejercicio 7.14 muestra la aplicación de tales pruebas sobre la variable `edad`. En específico, se utilizan las funciones `shapiro.test` y `ks.test` del paquete básico de R (`stats`). Su ejecución es bastante sencilla pues sólo requieren como argumento la variable sobre la cual se aplicará la prueba.

### Ejercicio 7.14

```
#Prueba de Shapiro Wilk (muestras pequeñas)
shapiro.test(CEP$edad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CEP$edad
## W = 0.97588, p-value = 1.037e-14
```

```
#Prueba Kolmogorov Smirnov (muestras grandes)
ks.test(CEP$edad, "pnorm", mean(CEP$edad, na.rm=T), sd(CEP$edad, na.rm=T))

## Warning in ks.test(CEP$edad, "pnorm", mean(CEP$edad, na.rm = T),
## sd(CEP$edad, : ties should not be present for the Kolmogorov-Smirnov test

##
## One-sample Kolmogorov-Smirnov test
##
## data: CEP$edad
## D = 0.056969, p-value = 0.0001935
## alternative hypothesis: two-sided
```

Para interpretar uno u otro estadístico, debemos hacerlo atendiendo a la cantidad de casos de la variable en cuestión. En este caso, la variable `edad` presenta 1.424 casos. Esto provoca que ninguno de los resultados sea confiable pues las pruebas tienden a fallar cuando no se cumple el supuesto del margen de casos tolerado.

No obstante, vale la pena recordar que para determinar normalidad univariada se busca un valor  $p$  mayor a 0,05 para no rechazar la hipótesis nula de igualdad entre la distribución normal teórica y la distribución empírica de datos que estamos evaluado. Si los resultados fueran válidos en términos estadísticos, en ambas situaciones deberíamos rechazar la normalidad pues se aprueba la hipótesis alternativa dado que ambas pruebas presentan valores  $p$  menores a 0,05.

### 7.1.5 Tablas para informes e interpretación de resultados

En el presente apartado se presenta una forma para poder exportar resultados construidos en R a un formato compatible con softwares de tipo planilla de datos como Excel de Microsoft Office o Calc de Libre u Open Office. Es importante recalcar que aquí se enseña una forma simple para exportar resultados, que no se apoya en paquetes especializados en tales procedimientos. Una aproximación general a formas más avanzadas de construcción de reportes de resultados con formato, se presenta en el capítulo 9 de este manual.

En términos generales, el método de tratamiento de datos que presentaremos a continuación supone tres operaciones: en primer lugar, los resultados calculados deben existir en forma de matriz de datos en el entorno de trabajo; en segundo término, tales objetos deben imprimirse a un archivo de tipo CSV (para lo cual emplearemos la función `write.csv2`), el cual podrá ser abierto con alguno de los softwares tipo planilla de cálculo ya mencionados; así, en tercer lugar, tales datos podrán manipularse y editarse, ajustando su formato según lo que se desee, siendo fácilmente exportables hacia procesadores de texto.

#### 7.1.5.1 Distribuciones de frecuencias

En primera instancia se guardarán como objetos las tablas de frecuencias ya existentes en nuestro entorno de trabajo de R (creadas en apartados anteriores de este capítulo). En el ejercicio 7.15 se observa la creación de tres objetos que contendrán, de forma respectiva, una tabla de frecuencias absolutas, una tabla de frecuencias relativas (porcentajes) y una tabla de frecuencia relativa acumulada (porcentajes) para la variable *evaluación de la economía chilena* (`eval_econ_factor`).

#### Ejercicio 7.15

```
f <- table(CEP$eval_econ_factor)
f_porc <- round((prop.table(tabla)*100),2)
f_porc_acum <- round(cumsum(prop.table(tabla)*100),2)
```




 Tabla 1	03-06-2018 20:29	Archivo de valores separados por comas de Microsoft Excel
 Tabla 2	03-06-2018 20:29	Archivo de valores separados por comas de Microsoft Excel
 Tabla 3	03-06-2018 20:29	Archivo de valores separados por comas de Microsoft Excel

Figura 7.1: Exportación de resultados a archivos CSV

Tabla 1.csv			Tabla 2.csv			Tabla 3.csv			Libro2					
A	B	C	A	B	C	A	B		B	C	D	E	F	
1	Var1	Freq	1	Var1	Freq	1		x						
2	1 Positiva	471	2	1 Positiva	33,4	2	Positiva	33,4						
3	2 Neutra	730	3	2 Neutra	51,77	3	Neutra	85,18						
4	3 Negativa	209	4	3 Negativa	14,82	4	Negativa	100						
5			5			5								
6			6			6								
7			7			7								
8			8			8								
9			9			9								
10			10			10								

Tabla 1. Distribución de frecuencias de la variable "evaluación de la economía chilena"			
Categoría/Indicador	Frecuencia absoluta	Porcentaje	Porcentaje acumulado
Positiva	471	33,4	33,4
Neutra	730	51,77	85,18
Negativa	209	14,82	100
Elaboración propia			

Figura 7.2: Tabla de frecuencias en archivo CSV

Luego, con la función `write.csv2` tales tablas pueden imprimirse, de forma individual, a archivos de tipo CSV. Como se observa en la continuación del ejercicio 7.15, al ejecutar esta función se indica el objeto a imprimir como planilla y luego con el argumento `file` se indica (entre comillas) el nombre del archivo a crear con su respectiva extensión (`.csv` en este caso).

### Ejercicio 7.15 (continuación)

```
write.csv2(f, file = "Tabla 1.csv")
write.csv2(f_porc, file= "Tabla 2.csv")
write.csv2(f_porc_acum, file= "Tabla 3.csv")
```

El resultado de la anterior operación será la creación de tres archivos de tipo CSV en la ubicación definida como carpeta de trabajo para la sesión de R. La aparición de tales archivos en la carpeta de trabajo que hemos definido para nuestra sesión de R se observan en la imagen a continuación.

A partir de estos archivos - como se observa en la siguiente imagen - es posible unificar y editar las tablas en un formato adecuado para su inserción en reportes de investigación académica o profesional. No se profundiza aquí en el manejo de datos en softwares tipo planilla de cálculo, pero las simples operaciones *copiar* y *pegar* permiten consolidar los diferentes resultados en una sola tabla. Las tres primeras tablas de la imagen 7.2 son los datos impresos por R en archivos CSV independientes; la cuarta tabla es una tabla configurada manualmente mediante Microsoft Excel, lista para ser copiada y pegada en cualquier procesador de texto.

Ahora bien, es importante recalcar cómo se debe realizar la interpretación de una tabla de resultados como la presentada (a continuación se incorpora en un formato adecuado para este documento interactivo).

Tabla 7.1: Distribución de frecuencias de la variable Evaluación de la Economía Chilena

Categoría/Indicador	Frecuencia absoluta	Porcentaje	Porcentaje acumulado
Positiva	471	33,4	33,4
Neutra	730	51,77	85,18
Negativa	209	14,82	100

Reiterando lo señalado para los resultados del ejercicio 7.7, se observa que una amplia mayoría de los casos presenta una evaluación neutra o positiva de la economía nacional con el 85,2% de los casos (en términos relativos), lo que equivale a 1.201 casos en términos absolutos. Dentro de ello destaca que un tercio (33,4% o 471 casos) de las personas declaran una opinión neutra. Finalmente, es posible señalar que sólo un 14,8% de las respuestas (209 casos) se concentran en una evaluación negativa de la economía.

### 7.1.5.2 Estadísticos descriptivos

A continuación, se explicarán dos formas para calcular de manera agregada diversos estadísticos descriptivos

**Ejercicio 7.16**

```
summary(CEP$satisfaccion_vida)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##    1.000   6.000   7.000   7.311   9.000  10.000      10
```

Ahora bien, para exportar estos resultados a una planilla de cálculo se deberán ejecutar diversas operaciones que se muestran en el siguiente ejemplo (7.17). Primero, los resultados del comando `summary` se guardan como un objeto específico (al cual se le asigna el nombre de `descriptivos`).

**Ejercicio 7.17**

```
#Guardar summary como objeto
descriptivos <- summary(CEP$satisfaccion_vida)
```

Luego, si se le aplican las funciones `names` y `as.numeric` a tal objeto (`descriptivos`) se obtiene un vector que contiene los encabezados de los resultados y otro vector que contiene su valores numéricos: en la continuación del ejercicio 7.17 se observa el resultado de aplicar ambas funciones sobre el objeto mencionado, que muestran en primer lugar los encabezados de la tabla y luego sus valores.

Finalmente, usando la función `as.data.frame` se configura como matriz de datos el resultado de unir como filas - mediante la función `rbind` (unir como filas, o *row bind* en inglés) el encabezado de los estadísticos descriptivos y sus valores numéricos. Tal objeto se nombra como `descr_sat_vida` y se configura como un objeto de tipo `data.frame`.

**Ejercicio 7.17 (continuación)**

```
#Ver nombres y valores del objeto
names(descriptivos)
```

```
## [1] "Min."    "1st Qu." "Median"  "Mean"    "3rd Qu." "Max."    "NA's"
```

```
as.numeric(descriptivos)
```

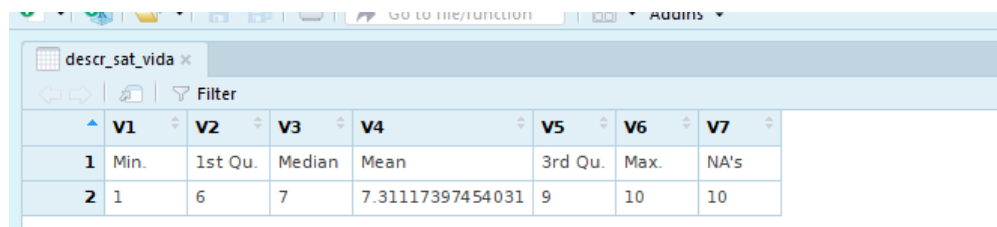
```
## [1] 1.000000 6.000000 7.000000 7.311174 9.000000 10.000000 10.000000
```

```
#Configurar como matriz de datos
descr_sat_vida <- as.data.frame(rbind(names(descriptivos), as.numeric(descriptivos)))
View(descr_sat_vida)
```

El resultado del ejercicio (como se observa en la imagen 7.3) muestra cómo ya hemos configurado el resultado en formato matriz de datos.

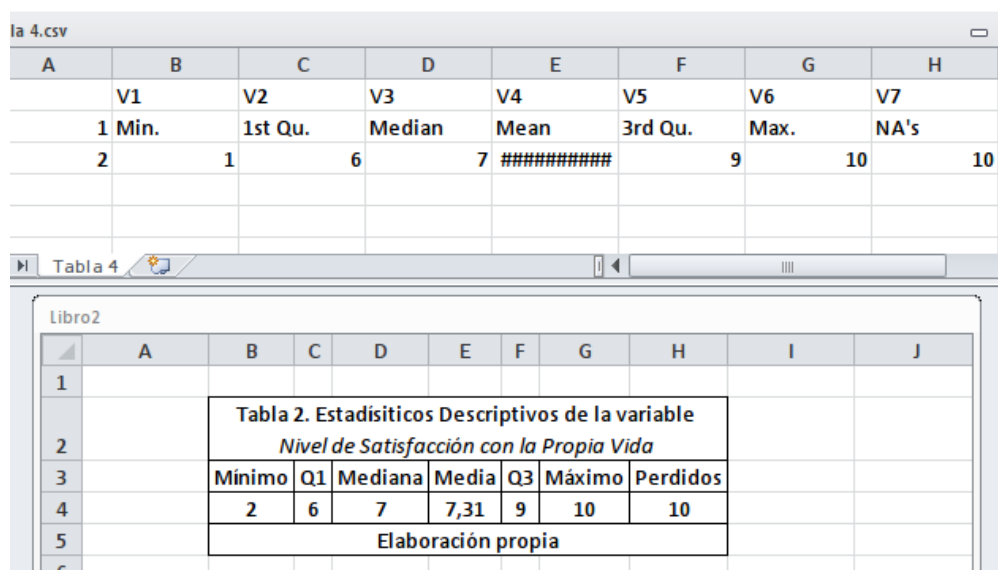
Así, a partir de esta matriz se puede crear un nuevo archivo tipo planilla de cálculo con estos resultados. Que quedará guardado en la carpeta de trabajo con el nombre de “Tabla 4.csv”.

**Ejercicio 7.17 (continuación)**



	V1	V2	V3	V4	V5	V6	V7
1	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
2	1	6	7	7.31117397454031	9	10	10

Figura 7.3: Comando summary configurado como matriz de datos



**la 4.csv**

	V1	V2	V3	V4	V5	V6	V7
1	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
2	1	6	7	7.31117397454031	9	10	10

**Libro2**

	A	B	C	D	E	F	G	H	I	J
1										
2		<b>Tabla 2. Estadísticos Descriptivos de la variable</b> <b>Nivel de Satisfacción con la Propia Vida</b>								
3		<b>Mínimo</b>	<b>Q1</b>	<b>Mediana</b>	<b>Media</b>	<b>Q3</b>	<b>Máximo</b>	<b>Perdidos</b>		
4		2	6	7	7,31	9	10	10		
5		<b>Elaboración propia</b>								
6										

Figura 7.4: Resultados de summary exportados a archivo CSV

```
#Exportar matriz a archivo CSV
write.csv2(descr_sat_vida, file = "Tabla 4.csv")
```

Como se observa en la imagen 7.4, la planilla superior es la exportada desde R, mientras la inferior es la que resulta luego de una edición simple enfocada en editar los encabezados de la tabla y editar el formato de la misma destacando todos los bordes de la tabla y centrando sus valores (entre otras ediciones).

Ahora bien, la función `summary` no calcula todos los estadísticos que pueden ser de interés. Para construir una tabla completamente personalizada es necesario calcular cada valor que sea de interés y disponerlo en una matriz de datos para así poder guardar una tabla en formato planilla de cálculo.

Como se observa en el ejercicio 7.18, cada valor construido se concatena como un sólo vector (`descriptivos_satvida`) de valores numéricos, que será una fila de la planilla a construir (la que contendrá los valores de cada estadístico). Adicionalmente, se construye un vector llamado `nombres` con los encabezados de cada estadístico a escribir en la tabla, que se utilizará como fila de títulos de la planilla de cálculo a exportar.

### Ejercicio 7.18

```
#Cálculo simple de estadísticos descriptivos
min <- min(CEP$satisfaccion_vida, na.rm = TRUE)
q1 <- quantile(CEP$satisfaccion_vida, probs = 0.25, na.rm = TRUE)
media <- mean.default(CEP$satisfaccion_vida, na.rm = TRUE)
```

A	B	C	D	E	F	G	H	I	J	K	L	M
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
nombres	Mínimo	Q1	Media	Media recortada	Mediana	Moda	Varianza	Desviación Estándar	Q3	Máximo	Simetría	Curtosis
descriptivos_satvida	1	6	7.31117397454031	7.390625	7	10	4.45936549978929	2.11172098057231	9	10	-0.533572100264213	-0.146125844618468

Figura 7.5: Tabla de estadísticos muestrales personalizada y exportada a CSV

Tabla 2. Estadísticos Descriptivos para la variable Nivel de Satisfacción con la Propia Vida											
Mínimo	Q1	Media	Media recortada	Mediana	Moda	Varianza	Desviación Estándar	Q3	Máximo	Simetría	Curtosis
1	6	7,31	7,39	7	10	4,46	2,11	9	10	-0,53	-0,15
Elaboración propia											

Figura 7.6: Tabla de estadísticos con edición de formato, lista para ser incorporada a procesador de texto

```
media_rec <- mean.default(CEP$satisfaccion_vida, trim = 0.025, na.rm = TRUE)
mediana <- median.default(CEP$satisfaccion_vida, na.rm = TRUE)
moda <- mfv(CEP$satisfaccion_vida)
var <- var(CEP$satisfaccion_vida, na.rm = TRUE)
desvest <- sd(CEP$satisfaccion_vida, na.rm = TRUE)
q3 <- quantile(CEP$satisfaccion_vida, probs = 0.75, na.rm = TRUE)
max <- max(CEP$satisfaccion_vida, na.rm = TRUE)
s <- skew(CEP$satisfaccion_vida)
c <- kurtosi(CEP$satisfaccion_vida)

#Valores de estadísticos como vector
descriptivos_satvida <- as.numeric(c(min, q1, media, media_rec, mediana, moda,
                                     var, desvest, q3, max, s, c))

#Encabezados de cada estadístico como un vector
nombres <- c("Mínimo", "Q1", "Media", "Media recortada", "Mediana", "Moda",
             "Varianza", "Desviación Estándar", "Q3", "Máximo", "Simetría", "Curtosis")
```

Con esta información ya es posible configurar una planilla de estadísticos descriptivos completamente personalizada. Para ello, como se muestra en la continuación del ejercicio 7.18, se configura una matriz de datos (función `as.data.frame`) a partir de los vectores de encabezados de los estadísticos (`nombres`) y los valores de tales coeficientes (`descriptivos_satvida`), que se almacena como un objeto de nombre `descr2`. Finalmente, esta matriz de datos se exporta a una planilla de cálculo mediante la función `write.csv2`.

### Ejercicio 7.18 (continuación)

```
descr2 <- as.data.frame(rbind(nombres,descriptivos_satvida))

write.csv2(descr2, file = "Tabla 5.csv")
```

Todo esto resulta en una planilla como la que se observa en la imagen 7.5, editable para la construcción de reportes formales.

En base a una edición muy simple en software tipo planilla de cálculo, se puede llegar a darle un formato como el que se observa en la imagen 7.6, útil para su presentación en reportes formales de resultados.

Ahora bien, es importante recalcar cómo se debe realizar la interpretación de una tabla de resultados como la presentada (a continuación se incorpora en un formato adecuado para este documento interactivo).

Tabla 7.2: Estadísticos Descriptivos de la variable Nivel de Satisfacción con la Propia Vida

Estadístico	Valor
Mínimo	1,0
Q1	6,0
Media	7,31
Media recortada	7,39
Mediana	7,0
Moda	10,0
Varianza	4,46
Desviación Estándar	2,11
Q3	9,0
Máximo	10,0
Simetría	-0,53
Curtosis	-0,15

En la tabla de estadísticos descriptivos presentada destacan diferentes elementos. En primera instancia, los valores mínimo (1) y máximo (10) indican que el rango de las respuestas observadas abarca todas las respuestas posibles de la pregunta del cuestionario.

En segunda instancia, el valor de la media (7,31) es bastante similar al de la mediana, lo que indica que la mitad de las personas de la distribución le asigna la nota 7 a la satisfacción con su propia vida, en la escala de 1 a 10 ya señalada.

En tercera instancia, se puede señalar - respecto a la dispersión de los datos - que la desviación estándar es de 2,11 unidades.

Finalmente, sobre la forma de la distribución se destacan dos elementos: dado que el coeficiente de simetría es negativo (-0,53), se está ante una variable de tipo asimétrica hacia la izquierda de la distribución de casos, en la cual los casos tienden a concentrarse hacia la derecha de la media, es decir, en las puntuaciones más altas de la distribución de respuestas; dado que el coeficiente de curtosis es negativo (-0,15), estamos ante una distribución platicúrtica, en la que existe una menor concentración de casos en torno a la media y presenta una forma más “achataada”.

## 7.2 Inferencia estadística univariada: de la estimación puntual al parámetro

En el apartado anterior se revisaron procedimientos para estimar de forma puntual estadísticos univariados provenientes de una muestra probabilística. En el presente apartado se explicará como calcular estimar intervalos de confianza para parámetros a partir de muestras probabilísticas. Concretamente se indicarán los procedimientos para calcular intervalos de confianza para medias y proporciones. Adicionalmente se expondrá como calcular estimadores insesgados a partir del uso de factores de expansión.

### 7.2.1 Cálculo de intervalos de confianza para proporciones

Para calcular el intervalo de confianza de una proporción es de utilidad la función `exactci` del paquete `PropCIs` (recordar instalarlo vía `install.packages("PropCIs")` de manera previa).

La información que necesita esta función es la cantidad de casos que coinciden con la condición de interés (la categoría de la variable cuya frecuencia relativa interesa) y la cantidad que representa a la totalidad de

casos de la muestra (*n muestral*). En este caso, a partir de la encuesta CEP se evaluará la hipótesis de si la mayoría de las y los chilenos declara tener una **opinión negativa** de la situación económica del país.

Como se observa a continuación en el ejercicio 7.19, solicitando una tabla de frecuencias es posible identificar la cantidad de casos probables que tienen una percepción negativa de la situación económica del país (730). Por otra parte, solicitando la cantidad de filas de la base de datos (usando la función `nrow`) se conoce el *n* total de casos (1.424). Con ambos valores es posible aplicar la función `exactci`; un tercer argumento a indicar es el nivel de confianza (`conf.level`), que se expresa en términos numéricos y proporcionales (para este caso, un 95% de confianza se expresa como *0.95*).

### Ejercicio 7.19

```
library(PropCIs)

table(CEP$eval_econ_factor)

##
## Positiva   Neutra Negativa
##      471      730      209

nrow(CEP)

## [1] 1424

exactci(x = 730, n = 1424, conf.level = 0.95)

##
##
##
## data:
##
## 95 percent confidence interval:
##  0.4863248 0.5389039
```

Así, el resultado de esta función es el límite superior e inferior del intervalo de confianza, construido a partir de la probabilidad ya indicada (*proporción de personas que declaran tener una percepción positiva de la situación económica del país*).

Con este resultado es posible concluir que, con un 95% de confianza, no es posible afirmar que la proporción de personas que declaran tener una percepción positiva de la situación económica del país sea más que la mitad de población nacional, debido a que el parámetro poblacional se sitúa entre el 48,6% y 53,9% de los casos.

## 7.2.2 Cálculo de intervalos de confianza para medias

Una segunda variante para el cálculo de parámetros poblacionales es la estimación del intervalo de confianza de una media. Para calcular el intervalo de confianza de una media es útil la función `ci.mean` del paquete `Publish` (recordar instalarlo vía `install.packages("Publish")` de manera previa). Su uso se detalla en el ejercicio 7.20.

La información que necesita esta función solamente es la variable a utilizar. En el caso que se observa a continuación, se calcula el intervalo de confianza para la media de la variable *nivel de satisfacción con la propia vida*.



El resultado es bastante sencillo e indica el valor del estadístico muestral (el valor de la *media* bajo la etiqueta *mean*) a la vez que indica el límite superior e inferior del intervalo de confianza para la media poblacional, o parámetro, bajo la etiqueta *CI-95%*; esto último indica el nivel de confianza utilizado para la construcción del intervalo.

### Ejercicio 7.20

```
library(Publish)

ci.mean(CEP$satisfaccion_vida) #Nivel de confianza por defecto.

## mean CI-95%
## 7.31 [7.20;7.42]
```

Como se observa en los resultados del ejercicio 7.20, al ejecutar el comando con su configuración por defecto, este utiliza un 95% de confianza. Con este valor se puede afirmar con un 95% de confianza que la media poblacional, es decir el promedio de satisfacción que los chilenos declaran tener respecto a su propia vida, se encuentra entre los valores 7,2 y 7,42, en una escala de 1 a 10.

Ahora bien, si al comando básico se le agrega el argumento **alpha** es posible definir el valor complementario al nivel de confianza, es decir la proporción de error aceptable para la estimación. En la continuación del ejercicio 7.20 se define el valor 0,2 o 20%.

### Ejercicio 7.20 (continuación)

```
ci.mean(CEP$satisfaccion_vida, alpha = 0.2) #Definición manual del nivel de confianza.

## mean CI-80%
## 7.31 [7.24;7.38]
```

Lo realizado en la continuación del ejercicio 7.20 implica que, observando los resultados, se puede afirmar con un nivel de confianza del 80% que el promedio de satisfacción que los chilenos declaran tener respecto a su propia vida, se encuentra entre los valores 7,24 y 7,38, en una escala de 1 a 10.

Un uso muy frecuente para este tipo de cálculos es evaluar si la diferencia entre una misma media para dos grupos es diferente, de manera estadísticamente significativa. Supondremos que se busca calcular el mismo indicador anterior pero efectuando el cálculo de intervalo de confianza para medias, diferenciando según hombres y mujeres.

Como se observa en el código a continuación (ejercicio 7.21), para indicar a la función cual es la variable de clasificación para construir los grupos, se debe indicar la variable de interés seguida a continuación de una virgulilla (mismo signo usado en la ñ, ~) que la separa de la variable de clasificación. Luego se debe indicar el conjunto de datos de donde provienen tales variables con el argumento **data**. Como antes, si no se le indica el valor del **alpha**, la función asume por defecto un cálculo con un 95% de confianza.

### Ejercicio 7.21

```
ci.mean(satisfaccion_vida~sexo_factor, data=CEP)

## sexo_factor mean CI-95%
## Hombre      7.46 [7.29;7.62]
## Mujer       7.22 [7.07;7.37]
```

Como se observa en los resultados, las medias muestrales permitirán afirmar que hombres y mujeres presentan una evaluación levemente diferente en relación a la satisfacción que declaran tener con su propia vida. Específicamente, la media de esta variables es de 7,46 para los hombres y de 7,22 para las mujeres. Sin embargo, al observar el comportamiento de tales mediciones a nivel poblacional, es posible afirmar que las variables no difieren de manera estadísticamente significativa. Para un nivel de confianza del 95%, la media de esta variable para los hombres se sitúa entre los valores 7,29 y 7,62 para los hombres y entre los valores de 7,07 y 7,37 para las mujeres. Los valores indican que los intervalos se intersectan, es decir, existe una elevada probabilidad de que estos parámetros sean muy similares e incluso iguales. Por lo tanto, no es posible afirmar que tales valores sean diferentes, de manera estadísticamente significativa, pues existe una alta posibilidad de que sean iguales.

### 7.2.3 Coeficientes de expansión para estimación de parámetros poblacionales

Las investigaciones sociales basadas en un diseño muestral probabilístico (aleatorio simple, estratificado, por conglomerado o diseños complejos) deben utilizar un ponderador en la estimación de variables de interés para alcanzar validez sobre la población objetivo. Lo anterior se relaciona con las probabilidades de selección de las distintas unidades de muestreo y da cuenta del número de personas de la población que representa cada individuo de la muestra. Este ponderador es conocido como factor de expansión.

Los factores de expansión buscan incorporar en las estimaciones puntuales surgidas desde la muestra, las características y el peso que tienen esos atributos en la población. De esa forma, se producen estimadores insesgados, es decir, que están centrados en el parámetro población. Así, denominaremos “sesgo de un estimador” a la diferencia entre la esperanza (o valor esperado) del estimador y el valor observado del parámetro a estimar. Decimos que un estimador es “insesgado” o “centrado” cuando su sesgo es nulo: o sea, cuando el valor esperado es igual al valor observado.

#### 7.2.3.1 Configuración de los datos a utilizar

Para hacer esto más claro, y a la vez indicar cómo se trabaja con coeficientes de expansión en R, lo explicamos con un ejemplo. Específicamente utilizaremos la Encuesta Nacional de Empleo, trimestre móvil Enero-Febrero-Marzo de 2019 publicada por el Instituto Nacional de Estadísticas de Chile. (2019a).

Según lo que se indica en el libro de códigos asociado a la base de datos, utilizaremos las siguientes variables para ejemplificar el uso de coeficientes de expansión.

Tabla 7.3: Variables a utilizar desde la ENE

Nombre de variable	Definición	Valores	Estado en base
fact	Factor de expansión	Peso de cada caso en la población	Creada por INE
edad	Edad de la persona entrevistada	Cantidad de años	Creada por INE
PET	Población en edad de trabajar	1 representa personas de 15 y más años - 2 representa a personas menores a 15 años	A crear

Como se observa en el siguiente ejemplo (7.22), primero se carga la base de datos con la función `read.csv2`, para luego recodificar la variable `edad` en la variable `PET`<sup>5</sup>. A esta última variable le agregamos etiquetas

<sup>5</sup>El número 1 representa a las personas mayores o iguales a 15 años, mientras que el número 2 representa a las personas menores a 15 años

mediante la función `factor` para poder identificar los casos en las salidas de resultados.

### Ejercicio 7.22

```
#Cargar ENE y guardar como objeto
ENE <- read.csv2("ENE 2019 02 EFM.csv", sep = ";", dec = ",")

#Recodificación edad en PET
library(dplyr)
ENE <- mutate(ENE, PET = car::recode(ENE$edad,
                                     "0:14=2; else = 1"))

# Etiquetado de categorías
ENE$PET <- factor(ENE$PET, labels = c(">=15", "<15"))
table(ENE$PET)

##
##   >=15   <15
## 86197 20254
```

Como se observa en los resultados del ejercicio 7.22, en la muestra se observa que existen 86.197 personas en edad de trabajar (con una edad igual o mayor a 15 años), mientras 20.254 personas no están en edad de trabajar (menores de 15 años).

Ahora bien, si se observan los resultados oficiales publicados por el INE (2019b), se observa que la cantidad de personas de 15 años o más, no es 86.197, sino 15.286.507 personas. Para lograr llegar a tal cifra se deben aplicar el factor - o coeficiente - de expansión de la base de datos, para luego de ello calcular la cifra de interés.

#### 7.2.3.2 Utilización del paquete *survey* para la utilización de coeficientes de expansión

Para aplicar coeficientes de expansión sobre nuestros resultados utilizaremos el paquete `survey` (Lumley, 2019).<sup>6</sup> Como se observa en el ejercicio a continuación, se debe crear un nuevo objeto que almacene la base de datos ponderada. En este caso llamaremos a tal objeto con el nombre `ENE_ponderada`, que almacenará el resultado de aplicar la función `svydesign`, con el argumento `data` indicando la base de datos a ponderar (ENE), con el argumento `id` indicando la unidad de muestreo que en este caso es el valor 1 pues cada caso se muestrea de forma equiprobable, y finalmente se indica el argumento `weights` indicando la variable que indica el peso que cada observación de la muestra representa en la población. Como puede verse, cada variable se indica luego de una virgulilla (~).

### Ejercicio 7.23

```
#Creación de base de datos ponderada
library(survey)
ENE_ponderada <- svydesign(data = ENE, id=~1, weights = ~fact)

#Características base ponderada
class(ENE_ponderada)

## [1] "survey.design2" "survey.design"
```

<sup>6</sup>Adicionalmente a la documentación del paquete disponible en el CRAN, existe una página con más ejemplos de cómo se utilizan sus principales funcionalidades, mantenida ya actualizada por su desarrollador (Lumley, 2019).

```
ENE_ponderada
```

```
## Independent Sampling design (with replacement)
## svydesign(data = ENE, id = ~1, weights = ~fact)
```

El resultado de aplicar la función `class` sobre el objeto que almacena la base de datos ponderada (`ENE_ponderada`) y ejecutar directamente tal objeto, es la información de que se trata de un objeto del tipo `survey.design` y que se trata de un diseño de muestreo con reemplazo, es decir, cada caso tuvo siempre la misma probabilidad de ser elegido (diseño *equiprobable*).

Ahora bien, sobre este objeto, podemos seguir utilizando funciones del mismo paquete `survey` para calcular algunos resultados. Retomando el ejemplo sobre la Población en Edad de Trabajar, con la función `svytotal` es posible calcular nuevamente las frecuencias de la variable `PET`. Como se observa en el ejercicio 7.24, esto se efectúa mediante la función `svytotal`, a la cual se le indica la variable de interés antecedida por una virgulilla (`~PET`) y el nombre del objeto que corresponde a uno del tipo `survey.design`.

### Ejercicio 7.24

```
svytotal(~PET, ENE_ponderada)
```

```
##           total      SE
## PET>=15 15286507 68710
## PET<15   3649759 42750
```

Como se observa en los resultados del ejercicio 7.24, ahora si logramos calcular las frecuencias reales de ocurrencia de la variable Población en Edad de Trabajar, pues la categoría igual o mayor a 15 años presenta 15.286.507 casos, equivalente a las cifras publicadas de manera oficial por el INE (2019b).

No se profundiza de manera adicional, pero el cálculo de otro tipo de resultados con una base ponderada es bastante similar a lo ya revisado en relación a frecuencias simples. Para mayores detalles revisar la página web mantenida por el equipo desarrollador del paquete *survey* (Lumley, 2019).

## Capítulo 8

# Construcción de gráficos usando RStudio: funcionalidades básicas y uso del paquete *ggplot2*

### 8.1 Funciones básicas para la construcción de gráficos

De modo general R, en su versión básica, incluye funciones para crear gráficos. Sin embargo, estas herramientas son bastante limitadas en cuanto a las posibilidades de edición que incluyen. Con todo, resultan válidas para un uso de análisis *exploratorio*. Esto es, un uso enfocado en la visualización de información que permita - dentro del contexto de un proceso de investigación - tomar decisiones para posteriores análisis estadísticos. Luego de explorar estas alternativas se profundizará en el uso de *ggplot2*, paquete especializado en el diseño de gráficos que permite una mejor visualización de resultados, sobre todo enfocados en el momento de *divulgación* de resultados de investigación (Field et al., 2012, 116-117).

#### 8.1.1 Gráficos de barras y gráficos circulares

De modo general, los gráficos de barras y circulares<sup>1</sup> se recomiendan para visualizar los valores de frecuencias absolutas o relativas de variables nominales u ordinales. Es decir, aquellas variables cuyos valores en la base de datos representan una simple clasificación de datos, sin que sea posible aplicar todas las propiedades aritméticas para tales números (Blalock, 1966, 26-37; Ritchey, 2008, 42-48).

Como un paso preliminar para la demostración de la construcción de gráficos mediante R configuraremos una nueva variable en la base de datos. Para esto, la variable Sexo (*CEP\$sexo*) será modificada en una nueva variable (*sexo\_factor*) agregando las etiquetas “Hombre” y “Mujer” a los valores numéricos, al configurarla como un vector de tipo factor. Realizamos esta operación para contar con una variable con valores numéricos y con sus respectivas etiquetas de respuesta, con el fin de observar si tal configuración de la variable es soportada por diferentes herramientas para construir gráficos.

#### Ejercicio 8.1

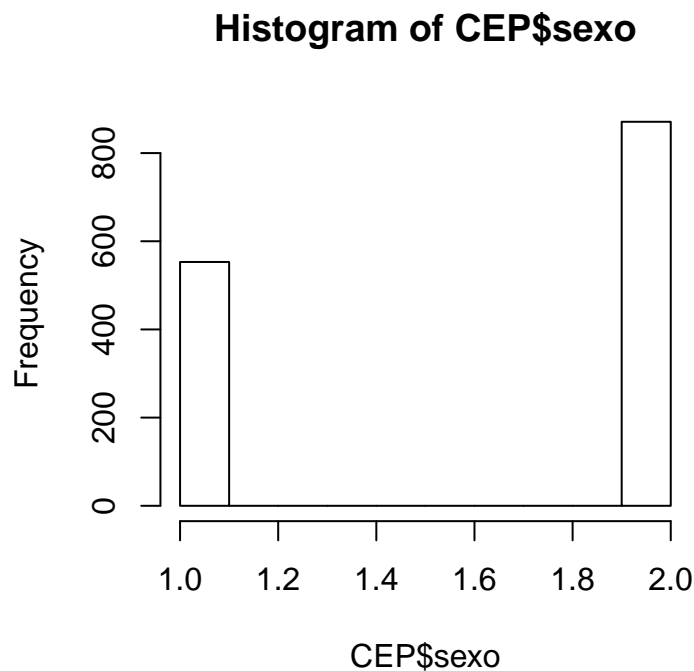
```
# Transformar variable sexo a factor (contar con valores y etiquetas)
CEP <- mutate(CEP, sexo_factor = factor(CEP$sexo,
                                         labels = c("Hombre", "Mujer")))
```

<sup>1</sup>Popularmente conocidos como gráficos “de torta”.

Un primer comando básico para la construcción de gráficos en R es `hist`. Este comando está pensado para construir histogramas. Como se observa a continuación, no resulta apropiado para variables nominales (aunque están codificadas como numéricas); tampoco soporta los formatos *character* ni *factor* como formato de entrada de los datos a graficar.

### Ejercicio 8.2

```
#Limitaciones de la función hist
hist(CEP$sexo)
```



```
hist(CEP$sexo_chr)
```

```
## Error in hist.default(CEP$sexo_chr): 'x' must be numeric
```

```
hist(CEP$sexo_factor)
```

```
## Error in hist.default(CEP$sexo_factor): 'x' must be numeric
```

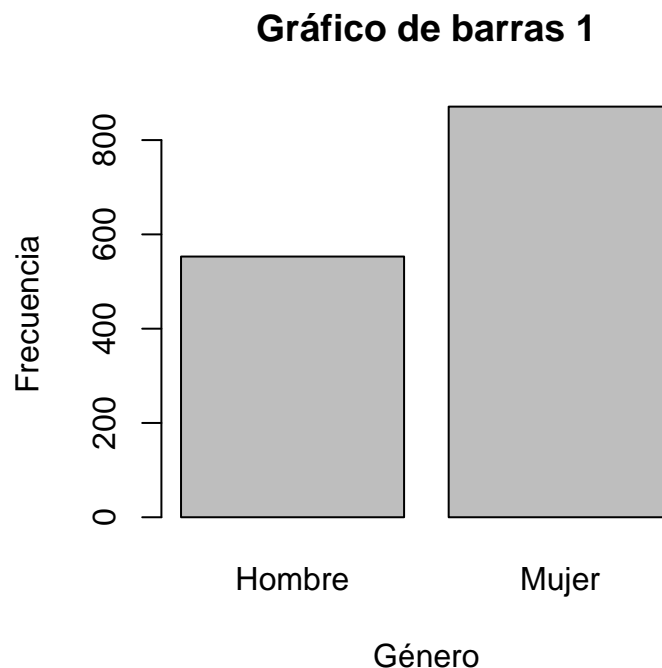
Los resultados de las líneas de código anterior muestran cómo solamente el primer gráfico es construido de manera adecuada, pero asume una continuidad de valores entre las categorías “hombre” y “mujer” (valores 1 y 2, respectivamente), lo que arroja un gráfico poco útil. Por otra parte, al indicar como variable de entrada un vector de tipo *character* o *factor*, el comando arroja error.

Por ello se sugiere el uso del comando `plot` para construir gráficos de barras. De manera general, este tipo de gráficos se usan para variables cualitativas (nominales u ordinales). Las categorías de una variable se sitúan en el eje X, mientras que la frecuencia absoluta o relativa (generalmente porcentajes) se sitúa en el eje Y (Ritchey, 2008, 83-85)

A continuación se muestra el uso del comando `plot` para construir un gráfico de barras. El primer argumento es la variable a graficar. Posteriormente se usan argumentos para agregar títulos: `main` sirve para agregar un título general; `xlab` sirve para agregar un título al eje X; `ylab` sirve para agregar un título al eje Y. Como se observa, este resulta bastante más adecuado para la visualización exploratoria de datos.

### Ejercicio 8.3

```
plot(CEP$sexo_factor, main = "Gráfico de barras 1",  
     xlab = "Género", ylab = "Frecuencia")
```



de barras-1.bb

Otra forma de presentar variables nominales son los gráficos circulares (o de “torta”); se trata de gráficos cuya división proporcional busca representar la distribución de categorías dentro de una variable nominal u ordinal: el área del círculo representa el 100% de los casos de una variable, mientras que el área de cada división del círculo, representa el porcentaje de casos en una categoría específica de esa variable (Ritchey, 2008, 80-83). Estos gráficos, a pesar de ser muy utilizados, no resultan tan recomendados pues tienden a distorsionar la percepción de la información presentada: cuando se trabajan variables con muchas opciones de respuesta la mirada tiende a distorsionar el tamaño relativo de las divisiones del círculo. No obstante pueden utilizarse cuando construimos gráficos con variables nominales de pocas categorías (entre dos y cinco), sobre todo para mostrar diferencias entre pocas categorías.

A continuación se muestra una configuración de elementos que servirán para construir el gráfico.

1. Primero se usa la tabla de frecuencias relativas construida en el apartado 7.1.2. Sobre ella se aplica la función `round`, que permite seleccionar la cantidad de decimales deseados para hacer la aproximación. Mediante la función `as.numeric` se guardan en un nuevo vector (*porcentajes*) los números que configuran los resultados de tal tabla.
2. Luego se construye un vector de caracteres (*etiquetas*) con las etiquetas de cada resultado.

3. A continuación se realizan dos operaciones sobre tal vector: usando el comando `paste` cada etiqueta se una con los valores porcentuales que serán cada proporción del gráfico; además, al final de cada valor se incorpora un signo `%` que quedará expresado en las etiquetas del gráfico. En el siguiente ejemplo se muestra el vector resultante de cada parte de la operación.

#### Ejercicio 8.4

```
#Valores que dividirán el gráfico
porcentajes <- as.numeric(round(((prop.table(table(CEP$eval_econ_factor)))*100),2))
porcentajes
```

```
## [1] 33.40 51.77 14.82
```

```
#Etiquetas para el gráfico
etiquetas <- c("Positiva", "Neutra", "Negativa")
etiquetas
```

```
## [1] "Positiva" "Neutra" "Negativa"
```

```
etiquetas <- paste(etiquetas, porcentajes)
etiquetas
```

```
## [1] "Positiva 33.4" "Neutra 51.77" "Negativa 14.82"
```

```
etiquetas <- paste(etiquetas, "%", sep = "")
etiquetas
```

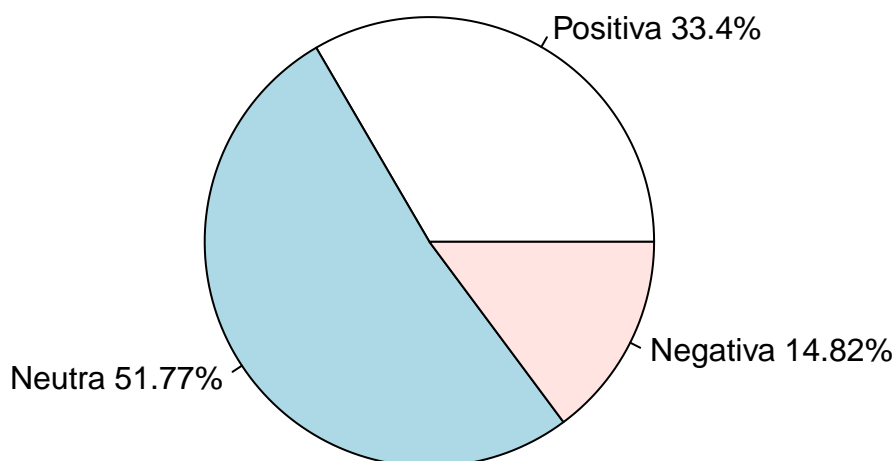
```
## [1] "Positiva 33.4%" "Neutra 51.77%" "Negativa 14.82%"
```

Considerando los dos elementos (*porcentajes* y *etiquetas*) se construye el gráfico circular. Como se observa a continuación, este se construye mediante la función `pie`. El primer argumento son los valores que demarcarán las divisiones del círculo que representa al 100% del área del círculo. Luego se indican los valores que determinarán la construcción de etiquetas. Posteriormente se indica mediante los argumentos `main` y `sub`, un título general para el gráfico y una descripción en su borde inferior, respectivamente. Ejecutando tal comando se obtiene el gráfico deseado.

#### Ejercicio 8.4 (continuación)

```
pie(porcentajes, etiquetas,
    main = "Gráfico de torta 1",
    sub = "Evaluación de la situación económica")
```



**Gráfico de torta 1**

### Evaluación de la situación económica

#### 8.1.2 Histogramas

El histograma es una representación gráfica que se utiliza para observar la forma de la distribución de variables cuantitativas (intervalo o razón). De manera similar al gráfico de barras, sobre el eje X se posicionan las puntuaciones de la variable, mientras que la frecuencia (absoluta o relativa) de cada valor se posiciona en el eje Y. Así, se construye un gráfico de columnas verticales; la principal diferencia entre un *gráfico de barras* y un *histograma* es que en este último las columnas se tocan entre sí, pues representa a un continuo de valores numéricos (Ritchey, 2008, 86-89).

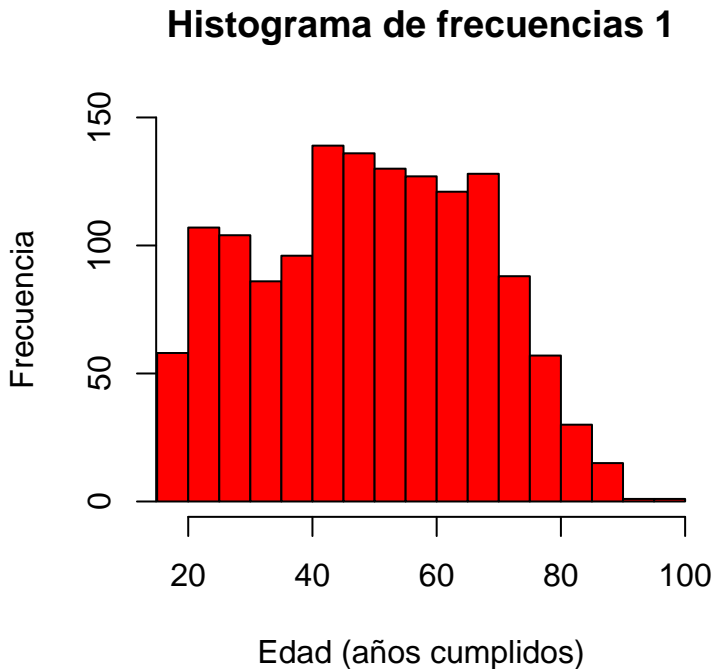
Con el siguiente comando se ilustra la construcción de un histograma usando R. El comando utilizado es `hist`; sus argumentos son `main`, `xlab` e `ylab` para incorporar un título general y para cada eje. Además podemos indicar el color de relleno de las barras mediante el argumento `color`, mientras que el color de los bordes de la barra se establece con el argumento `border`. Adicionalmente, los argumentos `xlim` e `ylim` permiten definir manualmente la extensión de los ejes (se indica un vector numérico con los valores mínimo y máximo).<sup>2</sup>

#### Ejercicio 8.5

---

<sup>2</sup>R incluye una paleta bastante grande de colores. Para una explicación general de cómo funcionan los colores en R se sugiere visitar este enlace o descargar la siguiente guía en inglés. Para descargar la plantilla general de colores, y asociar visualmente un color con su nombre estándar, presionar el siguiente enlace que efectúa la descarga de tal plantilla en formato PDF. Si se ejecuta el comando `colors()` se obtiene el listado general de posibilidades de colores con nombre, que incluye R.

```
hist(CEP$edad, main = "Histograma de frecuencias 1",
     xlab = "Edad (años cumplidos)",
     ylab = "Frecuencia",
     col = "red",
     border = "black",
     xlim = c(18, 97),
     ylim = c(0, 150))
```

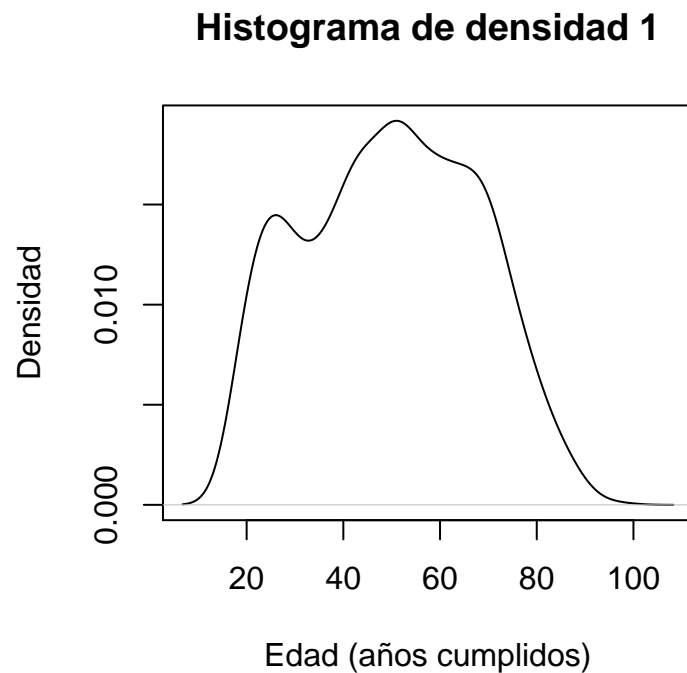


frecuencias variable edad-1.bb

Existe otro tipo de histograma denominado de “densidad”. Este gráfico no representa las barras de frecuencias de cada alternativa de respuesta de la variable, sino que aplica una *suavización* sobre los valores observados con el objetivo de conseguir una representación de la forma de la distribución que no se vea alterada por la cantidad de barras que incluya el histograma. La principal utilidad de este tipo de gráficos es poder observar la forma de la distribución de una variable, de la manera más precisa posible. En el ejemplo a continuación se muestra cómo construir un histograma de este tipo: en primer lugar se calcula la información de **densidad** de una variable con la función **density** asignando su resultado al objeto **densidad\_edad**; luego, mediante la función **plot** se gráfica esta información indicando título al gráfico y a los ejes de la forma que ya ha sido explicada.

### Ejercicio 8.6

```
densidad_edad <- density(CEP$edad)
plot(densidad_edad,
     main = "Histograma de densidad 1",
     xlab = "Edad (años cumplidos)",
     ylab = "Densidad")
```



densidad variable edad-1.bb

### 8.1.3 Diagrama de cajas

El último gráfico a construir con las funcionalidades básicas de R es el diagrama de cajas (*boxplot* en inglés). La línea central de este gráfico representa a la mediana, es decir al valor que señala el 50% de la distribución de la variable en estudio. La parte superior e inferior de la caja demarcan al tercer y segundo cuartil respectivamente (esto es, el 75% y 25% de los casos): así, la caja demarca al 50% central de los casos. Por otra parte, los extremos superior e inferior de la línea vertical, demarcan los valores máximo y mínimo de la distribución. Así, la distancia entre el extremo superior o inferior de tal línea y el borde superior o inferior de la caja, demarca la distribución de casos en el cuartil superior (25% de casos con puntuaciones más elevadas) y en el cuartil inferior (25% de casos con menores puntuaciones) de la distribución, respectivamente (Field et al., 2012, 151-152).

Para construir un gráfico de cajas el código de R es bastante simple. Basta con ejecutar la función `boxplot` indicando como primer argumento la variable de interés. A esto se le puede agregar un título (argumento `main`) y el argumento `outline` especificado como `TRUE`, para que si existen casos atípicos estos se grafiquen. A continuación se muestra un comando de este tipo y su resultado.

#### Ejercicio 8.7

```
boxplot(CEP$edad, main = "Gráfico de cajas 1",  
        outline = TRUE)
```

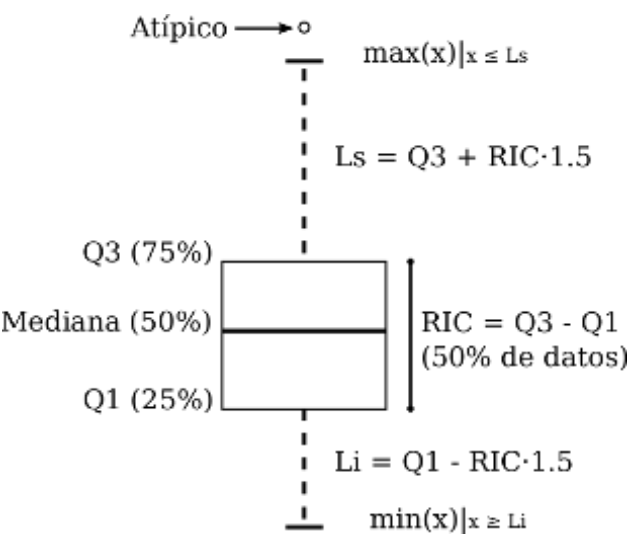
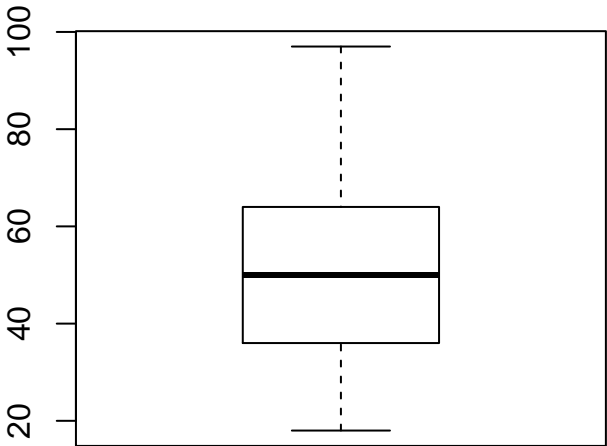


Figura 8.1: Gráfico de caja

### Gráfico de cajas 1



de cajas-1.bb

## 8.2 Introducción al uso del paquete *ggplot2*

### 8.2.1 Nociones generales

El paquete *ggplot2* es un paquete de R especializado en la construcción y diseño para la visualización de datos. En este sentido, se trata de un paquete cuyas funcionalidades van más allá de un uso puramente “científico” o *exploratorio* y se orienta a las diferentes dinámicas de divulgación de resultados de procesos de

investigación, esto incluye:

- **Divulgación de resultados de procesos de investigación científica para público especializado.** Esto puede referir a contextos académicos (publicaciones en revistas especializadas, libros, etc.) o profesionales (informes de investigación para actividades de consultoría en el ámbito público o privado, por ejemplo).
- **Divulgación de resultados de procesos de investigación científica para público no especializado.** Esto refiere por ejemplo a la difusión de información en contextos como medios de comunicación masivos (televisión, diarios en papel o digital) o redes sociales (twitter, facebook, instagram, etc.).

Se trata, en suma, de un conjunto de funciones altamente especializadas en la construcción de resultados visualmente atractivos para quien leerá la información. Debido a esta especialización, el nivel de configuración y trabajo de codificación para llegar a un resultado deseable, puede ser elevado.

En tal medida, para quien desempeña tareas de investigación social, se abre la tensión entre enfocar su trabajo en la construcción de productos para análisis exploratorio o enfocarse en diseñar visualizaciones de datos estéticamente atractivas para públicos amplios. Vale la pena señalar que el objetivo de este manual es enfocarse en un punto intermedio: manejar los elementos básicos de un paquete como *ggplot2* para no depender de que un tercer actor configure nuestros resultados básicos, pero sin profundizar su estudio hasta el punto de que nos aleje de nuestro principal objetivo: construir análisis sociológicamente relevantes y estadísticamente rigurosos.

Es por ello que en este manual centraremos los contenidos en una introducción general al uso de este paquete. Enfatizamos en que no es necesario convertirse en un experto, sino más bien manejar a cabalidad los fundamentos de su funcionamiento. Resultará central en esta dinámica manejar elementos de apoyo para realizar cuestiones más avanzadas.<sup>3</sup>

### 8.2.2 Gramática del paquete *ggplot2*

El paquete *ggplot2* presenta distintas características que lo distinguen:

1. En primer lugar, los objetos resultantes de la construcción de un gráfico no son una *imagen* sino un objeto de tipo *graphic* específico. Esto permite configurar un gráfico como cualquier otro elemento de R, directamente desde la sintaxis,
2. Debido a lo anterior, la editabilidad de los gráficos construidos es mayor. Definiendo el conjunto de información a visualizar, se pueden configurar diferentes tipos de gráficos.
3. En tercer lugar, puede señalarse que la estructura de este paquete presenta una gramática específica en relación a sus sintaxis. Como veremos a continuación, su sintaxis guarda directa relación con tres elementos que compondrán la estructura de cualquier visualización de datos: la información (*data*) a utilizar, la estética (*aesthetics*) o la definición de los ejes donde se posicionarán los datos a visualizar, y la geometría (*geometry*) o los elementos visuales que se posicionarán en la gráfica para representar los datos que interesa visualizar.

Resulta importante comprender que en el paquete *ggplot2* los gráficos se construyen en base a una serie de **capas de información** que superpuestas, configuran el resultado final. Andy Field et.al. (2012, 121-136) propone que los gráficos construidos mediante la función *ggplot* pueden entenderse como una serie de transparencias donde se imprime cierta información; esta información pueden ser números, títulos, barras, líneas, puntos, etc. Luego de definir tales capas de datos, lo que realiza la función *ggplot* es presentar el resultado de la superposición de tales capas de información como un elemento unitario.

---

<sup>3</sup>Una buena fuente de información, que incluye ejemplos interesantes de visualización de datos junto con sus respectivas sintaxis de configuración puede encontrarse en el apartado *ggplot2* de la Galería de Gráficos de R. Recomendamos explorar este tipo de recursos cuando se desee implementar una visualización cuyo código no se maneje.

Para estandarizar una gramática general para la construcción de gráficos mediante capas de información, el paquete se basa en un lenguaje estándar para la construcción de gráficos. Esta nomenclatura se basa en los planteamientos de un famoso libro titulado “The Grammar of Graphics (Statistics and Computing)” (Wilkinson et al., 2005), en el que sus autores definen una base común para la producción de gráficas basadas en información cuantitativa, aplicable a casi cualquier campo de producción de conocimiento.

Tabla 8.1: Elementos que componen un gráfico construido mediante la función *ggplot*.

Capa	Descripción
<b>Datos</b>	Conjunto de información que se representará de manera gráfica. En nuestro caso se trata de una o más variables, o una base de datos.
<b>Estética</b>	Escala en la cual se posicionará la información de interés. Refiere al posicionamiento de la información a representar sobre los diferentes ejes y dimensiones del gráfico resultante. Hablamos del posicionamiento de variables en los ejes X e Y como también de la posibilidad de indicar variables que pueden ser posicionadas como color de relleno dentro de los diferentes ejes, como una función de transparencia, etc.
<b>Geometría</b>	Formas, elementos visuales, que se emplearán para representar visualmente la información ya consignada en los <i>datos</i> y ubicada en las diferentes posiciones del gráfico mencionadas en la <i>estética</i> . <sup>4</sup> Cada especificación de geometría permite visualizar diferentes características de la(s) variable(s) y su distribución.

Es importante entender esta clasificación de elementos que componen un gráfico. Tal diferenciación de elementos comanda la estructuración de la sintaxis específica para crear gráficos mediante la función *ggplot*.

Como se observa a continuación, en la primera línea de argumentos de esta función (primer paréntesis) se definen los *datos* a visualizar y luego la *estética*. Luego, agregando un signo más al final de cada línea, se agregan líneas de comando adicionales que permiten agregar diferentes capas de información al gráfico final, cuyos *datos* y *estética* ya han sido definidos.

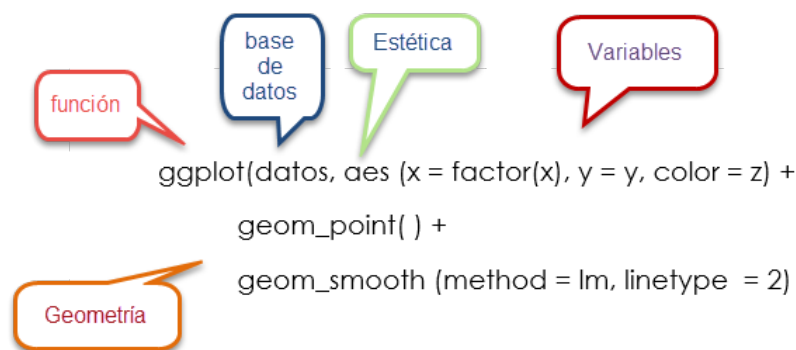
En la figura a continuación se observa una sintaxis que: i) posiciona una **variable X** en el eje X, configurándola como *factor*; que ii) posiciona una **variable Y** en el eje Y sin transformarla; y que iii) utiliza una tercera variable para el relleno de color de las figuras geométricas finales. En las dos líneas adicionales se establece: en primer lugar, el argumento `geom_point()` indica que uno de los elementos geométricos a posicionar son puntos, sin configuraciones adicionales; en segundo lugar, se posiciona - mediante el comando `geom_smooth(method=lm, linetype=2)` una línea de tendencia suavizada (mediante un procedimiento de regresión lineal) que mostrará la tendencia lineal de la nube de puntos.

**No olvidar que luego de la primera línea, las funciones de geometría deben agregarse indicando un signo +.**

En términos generales ésta es la estructura general para construir gráficos mediante *ggplot*. Puede que se introduzcan más o menos configuraciones dentro de cada elemento (*data*, *aesthetics*, *geom*) y que se agreguen más o menos capas de objetos geométricos sobre el gráfico a desplegar. No obstante, la lógica básica seguirá siendo siempre la misma.<sup>5</sup>

<sup>4</sup>Internamente, esta capa también puede contener propiedades *estéticas*, como la la transparencia, el color o tamaño de los objetos geométricos utilizados. Tal información *estética* interna a la geometría, no debe confundirse con la información *estética* general del gráfico.

<sup>5</sup>Para contar con un listado explicativo completo de las configuraciones posibles para la función *ggplot*, sugerimos revisar la Guía de Referencia para *ggplot2* elaborada y puesta en línea por su equipo de desarrollo. Desde esta página se puede acceder a la guía completa en línea de la familia de paquetes *tidyverse*, que incluye documentación de los siguientes paquetes: *dplyr*, *tidyr*, *tibble*, *purrr*, *readr* y, por supuesto, *ggplot2*.

Figura 8.2: Estructura básica de una sintaxis de la función `ggplot`

### 8.2.3 Construcción de gráficos usando la función *ggplot*: diagramas de barras, histogramas de frecuencia absoluta y relativa

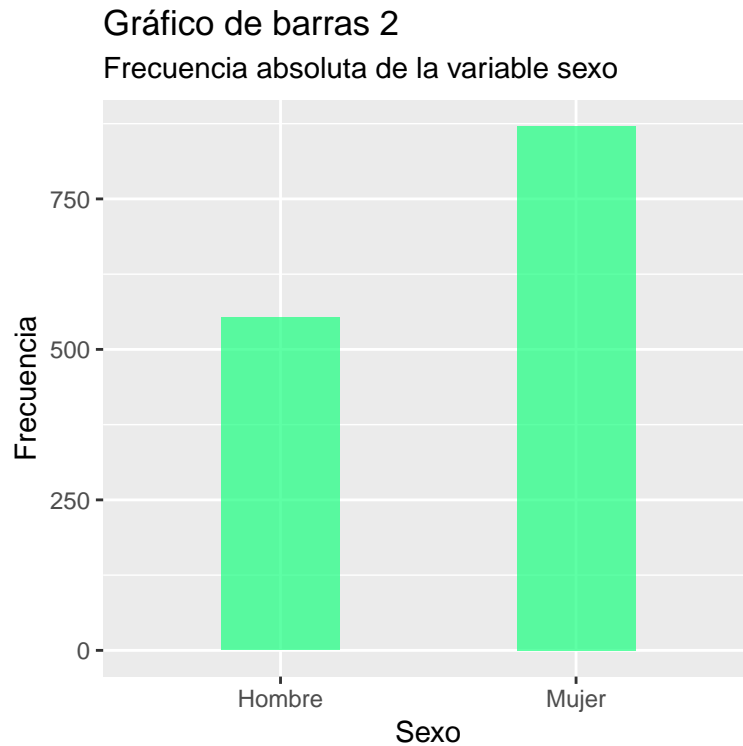
En este apartado se replicarán los gráficos construidos con las funcionalidades básicas de R, pero ahora aplicando la función `ggplot`. Se comienza por un diagrama de barras.

Como se observa en el código a continuación, este comando sigue la lógica y estructura de sintaxis ya explicada para el paquete `ggplot2`. En la primera línea de la función se definen los **datos** (`CEP`), y con el argumento `aes` (*estética*) se define qué variable irá en el eje X (`sexo_factor`). En este caso, al no definir una variable para el eje Y, el software efectúa un conteo simple de los casos que caen en cada categoría de la variable definida para el eje X. Luego, separado por un signo `+`, se define la figura geométrica o *geometría* a utilizar: en este caso se trata de un gráfico de barras (`geom_bar`); se agrega como argumento de esta subfunción un argumento `width` que permitirá definir el ancho de las barras, así como un argumento `fill` para definir el color de relleno de las barras, con la función `rgb`.<sup>6</sup> Finalmente los argumentos siguientes (también especificados luego de un signo `+`) definen diferentes elementos adicionales: título del eje X (`scale_x_discrete`), título del eje Y (`scale_y_continuous`) y título y subtítulo general para el gráfico (`labs`.)

#### Ejercicio 8.8

```
library(ggplot2)
#Gráfico de barras 2: sexo en frecuencias absolutas
ggplot(CEP, aes(x = sexo_factor)) +
  geom_bar(width = 0.4, fill=rgb(0.1,1,0.5,0.7)) +
  scale_x_discrete("Sexo") + # configuración eje X (etiqueta del eje)
  scale_y_continuous("Frecuencia") +
  labs(title = "Gráfico de barras 2",
       subtitle = "Frecuencia absoluta de la variable sexo")
```

<sup>6</sup>En este caso, se trata de la función `rgb`, que define concentraciones de los colores básicos rojo, verde y azul (*red, green and blue*).



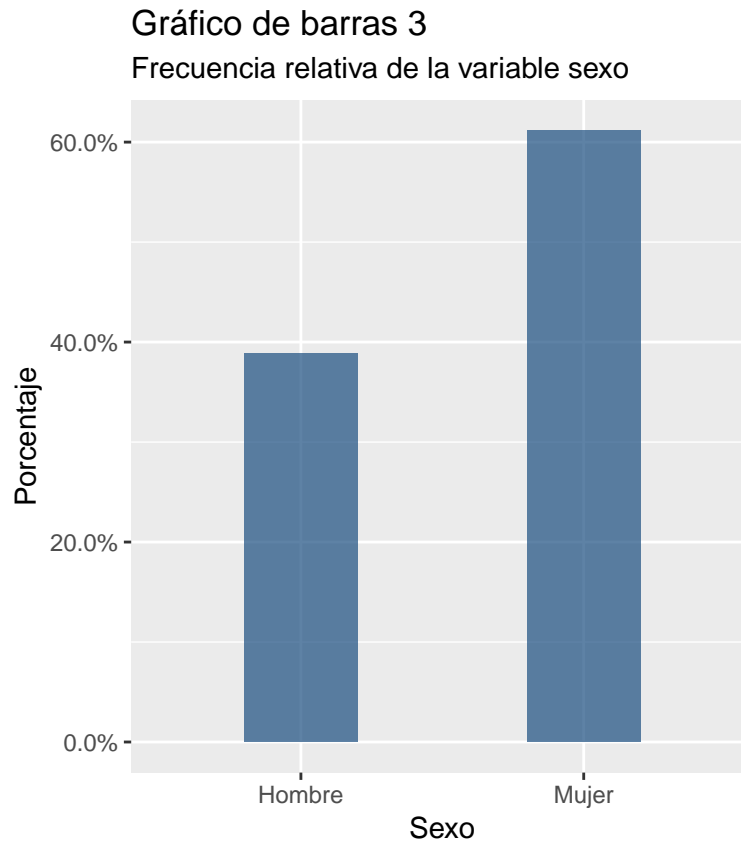
Si bien se trata de un código algo más complejo que la función básica que incluye el software, es posible afirmar que hay una mejora estética notable en comparación con los gráficos más básicos.

Una variante del gráfico anterior es construirlo para frecuencias relativas expresadas en porcentajes, uso muy común para la presentación de datos en contextos de divulgación de resultados.

### Ejercicio 8.9

```
ggplot(CEP, aes(x = sexo_factor)) +
  geom_bar(width = 0.4, fill=rgb(0.1,0.3,0.5,0.7), aes(y = (...count...)/sum(...count...))) +
  scale_x_discrete("Sexo") +      # configuración eje X (etiqueta del eje)
  scale_y_continuous("Porcentaje",labels=scales::percent) + #Configuración eje y
  labs(title = "Gráfico de barras 3",
        subtitle = "Frecuencia relativa de la variable sexo")
```





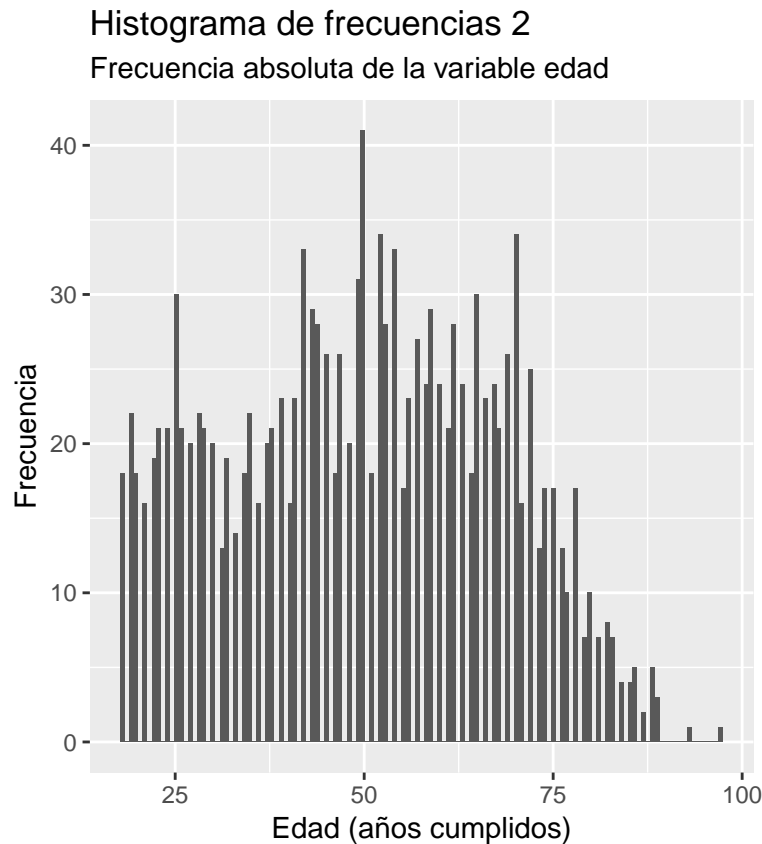
A continuación se indican las modificaciones realizadas al código anterior:

- Se modificaron algunos valores del argumento `fill` de la función `geom_bar` para alterar el color resultante de las barras.
- En la función `geom_bar` se agregó un argumento `aes` para editar el tipo de conteo ejecutado en el eje Y. Específicamente se le indica que el conteo hecho en Y es igual al conteo simple, dividido en la suma total de casos.
- Luego, se agrega un `scale_y_continuous` para editar los parámetros del eje vertical. Ello permite agregar el título de *porcentaje* al eje Y, a la vez que definir que las etiquetas del eje (`labels`) respondan a una escala de tipo porcentual (`label=scales::percent`).
- El resto de los comandos se mantuvo igual.

El siguiente comando muestra como construir un histograma de frecuencias absolutas. En este caso, luego de los datos a utilizar (CEP), se indica que en el eje X se posicione a la variable `edad` pero considerándola como variable numérica continua (`as.numeric`) pues originalmente es una variable de tipo `integer`, lo que impide posicionarla en el gráfico como variable continua. El resto de los comandos es similar a un gráfico de barras, sólo que ahora se le indica la función `geom_histogram` para construir el histograma; dentro de esta función se indica un `binwidth` para ajustar el ancho de la barra. Nótese que la configuración de ambos ejes indica que se trata de variables continuas (`scale_x_continuous`, `scale_y_continuous`), de lo contrario no podría configurarse el gráfico.

### Ejercicio 8.10

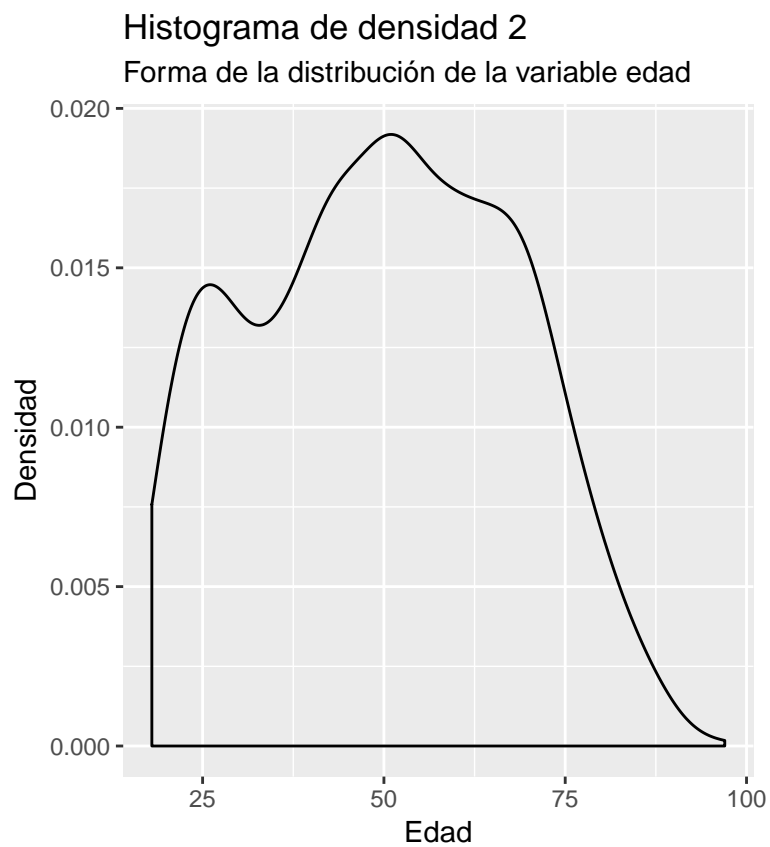
```
ggplot(CEP, aes(x = as.numeric(edad))) +
  geom_histogram(binwidth = 0.6) +
  scale_x_continuous("Edad (años cumplidos)") +
  scale_y_continuous("Frecuencia") +
  labs(title = "Histograma de frecuencias 2",
        subtitle = "Frecuencia absoluta de la variable edad")
```



Como fue señalado en el apartado anterior la manera que resulta más óptima para observar la forma de la distribución de una variable es la construcción de un histograma de densidad. A continuación se muestra la sintaxis que permite construir un histograma de densidad de la variable **Edad**. La estructura de la sintaxis sigue la misma lógica que los otros gráficos construidos con el paquete **ggplot2**; en este caso, la especificación de la geometría es el argumento **geom\_density**. El resultado permite observar de manera precisa la forma de la distribución de la variable **Edad**.

### Ejercicio 8.11

```
ggplot(CEP, aes(x = edad)) +
  geom_density() +
  scale_y_continuous("Densidad") +
  scale_x_continuous("Edad") +
  labs(title = "Histograma de densidad 2",
        subtitle = "Forma de la distribución de la variable edad")
```





## Capítulo 9

# Introducción al uso de RMarkdown para la compilación de resultados de RStudio en diferentes formatos

### 9.1 ¿Qué es RMarkdown y para qué sirve?

Por su amplio y fácil uso, las y los estudiantes, y posteriormente las y los académicos e investigadores en Ciencias Sociales, generalmente utilizan Microsoft Word para construir sus reportes de investigación.<sup>1</sup> Si bien es un software fácil de utilizar, presenta algunas limitaciones en el manejo de sus atributos estéticos y de contenido. Por una parte es difícil controlar las ediciones de su formato de presentación pues es bastante engorroso para dar formato a documentos extensos: obliga a invertir mucho tiempo haciendo clicks, indicando formatos que se pueden descalibrar al pasar a otro formato de texto o versión de software. Por otra parte, cualquier contenido que no sea texto resulta difícil de incorporar pues se generan incompatibilidades según las versiones del programa, a la vez que no se articula de manera sencilla con otras fuentes de información: por ejemplo los resultados contruidos a partir de un software de análisis estadístico.

A estas limitaciones, que han sido señaladas por otros autores (Miller, 2018) aquí se agrega otra. La interacción entre un software para análisis estadístico como R y Microsoft Word, es quizá una de las que presenta más complicaciones. Como ya fue visto en el capítulo 7 de este documento, para incorporar resultados de formato R a Microsoft Word se debe construir un tipo de datos *ad hoc* (matrices de datos) para poder *grabarlos* en un documento tipo planilla de cálculo. Recién desde allí se podrá editar el formato de tales resultados y copiarlos a Microsoft Word (por ejemplo, a un trabajo universitario, un reporte de investigación para consultoría, un borrador de un artículo académico o documento de trabajo, un libro, etc.)

¿Cuáles son, entonces, las características de RMarkdown que lo diferencian de esta forma de trabajo? En términos sencillos RMarkdown es un *procesador de texto* que ofrece además la posibilidad de incluir *trozos de código* desde R (u otros formatos). El principal beneficio de esta herramienta es que permite trabajar en un sólo documento tanto la *redacción* del contenido narrativo de reportes de investigación, como también la construcción y *presentación formal de resultados de análisis estadísticos*.<sup>2</sup>

Se distinguen entonces dos paradigmas de trabajo en los procesos de construcción de los diferentes formatos de presentación de resultados de investigación. Por un lado un *paradigma no integrado* en la construcción de informes de investigación: en este formato, el autor generalmente construye por separado los elementos de

---

<sup>1</sup>Debido a que es ampliamente utilizado se hace referencia al procesador de texto de Microsoft Office. Estas anotaciones también resultan aplicables a otros paquetes de software de oficina ya nombrados en este documento como Open Office y Libre Office.

<sup>2</sup>Para información de carácter general sobre RMarkdown como plataforma de trabajo (fichas técnicas, manuales, plantillas y tips de trabajo) sugerimos visitar la página oficial del proyecto.

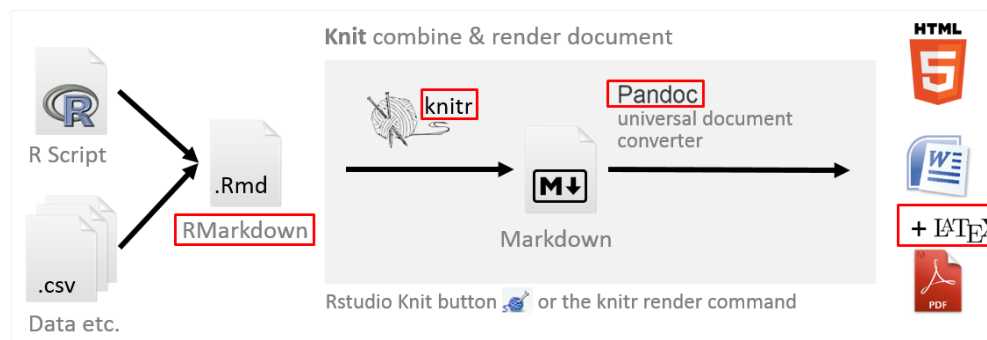


Figura 9.1: Elementos necesarios para construir reportes con RMarkdown

un reporte de resultados (texto, tablas de resultados estadísticos, gráficos, formato general del documento, referencias y listado de bibliografía utilizada). Por ejemplo, si construye una tabla de datos en cualquier software de análisis estadístico, luego la debe copiar y pegar en su reporte definitivo, editando allí su formato final. Si los datos o el sentido del análisis cambia, el autor debe repetir todo el proceso para actualizar la información su informe final. Por otro lado, en un *paradigma integrado* para la construcción de informes, como el que subyace a RMarkdown, existe un sólo lugar (archivo *RMarkdown*) donde se edita tanto el texto del reporte, como los códigos para construir los resultados a incluir, así como el formato general del documento y la gestión bibliográfica. Si el documento debe actualizarse, se efectúan en un sólo lugar todos los cambios y se re-compila el informe en el formato final deseado (sea PDF, Editor de Texto tipo Microsoft Word, diapositivas de presentación o un archivo dinámico tipo HTML) (Grolemund, 2014).

Ahora bien, ¿cómo funciona esta herramienta? Crear reportes desde RMarkdown combina diferentes aplicaciones computacionales. En tal sentido, es de las operaciones de uso de R que más precisa de la correcta instalación de una serie de elementos adicionales. En una breve síntesis, la construcción de informes de resultados desde R demanda la ejecución simultánea de cuatro elementos (Grolemund, 2014):

1. **RMarkdown.** Basado en el lenguaje *markdown* - funcionalidad que busca convertir rápida y fácilmente texto plano tipo bloc de notas a formato HTML - RMarkdown es un tipo de documento de RStudio que permite integrar texto con código de R.
2. **Knitr.** Este paquete integra en un sólo archivo *markdown* el texto ingresado en formato RMarkdown y los resultados de la ejecución de los códigos construidos mediante R.
3. **Convertor a formato final:**
  - **Pandoc.** Se trata de un paquete de R que convierte el formato *markdown* a alguno de los diferentes formatos de reporte ya señalados (HTML y editor de texto tipo Word).
  - **LaTeX.** Es una aplicación computacional en sí misma, enfocada en la *preparación de documentos* para su publicación con una alta calidad profesional del formato final. Está pensado para ser utilizado en procesos editoriales de alta complejidad y exigencia de calidad. Permite convertir los documentos *markdown* a PDF (Navarro, 2014).<sup>3</sup>

Todos estos elementos se ponen en juego en un flujo de trabajo (*workflow*) que, luego de una corta espera, generará un documento final con una elevada calidad final en su presentación de resultados. En la imagen a continuación (Workshop, 2016), se grafica tal flujo.

Para ejecutar correctamente un documento de RMarkdown y lograr construir un reporte final en el formato deseado, es preciso asegurar que todos estos elementos estén instalados en el computador. A continuación se presenta una breve tabla que resume los elementos a instalar y cómo hacerlo.

<sup>3</sup>Es importante señalar que *RMarkdown* simplifica el lenguaje LaTeX, haciendo más intuitivo su uso para un usuario que no está enfocado en la utilización profesional de sus prestaciones. Desde este lenguaje viene la impronta asimilada por RMarkdown de separar el **contenido** del **formato** de un documento.

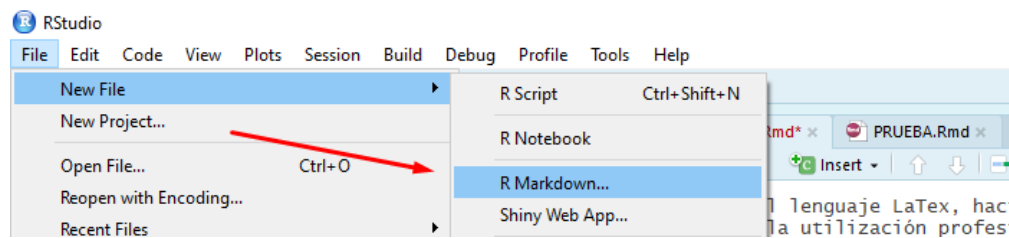


Figura 9.2: Apertura de nuevo documento Rmarkdown vía botonera

Tabla 9.1: Elementos necesarios para ejecución de RMarkdown: procedimientos de instalación

Elemento	Procedimiento de instalación	Código para instalación
<b>RMarkdown</b>	No es necesario ningún procedimiento adicional a la instalación de R y RStudio pues viene instalado con este último.	<i>No aplica</i>
<b>Knitr</b>	Debe instalarse como cualquier otro paquete de R, asegurando su disponibilidad para ser utilizado por RMarkdown al compilar documentos.	<code>install.packages("knitr")</code>
<b>Pandoc</b>	No es necesario ningún procedimiento adicional a la instalación de R y RStudio pues viene instalado con este último.	<i>No aplica</i>
<b>LaTeX</b>	Debe descargarse como un paquete de R. Para asegurar su disponibilidad para ser utilizado por RMarkdown al compilar documentos, debe instalarse con un comando adicional. <sup>4</sup>	<code>install.packages("tinytex")</code> <code>tinytex::install_tinytex()</code>

En síntesis: antes de proceder a la compilación de un documento RMarkdown, debemos asegurar la instalación de estos componentes en nuestros computadores.

## 9.2 Los diferentes elementos de una sintaxis de RMarkdown

En este apartado se indican las nociones generales a tener en cuenta para trabajar con una sintaxis de RMarkdown.

El procedimiento básico para abrir una sintaxis de RMarkdown es similar al usado para una sintaxis de R. Es posible crear un nuevo documento de RMarkdown usando la botonera superior de RStudio. Como se observa a continuación, yendo a *Archivo*, *Nuevo Archivo* y seleccionando la opción *RMarkdown*.

Una vez hecho tal procedimiento aparecerá una nueva pantalla de configuración general. Aquí se establece la configuración básica del reporte de resultados, indicando un título para el documento, el nombre del autor, así como el formato de informe a construir (en este ejemplo se selecciona el formato *PDF*).

<sup>4</sup>Existen diferentes versiones de LaTeX, de descarga gratuita. Sin embargo, estas distribuciones de LaTeX (por ejemplo MikTeX es una de las más conocidas para sistema operativo Windows) presentan diversas limitaciones y complicaciones de configuración que hacen difícil su uso integrado con RMarkdown. Por eso se sugiere usar TinyTeX, paquete diseñado para disminuir tales complicaciones y facilitar la integración de RMarkdown con las funcionalidades de LaTeX. Para mayores detalles se recomienda revisar la página web de su desarrollador.

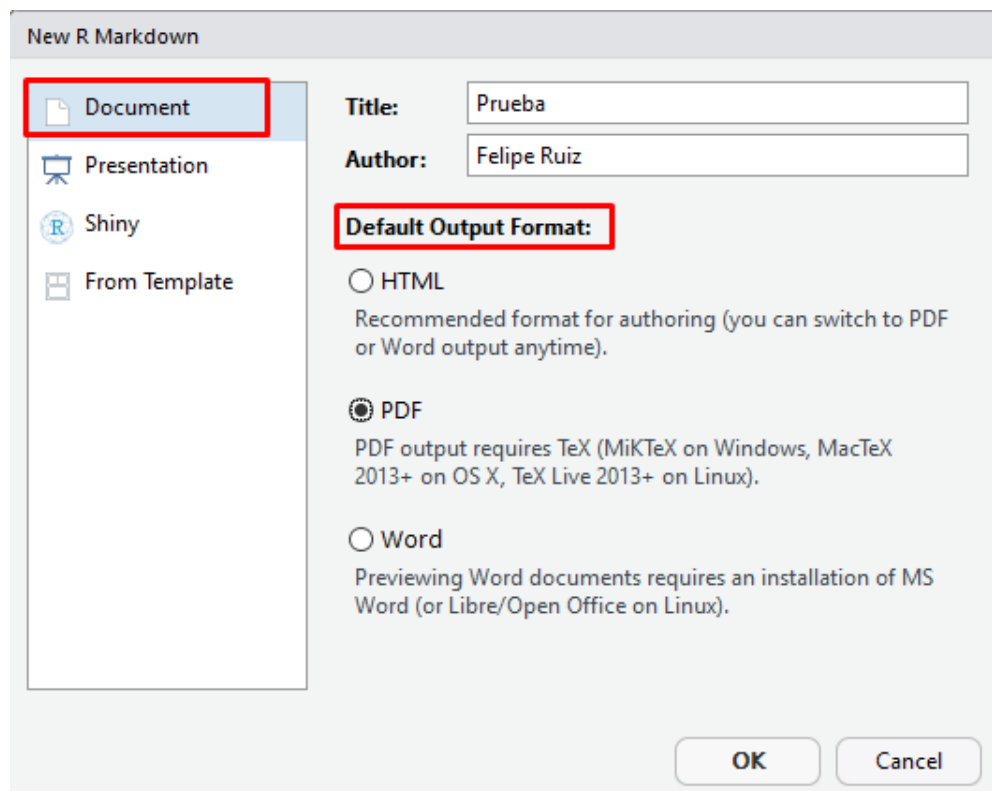


Figura 9.3: Ventana de preconfiguración del documento de RMarkdown

Si se especifican tales condiciones, luego de apretar *aceptar* se desplegará una nueva pestaña en R (con extensión *.Rmd*) que será la sintaxis de RMarkdown en la cual se editará el reporte de investigación. Como se observa en la imagen a continuación RStudio ofrece una sintaxis de RMarkdown preconfigurada.

De esta sintaxis conviene poner atención sobre tres tipos de elementos que resultan distintivos del trabajo con RMarkdown.

1. **Encabezado general de formato.** Al inicio del documento, y encerrado por tres guiones continuados, por arriba y por abajo de manera respectiva, se encuentra el encabezado que permite establecer los parámetros generales de formato que estructurarán el producto final (en términos técnicos se denomina como encabezado *YAML*). En la imagen destacada se observan los campos (en inglés): *título del documento*, *autor del documento*, *fecha* y *formato*.
2. **Texto.** A diferencia de una sintaxis de R, RMarkdown considera como formato primario de redacción el texto plano. Todo lo que se escriba - a no ser que se indique lo contrario - será considerado como texto en el documento final.
3. **Trozos de código de R.** Para indicar que se escribirá un código de R, las líneas de comando deben encerrarse entre tres cremillas seguidas de corchetes curvos con la letra *r* en su interior; luego de eso se escriben los comandos de R; al finalizar los comandos, se agrega una última línea con tres cremillas para indicar el cierre del trozo de código.

Si el documento de RMarkdown construido por defecto al abrir uno nuevo se compila en formato PDF, el inicio de este documento se verá como sigue. Se observa que los campos puestos en el encabezado han pasado a ser parte del título general del documento. A la vez, se distingue el texto plano del código de R que, en este caso, integra en el documento el resultado de una función `summary`:

En los siguientes apartados se explicará la configuración específica de diferentes elementos de formato básico para la edición de cualquier informe de resultados vía RMarkdown. Se ilustrará la visualización básica de



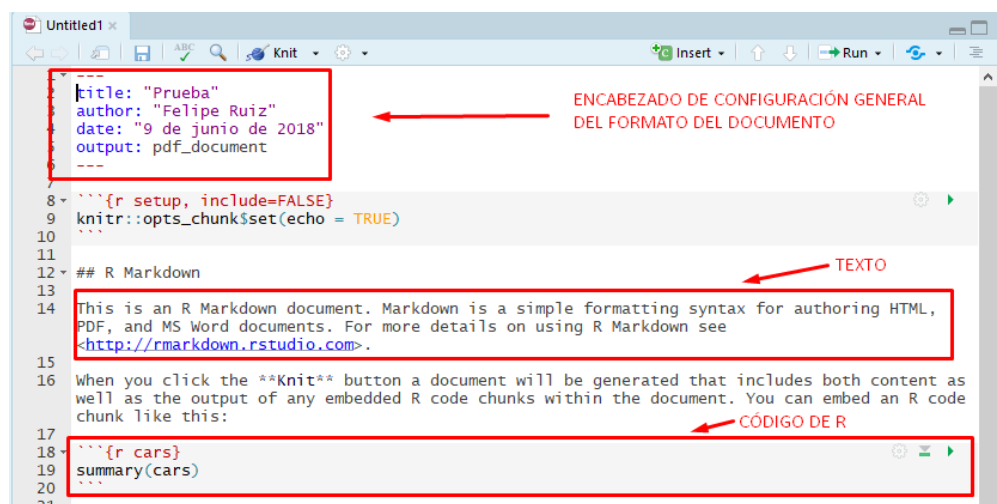


Figura 9.4: Sintaxis Rmarkdown con preconfiguración

## Prueba

*Felipe Ruiz*

*9 de junio de 2018*

### R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed      dist
## Min.   : 4.0    Min.   :  2.00
## 1st Qu.:12.0    1st Qu.: 26.00
## Median :15.0    Median : 36.00
## Mean   :15.4    Mean   : 42.98
## 3rd Qu.:19.0    3rd Qu.: 56.00
## Max.   :25.0    Max.   :120.00
```

Figura 9.5: Resultado de ejecución de plantilla básica de RMarkdown

los diferentes elementos indicados en RMarkdown. Todos los elementos indicados pueden ser encontrados de manera unificada en el siguiente enlace, donde nuestros lectores encontrarán una plantilla básica de RMarkdown disponible para descarga, así como el documento PDF que resulta de su ejecución.

## 9.3 Aspectos generales (formato de texto)

### 9.3.1 Formato general: encabezado YAML

Este encabezado determina los parámetros generales de formato para el reporte a compilar. En lenguaje computacional se entiende como los *metadatos* del documento; esto es, información que define el formato del archivo resultante, más no su contenido. En el ejemplo que aquí presentamos se observan las siguientes definiciones:

- **Título** (*title*). Texto entre comillas que servirá de título general al documento.
- **Subtítulo** (*subtitle*). Texto entre comillas que servirá de subtítulo para el título general del documento.
- **Autor** (*author*). Texto entre comillas para indicar el nombre del o los autores.
- **Fecha** (*date*). Campo para indicar la fecha. En el texto, con la expresión *today*, se solicita que imprima la fecha actual según el calendario del sistema operativo.
- **Bibliografía** (*bibliography*). Se indica el nombre del archivo que contiene los datos para construir el listado de referencia bibliográficas. Como será explicado más adelante, este archivo es de formato BibTeX y se construye usando un gestor de referencias como *Zotero*.
- **Formato de bibliografía** (*csl*). Nombre de un archivo de extensión *.csl* para indicar el formato de referencias en el cuerpo del texto y el listado de bibliografía al final del documento. La sigla refiere a *Estilo del Lenguaje de Referencias* (*Citation Language Style*) y permite definir si se usarán citas al estilo APA, ASA, Chicago, etc. Se adecúa a los diferentes requerimientos de referencias bibliográficas.
- **Color de los enlaces**. Mediante las opciones *linkcolor* y *urlcolor* se define el color asignado al indicador de notas al pie y a los enlaces a páginas web, respectivamente. En este caso se sugiere el color azul (*blue*).
- Los argumentos dentro del apartado *resultado* (*output*) son los siguientes:
  - **pdf\_document**: indica el formato preestablecido para compilar el documento. En este caso, se trata de un PDF. Puede ser *html\_document* o *word\_document*. El usuario puede escoger la modalidad que desee al compilar usando las opciones del botón *knit*; si se compila sin escoger ninguna opción, se compilará según el formato indicado en este encabezado.
  - **fig\_caption**: indica si las figuras deben incorporar leyendas.
  - **latex\_engine**: permite definir el motor de LaTeX utilizado para compilar los documentos.
  - **number\_sections**: si está definido como *yes* define que se numerarán los títulos y subtítulos a lo largo del documento, de manera automática y correlativa.
  - **toc**: es la abreviación de *table of contents*; si está definido como *yes* incorporará al inicio del documento una tabla de contenidos construida a partir de los tres primeros niveles de los títulos y subtítulos de sección utilizados.

#### Ejercicio 9.1

```
---
title: "Ejemplo Uso de RMarkdown"
subtitle: "Material de apoyo docente, asignatura Estadística Descriptiva - Semestre de otoño 2018"
author: "Giorgio Boccardo Bosoni y Felipe Ruiz Bruzzone"
date: "\today"
bibliography: bibliografia.bib
csl: apa.csl
linkcolor: blue
```

**Contents**

<b>1 Ejemplo de Título 1</b>	<b>1</b>
1.1 Ejemplo de Título 2 . . . . .	1

**1 Ejemplo de Título 1**

El texto antecedido de signo gato genera un título. El texto simple, indica cuerpo de texto para el documento. Al compilar el documento, el programa utiliza estas expresiones para construir la tabla de contenidos al inicio del documento,

**1.1 Ejemplo de Título 2**

Adicionando signos gatos, se construyen títulos de menor jerarquía. Por tanto, funcionan como subtítulos para la estructura interna de cada capítulo.

**1.1.1 Ejemplo de Título 3**

Figura 9.6: Compilación de Rmarkdown con índice, títulos y subtítulos

```
urlcolor: blue
output:
  pdf_document:
    fig_caption: yes
    latex_engine: xelatex
    number_sections: yes
    toc: yes
---
```

En definitiva, todos estos elementos configuran una importante información para indicar el formato general del documento que se busca compilar. No se trata específicamente del contenido a utilizar pero es, por así decirlo, toda la información de “contexto” que el software necesita manejar para construir un reporte final con el formato deseado.

### 9.3.2 Títulos, subtítulos y saltos de página

En el caso de RMarkdown los signos “gato” (#) indican que se está señalando un título o subtítulo de sección. Mientras más signos gatos se pongan se estará indicando que se trata de un título de menor jerarquía. Por otra parte, con el comando `\pagebreak` se pueden indicar saltos de página. Esto último sirve, por ejemplo, para comenzar cada sección del documento en una página nueva.

#### Ejercicio 9.2

```
# Título 1 (nivel de mayor jerarquía)
## Título 2 (nivel de segunda jerarquía)
### Título 3 (nivel de tercera jerarquía)

\pagebreak #Salto de página.
```

Si se utilizan tales instrucciones el documento final irá adquiriendo la siguiente forma:

### 9.3.3 Énfasis de texto

Otro elemento importante son los distintas marcas que permiten introducir diferentes énfasis en el texto. A lo largo del texto se pueden introducir las clásicas configuraciones de formato que, en los procesadores de texto más utilizados, se configuran mediante la selección del texto con el cursor y la utilización de uno o más botones. A continuación se muestra cómo se indican tales configuraciones en el texto plano (sin compilar):

#### Ejercicio 9.3

```
**negrita**

*cursiva*

_subrayado_

> Texto con un tabulado mayor al párrafo normal.

Texto que está hablando de un tema y quiere poner una nota al pie [1].

[1]: texto de la nota al pie

Para ingresar un enlace asociado a una palabra, se debe encerrar la palabra o palabras
que [queremos sea un enlace](www.url.com)
```

Considerando tales elementos, a continuación se presenta un texto con sus especificaciones de énfasis en formato RMarkdown y a continuación su resultado luego de la compilación.

#### Ejercicio 9.4

```
Así, si se aplica lo recién anotado, podemos configurar palabras en **negrita** y en
*cursiva*.

> Podemos incluir un párrafo con un tabulado mayor al del párrafo regular.

Asimismo, en cualquier parte del documento podemos incluir una nota al pie [1].

[1]: texto de la nota al pie. En cualquier parte del documento, aquí lo haremos en esta nota al pie, p
```

### 9.3.4 Listas

El proceso de construcción de listas de elementos en RMarkdown es muy sencillo. Pueden ocuparse números, para listas numeradas o signos asterisco (\*) o signos más (+), para definir listas no numeradas.

#### Ejercicio 9.5

```
#Construcción de una lista numerada

1. Elemento 1
2. Elemento 2
```

A lo largo del texto se pueden introducir las clásicas configuraciones de formato que, en los procesadores de texto más utilizados, se configuran mediante la selección del texto con el cursor y la utilización de uno o más botones. A continuación se muestra cómo se ven tales configuraciones en el texto plano (sin compilar):

Así, si aplicamos lo recién anotado, podemos configurar palabras en **negrita** y en *cursiva*.

Podemos incluir un párrafo con un tabulado mayor al del párrafo regular.

Asimismo, en cualquier parte del documento podemos incluir una nota al pie <sup>1</sup>.

---

<sup>1</sup>texto de la nota al pie. En cualquier parte del documento, aquí lo haremos en esta nota al pie, podemos definir enlaces que [redirigir a páginas web](#).

Figura 9.7: Compilación de documento RMarkdown con énfasis de texto

#### Construcción de una lista numerada

1. Elemento 1
2. Elemento 2

#### Construcción de una lista sin números

- Elemento 1
- Elemento 2

#### Lista sin números, con sublista numerada

- Elemento 1
- Elemento 2
  1. Sub elemento 1.
  2. Sub elemento2.

Para una sublista el tabulado se consigue con 4 espacios

Figura 9.8: Compilación de documento RMarkdown con listas

*#Construcción de una lista sin números*

```
* Elemento 1
* Elemento 2
```

*#Lista sin números, con sublista numerada (tabulado es de 4 espacios)*

```
+ Elemento 1
+ Elemento 2
  1. Sub elemento 1.
  2. Sub elemento2.
```

Al compilar los elementos recién expuestos, se obtiene el siguiente resultado:

### 9.3.5 Tablas

La construcción de tablas en RMarkdown se realiza separando los elementos de cada columna con un tabulador vertical (`|`). Cada fila de texto corresponderá a una fila de la tabla. Se puede definir la alineación del texto según la distribución de guiones debajo de los títulos. Si los guiones exceden hacia ambos lados del texto, este quedará centrado. Si excede hacia uno u otro lado quedará alineado hacia la izquierda o la derecha. Las reglas para destacar texto, incluir enlaces y notas al pie, también aplican las tablas. No es necesario separar cada fila con guiones, basta comenzar una nueva fila de texto.

#### Ejercicio 9.6

Table: Tabla simple

```

**Título Columna 1** | **Título columna 2** |
-----|-----|
*Texto 1* | Texto 2
*Texto 3* | Texto 4

```

A continuación se ve el resultado de la ejecución del comando anterior resultando una tabla simple en formato RMarkdown.

Tabla 9.2: Tabla simple

<b>Título Columna 1</b>	<b>Título columna 2</b>
<i>Texto 1</i>	Texto 2
<i>Texto 3</i>	Texto 4

### 9.3.6 Bibliografía

Este apartado asume que quien lee este documento tiene experiencia con gestores electrónicos de referencias bibliográficas como Zotero. Por lo tanto no se profundizará en explicaciones sobre cómo utilizarlo: se sugiere revisar esta documentación preparada por la unidad de *Información y Bibliotecas* de la Universidad de Chile, en torno al uso del software *Zotero* como gestor de referencias.

Para incorporar una bibliografía formal a un reporte, tanto en lo que respecta a referencias en el cuerpo del texto como para construir un listado de referencias bibliográficas utilizadas al final del documento, se deben manejar dos elementos:

1. *Zotero y Zotero connector*. Mediante este software se crea un listado bibliográfico compatible con LaTeX (archivo *.bib*) que - cargado a RMarkdown - permitirá insertar referencias y que luego se construya un listado bibliográfico. Ambos pueden descargarse desde este enlace.
2. Contar con un archivo *.csl* para indicarle el formato de cita al documento (*Citation Language Style*). Se puede descargar el adecuado para formato APA 6a Edición desde esta página. También se pueden descargar otros formatos como *Vancouver*, *Chicago* o *ASA* (*American Sociological Association*).

El primer elemento se construye exportando un listado de bibliografía desde Zotero al formato BibTex. Como se observa en la imagen a continuación es una exportación simple. Se opera haciendo click derecho sobre el conjunto de referencias bibliográficas que interesa, seleccionando la opción *exportar*. En la siguiente pestaña se selecciona el formato *BibTex* para luego introducir el nombre del archivo de extensión (*.bib*) a guardar. Importa guardarlo en la misma carpeta que funcione como carpeta de trabajo para el documento RMarkdown.

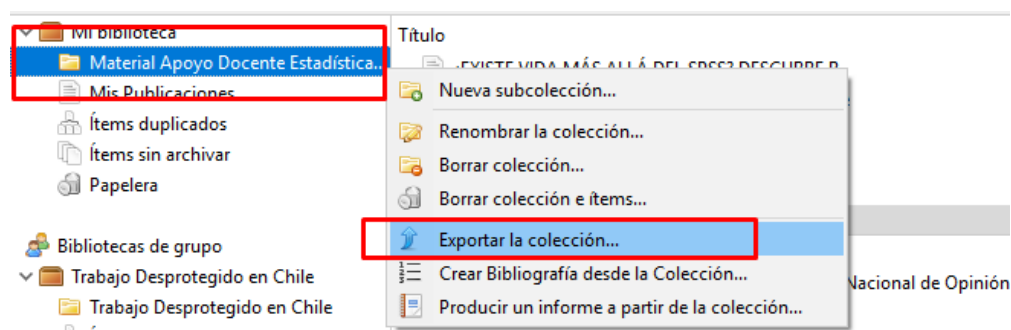


Figura 9.9: Construcción de archivo de referencia bibliográficas usando Zotero

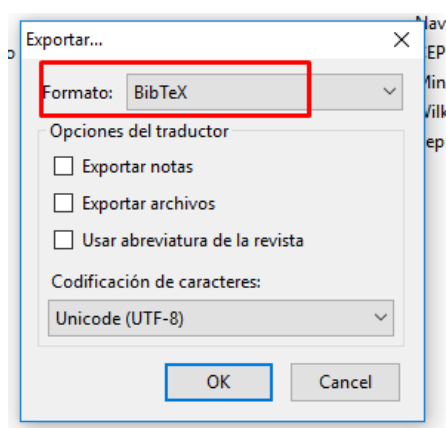


Figura 9.10: Elección de formato ".bibtex"

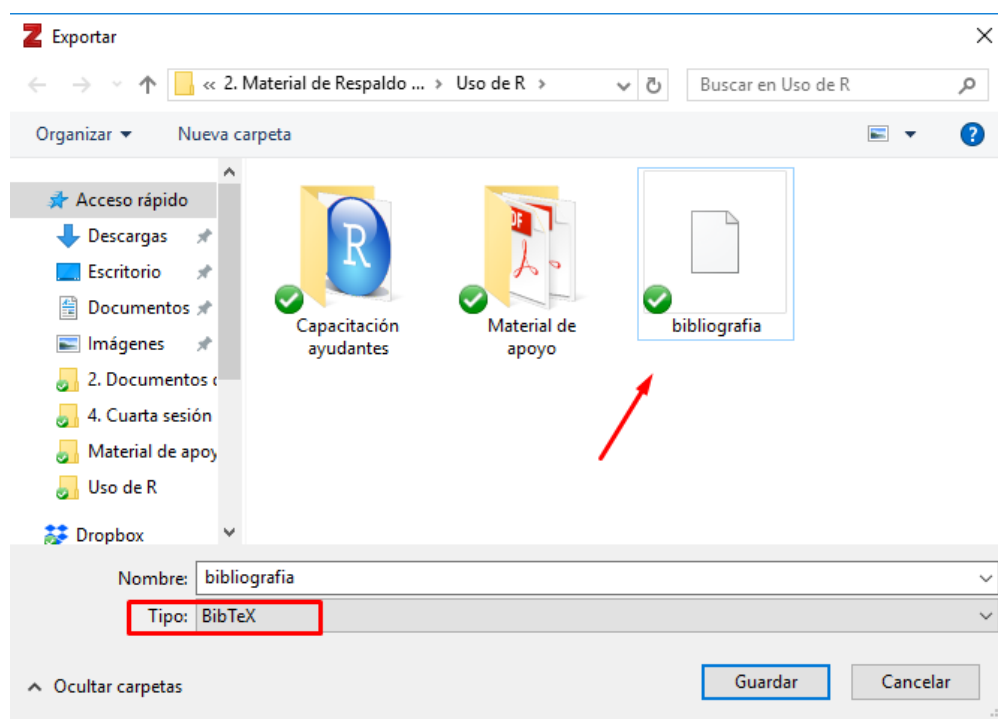


Figura 9.11: Elección de formato ".bibtex"

## Zotero Style Repository

Here you can find [Citation Style Language](#) 1.0.1 citation styles for use with [Zotero](#) and other CSL 1.0.1-compatible software. [Zotero wiki](#).

**Style Search** **Format:**

**Fields:**

☐ Show only unique styles

33 styles found:

- [American Psychological Association 5th edition](#) (2014-09-06 22:02:33)
- [American Psychological Association 6th edition](#) (2018-05-22 00:59:23)

Figura 9.12: Descarga de archivo CSL para formato de referencias

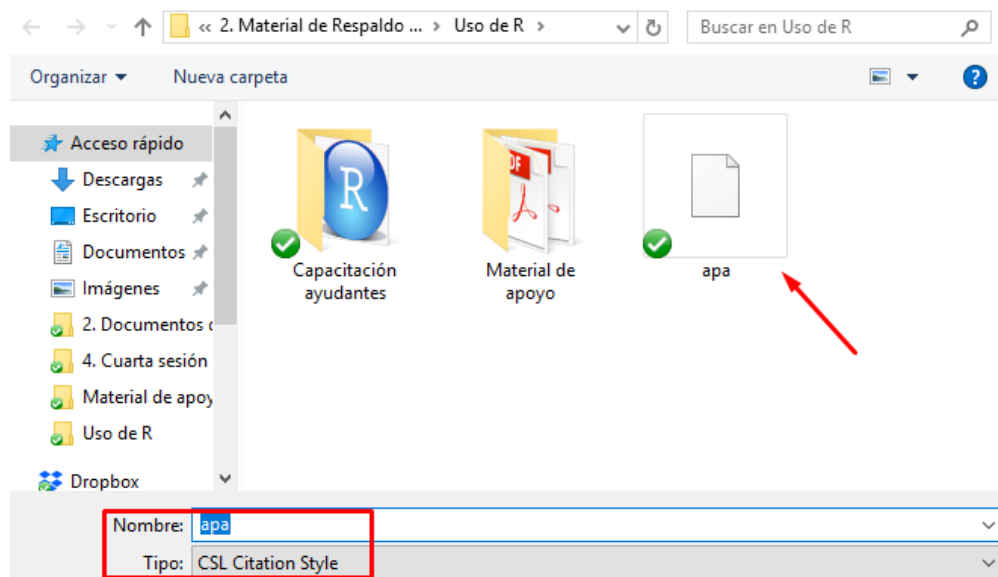


Figura 9.13: Archivo CSL para formato APA de referencias

El segundo elemento (archivo *CSL*) se puede descargar desde la página web de Zotero. Como se observa en las imágenes a continuación, mediante el buscador puede indicarse el estilo deseado (en este caso, APA). Luego se descarga la versión de formato de referencias que se adecúe al requerimiento específico. Para este material se utiliza la 6ta versión del formato APA de referencias bibliográficas, que además está recientemente actualizada. Tal archivo también debe quedar guardado en la carpeta de trabajo del documento RMarkdown.

En el encabezado del archivo RMarkdown se deben indicar los nombres de estos dos archivos para poder construir de forma adecuada las referencias bibliográficas. Ahora bien, ¿cómo agregar las referencias en el cuerpo del texto? Para determinar cómo se recomienda abrir con el *bloc de notas* el archivo *.bib* donde está el listado bibliográfico.

Si se observa la estructura interna del archivo CSL. Se notará que a lo largo del texto se van definiendo los diferentes campos de información que tiene cada referencia bibliográfica (autor, año, título, edición, tipo de referencia, etc.). Así, hay que encontrar el *identificador de la referencia*, luego del signo arroba (@). Abriendo un paréntesis de corchetes en el cuerpo del documento e incluyendo tal texto luego de una arroba, se incorporará una (Elousa, 2009) o más referencias bibliográficas (Elousa, 2009, 22; Grolemond, 2014).



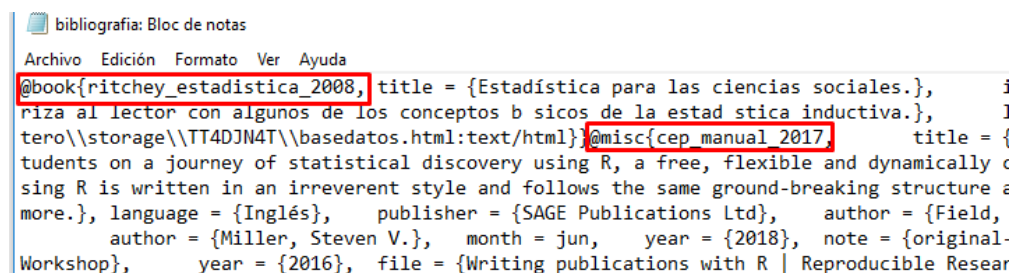


Figura 9.14: Estructura interna del archivo CSL

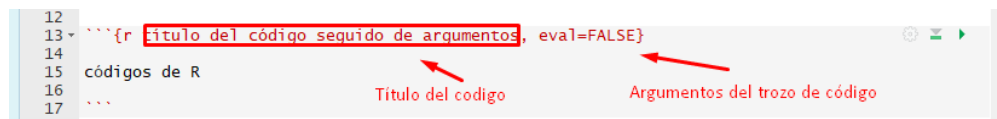


Figura 9.15: Trozo de código de R en documento RMarkdown

A continuación se observa el texto plano que define las referencias recién escritas:

### Ejercicio 9.7

Abriendo un paréntesis de corchetes en el cuerpo del documento e incluyendo tal texto luego de una arroba, se incorporará una `[@elousa_existe_2009]` o más referencias bibliográficas `[@elousa_existe_2009, 22; @grolemund_introduction_2014]`.

La bibliografía siempre se compila al final del documento. Eso significa que estará al final del último elemento escrito. El listado bibliográfico final se contruirá siempre con base a las citas realizadas en el cuerpo del texto y según todas las referencias presentes en el archivo de referencias exportado desde zotero. Podemos asegurarnos de que el último título del documento sea algo como *Referencias bibliográficas*, para que tal contenido tenga un encabezado apropiado.

## 9.4 Aspectos generales (configuración de trozos de código)

El tercer elemento de importancia (además del *encabezado de configuración* y del *texto* propiamente tal) son los *trozos de código de R* o *code chunks* (en inglés). Se trata de bloques de código delimitados por la siguiente estructura: el inicio de un código está delimitado por tres apóstrofes seguidos por un `r` entre corchetes curvos `{r}`, y su cierre por otros tres apóstrofes. Eso delimita lo que se ejecutará como código de computación, diferenciándolo respecto al texto simple. A continuación se observa un ejemplo de esto:

El título del código sirve para incorporarlo a la estructura de contenidos del archivo RMarkdown de manera similar a los títulos y subtítulos de secciones de texto. Por otra parte los argumentos sirven para configurar el comportamiento del código al momento de compilar el documento. Para introducir un trozo de código en nuestra sintaxis de RMarkdown se puede utilizar el botón *insert* existente en la botonera de la sintaxis RMarkdown, como se observa en la siguiente imagen: debemos seleccionar la primera opción que indica el lenguaje de programación R. Las otras opciones indican una de las potencias de RMarkdown, permite integrar en una sola plataforma diferentes lenguajes de programación.

Como no siempre se buscará que en el reporte final se despliegue la sintaxis original, o los mensajes y/o advertencias que reporta R luego de ejecutar un comando, es posible configurar la ejecución de cada trozo de código agregando diferentes opciones.

Por ejemplo, al abrir el siguiente código de nuestra sintaxis de Markdown con las siguientes opciones `{r, echo = FALSE, results = 'asis', message= FALSE}` les estamos indicando lo siguiente:

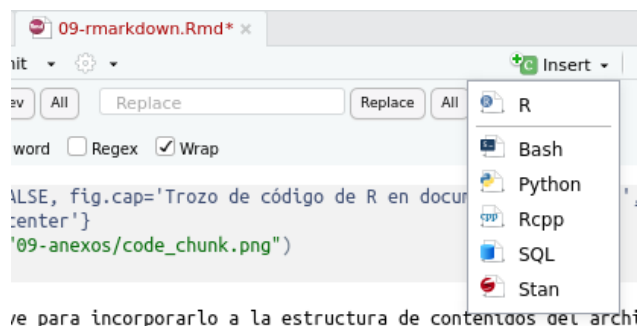


Figura 9.16: Insertar trozos de código en RMarkdown

- `echo = FALSE` significa que no se desplegará la sintaxis en el reporte, pero sí se ejecutará la operación y mostrarán los resultados.
- `results = 'asis'` indica que el resultado se exportará directamente al nuevo archivo, sin que sea configurado por *RMarkdown*. Esto es útil con funciones que formatean de manera inmediata los resultados al formato deseado.
- `message = FALSE` indica que no se mostrarán los mensajes de información en el informe final.

En la siguiente tabla se indican algunos de los argumentos de mayor utilidad para configurar trozos de código en RMarkdown.

Tabla 9.3: Algunos rgumentos para configurar los trozos de código en RMarkdown

Argumento	Valor por defecto	Detalle
<b>eval</b>	TRUE	Si se configura como FALSE, R sólo mostrará, pero no correrá el código
<b>include</b>	TRUE	Si se configura como FALSE, R no mostrará el código, pero correrá el comando y mostrará sus resultados.
<b>error</b>	TRUE	Si se configura como FALSE, R no mostrará los mensajes de errores que resulten de la ejecución del código.
<b>results</b>	—	Si se configura como <i>hide</i> , R no mostrará los resultados del código aunque lo ejecutará tras bambalinas. Si se configura como <i>delay</i> , R mostrará sólo el último resultado del trozo de código. Si se configura como <i>asis</i> R no configurará con <i>markdown</i> la estructura de los resultados, imprimiéndolos de manera directa en el reporte final. Esto resulta útil cuando usamos funciones específicas para presentar resultados en algún formato de reporte específico (Word, PDF, etc.)
<b>warning</b>	TRUE	Si se configura como FALSE R no mostrará los mensajes de advertencia que resulten de la ejecución del código.
<b>message</b>	TRUE	Si se configura como FALSE R no mostrará ningún tipo de mensaje que resulten de la ejecución del código.

Se indican sólo algunos de muchos argumentos posibles de utilizar. Para un listado completo de estas configuraciones posibles para los trozos de código, así como para la configuración y uso general de RMarkdown, se sugiere ver las diapositivas *RMarkdown Cheat Sheet* y *RMarkdown Reference Guide*. Estos materiales preparados por el equipo desarrollador de RStudio y RMarkdown resumen de manera muy sintética la mayor cantidad de argumentos y estructuras de código que pueden utilizarse en RMarkdown. Puede accederse a ello mediante la botonera de RStudio, como se muestra a continuación:

## 9.5 Presentación de resultados básicos en RMarkdown

Habiendo indicado una parte importante de los elementos básicos de redacción de RMarkdown, a continuación se explica como incorporar en un reporte que contenga tablas de frecuencias y de estadísticos descriptivos

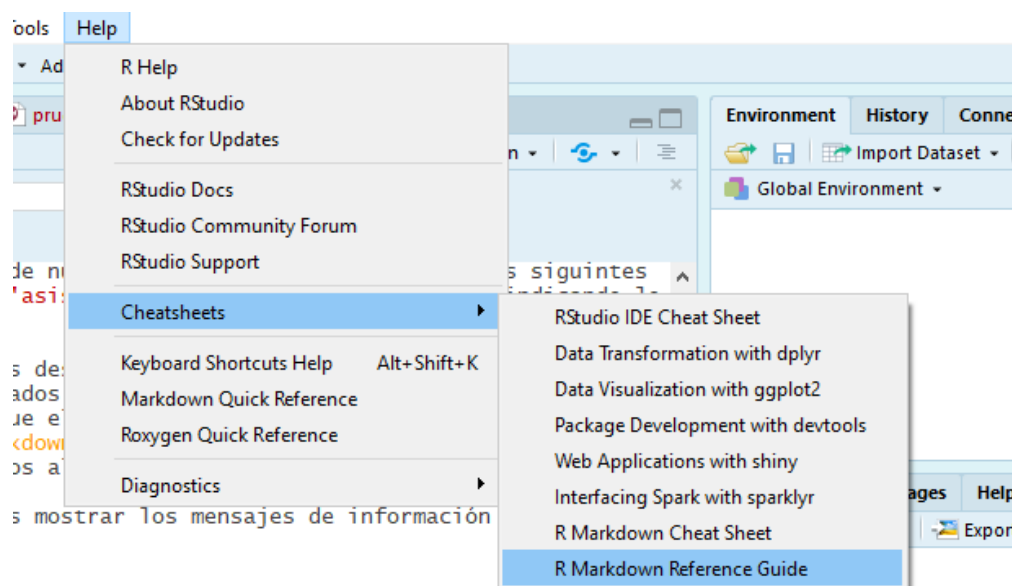


Figura 9.17: Uso de material de apoyo para RMarkdown

construyen y *cómo* son presentados (Comtois, 2018). Esto hace que las funciones de este paquete sirvan para construir resultados tanto para un **uso exploratorio** como para el diseño de su presentación en reportes de investigación, es decir un **uso orientado a la divulgación**.<sup>5</sup>

Debido a esta última característica las funciones de este paquete permiten obtener resultados en texto plano para ser visualizados en la consola de R (lo que sería el comportamiento predefinido para casi cualquier función de R), como también especificar formatos de presentación compatibles con RMarkdown para hacer posible su compilación en reportes.

### 9.5.1 Tablas de frecuencias

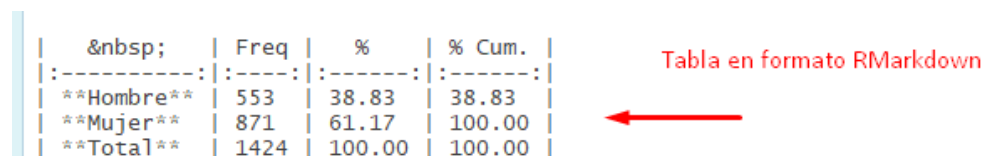
El primero comando a utilizar de este paquete será `freq`. Permite obtener una tabla de frecuencias absolutas y relativas para una variable de tipo nominal u ordinal. En el primer comando se utiliza el formato por defecto para la construcción de resultados en la consola de R (argumento `style = simple`). Con el argumento `justifiy` se indica la alineación del texto de la tabla, en este caso se configura un formato de texto **centrado** ("center"). Luego, con el argumento `omit.headings = TRUE` se indica que no se imprima un encabezado con la especificación de la tabla solicitada.

#### Ejercicio 9.8

```
library(summarytools)
freq(CEP$sexo_factor, style = "simple", justify = "center", omit.headings = TRUE)
```

```
##
##           Freq  % Valid  % Valid Cum.  % Total  % Total Cum.
## -----
##   Hombre    553    38.83    38.83      38.83    38.83
##   Mujer     871    61.17   100.00      61.17   100.00
```

<sup>5</sup>Para una introducción (en inglés) detallada al uso general de este paquete, se recomienda la lectura del siguiente documento elaborado Dominic Comtois para el sitio oficial de R (2018).



	&nbsp;	Freq	%	% Cum.
	***Hombre***	553	38.83	38.83
	***Mujer***	871	61.17	100.00
	***Total***	1424	100.00	100.00

Figura 9.18: Tabla de frecuencias en formato RMarkdown

```
##      <NA>      0      0.00      100.00
##      Total    1424    100.00    100.00    100.00    100.00
```

En la tabla resultante se presentan las frecuencias absolutas y relativas (expresadas como porcentaje) para la variable **sexo**, indicando también los porcentajes acumulados válidos (es decir, sin contar los NA's) y los porcentajes acumulados totales. Esta tabla es construida por defecto por la función, pero no aquella con un formato específico para RMarkdown.

En el siguiente trozo de código de R se configura este comando para su ejecución con RMarkdown. En primer lugar, destaca la configuración específica del trozo de código: mediante el argumento **results = 'asis'** se configura que el resultado sea impreso directamente al formato RMarkdown para que la tabla resultante esté construida en la lógica de construcción de tablas que ya ha sido presentada en este documento; en segundo lugar, con el argumento **warning = FALSE** se especifica que en el reporte no sean impresos los mensajes de alerta que pueda arrojar la ejecución del comando. También se puede incorporar el argumento **echo = FALSE** para evitar que el trozo de código se visualice en el reporte de resultados, dejando como resultado de esta operación sólo la compilación de la tabla de frecuencias; aquí se utiliza pues interesa mostrar el trozo de código con fines pedagógicos.

En segundo lugar, destacan las configuraciones adicionales establecidas para la ejecución del código con formato adecuado a RMarkdown. En primer lugar, el *estilo* definido para la tabla se ha indicado como "rmarkdown". Esto último permite que la tabla resultante sea como se observa en la siguiente imagen: integra los resultados a un tabulado de texto plano que permitirá contar con una tabla adecuadamente configurada al compilar el documento con RMarkdown.

Finalmente, el argumento **report.nas = FALSE** le indica al software que no incluya en la tabla el conteo de casos codificados como NA (perdidos), puesto que ya se sabe que esta variable no presenta casos de este tipo.

### Ejercicio 9.9

```
library(summarytools)
freq(CEP$sexo_factor, style = "rmarkdown", justify = "center", omit.headings = TRUE,
     report.nas = FALSE)
```

	Freq	%	% Cum.
<b>Hombre</b>	553	38.83	38.83
<b>Mujer</b>	871	61.17	100.00
<b>Total</b>	1424	100.00	100.00

Así, luego de incorporar el comando como trozo de código en un documento RMarkdown se obtiene una tabla con un diseño estético adecuado a RMarkdown que resulta apropiada como formato final para la presentación de resultados en un reporte de investigación.

### 9.5.2 Tabla de estadísticos univariados

Una segunda herramienta de utilidad del paquete **summarytools** es la función **descr** que permite obtener tablas de estadísticos univariados de nivel muestral, para variables continuas (de intervalo o razón). En el primer trozo de código se ejecuta la función en su configuración por defecto, indicando solamente argumentos adicionales para la *alineación del texto* y la presentación del *encabezado* (ya explicados para la función **freq**).

### Ejercicio 9.10

```
descr(CEP$edad, style = "simple", justify = "center", omit.headings = T)
```

	edad
Mean	49.87
Std.Dev	17.79
Min	18.00
Q1	36.00
Median	50.00
Q3	64.00
Max	97.00
MAD	20.76
IQR	28.00
CV	0.36
Skewness	0.01
SE.Skewness	0.06
Kurtosis	-0.93
N.Valid	1424.00
Pct.Valid	100.00

El resultado obtenido es una tabla vertical, con trece estadísticos univariados de tipo descriptivo. Este resultado podría configurarse para su presentación mediante RMarkdown simplemente indicando que utilizaremos el *estilo* de tabla *RMarkdown*. Sin embargo, como se observa en el siguiente trozo de código, se recomienda incorporar algunas configuraciones adicionales.

1. Indicar mediante el argumento `transpose = TRUE` que el formato final de la tabla sea construido horizontalmente; es decir, que tanto el título de cada estadístico como su valor sean ingresados como filas y no como columnas.
2. Indicar mediante un vector de valores alfabéticos el detalle específico de los estadísticos univariados descriptivos a incluir (y en qué orden) en la tabla de resultados.
3. Especificar una tabla construida en estilo *RMarkdown* con texto centrado y que omita el conteo de casos perdidos.

### Ejercicio 9.11

```
descr(CEP$edad, transpose = TRUE,
      stats = c("N.Valid", "min", "q1", "med", "mean", "sd", "q3", "max", "iqr"),
      style = "rmarkdown", justify = "center", omit.headings = T)
```

	N.Valid	Min	Q1	Median	Mean	Std.Dev	Q3	Max	IQR
<b>edad</b>	1424.00	18.00	36.00	50.00	49.87	17.79	64.00	97.00	28.00

De tal modo se construye una tabla de resultados adecuada para un reporte de investigación, con la información específica que interesa desplegar para análisis.

### 9.5.3 Tablas de estimación de parámetros

En este apartado se indica como construir tablas de presentación para los resultados de estimación de parámetros poblacionales a partir de resultados que ya han sido calculados en ejemplos anteriores.

#### 9.5.3.1 Intervalos de confianza para proporciones

Para el caso de construir una tabla para el intervalo de confianza de una proporción se considera lo ya ejecutado en el capítulo 7. Luego de contar con los resultados de tal ejercicio se procede a definir como vector simple a cada resultado. En el caso de los límites del intervalo de confianza simplemente se copia cada resultado (producto de las función `exactci`) y se almacena como un vector simple (`linf` y `lsup` respectivamente); lo mismo se realiza con el nivel de confianza (vector `nc`). Luego, mediante la función `cbind`, se configura una matriz (`data.frame`) con tales valores.

#### Ejercicio 9.12

```
library(PropCIs)
table(CEP$eval_econ_factor)

##
## Positiva    Neutra Negativa
##      471      730      209

nrow(CEP)

## [1] 1424

exactci(x = 730, n = 1424, conf.level = 0.95)

##
##
##
## data:
##
## 95 percent confidence interval:
##  0.4863248 0.5389039

#Definición de cada valor como vector simple
linf <- (0.4863248*100)
lsup <- (0.5389039*100)
nc <- 0.95*100

#Configuración de un data.frame a partir de los vectores creados
ICP <- cbind(linf, lsup, nc)
```

Con tales elementos se puede utilizar la función `kable` del paquete `knitr` para imprimir tal matriz de datos en formato RMarkdown. Así, se le indica el objeto a imprimir (ICP), con el argumento `caption` se le agrega un título a la tabla, mediante el argumento `align` se le indica la alineación del texto (en este caso, 'c' significa texto centrado), el argumento `digits` permite indicar la cantidad de decimales aceptados en el redondeo y finalmente el argumento `col.names` permite indicar un vector de caracteres con los títulos de cada columna. Especificar la opción `asis` en la configuración del trozo de código permite que los resultados se impriman sin ser editados en el formato RMarkdown.

Tabla 9.7: Tabla 1. Estimación de un intervalo de confianza para proporciones

Límite inferior	Límite superior	Nivel de confianza
48.63	53.89	95

**Ejercicio 9.12 (continuación)**

```
#Paquete necesario para imprimir tablas
library(knitr)
#Construcción de tabla de resultados con formato
kable(ICP, caption = "Tabla 1. Estimación de un intervalo de confianza para proporciones",
      align = 'c', digits = round(2),
      col.names = c("Límite inferior", "Límite superior",
                    "Nivel de confianza"))
```

**9.5.3.2 Intervalos de confianza para medias**

Algo similar a lo ya indicado se efectúa para construir una tabla de resultados de la estimación del intervalo de confianza de una media. En este caso se replica lo calculado en el capítulo 7 y los resultados de la función `ci.mean` se almacenan configurados como matriz de datos (usando la función `as.data.frame`) en un nuevo objeto denominado `ic`.

**Ejercicio 9.13**

```
# Intervalos de confianza
library(Publish)
#Nivel de confianza por defecto.
ci.mean(CEP$satisfaccion_vida)
```

```
## mean CI-95%
## 7.31 [7.20;7.42]
```

```
ic <- as.data.frame(ci.mean(CEP$satisfaccion_vida))
```

Sobre este objeto se aplica la función `kable` del paquete `knitr`. De forma similar al ejemplo anterior, se aplica la función señalada sobre el objeto `ic`; en este caso, se seleccionan las columnas 1, de la 3 a la 5, y la columna 2 (en ese orden). Con el argumento `caption` se le agrega un título a la tabla, mediante el argumento `align` se le indica la alineación del texto (en este caso, 'c' significa texto centrado), el argumento `digits` permite indicar la cantidad de decimales aceptados en el redondeo y finalmente el argumento `col.names` permite indicar un vector de caracteres con los títulos de cada columna. Especificar la opción `asis` en la configuración del trozo de código permite que los resultados se impriman sin ser editados en el formato RMarkdown.

**Ejercicio 9.13 (continuación)**

```
kable(ic[c(1,3:5,2)], caption = "Tabla 2. Estimación de intervalo de confianza para media",
      align = 'c', digits = round(2),
      col.names = c("Media", "Límite superior", "Límite inferior",
                    "Nivel de confianza", "Error estándar"))
```

Con tales configuraciones ya es posible contar con tablas de resultados configuradas para ser impresas vía RMarkdown, con un formato adecuado para su presentación en un reporte de resultados.

Tabla 9.8: Tabla 2. Estimación de intervalo de confianza para media

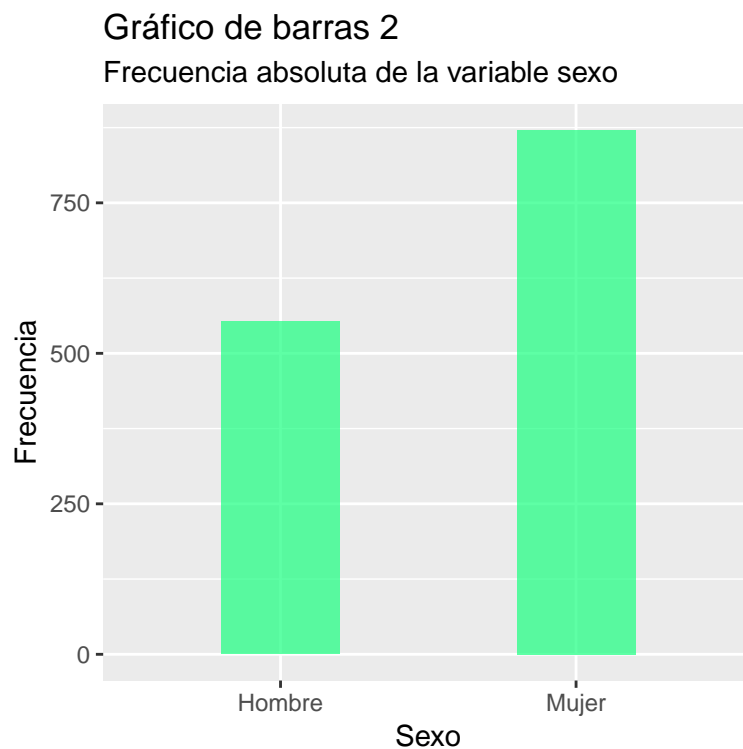
Media	Límite superior	Límite inferior	Nivel de confianza	Error estándar
7.31	7.2	7.42	0.05	0.06

### 9.5.4 Gráficos

La inserción de gráficos en un reporte a compilar vía RMarkdown es simple. Sólo se precisa incorporar alguna de las sintaxis para procesar gráficos ya construidas anteriormente en este material, en un trozo de código de R. En este caso se replica el gráfico construido en el capítulo 8. Vale la pena recordar que si se busca no compilar el trozo de código de R en el reporte, hay que agregar el argumento `echo = TRUE` a las especificaciones del trozo de código. Con los argumentos `fig.height`, `fig.width` y `fig.align` es posible configurar el ancho y alto máximos de la figura así como su alineación respecto a la página.

#### Ejercicio 9.14

```
library(ggplot2)
#Gráfico de barras 2: sexo en frecuencias absolutas
ggplot(CEP, aes(x = sexo_factor)) +
  geom_bar(width = 0.4, fill=rgb(0.1,1,0.5,0.7)) +
  scale_x_discrete("Sexo") +      # configuración eje X (etiqueta del eje)
  scale_y_continuous("Frecuencia") +
  labs(title = "Gráfico de barras 2",
        subtitle = "Frecuencia absoluta de la variable sexo")
```





## Capítulo 10

# Materiales de apoyo para el aprendizaje

### 10.1 Instalación de R y RStudio

- Página explicativa de la instalación de R en diferentes sistemas operativos Windows, Linux y MacOSX

### 10.2 Uso de RMarkdown

- Video (en inglés): cómo insertar bibliografía en RMarkdown, usando un archivo .bib existente:
- Página explicativa de diferentes herramientas de formato (metadatos) para RMarkdown, pensando especialmente en la construcción de reportes académicos.
- Enlaces de referencia para el uso de RMarkdown:
  - <https://www.r-bloggers.com/how-to-create-reports-with-r-markdown-in-rstudio/amp/>
  - [http://www.geo.uzh.ch/microsite/reproducible\\_research/post/rr-r-publication/](http://www.geo.uzh.ch/microsite/reproducible_research/post/rr-r-publication/)
  - [https://rmarkdown.rstudio.com/authoring\\_pandoc\\_markdown.html#images](https://rmarkdown.rstudio.com/authoring_pandoc_markdown.html#images)
- Plantillas de formato para pdfs contruidos desde RMarkdown

### 10.3 Uso de Zotero como gestor de referencias bibliográficas

- Página para descargar Zotero y Zotero connector
- Página para descargar diferentes formatos para referencias bibliográficas.

### 10.4 LaTeX y RMarkdown

- Explicación sobre programa TinyTex como plataforma LaTeX adecuada a RMarkdown



# Bibliografía



# Bibliografía

- Blalock, H. M. (1966). *Estadística social*. Fondo de Cultura Económica. Google-Books-ID: VD17NAAACAAJ.
- CEP (2017). Manual del Usuario Encuesta CEP N° 81. Estudio Nacional de Opinión Pública N° 51 – Tercera Serie, Septiembre-Octubre 2017.
- Comtois, D. (2018). Introduction to summarytools.
- Elousa, P. (2009). ¿EXISTE VIDA MÁS ALLÁ DEL SPSS? DESCUBRE R. *Revista Psicothema*, 21(4):652–655.
- Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications Ltd, London ; Thousand Oaks, Calif, edición: 1 edition.
- FSF (2019). ¿Qué es el software libre?
- Grolemund, G. (2014). Introduction to R Markdown.
- INE (2019a). Base de datos Encuesta Nacional de Empleo. Trimestre Enero-Febrero-Marzo de 2019. Archivo en formato CSV, Instituto Nacional de Estadísticas, Santiago de Chile.
- INE (2019b). Población total de 15 años y más por situación en la Fuerza de Trabajo, nivel nacional y regional, ambos sexos. Technical report, Instituto Nacional de Estadísticas, Santiago de Chile.
- Lumley, T. (2019). Survey analysis in R.
- Miller, S. V. (2018). A Pandoc Markdown Article Starter and Template. original-date: 2016-02-10T21:32:19Z.
- Navarro, J. (2014). LaTeX Fácil: Guía rápida de LaTeX.
- Ritchey, F. J. (2008). *Estadística para las ciencias sociales*. McGraw-Hill Interamericana de España S.L.
- van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition / Taylor & Francis Group*. Chapman and Hall/CRC, New York, second edition edition.
- Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'.
- Wilkinson, L., Wills, D., and Rope, D. (2005). *The Grammar of Graphics*. Springer, New York, edición: 2nd edition. 2005 edition.
- Workshop, R. R. (2016). Writing publications with R.