

INTRODUCCIÓN AL ANÁLISIS DE DATOS CON R Y R COMMANDER EN PSICOLOGÍA Y EDUCACIÓN

Jorge López Puga

2012

A Cristina, Álvaro, Julia y Ana María

Índice general

1. Instalación de R y R Commander	23
1.1. Instalación de R	23
1.2. Instalación de R Commander	28
2. Qué es R y para qué se utiliza	33
2.1. Sobre R	34
2.2. Preliminares sobre R	34
2.2.1. Introducción a los <i>objetos</i> de R	37
Vectores	37
Matrices	39
Factores	43
Listas	44
<i>Data frames</i> o bases de datos	45
Funciones	45
Modos y atributos de los objetos	46
2.2.2. Modelos estadísticos y gráficos	46
2.3. El paquete Rcmdr	47
2.3.1. El entorno gráfico de R Commander	47
2.3.2. Abrir archivos	50
2.3.3. Guardar archivos	55
3. Notas sobre la investigación en psicología y educación	57
3.1. Medición	58
3.2. Niveles o escalas de medida	61
3.2.1. Escalas nominales	62

ÍNDICE GENERAL

3.2.2. Escalas ordinales	63
3.2.3. Escalas de intervalo	64
3.2.4. Escalas de razón	65
3.2.5. Estadísticos admisibles en función del nivel de medida	66
3.3. Planificación y análisis estadístico	66
4. Estadísticos descriptivos	69
4.1. Estadísticos de tendencia central	70
4.2. Estadísticos de dispersión	73
4.3. Estadísticos de forma	77
5. Transformación de datos	79
5.1. Puntuaciones de escala	80
5.2. Recodificación de variables	82
5.3. Modificación del conjunto de datos	86
6. Inferencia sobre medias	89
6.1. El contraste de hipótesis	90
6.2. Contraste para una media	93
6.3. Contraste para dos medias	95
6.3.1. Medidas independientes	95
<i>t</i> de Student	95
Test de Wilcoxon	101
6.3.2. Medidas relacionadas	102
<i>t</i> de Student	102
Test de Wilcoxon	104
6.4. Contraste para más de dos medias	104
6.4.1. Análisis unifactorial de la varianza	104
6.4.2. Contraste de Kruskal-Wallis	108
7. Inferencias sobre proporciones	111
7.1. Inferencias sobre una variable	112
7.2. Inferencias sobre la relación entre dos variables	115
8. Correlación y regresión lineal	121
8.1. Correlación	123

ÍNDICE GENERAL

8.1.1. Coeficiente de Pearson	126
8.1.2. ρ de Spearman y τ de Kendall	129
ρ de Spearman	129
τ de Kendall	131
8.2. Introducción a la regresión lineal	132
9. Creación y manipulación de gráficas	137
9.1. Comandos de alto nivel	138
9.2. Comandos de bajo nivel	141
9.3. Personalización de parámetros gráficos	141
9.4. Facilidades que proporciona R Commander	142
Referencias	145

ÍNDICE GENERAL

Índice de figuras

1.1. Asistente de instalación de R	25
1.2. Licencia del programa R	25
1.3. Carpeta de destino para R	26
1.4. Componentes a instalar de R	26
1.5. Opciones de configuración de R	26
1.6. Accesos directos de R	27
1.7. Tareas adicionales	27
1.8. Progreso de instalación de R	27
1.9. Finalización de la instalación de R	28
1.10. R abierto	28
1.11. Selección de servidor	29
1.12. Selección de paquetes a instalar	30
1.13. Progreso de instalación de paquetes	30
1.14. Mensaje de paquetes faltantes	31
1.15. Modo de instalación de paquetes faltantes	31
1.16. Interface gráfica de R Commander	31
1.17. Selección de los plugins <i>HH</i> e <i>IP SUR</i>	32
1.18. Reinicio de R Commander	32
2.1. Estructura de una matriz tridimensional	41
2.2. Nuevo conjunto de datos en R Commander	50
2.3. Editor de datos en R Commander	51
2.4. Importar datos	53
2.5. Importar datos desde paquetes	54
4.1. Resúmenes numéricos	73

ÍNDICE DE FIGURAS

5.1. Calcular una nueva variable	81
5.2. Tipificar variables	84
5.3. Recodificar variables	86
5.4. Eliminar variables	87
6.1. Prueba t para una muestra	94
6.2. Creación de un histograma	96
6.3. Ejemplo de un histograma	97
6.4. Contraste para dos varianzas	99
6.5. Test t de medias independientes	100
6.6. Test t de medias dependientes	103
6.7. Anova unifactorial	106
6.8. Opciones del menú <i>Modelos</i>	109
7.1. Frecuencias y prueba χ^2 para una muestra	113
7.2. Frecuencias esperadas en la prueba χ^2 para una muestra	113
7.3. Prueba χ^2 para testar la independencia entre dos variables	117
7.4. Prueba χ^2 a partir de una tabla	118
8.1. Ejemplos de gráficos de dispersión	125
8.2. Matriz de correlaciones	127
8.3. Test de correlación	128
8.4. Regresión lineal	133
9.1. Menú del visor gráfico en R	138
9.2. Ejemplo de diagrama de caja en R	140

Prólogo

«*Hay tres clases de mentiras: las mentiras, las malditas mentiras y las estadísticas*»¹.

Esta cita, que se suele atribuir al escritor norteamericano Mark Twain —aunque realmente fue pronunciada por el político y escritor inglés Benjamin Disraeli— muestra, de alguna forma, el sentimiento popular que se tiene hacia la estadística.

Siempre que menciono a mis estudiantes esta frase suelo observar el esbozo de una leve sonrisa y un movimiento de asentimiento. Ciertamente, la ciencia estadística no suele tener demasiado «prestigio» —yo diría más bien credibilidad— entre la ciudadanía en general. Es fácil caer en la tentación de pensar que la estadística se utiliza para enmascarar la verdad, para manipular los resultados o, en el peor de los casos, que no sirve para nada ya que se suele errar habitualmente en sus predicciones —véase en las noches electorales—.

Sin embargo, la estadística, o mejor dicho, un buen uso de las herramientas que nos proporciona, es *fundamental* en nuestra Sociedad. Puede que alguno de los lectores se sorprenda al leer esta afirmación tan rotunda, pero le invito a que reflexione unos momentos y piense en cómo le afectaría personalmente que se hiciera un mal uso de la estadística.

¿No?, ¿piensa que a usted no? Le voy a ayudar un poco con algunos ejemplos que le pueden resultar cercanos.

Sin una validación estadística rigurosa, no tendríamos de medicamentos pues no seríamos capaces de evaluar su efectividad; una elaboración errónea del Índice de Precios al Consumo (IPC) o del Producto Interior Bruto (PIB) supondría un serio problema para la economía nacional y, en particular, de la suya;

¹«There are three kinds of lies: lies, dammed lies and statistics».

un censo mal elaborado trastocaría los planes de servicios que las administraciones han de prestar a la comunidad (limpieza, hospitales, colegios,...), y así un largo etcétera de situaciones en las que la estadística hace posible elaborar planes de previsión, proyecciones de futuro, etc.

Todo esto sin mencionar la enorme cantidad de datos con los que diariamente nos bombardean los medios de comunicación, y que hemos de ser capaces de interpretar correctamente para intentar no ser manipulados.

Así pues, considero de especial importancia inculcar una buena formación estadística en la población general y, en particular, en nuestros estudiantes universitarios, pues la mayoría de ellos, antes o después, se tendrán que enfrentar a situaciones en la que necesiten «extraer» la mayor y la mejor información posible de un conjunto de datos.

Por suerte —esa es mi opinión particular—, el acceso generalizado a los ordenadores ha posibilitado en acercamiento de la estadística a un público más amplio, pues ha permitido aplicar técnicas estadísticas complejas sin la necesidad de tener una base matemática muy potente.

No quiero que estas palabras últimas se malinterpreten. No disponer de una formación matemática de alto nivel no significa que «hacer estadística» consista en darle al botón del ordenador sin ton ni son. Suelo poner la siguiente comparación al respecto, se pueden hacer excelentes textos con un ordenador —véase este como ejemplo— sin necesidad saber como está montada la placa base o como funciona el procesador de su computadora. Son herramientas que se ponen a nuestra disposición y que hemos de utilizar correctamente.

Por lo tanto, la aplicación práctica de las técnicas estadísticas necesita de un conocimiento profundo de las posibilidades de las mismas. Si se quiere ir más allá, será necesario una formación más profunda, pero los paquetes estadísticos ponen a disposición del usuario la posibilidad extraer la información relevante —en el caso de que la hubiera— de los datos disponibles.

Este texto va en esa línea; con un lenguaje sencillo, adaptado a las personas a quienes va dedicado, se nos presentan diversas técnicas estadísticas muy útiles para abordar una gran cantidad de situaciones prácticas.

Además, hace que esto sea posible sin tener que gastar una gran cantidad de dinero, solamente la invertida en el ordenador que esté usted usando, pues el software aquí utilizado es del libre distribución —no me gusta la palabra gratis

ya que hay una gran cantidad de trabajo altruista invertido— surgido de un proyecto colaborativo en el que participa la comunidad estadística: el proyecto R (www.r-project.org).

Obviamente existe software comercial —muy popular— para la realización de estudios estadísticos, pero ninguno de estos programas tiene la flexibilidad y la potencia que tiene R. Ciertamente es que, en algunos casos, los resultados aparecen de una forma «menos vistosa» que en los programas comerciales, pero como contrapartida, R nos permite aplicar técnicas de última generación y, si tenemos la formación suficiente —realmente no hace falta ser un genio de la informática—, es posible implementar fácilmente nuestras propias metodologías.

Finalmente, quería dar las gracias a Jorge, autor de este texto, por invitarme a prologarlo. Es un orgullo para mí el poder hacerlo. Quisiera destacar el entusiasmo y pasión que Jorge le pone a las cosas que hace, y desde estas líneas le animo a seguir en esa línea. Nunca debemos perder la pasión por las cosas que hacemos, es el camino más corto a la felicidad.

Termino con otra cita, atribuída a un proverbio chino, «*conjeturar es barato; conjeturar erróneamente es caro*». Este libro le puede ayudar a no conjeturar erróneamente.

Fernando Reche Lorite

Profesor titular del área de Estadística e Investigación Operativa

Universidad de Almería, septiembre de 2012

Prefacio

Como señala De la Fuente (1998), podríamos destacar dos grandes problemas a los que nos enfrentamos cuando tratamos de gestionar asignaturas relacionadas con la estadística en las titulaciones de psicología y las asociadas a las ciencias de la educación. En primer lugar, los estudiantes no ven la utilidad práctica de los contenidos de éstas asignaturas. Ésto es, piensan que lo que ven en clase difícilmente tendrá aplicación en el contexto real de su desempeño profesional; que nunca van a tener que enfrentarse a una variable que se distribuya normalmente o que jamás tendrán que tomar decisiones que impliquen incertidumbre. Por otro lado, aunque los contenidos estadísticos forman parte del programa educativo en la enseñanza secundaria, también es cierto que muchos estudiantes tienen un bagaje sobre teoría estadística relativamente bajo.

Sin desligarme de lo anteriormente expuesto, me gustaría decir que el objetivo de este libro es tratar de solventar alguno de éstos problemas a los que se ha hecho referencia. Sin embargo, mis motivaciones para enrolarme en la ardua, que no desagradable, tarea de escribir este libro responden a otros factores. En primer lugar, desde un punto de vista pragmático, podría indicar que no existe ningún material a día de hoy (abril de 2012)², que yo sepa, destinado específicamente a alumnos de psicología y educación que presente el análisis de datos utilizando R o R Commander. Por consiguiente, creo que esta empresa que estoy empezando a construir está relativamente justificada. Más aún, si consideramos las interesantes ventajas (más abajo descritas) que supone el uso de estas herramientas



²Con posterioridad a haber escrito estas palabras (allá por el mes de julio de 2012 y durante mi asistencia al *V European Congress of Methodology*) supe de la existencia de un grupo de trabajo de la Universidad del País Vasco que está progresando en esta misma línea y que ha producido un libro (Elosua y Etxebarria, 2012) sobre esta temática altamente recomendable para usuarios del campo de las ciencias sociales.


estadísticas para los estudiantes.



En segundo lugar, me gustaría intentar sacarme una espinita que tengo clavada en relación a la elaboración de material didáctico para utilizar en mis clases. Aunque, como bien dice Andy Field (2009), la elaboración de libros que puedan usarse como material de apoyo en las clases no es una tarea que goce de gran prestigio y reconocimiento (los artículos publicados en revista indexadas en la *ISI Web of Knowledge* pesan más); creo que disponer de material específicamente desarrollado para entornos concretos de enseñanza-aprendizaje puede ser indudablemente útil desde un punto de vista didáctico. En mi caso concreto, había intentado previamente (sin ningún éxito) tratar de elaborar material didáctico para poder usarlo en mis clases de psicometría, asignatura que he estado impartiendo desde el año 2005, pero por diferentes circunstancias no he sido capaz de llevar a buen puerto esa encomiable tarea. Por tanto, este libro satisface mi deseo de escribir un libro que pueda ser útil a mis alumnos y alumnas.

Adicionalmente, he de decir que este libro está siendo concebido para ser un regalo con el que me gustaría obsequiar a mis actuales alumnos en la asignatura de psicometría. Y ésto es así porque me hubiese gustado elaborar material didáctico para mis alumnos durante el presente curso académico 2011/2012 en el idioma inglés (dado que la asignatura también está dentro del Plan de Fomento del Plurilingüismo) pero las circunstancias han frustrado mi intento, al menos, por el momento. Por ello, dados los problemas a que se enfrentan mis alumnos y alumnas cuando desarrollan trabajos prácticos en psicometría y que vengo observando en los últimos años, me agrada la idea de considerar que puedo compensarles con la producción de este manual.

En tercer lugar, creo que otro de los factores que ha desencadenado el inicio de éste trabajo podríamos encontrarlo en la satisfactoria experiencia que actualmente estoy disfrutando como tutor de alumnas ERASMUS en la asignatura «Tratamiento de Datos en Psicología». El caso es que ésta asignatura también se encuentra inscrita en el Plan de Fomento del Plurilingüismo que se está desarrollando en la Universidad de Almería y supuestamente debería de haberla impartido, en mayor medida, en el idioma inglés. Sin embargo, debido a circunstancias de diferente índole, me he visto obligado a impartir sólo una pequeña porción de la carga docente que oficial y originalmente tenía asignada. El caso es mucho más rico en matices dado que, a sabiendas de que la asignatura iba a impartirse (en


una gran proporción) en inglés, un grupo de alumnas provenientes de Polonia y Holanda se matricularon en la misma para satisfacer sus necesidades formativas. En esta situación, las alumnas y yo, acordamos celebrar una sesión semanal en la que pudiésemos ir trabajando con R ( de aquí en adelante) los contenidos que se estaban desarrollando en la asignatura los alumnos españoles. Así las cosas, cada vez me siento mejor usando  y más viable veo la idea de usar éste software como herramienta docente y analítica.





Por último, otro motivo por el que escribo éste libro es porque quiero homenajear a mi profesor, maestro y compañero Fernando Reche Lorite y, en parte, compensarle por no haberle elogiado más en los agradecimientos de mi tesis doctoral (López, 2009). Lo cierto es que gracias a Fernando supe de la existencia de  y aprendí a utilizarlo. También le debo el hecho de que haya organizado cursos de enseñanzas propias sobre L^AT_EX en los cuales he participado en dos ocasiones. No sólo por el hecho de haber tenido la oportunidad de adquirir ciertos conocimientos y/o competencias, sino porque mi forma de pensar y entender la informática cambió drásticamente desde que me empecé a familiarizar con estos entornos de trabajo.

Me gustaría, a continuación, destacar algunas de las ventajas que presentan  y R Commander ( de aquí en adelante) como herramientas aplicadas y aplicables al análisis de datos.

- En primer lugar está el tema de la *piratería*. Soy consciente de que vivimos en una cultura donde copiar ilegalmente música, películas y programas informáticos (entre otros) no está mal visto. Es más, está bien visto. Esto es, el que es capaz de *crackear* un programa informático para su beneficio es considerado como una persona exitosa, inteligente, como una especie de Robin Hood de las tecnologías de la información y la comunicación. Como señala Computer Music (1999), «algunos ven a las empresas de software como los malos de la película, que venden sus productos a precios abusivos mientras que los crackers (con su noble y desinteresado espíritu) nos ayudan a ganarles la partida» (p. 58). Y no sólo eso, sino que, pese a estar considerado como delito, la copia ilegal de material informático no genera el más mínimo remordimiento entre la comunidad universitaria (en ambos, profesorado y alumnado). Pues bien, yo soy crítico con esta situación.





Antes de nada, me gustaría aclarar que no me considero un moralista. Es







decir, yo mismo supongo que he copiado ilícitamente música, películas y programas informáticos (entre otras cosas) en el pasado y, aunque no digo que no vaya a volver a hacerlo en el futuro, también opino que «si hay una alternativa libre para ejecutar un proceso informático, ¿por qué cometer un delito copiando ilegalmente material protegido por la Ley de Propiedad Intelectual?» En éste contexto,  cobra protagonismo dado que al ser un programa libre podemos copiarlo, distribuirlo y/o modificarlo sin temor a incurrir en una falta legal.


- En segundo lugar, y no menos importante, habría que destacar, como señala Sáez (2010) y Elosua (2009), que  y  son *gratuitos*. Ésto no es, si se me permite la expresión, «moco de pavo». Con la situación económica que estamos atravesando creo que es de agradecer que se nos presenten alternativas que supongan el menor gasto económico posible. De esta manera, tanto la institución universitaria y el alumnado, así como cualquier usuario potencial, tendrían la posibilidad de ejecutar cálculos relativamente sofisticados que les supondría una pequeña inversión económica. De ésta manera, los alumnos podrían seguir ejecutando cálculos estadísticos al terminar sus estudios formales en la universidad sin piratear y sin gastar dinero en licencias de software. Por ejemplo, la Universidad de Almería tiene, en la actualidad, una licencia de servidor para que la comunidad universitaria pueda usar el programa SPSS³. Esto no está mal, excepto cuando el servidor de licencias falla por algún motivo. Otro problema aparece cuando tratamos de utilizar SPSS estando fuera de la universidad. Aunque se puede acceder al servicio estando conectado a Internet y disponiendo de una conexión VPN, el problema aparece cuando alguien (como es mi caso) no dispone de conexión a la red en su hogar.
- Otra ventaja que se deriva del uso de éstos programas informáticos, a mi modo de ver, está referida al aprendizaje que se desprende de su utilización. Esto es, aprender a usar  y  favorece que *se aprenda estadística*, entre otras cosas. Dado que en la mayoría de las situaciones el usuario tiene casi el control total sobre lo que está haciendo, ésto favorece que las personas que usan estos sistemas adquieran un conocimiento más profundo de las técnicas


³*Statistical Package for Social Sciences*




y métodos estadísticos que subyacen a las funciones implementadas en su código informático.


- En cuarto lugar, como señalan Elosua y Etxeberria (2012),  proporciona un considerable abanico de procedimientos y rutinas estadísticas que aún no están disponibles en los paquetes estadísticos comerciales.
- Para terminar, y haciendo gala del pretendido utopismo que me caracteriza y que raya en lo enfermizo, creo que  ha de ser considerado como, al igual que podríamos considerar al Dr. Valentino Rossi () *patrimonio de la humanidad*. Hasta no hace mucho, y dada la predilección que sentía mayormente por la edición de gráficos, pensaba que este calificativo había que dárselo a SPSS pero mis recientes experiencias con el programa me han hecho cambiar de opinión. En cierto modo, por todo lo anteriormente expuesto creo que  debería seguir siendo libremente accesible a todo el mundo. Deberíamos preservarlo, potenciarlo y mejorarlo. Más aún cuando algunos investigadores como Gred Guigerenzer consideran que el razonamiento probabilístico-estadístico debería trabajarse a edades tempranas en el colegio y que podría considerarse como una clave de éxito adaptativo en la sociedad contemporánea (Bond, 2009).

Dado que, en el monte «todo lo que reluce no es orégano», también me gustaría destacar algunas de las desventajas que presenta el uso de  y de . Siguiendo a Sáez (2010) podríamos destacar tres desventajas que se presentan al usar estos sistemas. En primer lugar, no tenemos un entorno tan *amigable* para ejecutar los cálculos como el que proporcionan otros paquetes estadísticos (por ejemplo, SPSS, SAS, Statgrapichs, etc.). Más bien, tenemos que escribir líneas de comandos, aunque  surgió para hacer más fluida la interacción con  al presentar un entorno gráfico típico de los programas al uso. Por otro lado, los resultados de los análisis no son tan fácilmente exportables a editores de texto (en muchas ocasiones no sólo consiste en *copiar-y-pegar*). Y, por último, también habría que destacar que algunas veces  se cierra sin motivo aparente, cosa que no suele pasar si utilizamos la consola de  directamente.




En mi opinión, para el usuario novel,  tiene un inconveniente importante (si es que se puede ver así, porque ésto también se puede ver como una ventaja): hay que indagar «mucho» para hacer ciertas cosas aparentemente sencillas. No

obstante, éste inconveniente se ve superado (o se superará con la práctica) por el gran control que se tienen sobre los gráficos y análisis que se ejecutan con .

Este manuscrito no es un texto completo ni exhaustivo⁴. Más bien, se puede considerar como una introducción y/o una guía para introducirse en el análisis de datos en los campos de la psicología y la educación. También puede considerarse, en algunos de sus pasajes, como una introducción o invitación a realizar cálculos y análisis más complejos utilizando  y . Voy a intentar desarrollar el contenido de este libro tratando de adaptarlo a la mayor parte del público y, por ello, trataré de explicar cada paso y análisis desde sus detalles más básicos o elementales. No obstante, también es cierto que en algunos pasajes presentaré la información de manera pseudo-telegráfica para favorecer que el usuario juegue un papel activo que le permita aprender de manera más profunda. Pero aunque voy a intentar desarrollar el contenido para todos los públicos sería conveniente enfrentarse al manual con nociones básicas de matemáticas y algo de estadística. Por tanto, este libro no pretende ser una receta que guíe a los usuarios de  por el sendero del análisis de datos. No es tampoco un libro de diseños de investigación, aunque en algunos casos se harán comentarios sobre los diseños que subyacen a tipos particulares de análisis de datos. Para las personas interesadas en aprender más sobre los diseños de investigación, recomiendo acceder al libro de León y Montero (2003)⁵.

Me gustaría destacar que, aunque no se van a tratar en este manual, existen paquetes de  específicos diseñados para ejecutar tareas y análisis estadísticos típicamente asociados a la psicología y al campo de trabajo de la educación. Por ejemplo, en el contexto de la medición psicológica se han desarrollado paquetes que permiten estimar diferentes modelos de medida basados en la Teoría de Respuesta al Ítem o trabajar con aspectos clave de la Teoría Clásica de Tests (de Leeuw y Mair, 2007; Mair y Hatzinger, 2007). Recomiendo al lector interesado que profundice en estos paquetes dado que le proporcionarán ideas y alternativas interesantes en sus proyectos de investigación.



A lo largo del libro el código fuente necesario para generar un gráfico o un

⁴Para un libro más completo y exhaustivo, aunque enfocado casi totalmente desde el punto de vista de la interfaz gráfica de , recomiendo trabajar con el manual de Elosua y Etxeberria (2012). También recomiendo el libro de Arriaza et al. (2008) para cubrir un curso introductorio de estadística con  y .

⁵Si te fijas en ésta referencia se ve claramente la poca creatividad que emana de mis castigadas neuronas cuando decidí dar un título a éste libro.



análisis aparecerá numerado y recuadrado de este modo:

```
1 Esto es un ejemplo #Esto es un comentario
2 de código fuente #Esto es otro comentario
```

así, el usuario podrá copiar-y-pegar y reciclar el código para aprender y para satisfacer sus necesidades. Para ir entrando en materia, me gustaría señalar que, como aparece en el recuadro precedente, todo lo que aparezca precedido de un símbolo de almohadilla (#) no será ejecutado por  o . Cuando anteponeamos el símbolo de almohadilla a una sección de código de programación informática decimos que *estamos comentando*, y esta porción de código es un *comentario* que se puede utilizar para aclarar aspectos funcionales del comando en particular que le precede. Dependiendo del lenguaje informático que estemos utilizando el símbolo que indica lo que es un comentario cambia. Así, por ejemplo, en la sintaxis de SPSS el símbolo del comentario es el * y en Visual Basic es el '. Por su parte, cuando te presente salidas de resultados verás cuadros como este:

```
1 Esto es un ejemplo
2 de salida del programa
```

Me gustaría resaltar, para terminar, que voy a intentar desarrollar todo este libro utilizando software libre (que no gratuito) y que trataré de depositarlo en el Repositorio de la Universidad de Almería bajo una licencia *Creative Commons* para que sea accesible a todo el mundo de manera libre (que no gratuita). En primer lugar, este libro está siendo compilado con \LaTeX y editado con TeXnicCenter (<http://www.texniccenter.org>). Las imágenes se generarán, a partir de capturas de pantalla, con el programa Gimp en su versión 2.6 (www.gimp.com).

Bueno, creo que eso es todo. Espero que disfrutes y que aprendas siguiendo este libro y utilizando  y .



Jorge López Puga

UNIVERSIDAD DE ALMERÍA

jpuga@ual.es

<http://www.ual.es/personal/jpuga>

Agradecimientos

En primer lugar, quiero agradecer a mi esposa Ana María que haya dedicado parte de su tiempo a revisar y corregir los numerosos errores que contenía éste manuscrito en versiones anteriores. Sin su ayuda el texto que aquí se presenta hubiese sido de menor calidad.

Por otro lado, le doy las gracias a mi hermano Víctor por haber dibujado la ilustración de la portada. Lo cierto es que está hecho todo un artista y a él le debo otra imagen que fue publicada en el artículo del *Boletín Matemático de la UAL* que redacté recientemente (López, 2012). La idea, en este caso, era representar cierta relación entre naturaleza y tecnología. En el caso concreto de la materia de éste libro, se trataba de representar cómo el desarrollo tecnológico, en concreto en el campo de la computación estadística, puede servirnos para conocer mejor la naturaleza y el universo.



Me gustaría agradecer al Profesor Fernando Reche Lorite su predisposición para escribir el Prólogo de este libro. Para mí es, sin duda, un gran honor y una gran satisfacción tanto por la forma como por el fondo.

Tengo también que dar las gracias a José Berenguel Sánchez de la Editorial de la Universidad de Almería por haberme ofrecido la posibilidad de publicar, aunque el libro estará a libre disposición en el repositorio de la Universidad de Almería, este manuscrito bajo el sello de esta editorial.


Por último, agradezco también a todos mis alumnos y alumnas el ánimo y la energía que me transmiten para que haga este tipo de cosas. Supongo que, en cierto modo, esto es para y por ell@s.

1







Instalación de R y R Commander

En primer lugar, he pensado que podría ser de utilidad dedicar algunas líneas a describir brevemente el proceso de instalación de R y R Commander. Algunas personas me criticarían diciendo que estoy haciendo un «guía-burros» pero lo cierto es que éstas instrucciones podrían servir para no desalentar a algunos potenciales usuarios a que usen  o . Por ello, describiré rápidamente cómo se instalan estos programas en nuestro equipo. Lamento decir que, afortunada o desafortunadamente, únicamente soy usuario de Microsoft Windows y, por tanto, voy a explicar el proceso a seguir para este sistema operativo. No obstante, supongo que los usuarios de otros sistemas operativos (MacOS X o Linux, por ejemplo) no encontrarán muchos problemas para seguir estas breves instrucciones en sus máquinas.

1.1. Instalación de R

Lo primero que tenemos que hacer es instalar . Para ello, sugiero seguir las siguientes instrucciones:

Capítulo 1 - Instalación de R y R Commander

1. Ir a la página oficial del proyecto  titulada *The R Project for Statistical Computing* y que se encuentra en la siguiente dirección de Internet: <http://www.r-project.org>.
2. Acceder al enlace que aparece a la izquierda de la página web llamado CRAN (*Comprehensive R Archive Network*).
3. Seleccionar el servidor más cercano a nuestra localización geográfica. Por ejemplo, si tratamos de acceder al recurso desde la Universidad de Almería, tendríamos que acceder al servidor de la Red de Investigación Nacional Española que se encuentra situado en Madrid cuya dirección web es <http://cran.es.r-project.org>.
4. Descargar e instalar . Dependiendo de nuestro sistema operativo tendremos que acceder a una de las opciones que se nos plantean. Dado que, como he indicado anteriormente, vamos a trabajar con Windows seleccionamos la opción correspondiente a (*Download R for Windows*).
5. Seleccionar el sub-directorio «base». Al hacer ésto estamos eligiendo descargar los paquetes y algoritmos básicos necesarios para que funcione . Creo que es conveniente señalar en este punto que  es un programa que funciona con base en lo que denominamos **paquetes**. Los paquetes son una especie de mini-programas informáticos que han sido desarrollados para llevar a cabo tareas concretas o específicas. El paquete **base** es el programa que contiene la información básica (los paquetes básicos) para que  funcione de manera genérica. Por poner otro ejemplo, éste libro está dedicado en su mayor parte a un paquete específico diseñado para  llamado **R_{gui}** que proporciona un entorno gráfico confortable para las personas que no estamos acostumbradas o habituadas a trabajar con código de programación.
6. Descargar el programa. Dependiendo de cuando descarguemos el programa accederemos a una versión más actualizada del mismo. Hoy, a día 28 de abril de 2012, la versión actualizada es la 2.15.0.
7. Una vez que tenemos el archivo con extensión «.exe» descargado en nuestro equipo tenemos que hacer doble clic sobre él.

1.1 - Instalación de R


- Tras aceptar, dependiendo de nuestro sistema operativo, las preguntas relativas a los controles de seguridad podremos elegir el idioma en el cual queremos instalar el programa.
- La ventana que aparecerá tras elegir el lenguaje en el que queremos instalar se parecerá a la que aparece en la Figura 1.1. Como verás, ésta ventana nos informa de que vamos a iniciar la instalación de una versión concreta del programa informático .



Figura 1.1: Asistente de instalación de R.

- Aceptación de la licencia del programa informático. Al pulsar en el botón «Siguiente» en el paso anterior, aparecerá una ventana similar a la que se muestra en la Figura 1.2. Como verás, consiste en un tipo de licencia de software tipo GNU (acrónimo que proviene de la expresión inglesa *GNU is Not Unix*) que caracteriza a los programas que se denominan de uso libre. Si tienes tiempo e interés te recomiendo que le eches un vistazo.

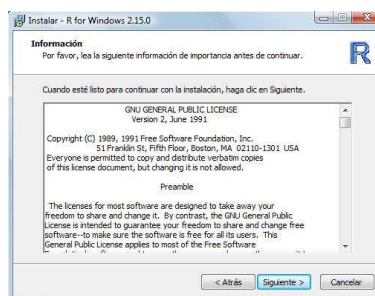



Figura 1.2: Licencia del programa R.

- Tras haber pulsado en el botón «Siguiente» aparecerá el típico cuadro de diálogo que solicita un destino de instalación para el programa (Figura 1.3). Indica dónde quieres instalar  y continúa con el proceso.

Capítulo 1 - Instalación de R y R Commander

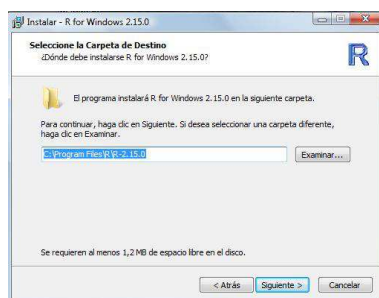


Figura 1.3: Carpeta de destino para R.

12. Selección de componentes. A continuación (Figura 1.4), y en función del tipo de ordenador que estemos utilizando, seleccionamos los componentes que queremos instalar.

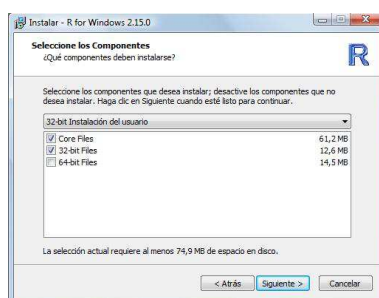


Figura 1.4: Componentes a instalar de R.


13. En la ventana que aparece seguidamente (Figura 1.5), podríamos elegir el modo de presentación de , esto es, si queremos disponer de una interface de una única ventana (SDI) o de ventanas separadas (MDI).



Figura 1.5: Opciones de configuración de R.

14. La siguiente ventana, como se puede apreciar en la Figura 1.6, simplemente sirve para indicar el lugar donde se crearán los acceso directos al programa.

1.1 - Instalación de R

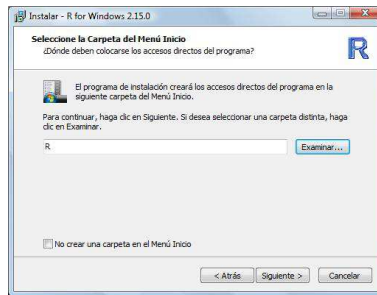


Figura 1.6: Accesos directos de R.

15. En la Figura 1.7 podemos seleccionar tareas adicionales que queremos que se lleven a cabo durante el proceso de instalación.

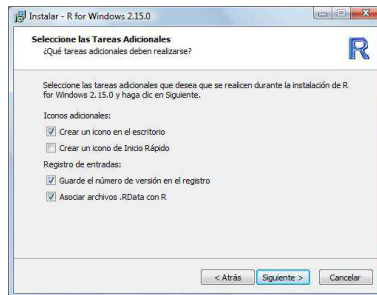


Figura 1.7: Tareas adicionales.

16. Tras pulsar el botón «Siguiente» en el paso anterior, el proceso de instalación comenzará y aparecerá una ventana similar a la que aparece en la Figura 1.8 donde se indica el progreso de instalación del programa.

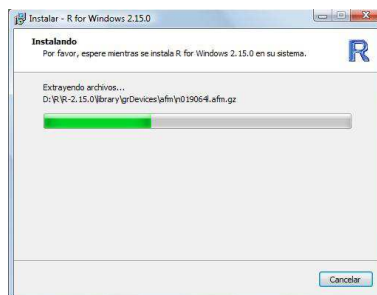


Figura 1.8: Progreso de instalación de R.



17. En el último paso que precede a la finalización de la instalación (Figura 1.9), sólo hay que pulsar en el botón «Finalizar».



Capítulo 1 - Instalación de R y R Commander



Figura 1.9: Finalización de la instalación de R.

1.2. Instalación de R Commander

Una vez que hemos instalado  correctamente, y tras haber seguido éstas instrucciones tan tediosas y exageradamente detalladas, tenemos que instalar el paquete sobre el que se basa la mayor parte de este manuscrito: . Para ello, recomiendo seguir los siguientes pasos.

1. En primer lugar, tenemos que abrir . Para ello, hacemos doble clic sobre el icono de  que se habrá creado en nuestro escritorio o sobre la pestaña correspondiente que nos aparecerá en la sección de *Programas* de nuestra barra de *Inicio*. El aspecto del programa, una vez abierto, será similar a lo que aparece en la Figura 1.10.

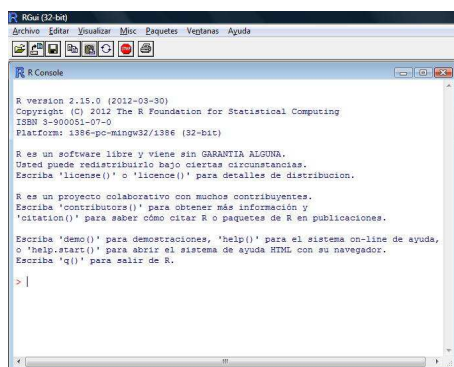



Figura 1.10: R abierto.

2. En el menú «Paquetes» seleccionamos la opción «Instalar paquete(s)...» para instalar los contenidos directamente desde Internet. Es decir, lo que vamos a hacer es decirle a nuestro ordenador que acceda a un servidor de descargas, que se descargue ciertos paquetes que se añadirán a  y que

1.2 - Instalación de R Commander

los instale en nuestro disco duro. Podríamos querer instalar los contenidos correspondientes desde archivos locales comprimidos en formato WinZip, para lo cual tendríamos que seleccionar la opción «Instalar paquete(s) a partir de archivos zip locales...» y seleccionar la ubicación y los archivos que serían objeto de instalación.


- Tras la acción anterior nos aparecerá una ventana similar a la que aparece en la Figura 1.11 donde se nos insta a seleccionar el servidor más cercano para ejecutar la descarga de los contenidos a instalar. Esto es algo parecido a lo que hacíamos cuando descargábamos .



Figura 1.11: Selección de servidor.

- Seguidamente, tendremos que seleccionar los paquetes que queremos descargar e instalar de una lista que se nos presentará en una ventana similar a la que aparece en la Figura 1.12. En éste punto, es crucial seleccionar el paquete `Rcmdr` que dará pie a la instalación de `Rcmdr`. Adicionalmente, recomiendo seleccionar el resto de paquetes que comienzan por esa misma secuencia de caracteres (`Rcmdr`) seguidos por la expresión `Plugin`. (Figura 1.12). Haciendo ésto estamos pidiendo que se descarguen y se instalen cier-

Capítulo 1 - Instalación de R y R Commander

tas aplicaciones que se acoplan a **Rcmdr** y que son de mucha utilidad en ciertas situaciones.

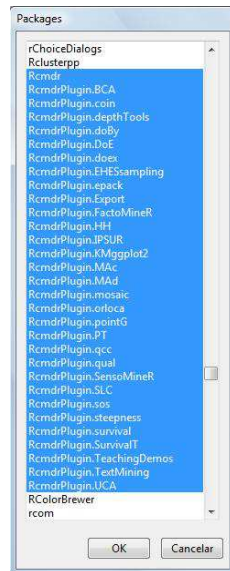


Figura 1.12: Selección de paquetes a instalar.

- Al pulsar en el botón «OK» en el paso anterior se comenzarán a instalar todos los paquetes seleccionados uno por uno desde Internet mostrando una barra de progreso para cada uno de ellos como se muestra en la Figura 1.13.

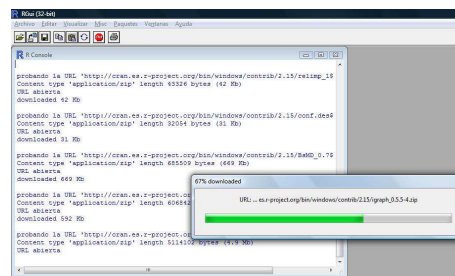


Figura 1.13: Progreso de instalación de paquetes.

- Aunque **Rcmdr** y todas sus aplicaciones asociadas han sido instaladas correctamente, todavía quedan por instalar algunos paquetes necesarios que hagan que **Rcmdr** funcione correctamente. Para culminar la instalación tenemos que ejecutar **Rcmdr** por primera vez e instalar ciertos componentes faltantes. Para ello, abrimos **Rcmdr**, escribimos lo siguiente en la consola de comandos y presionamos la tecla «Enter»:

1.2 - Instalación de R Commander

```
1 library(Rcmdr) #Éste comando ordena que se abra R Commander
```

7. A continuación, nos aparecerá un mensaje (Figura 1.14) advirtiéndonos de que faltan paquetes por instalar para que **Rcmdr** funcione correctamente. Hacemos clic en «Sí» para instalar los paquetes faltantes.

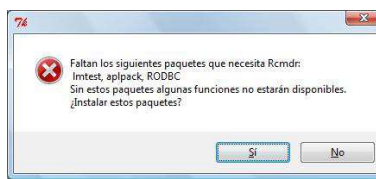


Figura 1.14: Mensaje de paquetes faltantes.

8. El asistente (Figura 1.15) nos preguntará si queremos instalar los paquetes faltantes desde el CRAN o si los queremos instalar desde un directorio local. Para instalarlos desde el CRAN pulsamos en «OK».



Figura 1.15: Modo de instalación de paquetes faltantes.

9. Cuando se instalen los paquetes faltantes se abrirá la interface gráfica de **Rcmdr** que tendría un aspecto similar al que aparece en la Figura 1.16.

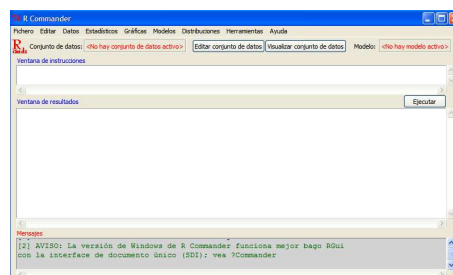


Figura 1.16: Interface gráfica de R Commander.

Capítulo 1 - Instalación de R y R Commander

- Una vez abierto **Rcmdr** recomiendo activar dos *Plugins* muy interesantes cada vez que lo utilizemos. Para ello accedemos al menú «Herramientas» y seleccionamos la opción «Cargar plugin(s) de Rcmdr». Marcamos los plugins *HH* e *IPSUR* y presionamos «Aceptar» (Figura 1.17).

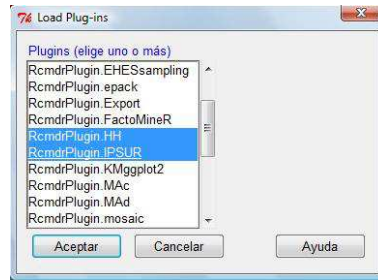


Figura 1.17: Selección de los plugins *HH* e *IPSUR*.

- Nos aparecerá un mensaje para reiniciar **Rcmdr** con el objetivo de que los Plugins seleccionados previamente estén disponibles. Hacemos clic en el botón «Sí» para que se reinicie **Rcmdr**.



Figura 1.18: Reinicio de R Commander.

2

Qué es R y para qué se utiliza

En este capítulo me gustaría introducir de manera general, no exhaustivamente, algunos de los rasgos que considero más destacables o relevantes de \mathbb{R} y de \mathbb{R}_{mir} . El lector ávido de conocimiento «útil», en cierto modo un tipo de lector pragmático, en el contexto del análisis de datos aplicado podría saltarse parte de, o todo, éste capítulo. El motivo es que una porción considerable de éste capítulo está destinada a describir algunos rasgos genéricos de \mathbb{R} y podría no verse, en principio, como algo útil. No obstante, he de decir que, desde mi punto de vista, el tiempo dedicado a la lectura de este capítulo no será mal invertido. Aunque se presentará información genérica sobre el funcionamiento de \mathbb{R} (especialmente en relación a tipos de objetos, funciones y manipulación de datos), ésta puede ser de utilidad cuando tratemos de personalizar análisis o de ejecutar repetidamente procedimientos similares. He de destacar que la primera parte de este capítulo se basa en el documento publicado por Venables, Smith, y the R Development Core Team (2011), por lo que se recomienda al lector interesado en profundizar en éste tema que acceda a esta referencia. Otra referencia que puede ser de extrema utilidad en este momento es el libro de Elosua (2011), donde se presenta una

introducción amigable al entorno de trabajo que proporciona \mathbb{R} . Adicionalmente, se introducirán algunas notas de considerable importancia en relación al uso de \mathbb{R} que se tratarán con posterioridad pero que pueden hacer más comprensible y productivo el seguimiento de éste manual.

2.1. Sobre R

\mathbb{R} es un entorno de trabajo basado en los entornos de programación S y S-PLUS desarrollados a principios de los años noventa del pasado siglo por Bill Venables y David M. Smith cuando se encontraban trabajando en la *University of Adelaide*¹. Desde entonces, \mathbb{R} se ha desarrollado muy rápidamente y ha acumulado, como se señalaba anteriormente, una gran cantidad de *paquetes* que ejecutan análisis estadísticos muy específicos (R Development Core Team, 2011).

Como señalan Venables et al. (2011), « \mathbb{R} es un entorno integrado de facilidades informáticas para la manipulación de datos, el cálculo y la generación de gráficos» (p. 2). La idea de considerar a \mathbb{R} como un «entorno» es conceptual y epistemológicamente interesante. Esto es, más que un colección de herramientas de análisis relativamente inflexibles, \mathbb{R} pretende convertirse en un sistema internamente coherente que se caracterizaría por un desarrollo basado en la contribución relativamente altruista de la comunidad científica.

Aunque este manual versa sobre análisis estadísticos relativamente sencillos, con \mathbb{R} se pueden, entre otras cosas, crear rutinas de análisis personalizadas, crear gráficos relativamente vistosos o trabajar con números complejos.

2.2. Preliminares sobre R

El lenguaje en que se basa \mathbb{R} es sensible a mayúsculas y minúsculas. Esto es, a es algo diferente a A. Por ello, cuando definimos objetos o variables debemos tener cuidado y recordar este hecho. Los símbolos permitidos para nombrar variables u objetos incluyen, normalmente, a todos los caracteres alfanuméricos². Incluso en algunos idiomas las tildes están permitidas. Adicionalmente, se pueden usar el punto «.» y el guión bajo «_». No obstante, hay que tener en cuenta que cuando

¹Universidad pública australiana fundada en 1874 y afincada en el sur del país.

²Esto es, de la A a la Z, de la a a la z y del 0 al 9.

comencemos a nombrar a un objeto empezando por punto el siguiente carácter no puede ser un número.

Los comandos más elementales que se pueden manejar son las *expresiones* o *asignaciones*³. Cuando se introduce una expresión en `R` y se presiona la tecla «Enter» de nuestro teclado, `R` la evalúa, la imprime y su valor se desvanece. Sin embargo, cuando tecleamos una asignación su contenido es evaluado, su valor es transferido (en caso de no haber errores en la sintaxis) y no se imprime en la consola. Por ejemplo, abre la consola de `R` y teclea lo siguiente:

```
1 3*8
```

Al presionar la tecla «Enter», `R` imprimirá el resultado de la operación 3×8 en color azul. Esto es un ejemplo de expresión que ha sido evaluada, impresa y volatilizada. Con «evaluada» me refiero a que se ha chequeado su validez desde el punto de vista del lenguaje que subyace a `R`, y con respecto a lo de la impresión me estoy refiriendo a que el resultado ha sido impreso en el monitor de nuestro ordenador. Si escribes lo siguiente en tu consola de comandos:

```
1 6_2
```

obtendrás un mensaje como este:

```
1 Error: inesperado entrada en "6_"
```

Lo que pasa es que al evaluar la expresión, `R` se ha cerciorado de que no es válida. Esto es, contiene una secuencia de caracteres que no puede interpretar. En concreto, no entiende que la expresión corresponda a algún tipo de cálculo numérico, que sea la aplicación de una función o que responda a algún tipo de operación permitida.


Por otro lado, la expresión 3×8 se ha volatilizado, lo que quiere decir que no se ha almacenado en ningún sitio⁴. Para que una expresión quede almacenada en

³Estas ideas ya han sido presentadas informalmente con anterioridad pero voy a tratar de retomarlas de una manera más formal aquí.

⁴Bueno, esto no es del todo cierto ya que sí que está almacenada en la memoria del ordenador. Por ejemplo, si pulsas en la tecla «▲» del cursor de tu teclado un par de veces, volverás a la expresión citada. Lo que pasa es que está almacenada de tal manera que no puede invocarse o utilizarse para que `R` trabaje con ella manipulándola.

Capítulo 2 - Qué es R y para qué se utiliza


la memoria interna del ordenador hay que hacer una asignación a un objeto.

Los elementos que se pueden crear y manipular con  se conocen como *objetos*. Estos objetos pueden ser variables, matrices de números, cadenas de caracteres, funciones u otras estructuras creadas a partir de éstos elementos individuales. Por ejemplo, el siguiente código sirve para crear un objeto llamado x que es un vector numérico que contiene los números del 1 al 10:


```
1 x <- 1:10
```

Al ejecutar este comando aparentemente no pasa nada. Sin embargo, lo que ha pasado es que la asignación `1:10` ha sido evaluada y almacenada en la memoria de nuestra computadora. Como te habrás dado cuenta, el par de caracteres «<-» han servido para asignar a x el conjunto de números que van del 1 a 10^5 . Si ahora escribimos x en nuestra consola de comandos y presionamos la tecla «Enter» nos aparecerá el contenido del objeto x ⁶:

```
1 [1] 1 2 3 4 5 6 7 8 9 10
```

La función `objects()` sirve para que  nos informe de los objetos que tenemos disponibles para operar sobre ellos. Al conjunto de los objetos almacenados en la memoria del ordenador en un momento dado se les denomina conjuntamente como *espacio de trabajo* o *workspace*. En nuestro caso, si tecleamos la función y la ejecutamos, la ventana de comandos ofrecerá el siguiente resultado:

```
1 [1] "x"
```

Todos los objetos creados en una sesión de trabajo pueden ser guardados en un archivo para que se puedan recuperar en una sesión posterior. Si así lo deseas, podrás guardar estos objetos en un archivo con extensión «.RData» en el *directorio de trabajo*. El directorio de trabajo, es una carpeta de tu ordenador donde  irá almacenando, por defecto, los archivos derivados de tus manipulaciones. Para saber cual es el directorio de trabajo donde serán guardados los archivos

⁵En algunos contextos el símbolo «=» es equivalente a «<-». Además, una asignación también se puede ejecutar en el otro sentido, esto es, con los símbolos «>».

⁶También se puede ejecutar la función `print()` para obtener el mismo resultado.

correspondientes puedes utilizar la función `getwd()`. Si la escribes y la ejecutas obtendrás algo parecido a esto:

```
1 [1] "D:/datos/Mis Documentos"
```

Para cambiar el directorio de trabajo puedes utilizar la función `setwd()`. Por ejemplo, si yo quisiera guardar los ficheros *producto* de una sesión en una carpeta llamada `Libro_R` que se encuentra en la carpeta `Mis Documentos`, tendría que ejecutar la siguiente sintaxis:

```
1 setwd("D:/datos/Mis Documentos/Libro_R")
```

Todos los comandos que han sido utilizados en una sesión también son susceptibles de ser guardados en un archivo llamado «.Rhistory». De esta manera puedes recuperar el trabajo de un día previo cargando esta secuencia histórica de comandos.

2.2.1. Introducción a los *objetos* de R

Esta sub-sección está dedicada a proporcionar una introducción muy somera de las *estructuras de datos* sobre las que opera \mathbb{R} y que se denominan genéricamente *objetos*. Por ello, se recomienda encarecidamente al lector interesado a que profundice en los conceptos e ideas que se exponen brevemente en lo que sigue dado que le ayudarán a optimizar su conocimiento del entorno de trabajo de \mathbb{R} .

Vectores

Los vectores son la estructura de datos básica y más simple con la que podemos operar en \mathbb{R} . Un vector es, en su definición más general, «un conjunto de números ordenados» (Venables et al., 2011, p. 7). Por ejemplo, consideremos que el vector x representa las estaturas, en centímetros, de seis jóvenes que forman un equipo de voleibol y que son: 174, 182, 181, 179, 188 y 185. Si queremos incorporar este vector en \mathbb{R} tendremos que utilizar la función `c()` y, como se ha comentado anteriormente, los símbolos de asignación «`<-`» o «`=`» del siguiente modo:

```
1 x <- c(174, 182, 181, 179, 188, 185)
```

Capítulo 2 - Qué es R y para qué se utiliza

Lo que hemos hecho ha sido *asignar* al objeto x un conjunto de valores. Esto también puede hacerse utilizando la función `assign()` del siguiente modo:

```
1 assign("x", c(174, 182, 181, 179, 188, 185))
```

Como se puede apreciar la diferencia entre ambas formas de asignación, entre otras, radica en que el primer método es un atajo frente a la segunda asignación sintáctica.

Como se ha dejado entrever más arriba el símbolo «:» se puede utilizar para generar secuencias ordenadas de números entre dos valores dados. Por ejemplo, el comando

```
1 y <- -5:5
```

generaría un vector de longitud 11 que contendría los números enteros comprendidos entre -5 y 5. Existe una función llamada `seq()` que permite generar vectores consistentes en secuencias numéricas limitadas por dos números y cuyos elementos son equidistantes unos de otros. La expresión más sencilla de la función funciona de igual modo a como funcionan los dos puntos para generar secuencias de números. Así, la expresión `seq(-5:5)` es equivalente a la previamente expuesta. No obstante, se puede añadir un argumento llamado `by` en la función que especifique la diferencia entre cada valor consecutivo del vector. Por defecto este parámetro está ajustado a 1 y por ello la diferencia entre cada valor consecutivo del vector es de una unidad. Ésto es, sería como una especie de frecuencia de muestreo que por defecto está ajustada a uno. ¿Qué pasaría si manipulásemos ese parámetro y lo hiciésemos valer 0,95⁷? Para hacerlo, tendremos que teclear en la ventana de comandos la siguiente expresión:

```
1 w <- seq(-5,5, by=0.95)
```

⁷Creo que éste es el primer lugar donde utilizo números decimales en este texto y he de advertir que, dado que estamos utilizando un software informático no desarrollado originariamente en español, tendremos que utilizar el punto como delimitador decimal «.». Sin embargo, trataré de ser lo más formal y escrupuloso posible a este respecto cuando utilice números decimales en el cuerpo del texto siguiendo, en la medida de lo posible, las directrices del idioma castellano o español.

cuyo resultado será:

```
1 -5.00 -4.05 -3.10 -2.15 -1.20 -0.25 0.70 1.65 2.60 3.55 4.50
```

Además de vectores numéricos, \mathbb{R} también puede manipular y trabajar con **vectores lógicos**. Un vector lógico contiene elementos que pueden ser verdaderos (TRUE), falsos (FALSE) o casos perdidos (NA⁸). Los vectores lógicos pueden usarse algunas veces para ejecutar operaciones aritméticas en cuyo caso el valor TRUE es truncado a 1 y el valor FALSE a 0.

Matrices

Si consideramos que los vectores son estructuras de datos unidimensionales (ya que solo tienen la dimensión de longitud) las matrices son generalizaciones multidimensionales de los vectores. Esto es, una matriz de dos dimensiones sería una especie de tabla consistente en vectores columna y vectores fila mientras que una matriz tridimensional consistiría en una especie de cubo que contendría vectores columnas, vectores fila y vectores que se proyectarían en la tercera dimensión (de profundidad, por ejemplo).

Veamos cómo funcionan las dimensiones de una matriz utilizando un vector de treinta elementos. En primer lugar, tendremos que crear un vector, llamémosle v , que contenga los números naturales comprendidos entre el 1 y el 30. Luego, generaremos una matriz bidimensional con 10 filas y 3 columnas utilizando la función `dim()`. Ésta sería la sintaxis:

```
1 v <- 1:30
2 dim(v) <- c(10,3)
```

Nuestra matriz consistiría en una tabla de diez filas y tres columnas (los números de fila y columna aparecen impresos entre corchetes) con los números de 1 al 30:

```
1      [,1] [,2] [,3]
2 [1,]    1  11  21
3 [2,]    2  12  22
```


⁸*Not Available* o No Disponible. Existe otro tipo de valor o caso perdido en \mathbb{R} que se representa como `NaN` que se refiere a la expresión *Not a Number* y que aparece, por ejemplo, en el caso que dividamos cero entre cero.

Capítulo 2 - Qué es R y para qué se utiliza

4	[3,]	3	13	23
5	[4,]	4	14	24
6	[5,]	5	15	25
7	[6,]	6	16	26
8	[7,]	7	17	27
9	[8,]	8	18	28
10	[9,]	9	19	29
11	[10,]	10	20	30

Sin embargo, si damos tres dimensiones al vector v obtendríamos una estructura tridimensional. Por ejemplo, consideremos la idea de redimensionar el vector v original en tres dimensiones. Algo que podríamos hacer para que resultase ilustrativo sería crear una especie cubo o «dado» de datos con parámetros 5, 2, y 3. Esto es, vamos a crear una estructura de datos consistente en una tabla de cinco filas y tres columnas que se proyecta en una tercera dimensión tres veces. Para ello, podemos utilizar la siguiente sintaxis:

```
1 v <- 1:30
2 dim(v) <- c(5,2,3)
```

Si escribimos v en el editor de comandos de  podremos visualizar el objeto tridimensional despelegado que tendría este aspecto:

```
1 , , 1
2
3      [,1] [,2]
4 [1,]    1    6
5 [2,]    2    7
6 [3,]    3    8
7 [4,]    4    9
8 [5,]    5   10
9
10 , , 2
11
12      [,1] [,2]
13 [1,]   11   16
14 [2,]   12   17
15 [3,]   13   18
16 [4,]   14   19
17 [5,]   15   20
18
19 , , 3
20
21      [,1] [,2]
22 [1,]   21   26
```



```

23 [2,] 22 27
24 [3,] 23 28
25 [4,] 24 29
26 [5,] 25 30

```

Como se puede observar aparecen tres tablas cada una de ellas con cinco filas y dos columnas. En la figura 2.1 aparece representada la matriz tridimensional que acabamos de generar de un modo gráfico que puede ayudar a aclarar su interpretación.

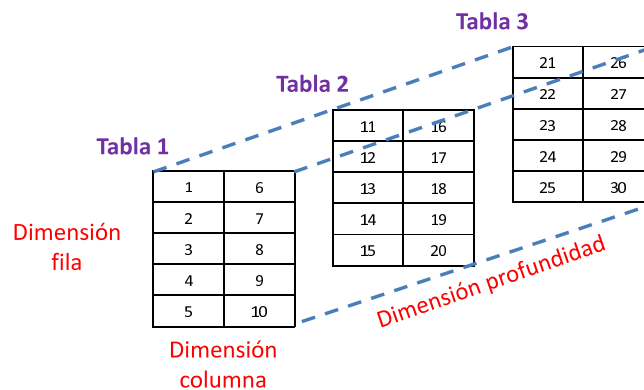


Figura 2.1: Estructura de una matriz tridimensional.

Un método destinado a crear matrices que simplifica todo lo anterior está basado en la utilización de la función `array()`. Por ejemplo, para construir una matriz que contenga los números del -5 al 10 con 4 filas y 5 columnas se puede utilizar la siguiente sintaxis:

```
1 z <- array(-5:10, dim=c(4,4))
```

Para terminar, me gustaría comentar una última función, `matrix()`, que también puede ser usada para crear matrices. Podríamos decir que la función tiene tres parámetros básicos: los datos, el número de columnas y el número de filas. Por ejemplo, para crear una matriz con tres columnas y con tres filas cuyos elementos sean el número 46 podríamos utilizar el siguiente código:

```
1 matrix(data=46, nr=3, nc=3)
```

Capítulo 2 - Qué es R y para qué se utiliza

donde `nr` (*number of rows*) se refiere al número de filas y `nc` (*number of columns*) al número de columnas y produciría el siguiente resultado:

```
1      [,1] [,2] [,3]
2 [1,]   46   46   46
3 [2,]   46   46   46
4 [3,]   46   46   46
```

Para simplificar la edición de sintaxis se pueden obviar los nombres de los parámetros de la función. Así, por ejemplo, si quisiésemos generar una matriz de 3 columnas y 4 filas que contuviese los números comprendidos entre 20 y 32 podríamos proceder del siguiente modo

```
1 matrix(20:31, 4, 3)
```

cuyo resultado sería:

```
1      [,1] [,2] [,3]
2 [1,]   20   24   28
3 [2,]   21   25   29
4 [3,]   22   26   30
5 [4,]   23   27   31
```

Si nos damos cuenta, en todas las matrices que hemos creado anteriormente los números se empiezan a ubicar en las celdas de la matriz por columnas (opción por defecto en la función). Sin embargo, podemos hacer que la matriz se rellene por filas. Para ello, hay que activar el parámetro `byrow`. Para crear la matriz anterior enumerada por filas utilizaríamos la sintaxis:

```
1 matrix(20:31, 4, 3, byrow=TRUE)
```

lo que produciría:

```
1      [,1] [,2] [,3]
2 [1,]   20   21   22
3 [2,]   23   24   25
4 [3,]   26   27   28
5 [4,]   29   30   31
```

Factores

Un factor es una especie de vector cualitativo que suele utilizarse como variable de agrupación cuando se llevan a cabo ciertos tipos de análisis estadísticos. Esto es, es un tipo de variable que almacena información categórica y que se puede utilizar para generar resúmenes numéricos respecto a otras variables cuantitativas.

Veamos en que consisten los factores con un ejemplo. Consideremos un grupo de trabajo en la universidad que consta de 14 miembros. Si registrásemos el color de ojos de los componentes del grupo tendríamos un factor (`ojos`) que podríamos incorporar a `R` del siguiente modo:

```
1 ojos <- c("Negros", "Marrones", "Azules", "Verdes", "Marrones",
2 "Azules", "Negros", "Marrones", "Marrones", "Azules", "Marrones",
3 "Azules", "Verdes", "Marrones")
```

Como se puede observar, cada elemento del objeto `ojos` está entrecomillado. Sin embargo, lo que tenemos por ahora es únicamente un vector de caracteres. Para convertirlo en un factor tenemos que utilizar la función `factor()` del siguiente de esta manera:

```
1 f_ojos <- factor(ojos)
```

Si ahora imprimimos el factor `f_ojos`, utilizando la función `print()` o escribiendo el nombre del objeto en la consola de comandos y presionando la tecla «Enter» de nuestro teclado, podremos observar que `R` presenta los factores de manera ligeramente diferente a como presenta a los vectores:

```
1 [1] Negros Marrones Azules Verdes Marrones Azules Negros Marrones
2 [9] Marrones Azules Marrones Azules Verdes Marrones
3 Levels: Azules Marrones Negros Verdes
```

Como se puede observar, `R` imprime los valores del factor obviando las comillas y, en la tercera línea del código anterior, añade información extra sobre los niveles⁹ (*Levels*) del factor. La función `levels()` también se puede utilizar para identificar cuáles son los niveles de un factor dado.

⁹También denominados como categorías o espacio de estados en otros contextos.


Capítulo 2 - Qué es R y para qué se utiliza

Para ilustrar el modo en que se pueden utilizar los factores propongo un ejercicio. Supongamos que conocemos el número de nominaciones como persona más guapa del grupo que ha recibido cada uno de los integrantes del grupo de trabajo anteriormente referido. Consideremos que el vector *nominaciones* representa el número de veces que una persona ha sido elegida por el resto de sus compañeras o compañeros como una persona bella:

```
1 nominaciones <- c(12, 10, 7, 8, 9, 6, 13, 10, 11, 6, 10, 8, 6, 9)
```

Si ahora quisiésemos saber cuál es el promedio de nominaciones positivas que recibe cada uno de los colores de ojos que hay en el grupo de trabajo podríamos utilizar la función `tapply()` de la siguiente manera:


```
1 tapply (nominaciones, f_ojos, mean)
```

Como se puede observar, la función `tapply()` tiene tres parámetros en este contexto separados por comas. Si traducimos la sintaxis anterior a lenguaje verbal podríamos decir que hemos pedido que se calculen las medias (*mean*) de nominaciones para cada uno de los colores de ojos que hay en el grupo docente. Como resultado  generaría el siguiente resultado:

```
1      Azules  Marrones   Negros   Verdes
2  6.750000  9.833333 12.500000  7.000000
```

Es decir, que el promedio de nominaciones para los ojos azules es de 6,75, para los ojos marrones 9,83 y así sucesivamente.

Listas

Las listas son una especie de generalización de los vectores que pueden contener elementos o *componentes* de naturaleza diversa. En muchas de las ocasiones  genera listas para informar sobre los resultados de análisis estadísticos. Por ejemplo, el resultado obtenido al utilizar la función `tapply()` que se ha introducido anteriormente es una lista.

Data frames o bases de datos

Los *data frames*, bases de datos o conjuntos de datos, son estructuras de datos análogas a las matrices. Por lo general, en el contexto que nos ocupa éstas matrices se interpretarán en el sentido en que hacen los programas estadísticos comerciales como SPSS. Esto es, cada fila corresponde a una observación, persona o participante y cada columna representa valores para una variable. Sin embargo, en contraposición a las matrices, los *data frame* pueden contener información de diversa índole¹⁰ y por tanto pueden contener tanto variables cuantitativas como cualitativas. En este libro se trabajará principalmente con este tipo de objetos cuando utilicemos [R](#).

Funciones

Una de las grandes ventajas que ofrece [R](#) es el hecho de que permite al usuario definir sus propias funciones. Ésto es, estructuras de cómputo programadas que realizan operaciones sobre estructuras de datos u otras funciones.

La manera general de definir una función toma la siguiente forma:

```
nombre-de-la-función <- function(arg-1, arg-2, ...) expresión
```

donde los elementos que aparecen en cursiva como *arg-n* se refieren a los argumentos de la función y la *expresión* se refiere a lo que hace la función. Por ejemplo, imaginemos que queremos crear una función que al darle dos números cualesquiera los multiplique y los divida por cinco. Para ello, tendríamos que definir la función del siguiente modo:

```
1 mifuncion <- function(x1, x2) x1 * x2 / 5
```

Si ahora queremos usar nuestra función tendremos que hacerlo de manera parecida a como hemos ido viendo hasta el momento con las funciones propias de [R](#). Por ejemplo, supongamos que queremos aplicar nuestra función a los números 55 y 28, tendríamos que proceder así:

```
1 mifuncion(55, 28)
```

Como habrás podido comprobar el *valor* generado por la función es 308.

¹⁰De hecho, los *data frame* son listas y un tipo particular de *clase* en [R](#).

Modos y atributos de los objetos

El *modo* de un objeto está referido al tipo básico de información que contiene. Consiste en una propiedad del objeto referida al tipo particular de información que contienen sus elementos particulares. Por ejemplo, un vector que contenga las estaturas de un grupo docente, como el que se ha descrito anteriormente, tendría un modo numérico (*numeric*), mientras que el vector que representa el color de los ojos de los integrantes del grupo sería un vector de caracteres (*character*). Para conocer cuál es el modo de un objeto se puede utilizar la función `mode(objeto)`.

Otra propiedad que se puede conocer fácilmente de un objeto es su longitud utilizando la función `length(objeto)`. Adicionalmente, podemos utilizar la función `attributes(objeto)` para conocer atributos adicionales del objeto.

2.2.2. Modelos estadísticos y gráficos

Ajustar modelos con \mathbb{R} es relativamente sencillo y rápido. Sin embargo, la salida que produce cuando se ajusta algún modelo es muy escueta. Por ello, es necesario utilizar *funciones extractoras* que suministren más información sobre el modelo estadístico generado.

Por su parte, las facilidades gráficas que proporciona \mathbb{R} han sido para mí, de manera más llamativa, el elemento que me atrajo a utilizar este software. Aunque al principio es duro enfrentarse con la edición de gráficos en este sistema, los resultados son dignos de resaltar. En primer lugar, habría que destacar que existen funciones gráficas de alto y de bajo nivel¹¹. Las *funciones gráficas de alto nivel* están diseñadas para crear gráficos completos. Por lo general, a no ser que se haya especificado de otro modo, los títulos de los ejes y de las etiquetas son generados automáticamente y cada vez que se ejecuta una función de alto nivel se genera un nuevo gráfico borrándose el previamente creado. Por su parte, las *funciones gráficas de bajo nivel* permiten personalizar gráficos cuando las funciones de alto nivel no han producido la salida gráfica del modo en que prefiere el usuario. Así, el usuario puede añadir puntos, líneas, textos y modificar un sinfín de cosas más utilizando estas funciones de bajo nivel.

¹¹ *High-level y low-level plotting commands.*

2.3. El paquete Rcmdr

El paquete **Rcmdr** (forma abreviada de escribir **R Commander**) consiste en una interfaz gráfica de usuario¹² que permite interactuar con **R** de un modo «amigable». O lo que es lo mismo, es un programa informático que permite interactuar con **R** utilizando las típicas ventanas y menús en que se basa el sistema operativo Windows. Esta interfaz fue desarrollada por John Fox de la McMaster University (Hamilton, Ontario, Canada). Con posterioridad, el paquete fue traducido al español por un grupo de docentes e investigadores de la Universidad de Cádiz bajo el proyecto «R-UCA Project» (<http://knuth.uca.es/R>).

De entre las múltiples interfaces gráficas que se han desarrollado para interactuar con **R** (Valero-Mora y Ledesma, 2012), **Rcmdr** es el más recomendable para usuarios nóveles por varios motivos. En primer lugar, como señala Elosua (2009), **Rcmdr** puede considerarse como el salto intermedio ideal entre los usuarios que utilizan programas estadísticos comerciales y el entorno de programación **R**. Y ello es así porque **Rcmdr** recuerda mucho a los paquetes estadísticos comerciales como SPSS en su modo de funcionamiento y presentación. Por otro lado, el uso de **Rcmdr** permite al usuario ir familiarizándose con la forma en que trabaja **R** dado que la sintaxis es generada e introducida en una parte de la interfaz gráfica.

2.3.1. El entorno gráfico de R Commander

Pues bien, tras haber proporcionado información sobre algunos de los elementos básicos de **R**, vamos a dedicar unas líneas a comentar, de manera genérica, los principales componentes de la interfaz gráfica de **Rcmdr**.

Como se puede apreciar en la Figura 1.16, en la parte superior de la interfaz gráfica tenemos el menú principal típico que aparece en la mayoría de programas creados para Microsoft Windows. En ésta sección tenemos las opciones de «Fichero», que servirá principalmente para abrir y/o guardar los archivos con los que estemos trabajando; «Editar», que contiene las opciones más usuales de la edición de documentos como las de cortar y copiar; «Datos», que nos permitirá gestionar y/o modificar bases de datos que contengan la información objeto de análisis; «Estadísticos», que contiene las opciones necesarias para deleitarnos con la generación y estimación de modelos y parámetros estadísticos; «Gráficas»,

¹²En inglés se denomina como *graphical user interface* o *GUI*.

Capítulo 2 - Qué es R y para qué se utiliza

que como su nombre indica nos proporciona un amplio surtido de posibilidades para representar gráficamente resúmenes de los datos contenidos en nuestras bases de datos; «Modelos», que nos permite comparar modelos estadísticos y/o estudiar la bondad de ajuste de los mismos; «Distribuciones», que nos permite generar gráficos y calcular parámetros relacionados con las distribuciones estadísticas más comunes; «Herramientas», que nos permite, entre otras cosas, cargar paquetes adicionales y/o *plugins*; y, finalmente, aparece la opción «Ayuda», donde se podrá encontrar información adicional sobre **R**.



Justo debajo del menú principal encontramos dos botones y dos listas desplegables. Las listas desplegables («Conjunto de datos:» y «Modelo:») no contendrán nada por el momento y aparecerán los siguientes mensajes en color rojo <*No hay conjunto de datos activo*> y <*No hay modelo activo*> respectivamente. La primera de éstas listas desplegables servirá para seleccionar una base de datos (de las múltiples que podemos tener cargadas) como candidata a ser analizada utilizando alguno de los procedimientos estadísticos que ofrece **R**. Por su parte, en la lista de «Modelos:» podremos seleccionar alguno de los modelos (de los diferentes que podemos haber creado) para aplicarle algún test de bondad de ajuste o para generar, entre otras, gráficos de diagnóstico.

Los botones «Editar conjunto de datos» y «Visualizar conjunto de datos» nos van a servir, como sus nombres indican, para cambiar algún dato, o datos, de la base de datos que tenemos activa y para ver el contenido de la base de datos activa¹³.

En **R** tenemos tres ventanas que nos servirán para diferentes propósitos. En primer lugar, y en la parte superior de la interfaz, tenemos la «Ventana de instrucciones» donde se escribirán los comandos que serán enviados a **R** y que **R** escribirá por nosotros para hacernos más llevadera nuestra interacción con **R**. En la parte central y ocupando la mayor porción de la interfaz tenemos la «Ventana de resultados» que vendría a equivaler a la consola de **R** y donde aparecerán, en color rojo, los comandos que vayamos ejecutando y, en azul, los resultados de los análisis que hayamos ordenado. Finalmente, en la parte inferior de la interfaz tenemos una ventana llamada «Mensajes» donde se nos mostrarán informaciones relevantes relacionadas con los procesos o cálculos que estemos realizando.

¹³En este segundo caso no podremos modificar ningún dato, sólo podremos *visualizar* la base de datos

Por ejemplo, cuando cometamos algún error en algún comando, ésta ventana nos informará de ello.

Entre la ventana de instrucciones y la ventana de resultados (en la parte derecha de la interfaz) aparece un botón llamado «Ejecutar». Éste botón servirá para enviar a  porciones de código concreto para que las ejecute. Cuando pulsamos en el botón «Ejecutar»¹⁴ se envía a  la línea de código donde se encuentra el cursor. Podemos también enviar varias líneas de código si las seleccionamos previamente con el botón izquierdo del ratón. Por ejemplo, si escribimos lo siguiente en la ventana de instrucciones:

```
1 3+5
```

nos aparecerá lo siguiente en la ventana de resultados; eso sí, salvando la diferencia de que lo que aparece en la línea 1 está en rojo y lo que aparece en la línea 2 se ve en negro:


```
1 > 3+5
2 [1] 8
```

Como se puede observar, lo que aparece en la ventana de resultados es el resultado del comando que hemos mandado ejecutar ($3+5$) junto con el comando mismo¹⁵. Por lo general, los comandos aparecerán en color rojo (precedidos por el símbolo $>$ ¹⁶) en la ventana de resultados y los resultados, propiamente dichos, aparecerán en color azul. Supongamos que escribimos lo siguiente en la ventana de comandos:

```
1 x <- 3+5
2 x + 6
```

Si seleccionamos ambas líneas de código en la ventana de instrucciones y presionamos el botón «Ejecutar», nos aparecerá algo similar a ésto en la ventana de resultados:

¹⁴También podemos obtener el mismo resultado presionando las teclas **Ctrl+R**.

¹⁵Como se indicó previamente, y para facilitar la interpretación de los códigos que aparecerán en este manual, cuando aparezcan cuadros de código como el que aparece aquí arriba se estará tratando de una salida o un resultado de .

¹⁶Denominado como *símbolo del sistema*.

Capítulo 2 - Qué es R y para qué se utiliza

```
1 > x <- 3+5
2 > x + 6
3 [1] 14
```

Lo que aparece en la línea 1 del código que aparece más arriba, como se indicó previamente, es lo que se denomina como **asignación**. Esto es, hemos creado un objeto que se llama x cuyo valor es $3 + 5$. Luego hemos ordenado, en la segunda línea, que se sume 6 a x . Como resultado, que aparece en la línea 3, tenemos 14; que es justamente el resultado de sumar $3 + 5 + 6$. Lo importante por ahora, ya que aprenderemos más cosas sobre asignaciones en sucesivas secciones, es darse cuenta de que podemos ejecutar varias líneas de código si previamente las seleccionamos y, seguidamente, pulsamos el botón «Ejecutar».

2.3.2. Abrir archivos

Por lo general, para poder aplicar cualquier tipo de análisis estadístico hay que tener una base de datos activa en la ventana «*Conjunto de datos*». En el caso de que haya varios conjuntos de datos cargados, los análisis se ejecutarán sobre el conjunto de datos activo. R permite varias formas de incorporar datos en R:

1. Se puede crear una base de datos partiendo desde cero accediendo al menú *Datos* → *Nuevo conjunto de datos...* Cuando ejecutamos el comando aparece un cuadro de diálogo (Figura 2.2) que nos demanda un nombre para la base de datos que vamos a crear y por defecto nos propone el título de «Datos».

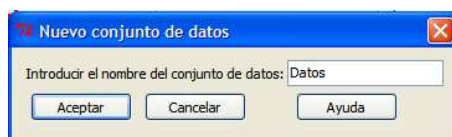


Figura 2.2: Nuevo conjunto de datos en R Commander.

Cuando pulsamos en el botón «Aceptar» nos aparecerá el editor de datos que podría describirse como una especie de tabla (Figura 2.3) con líneas de división rojas. Si nos fijamos, cada fila horizontal está enumerada y cada columna está etiquetada con la expresión «*var n* ». Si pulsamos en alguna de éstas etiquetas de las columnas nos aparecerá un cuadro de diálogo que sirve

para modificar ésta etiqueta (que será el nombre de la variable) y definir el tipo de variable que queremos introducir (que puede ser cuantitativa¹⁷ o cualitativa¹⁸).

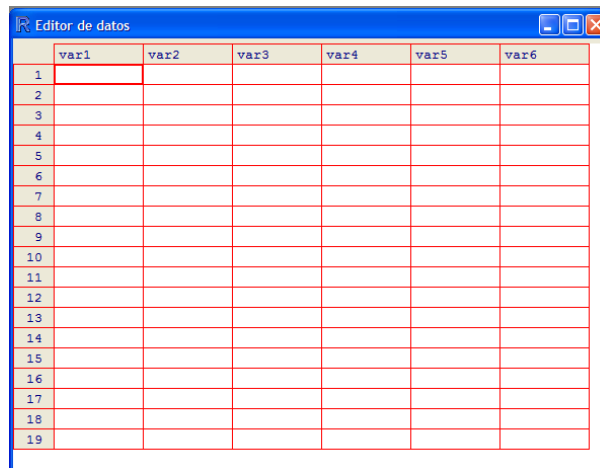


Figura 2.3: Editor de datos en R Commander.

Una vez denominadas las variables e identificada el tipo de información que contendrán se pueden ir introduciendo los valores en cada casilla como si estuviésemos utilizando una hoja de cálculo. No obstante, este procedimiento para incorporar datos a **R** sólo suele ser recomendado cuando tratemos de generar bases de datos relativamente pequeñas.

- Si tenemos bases de datos provenientes de otros programas informáticos o estadísticos también podemos importarlas con **R**. Por ejemplo, accediendo al menú *Datos* → *Importar datos*, tenemos la opción de importar archivos desde SPSS, Minitab, STATA, Excel, Access o dBase. Creo que es de particular importancia, en este punto, dedicar unas palabras a la primera de las opciones que nos encontramos en éste comando referida a la importación de datos *desde archivo de texto, portapapeles o URL...* A lo largo de mi experiencia trabajando con bases de datos que han sido analizadas estadísticamente, he llegado a la conclusión de que cuanto más sencilla sea la estructura de los datos mejor. Uno de los tipos de archivos más sencillos y versátiles que conozco es el archivo de texto plano (con extensión

¹⁷ *Numeric* o numérica.

¹⁸ *Character* o cualitativa.

Capítulo 2 - Qué es R y para qué se utiliza

«*.txt»). Mi experiencia con este tipo de archivos ha sido relativamente satisfactoria cuando he tratado de importar o exportar algún tipo de bases de datos en procesos de intercambio de datos entre unos programas y otros. En apariencia no son bonitos, pero creo que lo importante es que funcionen bien.

Creo que casi cualquier editor de textos puede generar y manipular documentos de texto plano y en Windows la herramienta que maneja estos archivos por defecto es el *Bloc de notas*. Los archivos de texto plano no contienen ningún tipo de floritura en sus caracteres (nada de negritas, cursivas, colores de letra, etc.), únicamente permiten la posibilidad de incorporar cierto tipo de caracteres alfanuméricos. Un ejemplo de archivo de texto sería el siguiente:

edad	color.ojos	nominaciones	nota.media	grupo
20	M	13	5,5	A
22	V	12	4,9	A
25	A	8	7	B
24	M	14	8,2	A
23	N	12	9,1	A
21	A	10	3,5	A
22	N	9	4,6	B
20	V	5	7	B
25	N	7	6,2	A

Si quisiéramos incorporar ésta base de datos a [R](#) podríamos seguir diferentes procedimientos. En primer lugar, podríamos copiar la tabla en el portapapeles desde éste documento PDF y seleccionar la opción *Datos → Importar datos → desde archivo de texto, portapapeles o URL...* Nos aparecerá un cuadro de diálogo como el que aparece en la Figura 2.4. Como se puede comprobar lo primero que se nos demanda es un nombre para el conjunto de datos y, dado que los nombres de las variables aparecen en la primera fila (edad, color.ojos, etc.) tendremos que dejar marcada la casilla de verificación que está activada por defecto. A continuación, tenemos que identificar la localización del archivo. Dado que hemos copiado los datos en el portapapeles tendremos que elegir esa opción. Posteriormente tenemos que indicar el elemento que separa los campos o variables (columnas) del

archivo de datos. En este caso tendremos que seleccionar la opción *Espacios en blanco*. Por último, dado que nuestra base de datos contiene una variable donde hay números decimales, tendremos que especificar que el separador decimal es la coma. Una vez especificadas todas las opciones correspondientes podemos clicar en el botón «Aceptar».

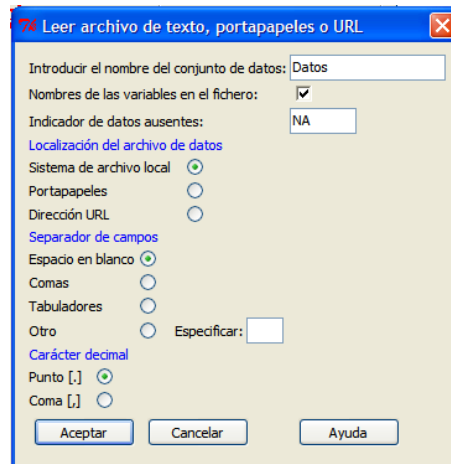



Figura 2.4: Importar datos.

Como habrás podido comprobar, el editor de sintaxis habrá escrito el siguiente código en el que se detallan las opciones que hemos definido en el cuadro de diálogo¹⁹:

```
1 Datos <- read.table("clipboard", header=TRUE, sep=" ", na.strings="NA", dec="
  ", strip.white=TRUE)
```

Si pulsamos en el botón «Visualizar conjunto de datos» de la interface gráfica podremos cerciorarnos de que la importación del archivo ha sido realizada correctamente. La importación de éstos datos podría haberse hecho de manera análoga utilizando un archivo de texto y seleccionando la opción correspondiente en el cuadro de diálogo. Para practicar éste segundo método de importación te recomiendo que utilices el archivo `texto-plano.txt` que acompaña a este manual.

3. Otra forma en que podemos encontrar los datos es en el formato propio de datos que maneja , esto es, en un archivo con extensión «*.RData».

¹⁹Importante es darse cuenta que se ha utilizado la función `read.table()`.

Capítulo 2 - Qué es R y para qué se utiliza

Como habrás podido comprobar en el editor de instrucciones, `Rcmdr` ha utilizado la función `load()` para abrir el archivo de datos indicando la ruta exacta donde se encuentra el archivo que deseas abrir.

- Una última forma de incorporar bases de datos en `Rcmdr` consiste en cargar un archivo contenido en algún paquete de `R`. Para ver un listado de las bases de datos que hay disponibles para poder ser cargadas desde `Rcmdr` puedes seleccionar el comando *Datos* → *Conjunto de datos en paquetes* → *Lista de conjunto de datos en paquetes*. Para cargar un conjunto de datos tendremos que seleccionarlo en el cuadro de diálogo que aparece al ejecutar el comando *Datos* → *Conjunto de datos en paquetes* → *Leer conjunto de datos desde paquete adjunto...* Como podrás comprobar en el cuadro de diálogo que aparece (Figura 2.5), a la izquierda aparece una lista de paquetes disponibles sobre los que podemos hacer clic para acceder a los archivos de datos específicos que contienen y que aparecerán en el cuadro de la derecha. Dado que las bases de datos contienen información que *a priori* puede no ser inteligible, el cuadro de diálogo da la opción de obtener información sobre un conjunto de datos particular pulsando sobre el botón «Ayuda sobre el conjunto de datos seleccionado».

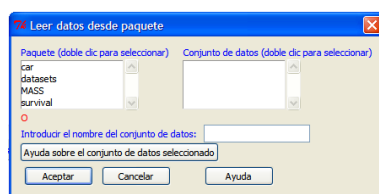


Figura 2.5: Importar datos desde paquetes.

Por último, otra opción es escribir directamente el nombre de la base de datos en la casilla de texto que se proporciona. En este caso lo que estamos haciendo es acceder a datos almacenados en el paquete `datasets` de `R`. Para hacerlo utilizando la sintaxis tenemos que utilizar la función `data()`. Por ejemplo, si quisiésemos cargar los datos contenidos en la base de datos `iris` que contiene información sobre diferentes características de 150 flores de tres especies diferentes (`setosa`, `versicolor` y `virginica`), tendríamos que escribir y ejecutar la siguiente sintaxis:

```
1 data(iris)
```

2.3.3. Guardar archivos

Un último elemento que me gustaría destacar llegados a este punto sobre el manejo básico de Rcmdr está referido a la grabación de archivos. Para guardar los archivos que hemos creado, importado o modificado con Rcmdr podemos utilizar, como mínimo, dos procedimientos a los que se accede desde el menú *Datos* → *Conjunto de datos activo*. Las dos últimas opciones de este comando están destinadas a guardar y a exportar los datos que estamos manejando. Si seleccionamos la opción *Guardar el conjunto de datos activo...* tendremos que seleccionar una ubicación de nuestro equipo para grabar el archivo en una extensión propia de R, por ejemplo, en «*.RData». Sin embargo, si seleccionamos *Exportar el conjunto de datos activo...* tendremos que especificar, en primer lugar, las características (separador de variables, codificación de los valores ausente, etc.) que deseamos para el conjunto de datos y, en segundo lugar, la extensión del archivo generado.

□ EJERCICIOS □

1. ¿Cómo podríamos generar una matriz con dos columnas y tres filas que contenga los valores 46, 34, 23, 56, 21, 90? Propón, al menos, tres posibilidades diferentes.
2. Crea una función que, dados cuatro números, genere una matriz 2×2 completada por filas.
3. Crea una pequeña base de datos, tres variables y diez casos, y guárdala en diferentes formatos (.txt, .RData o .csv).
4. Abre o importa las bases de datos que has generado previamente y comprueba que el proceso se ha ejecutado correctamente.

Capítulo 2 - Qué es R y para qué se utiliza

3

Notas sobre la investigación en psicología y educación

Este capítulo va a ser breve pero intenso, o al menos esa es mi intención. Aunque me voy a desviar ligeramente del tópico central de este manuscrito (el uso de \mathbb{R} y de \mathbb{R}_+ para el análisis de datos en psicología y educación), creo que unas pocas palabras genéricas sobre ciertos aspectos de la investigación científica no estarían de más.

La primera idea que me gustaría resaltar en este contexto es que, como indica Bachrach (1966/1994), «investigar no es sólo estadísticas» sino que, más bien, la estadística es «una herramienta de la actividad investigadora» (p. 17). Así, no podemos pensar que nuestra investigación es única y exclusivamente una estadística. Aunque tendremos que utilizar modelos estadísticos en nuestras investigaciones para contrastar hipótesis, tenemos que tener en cuenta que un estudio científico abarca más elementos que el análisis estadístico propiamente dicho.

Por otro lado, sin ánimos de entrar en una reflexión ética, moral o filosófica; también convendría señalar que los modelos estadísticos pueden usarse mejor o

peor. Como indica Jovel (1995), «es fácil mentir con la estadística, pero es aún más fácil mentir sin ella» (p. 10). Así, aunque los modelos estadísticos son potentes herramientas que ayudan al personal investigador a saciar sus necesidades de conocimiento sobre la Naturaleza, también es cierto que, en algunas ocasiones, pueden conducirnos a extraer conclusiones poco acertadas. Por ello, me gustaría alertar de que el mero hecho de utilizar modelos estadísticos no garantiza que nuestras conclusiones sean válidas desde el punto de vista científico. Por citar un ejemplo que suelo comentar con mi amigo Freddy Soto Bravo de la Universidad de Costa Rica, podría aludir a la representación gráfica de medias aritméticas en diagramas de barras y de cómo se puede generar una sensación de diferencia sustancial entre tratamientos o condiciones modificando la escala del eje de ordenadas. Por consiguiente, recomiendo a los usuarios de técnicas estadísticas en los campos de estudio que nos ocupan que traten de ser lo más asépticos y responsables posibles cuando usen modelos estadísticos para contrastar sus hipótesis de investigación.

3.1. Medición

Por suerte o por desgracia, la ciencia, tal y como es entendida hoy en día, no podría concebirse sin la idea de «medición». Independientemente del concepto de medición que tengamos en mente, parece poco sensato no asociar la medida a los números y a la cuantificación de un referente físico más o menos objetivable. Dado que dedicaré la próxima sección a tratar la relación entre los números y la realidad que representan, escribiré algunas líneas para comentar brevemente la idea del *referente físico* en los procesos de medición, particularmente en el campo de la psicología y la educación.

Bachrach (1966/1994) sugiere que las medidas psicológicas (o incluso cualquier tipo de medida) han de basarse en un fenómeno físico. Es decir, que cualquier medida psicológica ha de tener un referente físico que sea susceptible de ser medido objetivamente. Por ejemplo, si pretendemos medir el nivel de activación psicológica que tiene una persona podríamos utilizar un referente bio-físico como la respuesta electrogalvánica de la piel, la tasa cardíaca o la presión sistólica para estimar el grado de excitación psicológica que embarga a la persona. Sin embargo, el utilizar una medida física de un fenómeno no garantiza que dispon-

gamos de una estimación apropiada del fenómeno estudiado. Trataré de exponer un ejemplo aludiendo a las sofisticadas técnicas de neuroimagen que tanta fama y reconocimiento están recibiendo hoy en día en el contexto de la investigación psicológica.

Antes que nada, me gustaría aclarar que no es mi intención atacar destructivamente ningún área ni campo de trabajo. Esto es, no estoy en contra de la neurociencia, de la neurociencia cognitiva, de la psicobiología, de la psicofisiología, de la neuropsicología o de la psicofarmacología, por citar sólo algunas. Más bien al contrario. Yo comencé a estudiar psicología atraído por la *ciencia del cerebro*. Me fascinaba la idea de entender cómo nuestra «maquinaria» biológica era capaz de generar sensaciones, emociones, ideas, percepciones, aprendizajes y un sin fin de abstracciones complejas que permiten que seamos lo que somos (López, 2009). Sin embargo, si es cierto que tengo «clavada una espinita» al no haber sido lo suficientemente brillante como para dedicarme a ese campo de estudio que ha sido, y es, tan atractivo para mí. En cualquier caso, no voy hacer apología del la vocación frustrada y trataré de presentar lo más claramente posible el problema que se percibe cuando se tratan de medir fenómenos psicológicos utilizando técnicas de neuroimagen como la Resonancia Magnética Funcional o la (RMf) Tomografía por Emisión de Positrones (TEP).

Tanto la RMf como la TEP son técnicas de neuroimagen que permiten estudiar el funcionamiento el cerebro de una manera no lesiva. La TEP se sirve de moléculas marcadas radiactivamente (normalmente la 2-desoxi-D-glucosa o 2-DG) para identificar las neuronas activas del cerebro que están funcionando cuando los participantes experimentales realizan tareas cognitivas. Lo que se hace es inyectar una disolución de esta sustancia radiactiva y, tras un intervalo de tiempo, pedir al participante experimental que ejecute la tarea que implica procesos cognitivos como recordar, atender o aprender. Lo que hace la máquina de TEP es detectar en qué zona o zonas del cerebro se concentra la sustancia radiactiva. Por su parte, la RMf no requiere la administración de ninguna sustancia radiactiva sino que, más bien, es capaz de detectar variaciones en las concentraciones de oxígeno en diferentes partes del cerebro. Podríamos decir que ambos métodos de medición son técnicas metabólicas ya que estiman el grado en que las neuronas objetivo están metabolizando glucosa u oxígeno en un momento dado cuando la persona ha sido desafiada con una tarea psicológica. Dado que el

Capítulo 3 - Notas sobre la investigación en psicología y educación

consumo de glucosa y de oxígeno es un referente físico de la actividad neural, se puede concluir que cuando una neurona acumula glucosa radioactiva o metaboliza oxígeno está emitiendo potenciales de acción o impulsos nerviosos que denotan la activación funcionalmente relevante de tal célula. Sin embargo, pueden haber explicaciones alternativas (Bardin, 2012). Por ejemplo, la neurona que consume altas proporciones de oxígeno o de glucosa podría estar preparándose para una subsecuente síntesis de proteínas que le permita generar componentes celulares como canales iónicos u orgánulos celulares. Aunque ésta es una observación sin mayor importancia que cuestiona la validez de las medidas referidas a un componente objetivo de este tipo de técnicas de neuroimagen funcional, también es cierto que no suele comentarse en los libros de texto al uso que las describen (p. e., Carlson, 1993/2000) o se hace, a mi modo de ver, de manera superflua (p. e., Pinel, 2011).

Sin ánimo de crear discordia y evitando agravios comparativos, podríamos aludir a antecedentes históricos en los que la medición de fenómenos psicológicos que ha implicado un referente físico, en cierto modo, no han sido acertados del todo. La frenología es un ejemplo claro donde se utilizaba una medida física relativamente objetiva y que daba lugar a inferencias inválidas sobre el fenómeno medido (Hothersall, 1995/1997). Otro ejemplo de un uso, digamos, inapropiado de una medida física destinada a explicar un fenómeno psicológico lo representa el estudio del volumen craneal de diferentes razas y especies humanas usando perdigones o granos de mostaza descrito por Gould (1981).

Por todo lo anteriormente expuesto, creo que el hecho de que una medida tenga un referente físico relativamente objetivable no es requisito indispensable para considerarla de calidad. Más bien, ante cualquier medida deberíamos exigir, al menos, dos propiedades técnicas que garanticen la calidad de la estimación: **fiabilidad** y **validez**. La fiabilidad (entendida como estimación del grado de error que se comete al medir en términos de consistencia interna o de estabilidad temporal, entre otras) no parece ser la responsable de las vicisitudes anteriormente descritas. Más bien, el asunto que nos incumbe podría enfocarse desde el punto de vista de la valoración de la validez de la medida. En la actualidad, la validez de la medida se refiere al grado en que las inferencias que extraemos de una medida o puntuación son útiles en un contexto determinado y para un uso concreto (Cook y Beckman, 2006). Así, una medida podría ser válida para un uso y en un

contexto determinado mientras que la misma medida podría no serlo para otro uso o en otro contexto. Por ejemplo, si nos preguntamos sobre la validez de las medidas que generaba la frenología utilizando la pregunta ¿son útiles las medidas craneales para predecir el futuro laboral o el nivel de agresividad de una persona?, llegaríamos a la conclusión de que éstas medidas no gozaban de validez. En el caso de la TEP y la RMf cabría preguntarnos ¿son los cambios metabólicos de oxígeno y glucosa en el cerebro funcionalmente relevantes desde un punto de vista psicológico?

3.2. Niveles o escalas de medida

Como se habrá podido comprobar en la sección anterior, no puedo obviar mi experiencia como profesor de la asignatura *psicometría* en la licenciatura de psicología durante los últimos años. Pues bien, en esta sección voy a continuar aludiendo a conceptos e ideas que he venido tratando con relativa vehemencia en mis clases de esta asignatura. En concreto, voy a tratar de dar unas pinceladas sobre lo que propuso Stevens (1946) en un artículo sobre las escalas o niveles de medida que podríamos considerar como *una verdadera obra de arte*.

Como se ha visto en un capítulo anterior, **R** permite definir únicamente dos tipos de variables: las cuantitativas o numéricas y las cualitativas. Esta clasificación de los tipos de variables atiende a la diferenciación clásica estadística pero no es la única forma posible de clasificar variables. Por ejemplo, el paquete estadístico SPSS permite definir el nivel de medida de cada variable atendiendo a tres posibles valores: nominal, ordinal y escala. Yo no creo que la definición del tipo de una variable en un programa estadístico sea tan crucial. Esto es, no creo que tenga mayor importancia definir una variable de uno u otro modo. Ahora bien, lo importante de la definición del tipo de variable subyace en el tipo de análisis estadísticos que le podemos aplicar. Aunque en los programas estadísticos existen ciertos controles para evitar inconsistencias de cálculo (evitar que se calcule una media para una variable cualitativa que registra el color de ojos en una muestra), creo que es conveniente que se esté familiarizado con la idea de nivel o escala de medida de una variable para, en la medida de lo posible, adaptar los cálculos estadísticos permisibles para cada tipo de variable. Dado que la responsabilidad final de este asunto recae sobre las personas que realizan una investigación o

estudio científico, creo que sería deseable que el personal investigador estuviese familiarizado con estos conceptos dado que aplicar modelos estadísticos a variables que no satisfagan los criterios métricos correspondientes podría atentar seriamente contra las conclusiones que se extrajesen del análisis matemático.

Como se ha comentado anteriormente, las sub-secciones que aparecen a continuación tienen como objetivo introducir brevemente los tipos de escala o los niveles de medida propuestos por Stevens (1946). A grandes rasgos, podríamos decir que lo que persigue la propuesta de Stevens es tratar de establecer un conjunto de reglas por las se pueden asignar o asociar números a fenómenos observables. O dicho de otro modo, pretende identificar cuáles son las propiedades del fenómeno que están representadas por el número en cada caso. Como se verá a continuación, las cuatro escalas que se describen aquí tienen una especie de *estructura jerárquica* dado que las propiedades de una escala con un nivel de medida «inferior» son absorbidas o asumidas por la escala de un nivel superior. Es decir, por ejemplo, aunque la escala de intervalo se caracteriza por representar una propiedad particular de un fenómeno observable también asume las propiedades de la escala ordinal que le precede en la jerarquía.

3.2.1. Escalas nominales

El nivel más básico de medida estaría cubierto por la medición en una escala nominal. En este caso la única propiedad del número asignado a un fenómeno observable es la de igualdad-desigualdad. Esto es, al utilizar una medida a nivel nominal lo único que estamos haciendo es decir que cada categoría numérica es diferente a otra. Se podría decir que ésta forma de medir no dista mucho de lo que llamamos medición cualitativa ya que lo único que hacemos es identificar si las manifestaciones observables de un fenómeno son iguales o diferentes unas de otras.

Un ejemplo de variable nominal¹ podría ser el conjunto de los números del Documento Nacional de Identidad de una muestra de personas. En este caso cada uno de los números es como una especie de nombre que identifica inequívocamente a una persona concreta. Es decir, no puede haber dos personas con un mismo número ni una misma persona que tenga dos números diferentes. Además, dado

¹Fíjate que la palabra *nominal* recuerda a la palabra *nombre*.

que el número sólo representa la propiedad de igualdad-desigualdad, el hecho de que una persona tenga un número igual a 10.000.000 no implica que sea la mitad de persona que otra que tenga un número igual a 20.000.000. Lo único que indica el número es que ambas personas son diferentes porque tienen asignado un número diferente. Este sería un caso «extremo» de variable nominal en la que el mismo número nunca sería repetido pero en la práctica, las variables de tipo nominal suelen usarse para representar el color de los ojos (por ejemplo, 1 = azules, 2 = verdes, 3 = negros, etc.), el color del pelo o el estado civil.

En este tipo de escala se puede realizar un tipo de transformación que se denomina *permutación o transformación grupal simétrica* (Stevens, 1946). Esto es, podemos cambiar un número por otro cualquiera siempre y cuando se mantenga intacta la regla de asignación.

3.2.2. Escalas ordinales

La escala de tipo ordinal, además de conservar la propiedad de identificar igualdad-desigualdad heredada de la escala nominal, es capaz de representar el orden del fenómeno observado. De este modo, el número asignado a un caso o persona representa cierto aspecto de cantidad. En este caso podríamos decir que la regla de asignación del número al fenómeno vendría a ser algo así como «a más cantidad de fenómeno se asigna un número mayor».

Sin embargo, la relación que se establece entre el fenómeno y el número que lo representa no es de tipo lineal o directamente proporcional. Trataré de explicar este detalle tan interesante con un ejemplo que podría ser tomado de una investigación real. Consideremos una pregunta de una encuesta destinada a medir el grado con que una persona es favorable al uso de las centrales nucleares como fuente de energía eléctrica. Imaginemos que las personas que realizan la investigación han decidido proporcionar cuatro posibles alternativas de respuesta y que las codificarán utilizando números del siguiente modo: 1 = estoy totalmente en contra de las centrales nucleares, 2 = estoy ligeramente en contra de las centrales nucleares, 3 = estoy ligeramente a favor de las centrales nucleares, y 4 = estoy totalmente a favor de las centrales nucleares. Como se puede comprobar, el ítem al que se está aludiendo tiene como fin evaluar una actitud hacia un objeto (las centrales nucleares como fuente de energía) y, obviando por el momento las interesantes polémicas que circundan al estudio de las actitudes en psicología (p.

e., Allport, 1935; Ajzen y Fishbein, 1980, 2005), sugiero que nos centremos en la relación que se establece entre el número asignado a cada alternativa de respuesta y la descripción verbal de la actitud. Según lo que se ha definido previamente, cuando una persona puntúa 1 en la pregunta diríamos que su actitud hacia las centrales nucleares es negativa o desfavorable mientras que cuando puntúa un 4 diríamos que su actitud hacia estas factorías energéticas es favorable o positiva. En este sentido, podríamos decir que la persona que puntúa 2, en comparación con la que puntúa 1, tiene una actitud más favorable hacia las centrales nucleares. Del mismo modo, la persona que puntúa 3 mostraría una actitud más positiva que la que puntúa 1. Sin embargo, no podríamos afirmar que el cambio en actitud que se produce entre 1 y 2 sea el mismo que el que se produce entre 2 y 3 o entre 3 y 4. Lo único que podemos hacer es *ordenar* las respuestas de las personas en función de sus contestaciones, pero no podríamos hacer inferencias en relación a la cantidad de actitud que poseen cuando comparamos unas y otras respuestas. Del mismo modo, tampoco podríamos decir que la persona que puntúa 4 tiene una actitud el doble de positiva hacia las centrales nucleares cuando la comparamos con una persona que puntúa 2.

En las escalas ordinales se pueden realizar un tipo de recodificaciones llamadas *isotónicas* o de *preservación del orden* sin que alteremos las propiedades métricas de las variables utilizadas. Esto es, podemos cambiar los números de la escala por otros que sigan manteniendo la misma relación de orden previamente establecida.

3.2.3. Escalas de intervalo

Las escalas de intervalo suelen ser las más apreciadas en ciencias sociales en la actualidad dado que permiten la utilización de una mayor gama de técnicas estadísticas. La propiedad que añaden este tipo de escalas es que, como su nombre indica, el intervalo entre dos valores de la escala es ahora significativo pese a que el origen de la escala, el cero, es arbitrario. Es decir, la cantidad de fenómeno observado entre dos valores dados de la escala tiene sentido cuantitativo.

Uno de los ejemplos más utilizado de escala de intervalo es la temperatura medida con la escala de grados centígrados o Celcius. Consideremos las temperaturas dadas en grados centígrados de cuatro ciudades diferentes: la ciudad A con una temperatura de 10, la B con 12, la C con 18 y la D con una temperatura de 20 grados centígrados. Dado que la escala que estamos utilizando es de

intervalo podríamos decir que el incremento de temperatura que se produce entre las ciudades A y B es de la misma magnitud que el que se produce al comparar las ciudades C y D. Es decir, el intervalo tiene significado numérico real, no se trata sólo de ordenar las ciudades en función de si son más o menos cálidas. No obstante, no podemos afirmar que hace el doble de calor en la ciudad D cuando la comparamos con la ciudad A dado que el origen, el cero, de la escala Celcius es arbitrario. Es decir, cero grados centígrados no indican ausencia de temperatura sino que, más bien, indican la temperatura en la que el agua pasa de estado líquido a estado sólido.

En este tipo de escalas podemos realizar transformaciones lineales² de las puntuaciones ($x' = a + b \times x$) sin que se modifiquen sus propiedades métricas.

3.2.4. Escalas de razón

Para terminar, en la cúspide de la jerarquía, tenemos a las escalas de razón cuya característica primordial es que contienen lo que podríamos llamar *cero significativo*. Esto es, el cero en la variable que estemos midiendo representa ausencia de medida.

Por continuar con el ejemplo paradigmático que se ha introducido anteriormente para describir las escalas de intervalo (p. e., Pagano, 1998/1999), podríamos aludir a la escala de temperatura Kelvin. En esta escala los 0 grados Kelvin (aproximadamente unos -273 grados centígrados) se consideran como el «cero absoluto» dado que no se puede conseguir una temperatura más baja. En este caso sí que podríamos decir que la temperatura de 80 grados Kelvin es el doble de cálida que 40 grados Kelvin. En cierto modo, por ello este tipo de escalas han recibido el nombre de *razón* dado que las razones o proporciones son significativas. Es decir, tienen sentido matemático. Otras variables que podríamos definir como de razón podrían ser el número de hijos, la frecuencia de ocurrencia de un evento o la edad.

En este último caso sería en el único en que podríamos llevar a cabo transformaciones logarítmicas de las puntuaciones de las variables. Es decir, si utilizamos escalas de razón podremos realizar transformaciones de *similaridad* utilizando expresiones análogas a $x' = a \times x$.

²Más adelante trataremos este tipo de recodificación en el capítulo correspondiente.

3.2.5. Estadísticos admisibles en función del nivel de medida

Tal y como se ha indicado previamente, la decisión de qué análisis aplicar a los datos depende única y exclusivamente de la persona o personas que realizan la investigación y de quién ejecuta los análisis estadísticos. Sin embargo, convendría, llegados a este punto, hacer notar que estimar ciertos estadísticos en cierto tipo de variables podría no ser deseable. O, más bien, que las inferencias que extraeríamos de la estimación de ciertos estadísticos podría ser, al menos, confusa. Por ejemplo, ¿qué sentido tendría calcular la media aritmética en una variable que representase el color de ojos tras haber utilizado una escala nominal como la que se ha sugerido anteriormente? Más bien, cuando utilizamos un tipo de escala nominal únicamente se recomienda estimar frecuencias de aparición de cada nivel de la variable o porcentajes. También se podrían utilizar tests estadísticos, como el de χ^2 , diseñados para trabajar con frecuencias y proporciones.

Para el caso de las variables ordinales sería aconsejable utilizar estadísticos de posición como los cuantiles o estadísticos como el coeficiente de correlación de Spearman o Kendall. Por su parte, necesitaríamos, como mínimo, un nivel de medida de intervalo para poder utilizar exitosamente la media, la desviación típica o el coeficiente de correlación de Pearson. Por último, para poder aplicar el coeficiente de variación en los términos sugeridos por Stevens (1946) nuestras variables tendrían que haber sido medidas en una escala de razón.

3.3. Planificación y análisis estadístico

Para terminar este capítulo dedicado a tratar algunos aspectos que considero claves en el contexto de la investigación científica en psicología y educación, me gustaría dedicar algunas líneas al tema de la planificación de la investigación en relación con el análisis estadístico. En este sentido, creo que sin planificación cualquier análisis estadístico de los datos será prácticamente inútil, como un terreno baldío, estéril e infructífero.

Aunque bien es cierto que, como señala Bachrach (1966/1994), «no se investiga, por lo general, en la forma en que dicen que se hace los que escriben libros acerca de la investigación» (p. 22) cuando alude a que la planificación de la inves-

3.3 - Planificación y análisis estadístico

tigación no siempre se satisface en el desarrollo de un estudio científico; lo cierto es que sin un plan premeditado que guíe nuestra actividad científica estamos perdidos. Aunque el personal investigador no debe ser una mente obtusa, cerrada e inflexible donde no quepa la improvisación relativa; lo cierto es que las hipótesis deberían guiar la planificación de una investigación tal y como sugiere el método científico.

Mi recomendación en este contexto es «planificar siempre» y aunque, como suele ser lo normal, las cosas no vayan como se hayan planeado, sería deseable tener un plan que guíe nuestro quehacer como personas de ciencia. He tenido algunas experiencias en las que reputados investigadores e investigadoras me han planteado la posibilidad de analizar bases de datos para satisfacer tales o cuales objetivos una vez que la investigación ya ha sido llevada a cabo. En estos casos mi respuesta viene siendo la misma (eso sí, más enrabiada cada vez): el análisis de los datos debía de haber sido planificado antes de llevar a cabo la investigación. Y debería haber sido planificado en consonancia con las hipótesis de investigación que se derivan de una configuración particular de observaciones. Estas personas me han ofrecido la oportunidad de analizar para ellos bases de datos descomunales (que, por otro lado, harían las delicias de muchas personas de ciencia) pero me he sentido incapaz de hacerlo (además de las reticencias metodológicas que me suscitaba, hubiese sido como encontrar una aguja en un pajar). Aunque sus archivos de datos pudiesen ser deliciosos desde el punto de vista estadístico por el número de casos y de variables, lo cierto es que, como bien decía aquel anuncio publicitario de neumáticos, «la potencia sin control no sirve de nada» y los datos sin hipótesis son poco más que «agua de borrajas».



Por ello, sugiero fervientemente que se ponga un especial cuidado e interés en las fases de planificación de la investigación dado que todo el esquema científico de nuestro estudio dependerá de él. En la medida en que planificamos con la más estricta seriedad, mejor se desarrollará nuestro proceso de investigación (independientemente de si confirmamos o no nuestras hipótesis de trabajo). Es más, ahorraremos trabajo dado que no tendremos, o al menos con una probabilidad más baja, que volver sobre nuestros pasos para rehacer algo que no planificamos correctamente.

□ EJERCICIOS □

1. Reflexiona y pon ejemplos de escalas nominales, ordinales, de intervalo y de razón. Justifica tus respuestas.
2. Considera los valores de la siguiente variable de tipo nominal:
1, 2, 4, 5, 1, 2, 3, 4, 1, 2, 5, 3, 4, 1
¿cómo transformarías la variable para que la medida no se viese afectada?
3. Considera los valores de la siguiente variable de tipo ordinal:
12, 32, 12, 45, 12, 32, 12, 32, 45, 55, 21, 3, 32, 1, 3
¿cómo transformarías la variable para que la medida no se viese afectada?
4. Considera los valores de la siguiente variable medida en una escala de intervalo:
78, 95, 32, 14, 56, 47, 6, 3, 66, 23, 37, 85, 96, 41, 25
¿cómo transformarías la variable para que la medida no se viese afectada?
5. Considera los valores de la siguiente variable medida en una escala de razón:
7, 23, 4, 85, 0, 6, 45, 22, 87, 32, 325, 6, 88, 22, 47
¿cómo transformarías la variable para que la medida no se viese afectada?

4

Estadísticos descriptivos

En este capítulo vamos a entrar de lleno a analizar datos con  y . En concreto, vamos a abordar una parte crucial de la estadística: la estadística descriptiva. La estadística *descriptiva* pretende dar, como su nombre indica, una descripción de los datos contenidos en una muestra, mientras que la estadística *inferencial* pretende generalizar los resultados encontrados en una muestra a la población general de donde se tomó la muestra. Este capítulo está organizado en tres secciones destinadas a abordar genéricamente los análisis descriptivos más importantes desde el punto de vista de la tendencia central, de la dispersión y de la forma de variables individuales.

Para avanzar en este capítulo utilizaremos la base de datos desarrollada artificialmente y llamada `Econeg.RData` que acompaña a este libro. Se trata de un conjunto de datos procedente de una investigación destinada a conocer la relación que existía entre la creación de empresas y los valores ecológicos. El archivo tiene nueve variables y 201 casos. Esta es la descripción de las variables:

- *id*: es una variable destinada a identificar inequívocamente a cada caso,
- *sexo*: indica si la persona es mujer o hombre,

Capítulo 4 - Estadísticos descriptivos

- *estudios*: representa el tipo de estudios que estaba cursando la persona cuando participó en el estudio (Psi = Psicología, Emp = Empresariales, Inf = Informática),
- *emprende*: representa las respuestas (Sí o No) a la pregunta ¿consideras deseable crear una empresa propia al finalizar tus estudios universitarios?,
- *deseaEBT*: corresponde a la pregunta ¿consideras deseable crear una empresa de base tecnológica al terminar tus estudios universitarios?
- *curso*: se refiere al curso que estaban cursando los participantes en el momento de ser encuestados,
- *eco* y *antropo* son puntuaciones porcentuales de actitud ecocéntrica y antropocéntrica, respectivamente, derivadas de la escala construida por Thompson y Barton (1994).

4.1. Estadísticos de tendencia central

Los estadísticos univariados (para una única variable) de tendencia central¹ tienen como objetivo describir cómo es la variable de modo general. Es decir, pretenden resumir la variable en un solo índice o valor.

Una forma rápida y fácil de obtener un análisis preliminar de la base de datos que contenga un conjunto de estadísticos descriptivos para cada una de las variables sería utilizar la función `summary()`. Para que esta función actúe correctamente se ha de escribir, como argumento, el nombre del *data.frame* o de la base de datos que tenemos abierta con `Rcmdr`. Si ejecutamos el comando:

```
1 summary(Econeg)
```

obtendremos un conjunto básico de estadísticos descriptivos para cada una de las variables contenidas en la base de datos que tendrá un aspecto similar a este:

```
1
2      id          sexo      edad      estudios  emprende  deseaETB  curso
3 Min.   : 62.0  Hombre: 70   Min.   :18.00  Emp:66   No: 55   No:120   1º:55
4 1st Qu.:135.0  Mujer :131   1st Qu.:20.00  Inf:39   Sí:146  Sí: 81   2º:34
```

¹También denominados como estadísticos de posición.

4.1 - Estadísticos de tendencia central

5	Median	:222.0	Median	:22.00	Psi:96	3º:51
6	Mean	:228.4	Mean	:23.35		4º:10
7	3rd Qu.	:318.0	3rd Qu.	:24.00		5º:51
8	Max.	:442.0	Max.	:50.00		
9	eco		antropo			
10	Min.	: 0.0	Min.	: 0.00		
11	1st Qu.	: 0.0	1st Qu.	: 0.00		
12	Median	: 40.0	Median	: 10.00		
13	Mean	: 37.6	Mean	: 18.33		
14	3rd Qu.	: 50.0	3rd Qu.	: 30.00		
15	Max.	:100.0	Max.	:100.00		

Hay que decir que si queremos ejecutar este análisis desde la interface gráfica de **Rcmdr** tendremos que ejecutar la siguiente ruta: *Estadísticos* → *Resúmenes* → *Conjunto de datos activo*. No obstante, si queremos, podemos obtener estos estadísticos descriptivos para una sola variable del archivo de datos. Para ello tendremos que utilizar la notación `$`. Por consiguiente, tendremos que, si por ejemplo queremos obtener los descriptivos para la variable *edad*, escribir y ejecutar el siguiente comando:

```
1 summary(Econeg$edad) #Uso de la notación $
```

cuyo resultado será:

```
1   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2 18.00  20.00   22.00   23.35  24.00   50.00
```

Como se puede observar, la función nos proporciona el mínimo (**Min.**), el primer cuartil (**1st Qu.**), la mediana o segundo cuartil (**Median**), la media aritmética (**Mean**), el tercer cuartil (**3rd Qu.**) y el máximo (**Max.**) para cada variable numérica; mientras que para las variables cualitativas la función muestra la frecuencia de cada categoría. En este punto creo conveniente introducir las funciones `attach()` y `detach()` para poder analizar variables individualmente. Para hacer que las variables individuales sean «visibles» directamente a las funciones podemos utilizar la función `attach()` cuyo argumento será el nombre del conjunto de datos. Si ejecutamos el comando

```
1 attach(Econeg)
```

Capítulo 4 - Estadísticos descriptivos

podríamos utilizar la función `summary()` directamente sobre las variables individuales del conjunto de datos. Por ejemplo, si queremos obtener un análisis de la variable `edad` ahora tendremos que escribir:

```
1 summary(edad)
```

Para volver al estado inicial del archivo tendremos que utilizar la función `detach()` de manera análoga a como hemos utilizado la función `attach()`.

Para obtener un mayor conjunto de estadísticos descriptivos de posición podemos acceder al menú *Estadísticos* → *Resúmenes* → *Resúmenes numéricos...* de **R**. En el cuadro de diálogo que aparece (Figura 4.1) tendremos que seleccionar, en primer lugar, una o varias variables numéricas (no aparecen las cualitativas). Por defecto nos aparece marcada la media² y podemos solicitar que se nos calculen tantos cuantiles como nos apetezca. Habrá que indicarlo en el cuadro de texto «*cuantiles:*» y se nos facilitarán siempre y cuando la casilla de verificación «*Cuantiles*» esté activada. Para especificar los cuantiles tenemos que separarlos por comas y escribirlos utilizando el formato «*.xx*», donde *xx* se refiere a un número comprendido entre 00 y 99. Por ejemplo, si queremos obtener el percentil 30 (P_{30}) y el decil octavo (D_8 que equivale al percentil 80) tendremos que escribir «.3, .8» en el cuadro de texto. También podemos obtener cualquier cuantil utilizando la función `quantile()`. Como argumentos habría que especificar, en primer lugar, la variable sobre la que queremos realizar el análisis y, en segundo lugar separado por coma, los valores de los cuantiles que queremos estimar. Por ejemplo, para estimar los percentiles 40 y 79 de la variable `edad` tendríamos que escribir:

```
1 quantile(edad, probs = c(.4, .79))
```

Un par de funciones sencillas de recordar en este contexto son la función `mean()` y `median()` que estiman la media y la mediana de una variable respectivamente. Como argumento de estas funciones hay que indicar la variable de interés. Hay que destacar que la función `mean()` puede contener un argumento (`trim`) que permita obtener medias recortadas. Estimar la media recortada consiste en calcular la media tras eliminar una proporción de los valores más altos y

²Por el momento podemos desmarcar la desviación típica ya que será objeto de discusión de la siguiente sección.

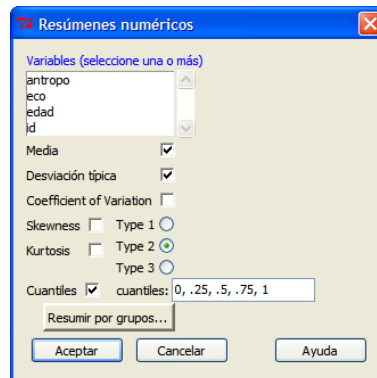



Figura 4.1: Resúmenes numéricos en Rcmdr.



más bajos de una variable. Esta forma de calcular la media es muy útil cuando nuestras variables contienen datos muy extremos o alejados de la gran masa de datos. El argumento `trim` está referido a la proporción de datos bajos y altos que se eliminan antes de calcular la media aritmética y puede tomar valores comprendidos entre 0 y 0.5. Un ejemplo para estimar la media en la variable *edad* eliminando un 2,5% de los datos más altos y más bajos de la variable sería el siguiente:

```
1 mean(edad, trim=0.025)
```

4.2. Estadísticos de dispersión

Cuando hablamos de dispersión estadística nos estamos refiriendo al grado en que los valores de una variable se concentran, o no, alrededor de un valor de la variable. De manera genérica, podríamos decir que la dispersión en una variable estadística tiene que ver con su homogeneidad, con el grado en que los valores de la variable están concentrados en una parte de la misma. Los estadísticos de dispersión más conocidos y comúnmente utilizados son la varianza y la desviación típica. Sin embargo, antes de tratar estos estadísticos, dedicaré algunas líneas a presentar otros parámetros diseñados para evaluar la dispersión en variables numéricas que, aunque son menos usados, suelen ser convenientes en determinadas situaciones reales o aplicadas. Para cada estadístico proporcionaré su ecuación matemática y el código fuente que permite estimar el valor del parámetro en . En el código

Capítulo 4 - Estadísticos descriptivos


fuente de cada función aparecerá, precedido del símbolo #, el modo en que se debe usar la función. Aunque lo volveré a explicar más abajo con un ejemplo concreto se tendría que copiar el código fuente en el editor de instrucciones de  o  y ejecutarlo³. Seguidamente se puede invocar a la función del modo en que se describe en el código fuente⁴.

Uno de los estadísticos de dispersión más sencillos que se pueden calcular es la amplitud o rango (Rg). La amplitud de una variable es el resultado de restar el menor valor de la variable al mayor (ecuación 4.1).

$$Rg = \max_x - \min_x \quad (4.1)$$


Un mayor valor de amplitud indicará mayor dispersión en los datos mientras que un valor más pequeño indicará menor dispersión en los datos. Una función que puedes utilizar para calcular el rango de una variable es la siguiente:


```
1 # Amplitud o Rango
2 # Uso: rango(variable)
3 rango <- function(x) max(x) - min(x)
```

Pega y ejecuta esta función en tu editor de instrucciones de . Seguidamente podrás utilizarla especificando como argumento alguna variable de tu conjunto de datos. Por ejemplo, para estimar el rango de la variable edad habría que escribir:

```
1 rango(edad)
```

La mediana de las desviaciones absolutas respecto de la mediana (MAD) es un estadístico muy estable dado que se basa en el cálculo de la mediana de las desviaciones relativas de cada valor sobre la mediana de la variable. Su ecuación es

³Recuerda que para ejecutar bloques de código en el editor de instrucciones de  hay que seleccionarlo previamente.

⁴Tengo que destacar que las ecuaciones que presento podrían haberse escrito de una manera más sencilla y eficiente. No obstante, pido disculpas a los programadores más experimentados pero creo que las ecuaciones escritas de este modo pueden ayudar a los estudiantes a familiarizarse con los procesos de programación en .

$$MAD = Md|x_i - Md_x|, \quad (4.2)$$

donde Md se refiere a la mediana. La función que sirve para calcular el MAD es la siguiente:

```
1 # Mediana de las Desviaciones Absolutas respecto de la Mediana
2 # Uso: mad(variable)
3 mad <- function(x) median(abs(x-median(x)))
```

La amplitud intercuartílica (A_Q) es un parámetro que representa la diferencia entre el tercer y el segundo cuartil de una variable. Si se divide entre dos se obtiene lo que se denomina como *amplitud semi-intercuartílica*. Su ecuación se concreta como

$$A_Q = P_{75} - P_{25}, \quad (4.3)$$

donde P_{75} se refiere al valor del percentil 75 o tercer cuartil (Q_3) y P_{25} está referido al percentil 25 o primer cuartil (Q_1). La función que he escrito para \mathbb{R} sería la siguiente:

```
1 # Amplitud Intercuartilica
2 # Uso: aq(variable)
3 aq <- function(x) {
4   p75 <- quantile(x, probs=0.75, names=FALSE)
5   p25 <- quantile(x, probs=0.25, names=FALSE)
6   p75-p25
7 }
```

El coeficiente de variación cuartílico (CV_Q) es un parámetro que representa el cociente entre una medida de dispersión (la amplitud semi-intercuartílica) y una medida de tendencia central (el promedio de cuartiles). Su ecuación (4.4) y el código fuente de la función para su utilización en \mathbb{R} se presentan a continuación:

$$CV_Q = \frac{P_{75} - P_{25}}{P_{75} + P_{25}}, \quad (4.4)$$

Capítulo 4 - Estadísticos descriptivos

```
1 # Coeficiente de Variación Cuartílico
2 # Uso: cvc(variable)
3 cvc <- function(x) {
4   p75 <- quantile(x, probs=0.75, names=FALSE)
5   p25 <- quantile(x, probs=0.25, names=FALSE)
6   n <- sum(p75, -p25)
7   d <- sum(p75, p25)
8   n/d
9 }
```

En contraposición a los estadísticos de dispersión que hemos visto hasta el momento, el cálculo de la desviación típica se puede realizar directamente desde la interface que tenemos cargada de **R**. Para ello tenemos que acceder al cuadro de diálogo que aparece en la Figura 4.1 seleccionando la ruta *Estadísticos* → *Resúmenes* → *Resúmenes numéricos...* del menú. Además de estimar la desviación típica de cualquier variable numérica también podemos estimar el coeficiente de variación. La desviación típica también se puede obtener directamente utilizando la función `sd()`.

La desviación típica (s_x) es la raíz cuadrada de la varianza ($s_x = \sqrt{S_x^2}$) mientras que la varianza (S_x^2)⁵ es el cuasi-promedio de las desviaciones cuadráticas de cada valor de la variable respecto de la media. Matemáticamente, la ecuación de la varianza quedaría expresada como

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}. \quad (4.5)$$

Por último, dado que la varianza y la desviación típica son parámetros que se ven afectados por las unidades de medida de la variable sobre la que realizamos el análisis, el coeficiente de variación (CV_x) se utiliza como medida *adimensional* de dispersión. El coeficiente de variación resulta de dividir la desviación típica por el valor absoluto de la media en una variable. Esto es,

$$CV_x = \frac{s_x}{|\bar{x}|}. \quad (4.6)$$

⁵Se puede obtener directamente utilizando la función `var()`.

Dado que la media esta en valor absoluto el coeficiente de variación siempre será positivo. Este estadístico se puede utilizar para comparar las dispersiones que presentan varias variables que han sido medidas en unidades de medida muy dispares como, por ejemplo, el número de hijos y los centímetros cúbicos del automóvil familiar.

4.3. Estadísticos de forma

El último conjunto de estadísticos que voy a presentar están referidos a los estadísticos de forma, en concreto a los estadísticos de curtosis o apuntamiento y a los de asimetría o sesgo.

Para estimar los estadísticos de curtosis y asimetría de una o más variables tenemos que acceder al cuadro de diálogo que venimos utilizando en este capítulo (*Estadísticos* → *Resúmenes* → *Resúmenes numéricos...*) y marcar las opciones de «asimetría» y «apuntamiento». Adicionalmente, podemos seleccionar tres tipos de índices de asimetría y apuntamiento. En cualquier caso, en el manual de Solanas, Salafranca, Fauquet, y Núñez (2005) puedes encontrar las ecuaciones que definen a estos y otros índices de asimetría y apuntamiento, la interpretación de estos índices es relativamente sencilla y suelen coincidir en sus estimaciones a nivel general.

La curtosis tiene que ver con la dispersión de la variable. De modo genérico podemos encontrarnos con tres tipos de variables caracterizadas por el valor que generan de curtosis: a) variables *mesocúrticas*, en las que el valor de curtosis es cero o un valor muy cercano a este valor; b) variables *leptocúrticas*, que generarían valores mayores que cero; y c) variables *platicúrticas*, que generarían valores inferiores a cero. El análisis gráfico de la curtosis se puede evaluar inspeccionando el histograma de la distribución de frecuencias de la variable. En este caso, se suele aludir a la comparación de la variable de interés con la distribución normal. Las variables mesocúrticas tendrían un perfil de frecuencias semejante a la distribución normal mientras que las variables leptocúrticas y platicúrticas serían más apuntadas o más aplanadas, respectivamente, cuando se comparan con una distribución normal.

Por su parte, la asimetría se refiere al grado con que los valores se acercan o alejan del valor intermedio (respecto de su rango) de la variable. De nuevo

Capítulo 4 - Estadísticos descriptivos

tenemos tres casos: a) que la variable sea simétrica, lo que indica que la distribución de frecuencias se reparte equitativamente entre un lado y otro del centro de gravedad de la distribución en cuyo caso el valor del estadístico es cero; b) que la distribución sea asimétrica positiva, que se produce cuando el valor de asimetría es positivo y que gráficamente produce histogramas con más valores a la izquierda del centro de gravedad de la variable; y c) que la variable sea asimétrica negativa, en cuyo caso habrá más valores a la derecha del centro de la variable y se obtendrá un valor de asimetría inferior a cero.

□ EJERCICIOS □

1. Calcula los siguientes percentiles 23, 36, 48, 76 y 92 de las variables edad, eco y antropo de la base de datos Econeg.
2. Calcula las medias recortadas al 90 % de las variables edad, eco y antropo de la base de datos Econeg.
3. Calcula la media de la edad para cada uno de los grupos definidos por los tipos de estudios que hay en la muestra.
4. Calcula la amplitud, la mediana de las desviaciones absolutas respecto de la mediana, la amplitud intercuartílica, el coeficiente de variación cuartílico, la varianza, la desviación típica y el coeficiente de variación de las variables edad, eco y antropo.
5. Crea una función que estime la amplitud semi-intercuartílica.
6. Crea una función para estimar la varianza de una variable usando la ecuación que aparece más arriba.
7. Indica que tipo de distribuciones tienen las variables edad, eco y antropo en función de su asimetría y curtosis.

5

Transformación de datos

Este capítulo está dedicado a mostrar cómo se pueden modificar variables que contenga nuestro conjunto de datos para adaptarlas a los análisis en los que estamos interesados. Entre otras cosas, se verá cómo se pueden generar puntuaciones de escala a partir de un conjunto de variables utilizando diferentes métodos, cómo se pueden recodificar variables atendiendo a ciertas condiciones y cómo se pueden modificar los archivos de datos según nos convenga.

Para este capítulo vamos a utilizar una base de datos llamada `escala.RData` que acompaña a este manual. Únicamente contiene once variables y 85 casos. La primera variable (`id`) es una variable, como suele ser habitual, de identificación que localiza inequívocamente a cada participante. Las otras diez variables (`in`) recogen las respuestas de cada participante a diez ítems destinados a evaluar creatividad. Cada una de estas diez variables tienen un máximo de 5 y un mínimo de 1; donde 5 es una etiqueta numérica que representa que el participante estaba «muy de acuerdo» con lo que expresaba el ítem, el 4 representa que el participante estaba «de acuerdo» con lo que indicaba el ítem, el 3 indica que el participante no estaba «ni de acuerdo ni en desacuerdo» con el ítem, el 2 representa estar «en

desacuerdo», mientras que el 1 se refiere a un estado de «total desacuerdo» con lo expresado en la declaración.

5.1. Puntuaciones de escala

En muchas situaciones aplicadas tenemos la necesidad de medir un constructo psicológico como la emoción, la personalidad, el optimismo, la satisfacción o la motivación. Lo que ha venido haciendo la psicología en los últimos tiempos ha sido desarrollar tests que permitan estimar la cantidad de este tipo de constructos que tienen las personas. Un *constructo* es, ni más ni menos, una construcción verbal referida a un conjunto de observaciones que tienen cierta consistencia. Por ejemplo, el constructo *liderazgo* sólo tiene sentido cuando resume ciertas observaciones que la sociedad entiende como definitorias de un fenómeno abstracto o latente y que sirve para explicar la razón por la cual algunas personas tienen unas características especiales en su contexto social. No vamos a entrar aquí en discusiones o polémicas relacionadas con las definiciones de constructos, bien semánticas o sintácticas, ni con su estatus científico. Únicamente nos limitaremos a adquirir una serie de competencias básicas relacionadas con la estimación de puntuaciones de test al amparo de lo que es ampliamente aceptado por la comunidad científica.

Cuando se desarrolla, o se utiliza, un test o una escala para medir un constructo psicológico tenemos que obtener la puntuación del constructo o de la escala utilizando alguna *función de corrección*. Los elementos que formarán parte de esta función de corrección serán los ítems de la escala. Un *ítem* es cada una de las partes de información que contiene un test y que recogen información sobre el constructo que nos interesa. Por ejemplo, en una escala sobre actitudes hacia el medio ambiente, cada una de las declaraciones sobre las que tendremos que emitir un juicio (por ejemplo, el grado con que estamos de acuerdo con la declaración) serán ítems.

El método más común y extendido que se utiliza para estimar la puntuación total (X_T) de una escala consiste en sumar las puntuaciones parciales de cada ítem (x_i) en una nueva variable. Formalmente podríamos expresar esta idea con la ecuación

$$X_T = \sum_{i=1}^n x_i. \quad (5.1)$$

Para realizar esta operación sobre el archivo `escala` que utilizaremos en este capítulo tendríamos que escribir:

```
1 escala$Xt <- with(escala, i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10)
2 # Observa cómo se ha utilizado la notación $
3 IGU: i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10
```

Este y otros cálculos que vamos a ejecutar de aquí en adelante sobre las variables se pueden conseguir reduciendo la cantidad de código que tenemos que escribir accediendo al menú *Datos* → *Modificar variables del conjunto de datos activo* → *Calcular una nueva variable...* En el cuadro de diálogo que aparece (Figura 5.1) tenemos, en la parte superior izquierda, un cuadro que contiene las variables numéricas del conjunto de datos. A su vez, también hay un cuadro de texto llamado «Nombre de la nueva variable» en el que tendremos que especificar cómo llamaremos a la variable que vamos a crear y en el cuadro de texto «Expresión a calcular» tenemos que detallar la ecuación de cálculo que queremos aplicar sobre las variables. Por lo tanto, cuando utilice códigos en lo sucesivo especificaré el código estándar que tenemos que utilizar para ejecutar una función desde el editor de instrucciones y el código que habría que utilizar desde el cuadro de diálogo. El código que habría que utilizar desde el cuadro de diálogo irá marcado (como he hecho anteriormente) con la expresión `IGU`, como abreviatura de la expresión *Interfaz Gráfica de Usuario*.

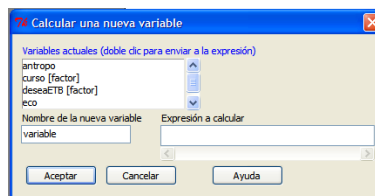


Figura 5.1: Calcular una nueva variable en Rcmdr.

Otra posibilidad que también se suele utilizar es calcular el promedio respecto a todos los ítems ($X_{\bar{T}}$). O sea, aplicar la ecuación

$$X_{\bar{T}} = \frac{\sum_{i=1}^n x_i}{n}, \quad (5.2)$$

donde n se refiere al número de ítems del test. En este caso, podríamos ejecutar el siguiente código:

```
1 escala$Xt_m <- with(escala, (i1+i2+i3+i4+i5+i6+i7+i8+i9+i10)/10)
2 # Observa cómo se ha utilizado la notación $
3 IGU: (i1+i2+i3+i4+i5+i6+i7+i8+i9+i10)/10
```

Realmente se podría utilizar cualquier otra función u operación matemática que especifique el modo de obtener la puntuación total de la escala. Para ello se podrían utilizar los operadores matemáticos o trigonométricos como los que se listan a continuación:

```
1 +      # Suma
2 -      # Resta
3 *      # Multiplicación
4 /      # División
5 abs()  # Valor absoluto
6 sin()  # Función que extrae el seno de un valor
7 asin() # Inverso de la función seno
8 cos()  # Función que extrae el coseno de un valor
9 acos() # Inverso de la función coseno
10 tan()  # Función tangente de un valor
11 atan() # Inversa de la función tangente
12 sqrt() # Raíz cuadrada de un valor
13 ^      # Eleva un valor a la potencia indicada
```

5.2. Recodificación de variables

En algunas ocasiones es necesario que transformemos las puntuaciones directas de un test (lo que hemos calculado en la sección anterior) a otro tipo de puntuaciones para realizar ciertos cálculos o para presentar las puntuaciones del test en cierto formato. Uno de los formatos de transformación más sencillo, y que se suele utilizar para realizar análisis multinivel (Field, 2009; Pardo, Ruiz, y San Martín, 2007), lo representan las puntuaciones diferenciales respecto de la media ($Diff_{\bar{x}}$):

$$Dif_{\bar{x}} = x_i - \bar{x}. \quad (5.3)$$

Lo que conseguimos con este tipo de transformación es centrar la variable en relación a la media aritmética. De este modo, la persona que tuviese una puntuación igual a la media obtendrá ahora una puntuación de 0. Las puntuaciones originales de la variable que estuviesen por encima de la media obtendrán valores positivos en la nueva variable y las que estaban por debajo de este parámetro tendrán ahora valores negativos. Para obtener nuestra puntuación diferencial en la puntuación total del test respecto de la media en el archivo que estamos utilizando tendríamos que ejecutar la siguiente sintaxis:

```

1 escala$dif.m <- with(escala, Xt - mean(Xt,na.rm=TRUE))
2 # Notar que se ha utilizado la notación $ para referirse a la nueva variable
   'dif.m' y que se ha utilizado el parámetro 'na.rm' para la función que
   estima la media dado que existen casos perdidos (NAs) en la variable 'Xt'
3 IGU: Xt - mean(Xt,na.rm=TRUE)

```

Por su parte, las puntuaciones típicas comparan la distancia de cada valor de la variable con la media en relación a la desviación típica. Este tipo de transformación es muy utilizada, como veremos más abajo, para generar puntuaciones del test que tengan unos parámetros concretos de tendencia central y de dispersión. Las puntuaciones típicas de una variable (z_i) tienen una desviación típica de 1 y una media 0 y se calculan aplicando la ecuación

$$z_i = \frac{x_i - \bar{x}}{s_x}. \quad (5.4)$$

Para calcular la puntuación típica de nuestra variable Xt tendríamos que escribir el siguiente código:

```

1 escala$z.i <- with(escala, (Xt - mean(Xt,na.rm=TRUE))/sd(Xt,na.rm=TRUE))
2 # Notar que se ha utilizado la notación $ y el parámetro 'na.rm' en las
   funciones de la media y de la desviación típica.
3 IGU: (Xt - mean(Xt,na.rm=TRUE))/sd(Xt,na.rm=TRUE)

```

Capítulo 5 - Transformación de datos

Adicionalmente, **Rcmdr** trae incorporada una función propia para estimar puntuaciones típicas. Si accedemos a la ruta de menú *Datos* → *Modificar variables del conjunto de datos activo* → *Tipificar variables...*, aparecerá un cuadro de diálogo (Figura 5.2) en el que podremos señalar las variables que queremos tipificar directamente.

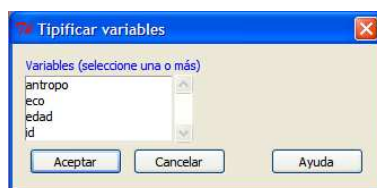


Figura 5.2: Tipificar variables en Rcmdr.

En algunas ocasiones las puntuaciones tipificadas se utilizan para calcular lo que se denominan como *escalas derivadas* (García, De la Fuente, y Martín, 1998). Un ejemplo de ello lo encontramos en los test de inteligencia donde, en algunos de ellos, se suele decir que la media de inteligencia es 100 con una desviación típica de 15. Lo que se ha hecho en este caso, tras haber estimado las puntuaciones del cociente de inteligencia en una muestra lo suficientemente grande y haber tipificado los valores observados, ha sido realizar una transformación lineal de la puntuación típica del conjunto de la muestra. De esta manera, se genera una distribución normalizada y tipificada con la media y desviación típica deseadas. Las escalas derivadas (T_x) se obtienen aplicando una ecuación lineal de la forma

$$T_x = a + z_{x_i} \times b, \quad (5.5)$$

donde a y b son constantes referidas a la nueva media y desviación típica respectivamente. Por ejemplo, supongamos que queremos obtener una puntuación derivada de nuestra variable tipificada que tuviese una media de tres y una desviación típica de cinco. Para ejecutar este cálculo con **Rcmdr** habría que utilizar el siguiente código:

```
1 escala$Tx <- with(escala, 3 + z.i * 5)
2 # Observa el uso de $
3 IGU: 3 + z.i * 5
```

5.2 - Recodificación de variables

Aparte de las transformaciones con sentido estadístico que hemos estado viendo hasta ahora, hay situaciones en las que se requiere que una variable concreta sea recodificada de una manera diferente a como ha sido recogida. Por ejemplo, en algunas estadísticas se nos suele decir que el porcentaje de personas que bebe alcohol masivamente los fines de semana con una edad comprendida entre los 18 y los 24 años es del 25% mientras que el porcentaje de personas que realiza la misma actividad con edades comprendidas entre los 25 y los 35 años es del 13%. El caso es que aunque en la investigación se recogiese la edad en términos exactos (cosa que recomiendo encarecidamente) lo que se ha hecho ha sido recodificar la variable original en diferentes intervalos que pueden ser interesantes desde el punto de vista de la investigación. Consideremos que, pensando en nuestro archivo de datos sobre creatividad, en investigaciones anteriores se identificó que la puntuación de la escala puede ser clasificada en tres grupos: personas con bajos niveles de creatividad (puntuación entre 10 y 31), personas con niveles intermedios de creatividad (personas con puntuación entre 31 y 34) y personas con altos niveles de creatividad (desde 34 a 50). Para obtener una nueva variable que implicase tales recodificaciones tendríamos que ejecutar el siguiente código:

```
1 escala$gru.3 <- recode(escala$Xt,  
2 '10:31 = "Baja";  
3 31:34 = "Media";  
4 34:50 = "Alta"',  
5 as.factor.result=TRUE)
```

Con este código se genera una nueva variable llamada *gru.3* a partir de la variable *Xt* con las características que se especifican entre las líneas de 2 a 4. En la línea 5 la sintaxis acaba activando un parámetro (`as.factor.result`) que transforma la nueva variable en una variable cualitativa o factor. Otro detalle que destacar del código anterior, y de \mathbb{R} en general, es que las etiquetas de la nueva variable tipo factor (Baja, Media y Alta) aparecen entre comillas. El código presentado anteriormente ha sido generado por una opción que proporciona la interfaz de \mathbb{R} . Para acceder a ella hay que ejecutar la ruta *Datos* \rightarrow *Modificar variables del conjunto de datos activo* \rightarrow *Recodificar variables...* del menú. Nos aparecerá un cuadro de diálogo (Figura 5.3) donde tendremos, en primer lugar, que especificar la variable o variables que queremos recodificar. Más abajo tenemos un cuadro de texto destinado a que se especifique el nombre de la

Capítulo 5 - Transformación de datos

nueva variable recodificada. También tenemos la opción de hacer que la variable creada sea un factor. Por último, tenemos que introducir las características de la recodificación que queremos ejecutar en el cuadro de texto «Introducir directrices de recodificación». Por ejemplo, en el caso del ejemplo que hemos presentado previamente se ha introducido el siguiente código:

```
1 10:31 = "Baja"  
2 31:34 = "Media"  
3 34:50 = "Alta"
```

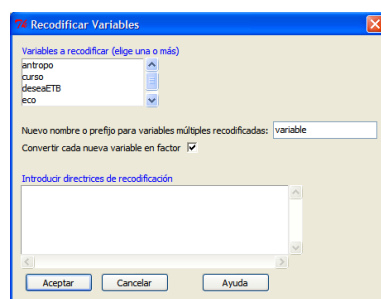


Figura 5.3: Recodificar variables en Rcmdr.

5.3. Modificación del conjunto de datos

Para terminar con este capítulo, dedicaremos algunas líneas a familiarizarnos con comandos y procedimientos que sirven para modificar nuestros archivos de datos de tal manera que los haga más prácticos o útiles.

En algún momento tendremos que eliminar alguna variable o variables que ya no necesitamos, bien porque nos hemos equivocado en alguna transformación o recodificación o porque algún sistema de recogida automática nos la ha incluido por defecto pese a que no es relevante para nuestra investigación. Para eliminar variables de nuestro conjunto de datos tenemos que hacer una especie de asignación en la que el objeto de la misma es la variable que queremos eliminar y que consiste en asignar el valor NULL. Por ejemplo, si quisiésemos eliminar la variable tipificada (calculada manualmente) y la puntuación diferencial que hemos creado nosotros mismos previamente tendríamos que ejecutar el siguiente código:

5.3 - Modificación del conjunto de datos

```
1 escala$z.i <- NULL
2 escala$dif.m <- NULL
```

Para ejecutar estos cambios en nuestro archivo también podemos hacerlo a través de la interfaz de **Rcmdr**. Para ello, tendremos que acceder al menú *Datos* → *Modificar variables del conjunto de datos activo* → *Eliminar variables del conjunto de datos...* y nos aparecerá el cuadro de diálogo que aparece en la Figura 5.4. Lo único que tenemos que hacer es seleccionar la variable o variables que queremos eliminar y pulsar en el botón «Aceptar».

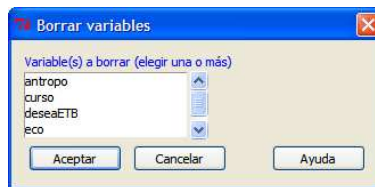


Figura 5.4: Eliminar variables en Rcmdr.

También se pueden eliminar casos que no nos interese conservar o eliminar aquellos casos que contengan datos perdidos. Para ello, podemos acceder a los menús *Datos* → *Conjunto de datos activo* → *Borrar fila(s) del conjunto de datos activo...* o a *Datos* → *Conjunto de datos activo* → *Eliminar casos con valores omitidos*. No obstante, en el caso de eliminar registros con datos perdidos u omitidos recomiendo hacer, previamente, un estudio multivariante de los datos perdidos con el fin de proceder a su imputación o a su eliminación dependiendo de las conclusiones a las que lleguemos (Hair, Anderson, Tatham, y Black, 1998).

□ EJERCICIOS □

1. Imagina que hay que aplicar la siguiente ecuación para obtener la puntuación total del test de creatividad (P_C) que hemos utilizado en este capítulo:

$$P_c = \frac{\sum_{i=1}^n x_i}{\sqrt{\bar{X}_T}}.$$

Escribe el código necesario para obtener la puntuación del test.

2. Elimina la variable que has creado en el ejercicio anterior y obtén otra que sea el cuadrado del promedio de los ítems del test.

6

Inferencia sobre medias

En este capítulo vamos a tratar un tipo de análisis estadístico que está muy extendido en el campo de la psicología científica y en el de la educación. Hasta ahora nos hemos dedicado a calcular estadísticos sin intención de generalizarlos a la población de la que hemos extraído la muestra. Nos hemos limitado a describir las variables que contenían nuestros conjuntos de datos. A partir de ahora, vamos entrar en un campo de la estadística llamado *estadística inferencial* que pretende dar un paso más allá de lo observado en la muestra. En vez de limitarse a describir lo que pasa en las variables de nuestros archivos, como hace la *estadística descriptiva*, cuando hacemos estadística inferencial tratamos de estimar el grado en que lo que hemos observado en una muestra lo podemos generalizar a la población de la que se extrajo el conjunto de datos.

Para enfrentarnos a este capítulo vamos a utilizar una base de datos llamada `Perros.RData` que acompaña a este libro. El archivo contiene 8 variables y 300 registros. Son de una investigación que se preocupó por estudiar si el hecho de tener perro (*perro*) o vivir en una casa de ciertas proporciones (*tipo.casa*) tenía influencia en el promedio de días de catarro que experimentaba una persona a

lo largo del invierno (*d.cata*), en el número de problemas intestinales evaluados que una persona sufría durante un año (*intestinal*) o sobre el número de veces que una persona padecía de problemas de urticaria en un periodo de seis meses (*d.urticaria*). Todas estas variables fueron promediadas tomando el mes como unidad de muestreo tras haberse observado al conjunto de personas de la muestra durante tres años. Adicionalmente, el equipo de investigación desarrolló dos vacunas: una para prevenir los problemas catarrales a lo largo del invierno y otra para evitar afecciones intestinales durante el verano. La base de datos también recoge información sobre el efecto de las vacunas contra esos dos tipos de enfermedades al año siguiente a su inoculación en las variables *d.cata.V* e *intestinal.V* que fueron administradas tras los tres primeros años de estudio observacional. Como suele ser habitual, y recomendable, la variable *id* es un código que identifica inequívocamente a cada participante del estudio.

6.1. El contraste de hipótesis

Antes de adentrarnos en el análisis estadístico inferencial conviene dedicar unos minutos a afianzar y reflexionar sobre lo que vamos a hacer. No es este el lugar apropiado para hacer una exposición detallada del proceso de inferencia estadística o del procedimiento de contraste de hipótesis. Para ello puedes encontrar muy buenos manuales que te ayudarán como el de Pagano (1998/1999) o el de León y Montero (2003). No obstante, sí que creo interesante incidir en ciertos elementos que nos encontraremos antes y después del análisis de datos y que serán de vital importancia para entender lo que estamos haciendo.

Siguiendo las directrices sugeridas por el método científico, cada análisis estadístico inferencial implica un contraste de hipótesis. Como sabrás, una *hipótesis* no es nada más que una proposición no corroborada, pero sobre la que tenemos sospechas razonables de que sea cierta. Normalmente, o deseablemente, las hipótesis científicas no surgen de la nada, «como por arte de magia». Independientemente de que, como yo creo, la ciencia es una especie de arte, las hipótesis científicas surgen en la mente de las personas de ciencia tras muchos años de estudio y de observación informal o sistemática. Bueno, también es cierto que hay personas de ciencia brillantísimas que llegan a desarrollar hipótesis científicas de un día para otro y que son capaces de generar conocimiento como si fuesen caudales ingentes

de creatividad que rayan en lo fantástico. En cualquier caso, la hipótesis se ha convertido hoy en día en el «caballo de batalla» de la ciencia que nos ayuda a entender y a controlar la naturaleza.

Cuando realizamos análisis estadísticos inferenciales solemos establecer un *contraste de hipótesis* que contiene dos tipos básicos de hipótesis que denominamos hipótesis nula e hipótesis alternativa. La **hipótesis nula** o H_0 siempre suele estar expresada en términos de igualdad (esto es, utilizando el símbolo $=$), mientras que la **hipótesis alternativa** o H_1 está expresada en términos diferenciales (utilizando el símbolo \neq) o direccionales (utilizando los símbolos $<$ o $>$). Como verás a continuación, una regla mnemotécnica para recordar el significado de la hipótesis nula consiste en asumir que lo que especifica esta hipótesis «nulifica», por ejemplo, el efecto de cualquier tratamiento experimental. Por su parte, el hecho de que la hipótesis alternativa sea diferencial o direccional depende de nuestras hipótesis de investigación y, como verás, tiene importantes repercusiones desde el punto de vista de la toma de decisiones que implica el proceso de contraste de hipótesis.

Lo que estamos haciendo siempre que llevamos a cabo inferencias estadísticas es tomar decisiones sobre la hipótesis nula. Es decir, siempre estamos evaluando la verosimilitud de ésta hipótesis de igualdad. La decisión de mantener, o no, la hipótesis nula depende de un estadístico de contraste observado que calcularemos sobre los datos de la muestra y que compararemos con otro estadístico de contraste teórico. Con base en esta comparación podremos estimar una probabilidad que denominamos *p*-valor¹ y que se refiere a

$$p(\text{Rechazar } H_0 | H_0 \text{ es cierta}), \quad (6.1)$$

o lo que es lo mismo, este valor indica la probabilidad de rechazar la hipótesis nula cuando en realidad es cierta. Sin enredarme mucho en las palabras y para ser lo más funcional posible, podríamos decir que cuanto más pequeño sea este *p*-valor menor serán las posibilidades de que nos equivoquemos al rechazar una hipótesis nula correcta. Muchas veces se suele decir, al hilo de este asunto, que existe una regla de decisión por la cual debemos discernir sobre si rechazar o no

¹También llamado en otros contextos como *nivel de significación* o incluso α , aunque éste último tiene otras connotaciones.

Capítulo 6 - Inferencia sobre medias

la hipótesis nula que viene a decir que «si p es inferior o igual a 0,05, entonces rechazamos H_0 ». Y se dice esto porque si tenemos un valor tan pequeño de p el riesgo de tomar una decisión acertada es muy grande, del orden del 95 % o mayor ($[1 - p] \times 100$).

Lo que pretendo hacer con este incompleto y telegráfico resumen del proceso de contraste de hipótesis es llamar tu atención sobre tres elementos que considero cruciales para manejarse con el análisis de datos cuando utilizamos programas informáticos: la especificación del contraste de hipótesis, el estadístico de contraste implicado en la toma de decisiones y el valor de significación del test de hipótesis. En lo que sigue a continuación trataré de hacer explícito el contraste de hipótesis que se está testando en cada caso, aunque espero que vayas aprendiendo a deducir los contenidos de las dos hipótesis dependiendo del contexto de análisis en el que estemos.

Antes de entrar en materia me gustaría aclarar otro aspecto interesante de algunas pruebas o tests estadísticos referido a los supuestos que subyacen en cada uno de ellos. El caso es que hay algunas pruebas estadísticas que requieren que se cumplan ciertas condiciones para que sean válidas. Esto es, algunos de los tests estadísticos tienen que conformarse a unas normas para que las inferencias que extraigamos de ellos sean útiles. A estos tests o pruebas se les denomina *paramétricas*. Por ejemplo, para que podamos aplicar con éxito el test t de Student para grupos independientes, una de las variables (la variable dependiente, desde el punto de vista del diseño) tiene que haber sido medida en una escala, como mínimo, de intervalo y que la distribución muestral de la diferencia entre las medias de los grupos ha de distribuirse siguiendo una distribución normal (para una explicación didáctica de la idea de distribución muestral puede consultarse a Field (2009)). Por su parte, aquellas técnicas estadísticas que implican la presencia de unas condiciones más relajadas en nuestros datos se les llaman pruebas *no paramétricas*. En lo que sigue a continuación se presentará una técnica paramétrica y otra no paramétrica paralela que servirá para alcanzar el mismo objetivo analítico.

6.2. Contraste para una media

La situación más sencilla que nos podemos encontrar cuando hablamos sobre contraste de medias es aquella en la que se nos presenta una variable y estamos interesados en saber si la media de esa variable es estadísticamente diferente de un valor dado. Por ejemplo, el promedio de infecciones intestinales producidas en verano en la muestra del archivo de este capítulo es de 3,1 y, por estudios previos, sabemos que el promedio poblacional de infecciones gastrointestinales en nuestro país es de 3. Por tanto, podríamos preguntarnos si la media de nuestra población, estimada usando la muestra de la que disponemos, difiere estadísticamente del valor 3 como apuntan los estudios previos. En términos formales podríamos decir que el contraste de hipótesis quedaría expresado así

$$H_0 : \mu = 3$$

$$H_1 : \mu \neq 3$$

donde μ se refiere a la media poblacional.

Para ejecutar este contraste de hipótesis con **R** tendríamos que utilizar el siguiente código²:

```
1 t.test(intestinal, alternative='two.sided', mu=3, conf.level=.95)
```

o, alternativamente, accediendo al menú *Estadísticos* → *Medias* → *Test t para una muestra...* Al utilizar la interfaz gráfica nos aparecerá un cuadro de diálogo como el que aparece en la Figura 6.1. En primer lugar nos aparece una lista con las variables que son susceptibles de ser elegidas para este tipo de análisis. Justo debajo de esa lista nos aparecen tres alternativas para definir la hipótesis alternativa: la de desigualdad (expresada en \mathbb{R} como \neq), la de menor que ($<$) y la de mayor que ($>$). En el cuadro de texto *Hipótesis nula: mu =* tendremos que escribir el valor del parámetro que estamos interesados en contrastar (3 en nuestro ejemplo). Este cuadro de diálogo también nos da la posibilidad de especificar un porcentaje para la estimación de un intervalo de confianza en el que, con una

²Siempre y cuando hayamos usado la función `attach()`. En caso contrario habrá que utilizar la notación `$` como ya sabes.

Capítulo 6 - Inferencia sobre medias

probabilidad igual a la indicada, se encontrará la media poblacional. Por defecto el programa estimará un intervalo de confianza para la media muestral del 95 %, o lo que es lo mismo, que calculará dos valores entre los cuales estará el parámetro poblacional con un grado de error del 5 %.

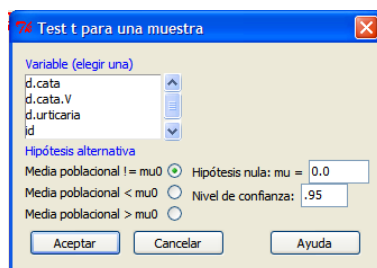


Figura 6.1: Prueba t para una muestra en Rcmdr.

La salida del contraste para una media que hemos especificado más arriba sería la siguiente:

```
1  One Sample t-test
2
3  data:  intestinal
4  t = 1.0202, df = 299, p-value = 0.3085
5  alternative hypothesis: true mean is not equal to 3
6  95 percent confidence interval:
7  2.900909 3.312424
8  sample estimates:
9  mean of x
10 3.106667
```

En el renglón 3 se nos indica la variable sobre la que hemos realizado el análisis por si hemos cometido algún error. En la línea 4 aparece, en primer lugar, el valor del estadístico de contraste t de Student, los grados de libertad (df^3) y el valor de p . En la línea 5 de la salida se nos informa del valor de la hipótesis alternativa mientras que en la séptima línea nos aparece el intervalo de confianza para la media poblacional. Finalmente, en la línea 10 la salida nos presenta la media muestral. Como habrás concluido ya, la media poblacional de la cual se extrajo la muestra no es estadísticamente diferente de 3 dado que el valor de p no es inferior a 0,05. ¿Qué pasaría si el valor de contraste para la media fuese 0?

³Del inglés *degrees of freedom*.

6.3. Contraste para dos medias

En numerosas ocasiones tenemos la necesidad de contrastar si existen diferencias en una variable (por ejemplo, cantidad de cigarrillos fumados en un día) entre dos grupos (uno que ha seguido un tratamiento terapéutico y otro que ha seguido otro diferente). En estas situaciones, dado que estamos hablando de dos grupos de personas diferentes, decimos que tenemos que realizar un *contraste de medias para grupos independientes*. En otras ocasiones necesitamos comprobar si un tratamiento experimental ha surtido efecto en las mismas personas tras haber registrado una línea base (por ejemplo, cuando tras evaluar el grado de hábito tabáquico pedimos a un conjunto de personas que siga un tratamiento concreto para superar su trastorno). Esta vez, dado que estamos comparando dos medias generadas por el mismo grupo de personas (la media de hábito antes del tratamiento y la media de consumo después de la implementación del tratamiento) decimos que nos enfrentamos a un *contraste de medias relacionadas*.

En esta sección se explica cómo hacer estos tipos de contraste utilizando R y se presentan tanto una alternativa paramétrica (la t de Student) como sus contrapartidas no paramétricas (con el test de Wilcoxon).

6.3.1. Medidas independientes

t de Student

Imaginemos que estamos interesados en saber si existen diferencias en el promedio de catarros que contraen las personas que tienen perros y las personas que no tienen. En primer lugar, lo más sensato sería estimar el promedio de catarros que afecta a uno y otro grupo de personas. Si utilizamos la función de resúmenes numéricos que vimos anteriormente segmentando por la variable `perro` obtendríamos que las personas que no tienen perro contraen una media de 1,86 catarros por invierno mientras que las personas que sí tienen lo hacen con una frecuencia promedio de 3,6.

Dado que la prueba t de Student es paramétrica lo primero que sería recomendable hacer es testar si los supuestos que subyacen a su utilización se cumplen. Un requisito clave para que la técnica se pueda aplicar con garantías es que la distribución muestral de la variable dependiente se distribuya normalmente. Aunque la normalidad de la distribución muestral se supone implícita para muestras

Capítulo 6 - Inferencia sobre medias

grandes, algunos autores sugieren que se teste la normalidad de los datos brutos (p. e., Field, 2009). Para testar la normalidad de la variable *d.cata* en nuestro archivo podemos utilizar un histograma (para inspeccionar gráficamente la distribución de frecuencias) y realizar un contraste de hipótesis sobre su normalidad. Aunque retomaremos el tema de los gráficos en un capítulo posterior, vamos a generar un histograma para inspeccionar visualmente el grado en que la variable se distribuye normalmente. Para ello, podemos, como viene siendo habitual, utilizar la ventana de instrucciones y ejecutar una sintaxis. En este caso tendríamos que escribir y ejecutar:

```
1 Hist(d.cata, scale="frequency", breaks="Sturges", col="darkgray")
```

o, alternativamente, acceder al menú *Gráficas* → *Histograma...* En el cuadro de diálogo que aparece (Figura 6.2) tendremos que elegir la variable *d.cata* y dejar todas las opciones que nos aparecen por defecto (más tarde volveremos con este asunto y veremos cómo podemos modificar las propiedades del histograma cuando tratemos el tema de los gráficos). Al ejecutar el proceso nos aparecerá un gráfico como el que aparece en la Figura 6.3.

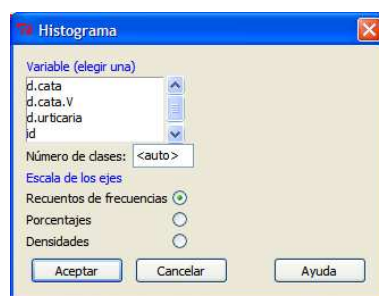


Figura 6.2: Creación de un histograma en Rcmdr.

Por lo que podemos observar en la Figura 6.3 da la sensación de que la variable se distribuye de manera parecida a como lo hace una normal. Sin embargo, aseverar esto basándonos en un gráfico sería demasiado arriesgado. Por ello, tendríamos que realizar un test sobre la normalidad de la variable. Rcmdr incorpora un procedimiento que permite testar el grado con que una variable se distribuye como si fuese una variable normal: el test de Shapiro-Wilk. Para ejecutarlo sobre la variable que nos ocupa podemos utilizar la función `shapiro.test()` haciendo que el argumento de la misma sea el nombre de la variable o, alternativamente, utilizar

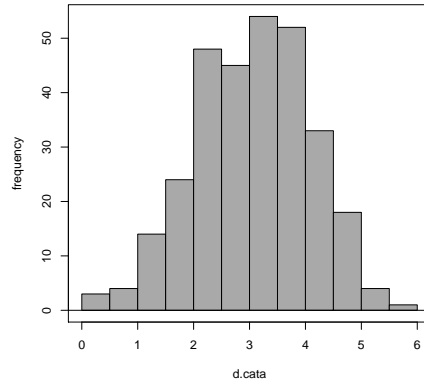


Figura 6.3: Ejemplo de un histograma en Rcmdr.

el menú *Estadísticos* → *Resúmenes* → *Test de normalidad de Shapiro-Wilk...* de la interfaz gráfica. Antes de entrar a analizar la salida que proporciona el programa es conveniente que nos detengamos a clarificar cuál es el contraste de hipótesis que se lleva a cabo en este caso. Bien, pues el contraste que se está evaluando cuando ejecutamos este análisis consiste en testar la hipótesis nula de normalidad. Es decir, el contraste podría expresarse del siguiente modo

$$H_0 : x = N(\mu, \sigma),$$

$$H_1 : x \neq N(\mu, \sigma),$$

donde la hipótesis nula indica que la variable de interés se distribuye normalmente con media μ y varianza σ mientras que la hipótesis alternativa indica lo contrario. Pues bien, si prestamos atención a la salida del análisis podemos ver que tenemos un estadístico de contraste llamado W y un valor p . Si interpretáramos el valor de p como ya hemos explicado tendríamos que aceptar la hipótesis nula y concluir que la variable se distribuye normalmente.

El otro supuesto que deben cumplir los datos para que las inferencias que extraigamos del test t para grupos independientes es el de homocedasticidad o igualdad de varianzas grupales. Dicho de otro modo, las varianzas en la variable dependiente han de ser las mismas para los grupos que definen la variable independiente. En nuestro ejemplo la varianza de los catarros debería ser la mis-

Capítulo 6 - Inferencia sobre medias

ma para las personas que tienen perros y para las que no. Aunque existe una corrección estadística para los casos en los que no se satisface éste supuesto, es recomendable valorar la posibilidad de que las varianzas de los grupos implicados en el análisis sean diferentes. Para ejecutar este análisis, en **R** tenemos que acceder al menú *Estadísticos* y seleccionar la opción *Varianzas*. Tenemos tres opciones: una prueba basada en la F de Snedecor específica para comparar las varianzas de dos grupos, y las pruebas de Barlett y de Levene que permiten que haya variables de agrupación con más de dos grupos. Aunque podríamos utilizar cualquier procedimiento de los tres y llegaríamos a conclusiones muy parecidas vamos a aplicar el primero de ellos, el de la F para dos muestras, por ser el que más se ajusta a la situación de este ejemplo. Lo primero que debemos plantearnos, como suelo recomendar, es saber cuál es el formato del contraste de hipótesis. En este caso, dado que lo que pretendemos evaluar es si dos varianzas son iguales o no vamos a utilizar la fracción entre ambas para confirmar o desconfirmar este hecho. Dado que si dividimos un número por él mismo obtendríamos el valor 1, este contraste de hipótesis se plantea en estos términos

$$H_0 : \frac{\sigma_1}{\sigma_2} = 1,$$
$$H_1 : \frac{\sigma_1}{\sigma_2} \neq 1.$$

La sintaxis que necesitamos ejecutar par obtener el análisis es esta:

```
1 var.test(d.cata ~ perro, alternative='two.sided', conf.level=.95, data=Perros)
```

Como habrás comprobado se utiliza una función llamada `var.test()` que abrevia la expresión inglesa *variance test* (test de varianza) y cuyo primer argumento es la variable dependiente o explicada y la variable independiente o explicativa respectivamente separadas por el símbolo `~`. Luego aparece un parámetro que define el tipo de hipótesis alternativa testada, la probabilidad para la estimación del intervalo de confianza para el resultado de la fracción y, por último, tenemos el parámetro `data` que especifica el objeto donde se encuentran las variables de interés. Si accedemos a la interfaz gráfica podemos ver que todas esas opciones están disponibles en el cuadro de diálogo que nos aparece (Figura 6.4).

6.3 - Contraste para dos medias

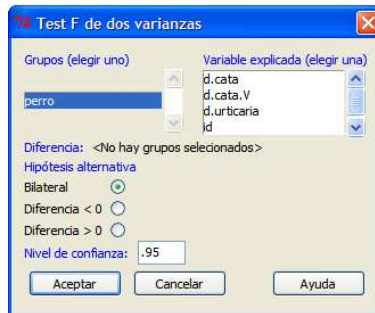


Figura 6.4: Contraste para dos varianzas en Rcmdr.

Como habrás podido comprobar la estimación de la fracción de las varianzas es de 0,55, el valor observado del estadístico F de Snedecor es el mismo valor, y que le corresponde un p -valor inferior a 0,005; luego la decisión más sensata sería rechazar la hipótesis nula y asumir que las varianzas no son iguales en el grupo de personas que tiene y no tienen perro.

Una vez sabido esto podemos realizar el test de comparación de medias para muestras independientes. Sin embargo, primeramente me gustaría hacer explícito el contraste de hipótesis que subyace en este test. Aunque normalmente se dice que el test t contrasta la hipótesis de que dos medias son iguales ($\mu_1 = \mu_2$), lo cierto es que si nos ponemos estrictos desde el punto de vista formal no es exactamente así, aunque en el fondo queramos decir lo mismo. Lo cierto es que lo que se contrasta en el test es si la diferencia entre medias es igual a 0, esto es

$$H_0 : \mu_1 - \mu_2 = 0,$$

$$H_1 : \mu_1 - \mu_2 \neq 0.$$

De este modo, si las dos medias son iguales, la diferencia entre ellas será cero mientras que si la primera es mayor o es menor que la segunda la diferencia entre ambas será positiva o negativa respectivamente. Aclarado este punto, para obtener el test tendríamos que acceder al menú *Estadísticos* \rightarrow *Medias* \rightarrow *Test t para muestras independientes...* o ejecutar la sintaxis:

```
1 t.test(  
2 d.cata~perro,  
3 alternative='two.sided',  
4 conf.level=.95,  
5 var.equal=FALSE,
```

Capítulo 6 - Inferencia sobre medias

```
6 data=Perros
7 )
```

Como se puede observar, he desglosado la función `t.test()` (líneas 1 y 7) para ir comentando los argumentos que también se podrían haber manipulado utilizando el cuadro de diálogo que aparece en la Figura 6.5. El primer argumento de la función consiste en el par definido por la variable explicada y la variable explicativa o de agrupación separados por el símbolo `~`, seguidamente aparece un parámetro (`alternative`) que sirve para indicar el formato de la hipótesis alternativa, luego tenemos la opción de manipular la probabilidad asociada al intervalo de confianza para la diferencia de medias (`conf.level`), seguidamente aparece el parámetro donde debemos indicar si las varianzas son iguales o diferentes (`var.equal`) en cuyo caso nosotros tenemos que indicar que no lo son (`FALSE`) dados los resultados para el contraste de varianzas que hemos realizado previamente, y por último aparece el parámetro que nos permite identificar la base de datos de la que se han tomado las variables. Creo que el cuadro de diálogo es auto-explicativo y que recoge exactamente la misma información que la que se ha planteado al hablar de la sintaxis.

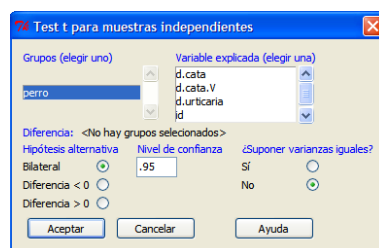



Figura 6.5: Test t de medias independientes en Rcmdr.

Como se puede comprobar a continuación, la salida que nos proporciona  es el test t adaptado por Welch para el caso en que no se cumple el supuesto de homocedasticidad:

```
1 Welch Two Sample t-test
2
3 data: d.cata by perro
4 t = -24.1776, df = 233.271, p-value < 2.2e-16
5 alternative hypothesis: true difference in means is not equal to 0
6 95 percent confidence interval:
7 -1.880392 -1.597024
```

6.3 - Contraste para dos medias

```
8 sample estimates:
9 mean in group No mean in group Sí
10      1.860523      3.599231
```

En la línea 4 de la salida aparecen el estadístico de contraste t de Student con sus grados de libertad ajustados para tolerar la no igualdad de varianzas y el valor de p . El valor de p en este caso está expresado en términos exponenciales y direccionales. Por direccionales me quiero referir a que no se nos da el valor exacto del parámetro sino que, mas bien se nos dice que es menor que un valor dado. Por su parte, el valor dado está expresado en notación científica utilizando potencias con base 10. En este caso el valor $2.2e-16$ que se proporciona equivaldría a $2,2 \times 10^{-16}$. O lo que es lo mismo, 0,00000000000000022. Es decir una probabilidad tan baja que invita a rechazar la hipótesis nula con poco resentimiento. En la línea 7 tenemos el intervalo de confianza para la diferencia entre las medias que aparece en términos negativos porque la segunda de las medias es mayor que la primera⁴ (línea 10).

Test de Wilcoxon

Una alternativa no paramétrica a la prueba t de Student que sirve para evaluar si existen diferencias entre dos grupos en una variable que alcanza, como mínimo, el nivel de medida ordinal y que no se distribuye normalmente es el test de Wilcoxon. Aunque no se basa en la media y, por tanto, esta subsección no debería de formar parte de una sección llamada «Contraste para dos medias»; desde un punto de vista didáctico puede ser ilustrativo contraponer ambas técnicas y presentarlas como alternativas para alcanzar objetivos cualitativamente similares.

El test de Wilcoxon se basa en la mediana en vez de basarse en la media y contrasta la hipótesis nula de que no existen cambios, diferencias o variaciones significativas respecto de la mediana de la variable explicada o dependiente en los dos grupos que evaluamos. Por ejemplo, consideremos que estamos interesados en estimar si existen diferencias en el número de problemas intestinales sufridos por una persona (*intestinal*⁵) en función de si la persona tiene perro o no. Para

⁴Al igual que sucede con el signo del estadístico t de contraste.

⁵Variable que no se distribuye normalmente. Puedes comprobarlo realizando la prueba de la normalidad Shapiro-Wilk o trazando el histograma correspondiente como hemos hecho anteriormente.

Capítulo 6 - Inferencia sobre medias

ejecutar el análisis podemos acceder al menú *Estadísticos* → *Tests no paramétricos* → *Test de Wilcoxon para dos muestras...* de la interfaz gráfica o ejecutar la siguiente sintaxis:

```
1 wilcox.test(intestinal ~ perro, alternative="two.sided", data=Perros)
```

Como se puede apreciar en la línea de sintaxis anterior, para ejecutar el análisis tenemos que acceder a una función llamada `wilcox.test()` cuyos argumentos son las variables implicadas, el tipo de hipótesis alternativa y el conjunto de datos donde se encuentran las variables. La salida que proporciona este análisis sería esta:

```
1  Wilcoxon rank sum test with continuity correction
2
3 data:  intestinal by perro
4 W = 9514, p-value = 0.8716
5 alternative hypothesis: true location shift is not equal to 0
```

Como se puede apreciar en la línea 4 de la salida el estadístico de contraste de Wilcoxon (W) no alcanza a ser estadísticamente significativo ($p = 0,87$) y, por tanto, tendríamos que aceptar la hipótesis nula que indica que las diferencias entre ambos grupos son estadísticamente significativas.

6.3.2. Medidas relacionadas

t de Student

Otro diseño básico en el que podemos estar interesados y que implicaría a dos medias sería el de grupos relacionados. En este caso las mismas personas son evaluadas en diferentes momentos temporales (t_1 y t_2). El objetivo sería evaluar si existen diferencias estadísticamente significativas entre las dos mediciones de la misma variable. Para seguir con el ejemplo que nos brinda el conjunto de datos que estamos utilizando en este capítulo, podríamos evaluar si la administración de la vacuna tiene algún efecto sobre el número de catarros que sufre una persona después del tratamiento (`d.cata.V`).

El contraste de hipótesis que se realiza con este test es el siguiente

6.3 - Contraste para dos medias

$$H_0 : \mu_{t_1} - \mu_{t_2} = 0,$$
$$H_1 : \mu_{t_1} - \mu_{t_2} \neq 0.$$

No obstante, antes de ejecutar el test sería recomendable que se contrastase uno de los supuestos específicos que implica el contraste t para medias relacionadas y que se refiere a la normalidad de la distribución muestral de la diferencia entre las variables. Algo que se podría hacer para contrastar este supuesto podría ser calcular una nueva variable que fuese el resultado de restar la puntuación de `d.cata` a la puntuación de `d.cata.V`. Seguidamente habría que testar la hipótesis de normalidad en esta nueva variable.

Para ejecutar un contraste de medias para grupos relacionados tendremos que seguir la siguiente ruta del menú de **Rcmdr**: *Estadísticos* → *Medias* → *Test t para datos relacionados...* En el cuadro de diálogo que nos aparece (Figura 6.6) podemos seleccionar la variable medida en el momento t_1 (en el cuadro de la izquierda) y la medida de la misma variable en el momento t_2 (en la lista de la derecha). También podemos, como viene siendo habitual, modificar el tipo de hipótesis alternativa que queremos testar y el porcentaje de seguridad del intervalo de confianza.

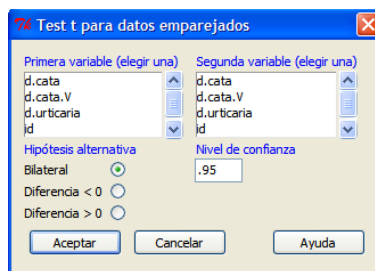


Figura 6.6: Test t de medias dependientes en Rcmdr.

La salida que ofrece el programa es análoga a la que obtenemos cuando ejecutamos el análisis para muestras independientes y su interpretación es idéntica por lo que no invertiré más tiempo en comentarla.

Test de Wilcoxon

Para terminar esta sección, presentaremos una alternativa no paramétrica a la prueba t de Student para valorar si dos medidas repetidas varían sistemáticamente desde el primer momento de medición al segundo. Para ello, utilizaremos nuevamente la prueba de Wilcoxon pero, adaptada esta vez, al caso de medidas repetidas.

Como se indicó anteriormente esta prueba se basa en la mediana y en las posibles diferencias que se establecen entre las variables en relación a sus respectivas medianas. Para ejecutar el análisis tenemos que utilizar la misma función que utilizamos para el caso de grupos independientes (`wilcox.test()`) aunque, en este caso, tendremos que añadir un parámetro que indique que son muestras relacionadas o pareadas (`paired=TRUE`). Para ejecutar el análisis tendremos que acceder al menú *Estadísticos* → *Tests no paramétricos* → *Test de Wilcoxon para muestras pareadas...* La salida de este análisis y su interpretación es semejante a los análisis previamente presentados.

6.4. Contraste para más de dos medias

Aunque las situaciones en las que se presentan dos grupos de medidas son más comunes de lo que pensamos en la realidad de la investigación, también es cierto que en muchos casos necesitamos más de dos grupos para poner a prueba nuestras hipótesis de trabajo.

En lo que sigue a continuación se darán unas pinceladas sobre cómo estimar análisis unifactoriales de la varianza y se presentará una alternativa no paramétrica (el test de Kruskal-Wallis) para contrastar la hipótesis de diferencias entre grupos respecto a una variable explicada que no supera los supuestos del modelo lineal general.

6.4.1. Análisis unifactorial de la varianza

Cuando hablamos de comparación de medias utilizando un análisis unifactorial de la varianza podríamos decir que estamos hablando de «palabras mayores». Y esto es así porque cuando ejecutamos un análisis de la varianza o ANOVA estamos estimando un modelo estadístico más sofisticado que los modelos estadísticos

anteriores y tanto \mathbb{R} como \mathbb{R}_{mult} tratan esta sofisticación de un modo cualitativamente diferente a como lo ha hecho en casos anteriores. Lo cierto es que podríamos dedicar todo un capítulo de este libro a tratar pormenorizadamente este tipo de análisis pero no lo voy a hacer así. Más bien, voy a dar unas pequeñas guías para que la persona interesada trate de avanzar en la utilización de la técnica. Esto se podría considerar como una pequeña introducción al uso de modelos estadísticos con \mathbb{R}_{mult} dado que el modo en que funcionan otros modelos estadísticos avanzados en este entorno de programación es bastante parecido. Por ello, recomendando que la persona interesada en continuar aprendiendo sobre la ejecución de análisis de varianza con \mathbb{R} y \mathbb{R}_{mult} consulte manuales especializados en el análisis multivariante de datos como los de Field (2009), León y Montero (2003) o Hair et al. (1998).

En primer lugar, habría que señalar que el análisis de varianza es una técnica paramétrica y, por tanto, se han de cumplir ciertos requisitos en los datos para que nuestras inferencias gocen de validez técnica. En segundo lugar, ha de existir *homocedasticidad* para cada grupo definido por la variable independiente en las medidas de la variable dependiente. O lo que es lo mismo, la varianza en cada condición del factor ha de ser similar. También se supone que las observaciones han de ser *independientes* y que el nivel de medida de la variable dependiente sea, al menos, de intervalo. Por último, pero no menos importante, la variable explicada o dependiente se ha de distribuir normalmente en cada uno de los grupos que define el factor o variable explicativa.

Supongamos que queremos saber si existen diferencias estadísticamente significativas en el número de catarros que sufre una persona dependiendo del tipo de casa donde vive. Dado que la variable *tipo.casa* tiene tres niveles no podemos utilizar la *t* de Student pero sí que podemos realizar un análisis de la varianza. El contraste de hipótesis que se está testando en este caso es el siguiente



$$H_0 : \mu_{\text{Grande}} = \mu_{\text{Mediana}} = \mu_{\text{Pequeña}},$$

$$H_1 : \mu_{\text{Grande}} \neq \mu_{\text{Mediana}} \neq \mu_{\text{Pequeña}}.$$

O lo que es lo mismo, la hipótesis nula indica que no existen diferencias estadísticamente significativas respecto a la media de catarros en función del tipo

Capítulo 6 - Inferencia sobre medias

de casa, mientras que la hipótesis alternativa indica que existen diferencias en el promedio de catarros dependiendo del tipo de casa donde se viva.

Para ejecutar el ANOVA unifactorial tenemos que acceder al menú *Estadísticos* → *Medias* → *ANOVA de un factor...* del menú gráfico de la interfaz. En el cuadro de diálogo que nos aparece (Figura 6.7) tendremos que especificar un nombre para el modelo en el cuadro de texto que aparece en la parte superior de la ventana.  almacenará el modelo en la memoria del equipo para que podamos analizarlo más detenidamente con posterioridad. Yo dejaré el nombre que asigna  por defecto: `AnovaModel.1`. Seguidamente tendríamos que elegir la variable de agrupación (*tipo.casa* en nuestro ejemplo) de la lista que aparece en el cuadro de la izquierda mientras que tendremos que elegir la variable dependiente o explicada de la lista que aparece en el cuadro de la derecha (*d.cata*). Por último, podríamos pedir que se realizasen comparaciones por pares ente los grupos que define la variable de agrupación al marcar el cuadro de verificación que aparece en la base del cuadro de diálogo. Sin embargo, esto no tiene sentido si no existen diferencias estadísticamente significativas entre los grupos de la variable explicativa.

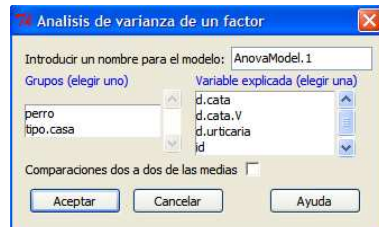


Figura 6.7: Análisis de la varianza (ANOVA) unifactorial en Rcmdr.

Como verás cuando ejecutes el análisis, aparecerán tres líneas nuevas de código en la ventana de instrucciones:

```
1 AnovaModel.1 <- aov(d.cata ~ tipo.casa, data=Perros,)  
2 summary(AnovaModel.1)  
3 numSummary(Perros$d.cata, groups=Perros$tipo.casa, statistics=c("mean", "sd"))
```

El código de la primera línea simplemente habrá servido para crear un objeto que contiene información sobre el modelo estimado utilizando la función del ANOVA (`aov()`) cuyos argumentos son la variable explicada y la variable explicativa separadas por el símbolo `~` y el parámetro que indica el conjunto de datos que contiene las variables de interés. En la línea 2 se solicita un resumen del modelo

6.4 - Contraste para más de dos medias



estimado y en la línea 3 se pide un resumen numérico como los que ya hemos trabajado previamente donde se pide la media y la desviación típica para cada grupo definido por la variable independiente. El resumen del modelo aparece en la salida del programa de este modo:

```
1           Df Sum Sq Mean Sq F value Pr(>F)
2 tipo.casa    2    0.99   0.495   0.465  0.629
3 Residuals  297  316.21   1.065
```

Un poco más abajo aparecen descriptivos básicos de los grupos (media, desviación típica y tamaño del grupo) estimados con la función que hemos utilizado previamente para obtener resúmenes numéricos de las variables:

```
1           mean          sd % data:n
2 Grande  2.978446  1.1201276  0      71
3 Mediana 3.131516  1.0139890  0     99
4 Pequeña 3.050613  0.9946466  0    130
```

Como se puede observar en las tablas anteriores el estadístico de contraste (F de Snedecor) no es lo suficientemente grande (0,467) teniendo en cuenta los grados de libertad del modelo (2) y, por tanto, no podemos rechazar la hipótesis nula de que las medias son iguales en los tres grupos ($p = 0,63$). Este hecho también se puede ver informalmente en la tabla de resumen donde aparecen las medias de los tres grupos de participantes y donde se aprecian pocas diferencias entre las medias.

Como se ha señalado anteriormente, esta sección no está dedicada a explicar los fundamentos de análisis de varianza sino que, más bien, pretende introducir al lector en la utilización de  y de  cuando realice sus análisis de varianza (en este caso unifactoriales). Por este motivo, me permitiré la libertad de mostrar algunas funcionalidades que se podrían utilizar con los ANOVAS estimados con éstos entornos de trabajo. Para obtener información adicional del modelo estimado podemos ejecutar la función `summary.lm()` del siguiente modo:

```
1 summary.lm(AnovaModel.1)
```

Como se puede ver en la sintaxis que aparece a continuación, ahora dispondremos de información más detallada del análisis de varianza tratado desde el modelo

Capítulo 6 - Inferencia sobre medias

lineal general en términos de una regresión múltiple (Field, 2009). En primer lugar se nos presentará un resumen superfluo de los residuos (líneas 6 y 7) y seguidamente tenemos una tabla de coeficientes de regresión con sus respectivos tests de significatividad para cada uno de los niveles de la variable explicativa (líneas desde la 8 a 12). En la línea 17 tenemos el valor de dos coeficientes de determinación múltiples (R^2), mientras que en la línea 18 aparece el test de significatividad para el modelo completo:

```
1 Call:
2 aov(formula = d.cata ~ tipo.casa, data = Perros)
3
4 Residuals:
5      Min       1Q   Median       3Q      Max
6 -2.82006 -0.72696  0.05078  0.72165  2.70385
7
8 Coefficients:
9              Estimate Std. Error t value Pr(>|t|)
10 (Intercept)      2.97845    0.12246  24.323  <2e-16 ***
11 tipo.casa[T.Mediana]  0.15307    0.16047   0.954   0.341
12 tipo.casa[T.Pequeña]  0.07217    0.15227   0.474   0.636
13 ---
14 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
15
16 Residual standard error: 1.032 on 297 degrees of freedom
17 Multiple R-squared:  0.003121, Adjusted R-squared:  -0.003592
18 F-statistic: 0.4649 on 2 and 297 DF,  p-value: 0.6286
```

Como el modelo estadístico está almacenado en memoria y activado por [Rcmdr](#), podemos ejecutar los análisis y obtener los gráficos adicionales referidos a ese modelo desde el menú general *Modelos* de la interfaz gráfica (Figura 6.8). Así, podemos estimar intervalos de confianza para los parámetros del modelo activo al nivel de confianza deseado, ejecutar análisis detallados sobre el funcionamiento del modelo o generar gráficos para estudiar si se acomoda a los supuestos subyacentes a la técnica estadística.

6.4.2. Contraste de Kruskal-Wallis

El test de Kruskal-Wallis es una alternativa no paramétrica al análisis de varianza unifactorial donde se estudia si existen diferencias estadísticamente significativas entre k grupos independientes en relación a una variable dependiente que ha sido medida en una escala, como mínimo, ordinal. Dado que la prueba no requiere

6.4 - Contraste para más de dos medias

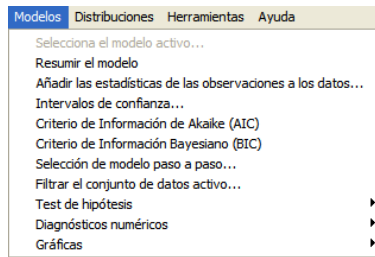


Figura 6.8: Opciones del menú *Modelos* en Rcmdr.

la normalidad en los datos ni la igualdad de varianzas en los grupos, se puede utilizar cuando los datos no se adecuan a ser analizados bajo la óptica de un ANOVA.

Para ejecutar el test de Kruskal-Wallis con **R**, tendremos que utilizar la función `kruskal.test()`. Por ejemplo, supongamos que estamos interesados en saber si existen diferencias estadísticamente significativas en el número de problemas intestinales sufridos tras la vacunación en función del tipo de casa en que vive el participante. Para ejecutar el análisis con la sintaxis tendríamos que escribir y ejecutar el siguiente código:

```
1 kruskal.test(intestinal.V ~ tipo.casa, data=Perros)
```

Si ejecutamos el análisis por medio de la interfaz (*Estadísticos* → *Test no paramétricos* → *Test de Kuskal-Wallis...*) se nos proporcionarán las medianas de cada grupo estudiado en la variable dependiente y, al igual que con la sintaxis, se nos generará la siguiente salida:

```
1 Kruskal-Wallis rank sum test
2
3 data: intestinal.V by tipo.casa
4 Kruskal-Wallis chi-squared = 266.1861, df = 2, p-value < 2.2e-16
```

Como se puede apreciar en la línea 4 se nos proporciona un estadístico de contraste basado en χ^2 , sus grados de libertad y su nivel de significación. Dado que el contraste de hipótesis que se está testando con este análisis es análogo a la que se testa con el ANOVA unifactorial, podríamos rechazar la hipótesis nula y concluir que existen diferencias estadísticamente significativas en la cantidad de

Capítulo 6 - Inferencia sobre medias

infecciones intestinales tras la vacunación en función del tipo de casa donde vivan los participantes.

▣ EJERCICIOS ▣

1. Considera el siguiente vector de datos:

4, 5, 3, 4, 2, 5, 4, 7, 8, 7, 2, 3, 1, 9, 5, 4

Contrasta las hipótesis de que la media del vector es igual a tres, inferior a tres y mayor que tres. Expón tus conclusiones.

2. ¿Existen diferencias estadísticamente significativas entre el número de catarras que se producen tras la vacunación entre las personas que tienen y no tienen perro? Justifica el uso de la técnica estadística que has utilizado.
3. ¿Existen diferencias estadísticamente significativas en el número de infecciones intestinales que se producen tras la vacunación entre las personas que tienen y no tienen perro? Justifica la utilización del test estadístico que has usado.
4. ¿Existen diferencias estadísticamente significativas en el promedio de catarras tras la vacunación en función del tipo de casa en que viven los participantes? Justifica el uso del test que has utilizado.
5. ¿Existen diferencias estadísticamente significativas en el número de infecciones intestinales que se sufren tras la vacunación en función del tipo de casa donde se vive? Justifica la elección de la técnica que has utilizado para realizar el contraste.

7

Inferencias sobre proporciones

En este breve capítulo se van a presentar un par de procedimientos que son útiles cuando trabajamos con variables cualitativas u ordinales con pocos niveles. En ambos casos estaremos trabajando con proporciones y utilizaremos el estadístico χ^2 o *ji-cuadrada*¹.

Para trabajar con este capítulo se proporciona una base de datos llamada `coches.RData` que contiene 5 variables y 1000 filas. Es un conjunto de datos que recoge información sobre si el riesgo de sufrir accidentes de tráfico (*accidente*) es alto o bajo en relación a otros tres factores: la **velocidad** promedio a la que se circule (que puede ser alta o baja), si se consume habitualmente **alcohol** o no, y el tipo de **coche** que maneje el conductor (deportivo, familiar o de transporte).

¹Aunque lo cierto es que a este estadístico se le llama chi-cuadrado (incluso yo mismo lo hago así en la mayor parte de las veces), lo cierto es que sería más correcto llamarlo ji-cuadrada por, al menos, dos motivos. En primer lugar, χ es la decimosegunda letra del alfabeto griego y se le denomina *ji* en castellano (Pabón, 1997). No obstante, en la mayoría de los paquetes estadísticos se le llama *Chi* (supongo que por la influencia anglosajona). Por otro lado, dado que es una letra (en femenino), tendríamos que apellidarla como *cuadrada* y no como *cuadrado*. Sin embargo, también se puede entender que con la expresión *ji-cuadrado* se podría estar acortando la expresión *ji al cuadrado*.

Como siempre la primera columna del archivo es simplemente un código que identifica a cada registro.


7.1. Inferencias sobre una variable

Una de las preguntas más sencillas que nos podemos hacer sobre una variable cualitativa de tipo nominal, o sobre una variable ordinal con pocos niveles, es si las proporciones estimadas para cada categoría de la variable son estadísticamente significativas. Para estimar si una proporción observada empíricamente es diferente a una proporción teórica se puede utilizar el test de χ^2 .

Por ejemplo, supongamos que, el año pasado, la tasa de conductores que fueron parados por la policía y que habían consumido alcohol fue del 50%. Imagina que tras estos datos alarmantes las autoridades en seguridad vial decidieron llevar a cabo una campaña publicitaria para reducir el consumo de alcohol en los conductores haciendo ver el riesgo que implica conducir en estado de embriaguez. La base de datos `coches.RData` contiene una variable (*alcohol*) que registra el número de personas que han sido detenidas por la policía y que han dado positivo en una prueba de alcoholemia a los 12 meses de la difusión de la campaña contra el alcohol que pusieron en marcha las autoridades en seguridad vial.

Antes de realizar ningún contraste de hipótesis tendríamos que calcular las frecuencias empíricas para la variable de interés; esto es, averiguar el porcentaje de personas que han sido detenidas y que dieron positivo en el control de alcoholemia. Para conocer la frecuencia absoluta de personas que han consumido alcohol recientemente podemos, como ya hemos comentado anteriormente, aplicar la función `summary()` a la variable *alcohol*. Al ejecutarla veremos que tenemos 648 personas que no dieron positivo mientras que 352 fueron acusados de haber consumido cantidades de alcohol que superaban los límites legales. Aunque sería fácil obtener el porcentaje que representa cada frecuencia en este caso (sólo con dividir por 10 dado que el tamaño muestral es 1000), en otros casos tendríamos que realizar más cálculos y podríamos pensar que `R` tampoco es de tanta utilidad. Por ello, `R` incorpora una opción que nos permite conocer las frecuencias absolutas y relativas (en términos porcentuales) de cada categoría de una variable cualitativa. Si accedemos a la ruta *Estadísticos* → *Resúmenes* → *Distribución de frecuencias...* del menú, nos aparecerá un cuadro de diálogo como el que aparece

7.1 - Inferencias sobre una variable

en la Figura 7.1. Como se puede apreciar, en la parte de la izquierda aparece una lista con las variables del archivo. Al seleccionar una variable y tras pulsar el botón «Aceptar»  calculará las frecuencias absolutas y los porcentajes para cada nivel de la misma. Adicionalmente, si queremos ejecutar (como es nuestro caso) un test sobre la bondad de ajuste ji-cuadrado para una variable, tendríamos que marcar la casilla de verificación que aparece en la base del cuadro de diálogo. Al pulsar en el botón «Aceptar» en este segundo caso, nos aparecerá un pequeño cuadro de diálogo (Figura 7.2) donde tendremos que especificar cuáles son las frecuencias teóricas o hipotéticas que consideramos para la variable analizada. Dado que nosotros queremos saber si las frecuencias observadas son diferentes al 50% tendremos que dejar la opción que nos aparece por defecto ($\frac{1}{2}$) intacta.

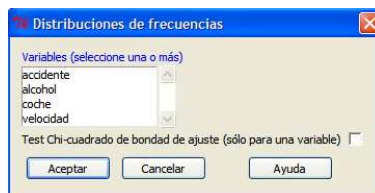


Figura 7.1: Frecuencias y prueba χ^2 para una muestra en Rcmdr.

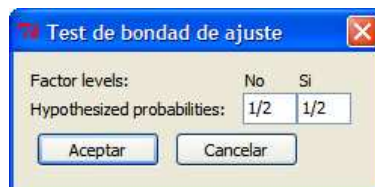


Figura 7.2: Frecuencias esperadas en la prueba χ^2 para una muestra en Rcmdr.

La salida del análisis ejecutado sería algo así:

```
1 > .Table <- table(alcohol)
2
3 > .Table # counts for alcohol
4
5 No Si
6 648 352
7
8 > round(100*.Table/sum(.Table), 2) # percentages for alcohol
9
10 No Si
11 64.8 35.2
12
13 > .Probs <- c(0.5,0.5)
```

Capítulo 7 - Inferencias sobre proporciones

```
14  
15 > chisq.test(.Table, p=.Probs)  
16  
17 Chi-squared test for given probabilities  
18  
19 data: .Table  
20 X-squared = 87.616, df = 1, p-value < 2.2e-16  
21  
22 > remove(.Probs)  
23  
24 > remove(.Table)
```

Tratemos de analizar que ha ido pasando en cada línea del código. En la línea 1 se ha utilizado la función `table()` sobre la variable `alcohol`. Esto genera una tabla de frecuencias para esta variable. Sin embargo, en vez de pedir que se muestre en ese momento, se ha creado un objeto llamado `.Table` que contiene esta información. En la línea 3 se pide que se muestre el objeto `.Table` y en las líneas 5 y 6 se muestra la tabla de frecuencias para la variable. En la línea 8 se realiza el cálculo necesario para que la tabla de frecuencias se transforme en una tabla de porcentajes²; esto es, se dice que se multiplique cada elemento de la tabla por 100 y que luego se divida por la suma de toda la tabla. En la línea 13 se genera un vector (`.Probs`) que contiene las probabilidades hipotetizadas para cada categoría de la variable cualitativa mientras que en la línea 15 se utiliza la función `chisq.test()` para ejecutar el test de bondad de ajuste. En la línea 20 nos aparece el resultado del test con el valor de χ^2 , sus grados de libertad y el p -valor correspondiente. Por último, en las filas 22 y 24, se borran de la memoria de nuestro ordenador el vector de probabilidades teóricas y la tabla de frecuencias que se han generado previamente.

Dado que la hipótesis que se contrasta podría expresarse del siguiente modo

$$H_0 : \pi_{Teo} = \pi_{Obs},$$

$$H_1 : \pi_{Teo} \neq \pi_{Obs},$$

o lo que es lo mismo, que la hipótesis nula estipula que las proporciones teóricas

²Se ha utilizado la función `round()` que sirve para redondear los números decimales. En este caso se ha limitado el número de decimales a dos dígitos.

o hipotéticas (π_{Teo}) predichas son iguales a las observadas (π_{Obs}) mientras que la hipótesis alternativa indica que no son iguales; podríamos decir que la proporción de personas que han sido «cazadas» conduciendo con algunas copas de más ha variado respecto a lo que esperábamos. En concreto, parece que la tasa de personas que no ha consumido alcohol ha aumentado hasta casi un 65 %.

7.2. Inferencias sobre la relación entre dos variables

En algunas situaciones estamos interesados en saber si dos variables de tipo cualitativo están relacionadas. Para estimar si existe independencia estadística se suele usar el estadístico χ^2 sobre tablas de contingencia. Una tabla de contingencia no es más que una tabla donde se cruzan todos los posibles valores, categorías o niveles de dos (o más) variables y que contiene la frecuencia absoluta o relativa de la ocurrencia de cada combinación de niveles. Por su parte, el estadístico χ^2 como test de la independencia entre dos variables ha recibido mucha atención en el campo de los algoritmos destinados a «descubrir» la estructura causal en un conjunto de datos (p. e., Scheines, Spirtes, Glymour, Meek, y Richardson, 2005; Spirtes, Glymour, y Scheines, 2000; Spirtes, Scheines, Glymour, Richardson, y Meek, 2004); esto es, a identificar qué variables causan la modificación de otras variables. En esta sección no profundizaremos en este tema tan interesante ya que simplemente nos limitaremos a testar la hipótesis nula de independencia entre dos variables utilizando tablas de contingencia bidimensionales.

Por ejemplo, consideremos que estamos interesados en saber si existe relación estadísticamente significativa entre el consumo de alcohol y el riesgo de sufrir un accidente de tráfico. Una de las primeras cosas que podríamos hacer con nuestra base de datos para tratar de responder a nuestra pregunta podría ser crear una tabla de contingencia entre las dos variables. Para crear una tabla de contingencia entre las dos variables podríamos utilizar la función `table()`³ que se ha introducido anteriormente considerando cada una de las variables como argumentos de la función separados por comas. Esto es, tendríamos que escribir:

³Posteriormente veremos que **R** utiliza otra función (`xtabs()`) cuando tratamos de obtener lo mismo utilizando la interfaz gráfica.

Capítulo 7 - Inferencias sobre proporciones

```
1 table(accidente, alcohol)
```

lo cual produciría:

```
1           alcohol
2 accidente No  Si
3     Alto 108 308
4     Bajo 540  44
```

Como se puede apreciar en la salida que ofrece el programa parece ser que existe cierta relación entre el consumo de alcohol y el riesgo de sufrir un accidente. Como se puede observar, cuando se consume alcohol lo más frecuente es que se tenga un riesgo alto de sufrir un accidente (308 personas consumieron alcohol y mostraron un alto riesgo de sufrir un accidente de coche), mientras que cuando no se consume alcohol lo más frecuente es que el riesgo de sufrir un accidente de tráfico sea bajo (540 personas mostraron bajo riesgo de sufrir un accidente tras no haber bebido alcohol). Sin embargo, estos datos son descriptivos de la relación que se establece entre estas dos variables. Si quisiésemos estimar si existen diferencias estadísticamente significativas en el riesgo de sufrir un accidente tras haber consumido alcohol podríamos utilizar el estadístico ji-cuadrado. Para ejecutar el test con **R** tendríamos que acceder al menú *Estadísticos* → *Tablas de contingencia* → *Tabla de doble entrada...* Al ejecutar el comando aparecerá el cuadro de diálogo que aparece en la Figura 7.3. Como se puede ver, hay dos listas (una para la variable *fila* y otra para la variable *columna*) de las que tendremos que elegir una variable en cada caso para generar la tabla de contingencia. En la parte de *Calcular porcentajes* podemos pedir que nos calculen los porcentajes totales, por filas, por columnas o (la opción por defecto) que se calculen únicamente las frecuencias absolutas. En la sección *Test de hipótesis* tendremos que seleccionar la casilla de verificación referente a la prueba de χ^2 .

La salida que habrá generado el programa será similar a esta:

```
1 > .Table <- xtabs(~accidente+alcohol, data=coches)
2
3 > .Table
4           alcohol
5 accidente No  Si
6     Alto 108 308
```

7.2 - Inferencias sobre la relación entre dos variables

```
7      Bajo 540  44
8
9  > .Test <- chisq.test(.Table, correct=FALSE)
10
11 > .Test
12
13      Pearson's Chi-squared test
14
15 data:  .Table
16 X-squared = 471.0715, df = 1, p-value < 2.2e-16
17
18
19 > remove(.Test)
20
21 > remove(.Table)
```

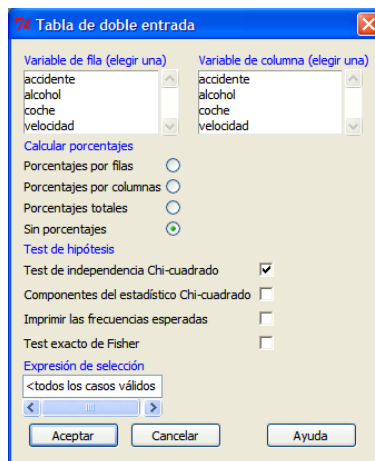


Figura 7.3: Prueba χ^2 para testar la independencia entre dos variables en Rcmdr.

En la línea 1 del código se crea la tabla de contingencia que relaciona a ambas variables mientras que en la línea 3 se solicita imprimir la tabla que aparece entre las líneas 4 y 7 (observa que los datos son los mismos a los generados con la función `table()`). En la línea 9 se solicita que se cree un objeto que contenga el test sobre la tabla de contingencia generada previamente mientras que en la línea 11 se ordena imprimir el resultado de la prueba. En la línea 16 aparece el valor de χ^2 con sus grados de libertad y su nivel de significación. Por último, en las líneas 19 y 21 se eliminan la tabla de contingencia y el objeto que contenía los resultados del análisis.

Dado que, como se indicó anteriormente, la hipótesis nula que se testa en esta

Capítulo 7 - Inferencias sobre proporciones

prueba es la de independencia entre las dos variables; una vez visto el estadístico de contraste y su p -valor asociado tendríamos que rechazar la hipótesis nula y aceptar la alternativa que indicaría existencia de relación entre el consumo de alcohol y el riesgo de sufrir accidentes.

Existe otro modo de ejecutar un análisis sobre la relación entre dos variables cualitativas utilizando χ^2 con **Rcmdr**. En este segundo caso, lo que tendríamos que hacer es introducir manualmente la tabla de contingencia y pedir que se ejecute el test sobre este objeto. Esta opción es muy útil cuando no disponemos de los datos brutos originales y únicamente tenemos la tabla de contingencia con las frecuencias absolutas o relativas. Para realizar el test de esta manera tenemos que acceder al cuadro de diálogo que aparece en la Figura 7.4 seleccionando el comando *Estadísticos* \rightarrow *Tablas de contingencia* \rightarrow *Introducir y analizar una tabla de doble entrada...* Lo primero que encontraremos en el cuadro de diálogo son dos barras de desplazamiento con las que podremos definir las características, número de filas y columnas, de la tabla de contingencia que queremos analizar. Seguidamente tendremos que introducir las frecuencias para cada casilla de la tabla que hemos definido previamente. A continuación tenemos las mismas opciones que existen en el cuadro de diálogo destinado a ejecutar el análisis sobre los datos brutos.

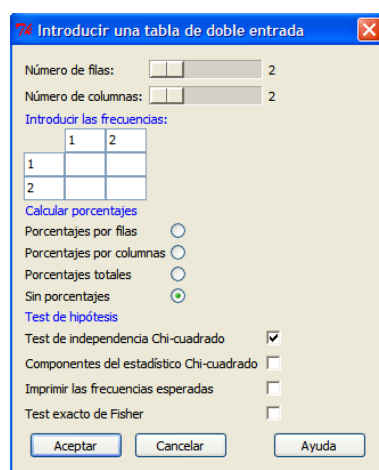


Figura 7.4: Prueba χ^2 para testar la independencia entre dos variables a partir de una tabla en Rcmdr.

EJERCICIOS



1. Un estudio que se realizó el año pasado indicó que la proporción de coches familiares que se detuvieron en los controles de alcoholemia fue del 56 %, mientras que se pararon un 19 % de coches deportivos y un 27 % de vehículos destinados al transporte. ¿Ha variado, en relación a lo que pasaba el año pasado, la proporción del tipo de coches que se han parado en los controles de alcoholemia? Justifica tu respuesta.
2. ¿Existe relación entre el tipo de coche y la velocidad a la que se circula por la carretera? ¿Existe relación entre el consumo de alcohol y el tipo de coche que se conduce? Justifica tus respuestas.
3. Un centro de enseñanza secundaria ha estudiado, durante todo un curso, la posible relación que se establece entre el rendimiento deportivo y el cociente de inteligencia. La tabla de contingencia que aparece más abajo resume parte de los datos que han obtenido. ¿Podría decirse que existe relación estadísticamente significativa entre el rendimiento deportivo y el cociente de inteligencia? Justifica tu respuesta. *La tabla informa sobre el número de personas que caen dentro de cada categoría definida por los niveles que contienen las variables utilizadas.*

		Cociente de Inteligencia		
		Bajo	Medio	Alto
Rendimiento Deportivo	Bajo	50	30	26
	Medio	13	58	15
	Alto	10	39	18

Capítulo 7 - Inferencias sobre proporciones

8

Correlación y regresión lineal

Aunque lo hemos estado haciendo a lo largo de los capítulos anteriores de manera implícita, en este capítulo vamos a tratar uno de los modelos estadísticos más afamados dentro de lo que denominamos como *modelo lineal general*: la regresión lineal. Previamente se indicará cómo estimar coeficientes de correlación lineales que servirán para introducir el análisis de regresión lineal simple y múltiple. Sin embargo, este capítulo no será un documento pormenorizado del modelo de regresión lineal sino que, más bien, como viene siendo habitual, se podría considerar como una pequeña introducción a la utilización de  y  para estimar este tipo de modelos. Recomiendo que se acceda a manuales especializados para profundizar en los supuestos y componentes del modelo. Por ejemplo, para una introducción «amigable» sobre este asunto recomiendo el libro de Pagano (1998/1999), para una exposición detallada del modelo sugiero que se consulte el manual de Hair et al. (1998), mientras que para revisar una de sus implementaciones informáticas se podría consultar el manuscrito de Field (2009).

Para este capítulo vamos a utilizar una base de datos llamada `ecopaz.RData`. El conjunto de datos contiene nueve variables que recogen información sobre 174

Capítulo 8 - Correlación y regresión lineal

países relacionada con datos económicos y de bienestar social. La variable *pais* es, en este caso, el código que identifica inequívocamente a cada registro de la base de datos. Estas son las variables contenidas en la base de datos¹ junto con una pequeña descripción de su significado:

- *IPG*: es el *Índice de Paz Glogal* (o *Global Peace Index*) reportado el 10 de junio del año 2010. Para estimar este índice se utilizan parámetros de violencia, criminalidad, gasto militar o información sobre conflictos bélicos sobrevenidos en cada país. Los países considerados más pacíficos tienen asignada una puntuación más baja.
- *SWL*: es el *Índice de Satisfacción Vital* (o *Satisfaction with Life Index*) que fue creado por el psicólogo social Adrian G. White de la *University of Leicester*. Este índice representa un intento por estimar el grado de felicidad de los países del mundo basado tanto en preguntas directas a los ciudadanos sobre su felicidad así como tomando en cuenta parámetros de desarrollo económico y bienestar social. Cuanto mayor es el índice mayores niveles de felicidad promedio experimentan los ciudadanos del país.
- *IDH*: es el *Índice de Desarrollo Humano* publicado el 2 de noviembre de 2011 al auspicio del Programa de las Naciones Unidas para el Desarrollo (PNUD). Este índice se basa en tres parámetros básicos para evaluar el grado de desarrollo humano en los ciudadanos del país: duración de la vida en condiciones saludables, acceso a educación y grado en que la vida se disfruta dignamente. Cuanto mayor es el índice mayor es el grado de desarrollo del país.
- *PIB*: es una estimación del *Producto Interior Bruto* entre 2005 y 2010 realizado por el Banco Mundial y está medido en millones de dólares americanos.
- Las variables *orden_IPG*, *orden_SWL*, *orden_IDH*, y *orden_PIB* representan las posiciones de cada país en el *ranking* de cada una de las variables a las que se refieren. En cada caso un valor más pequeño indican mayores valores

¹Esta base de datos ha sido elaborada manualmente tomando los datos de Wikipedia www.wikipedia.org y, por tanto, los errores que se hayan podido producir (y los cuales lamentaría muchísimo) habrán sido debidos a la manipulación de datos que he llevado a cabo. Si detectas alguna errata en los datos agradecería enormemente que me informases sobre ello. ¡Gracias!

en sus correspondientes parámetros excepto para el *IPG* donde un valor más bajo corresponde también a un valor más bajo en el *ranking*.

8.1. Correlación

Como he comentado anteriormente, he sido profesor de psicometría en la Universidad de Almería durante los últimos siete años y no deja de sorprenderme el ver como algunas personas que llegan al tercer curso de los estudios en psicología siguen teniendo concepciones poco acertadas de lo que es la *correlación*. Soy consciente de que muchas personas (quizá cada vez menos) comienzan a estudiar psicología pensando que es una carrera donde la estadística o las matemáticas no tienen cabida. No obstante, también es cierto que la idea de correlación ha estado estrechamente ligada al desarrollo de la psicología y, por tanto, en cierto modo, merece una atención especial.

Los errores más comunes que encuentro en mis alumnos y alumnas están relacionados con la interpretación del coeficiente de correlación. Por ejemplo, en ocasiones piensan que cuando un índice de correlación es negativo ésto es indicativo de ausencia de relación entre las variables. Por ello, cuando tengo la oportunidad de hacerlo, trato de explicar lo que he venido a denominar como *interpretación bidimensional* de un coeficiente de correlación.

La idea de la interpretación bidimensional del coeficiente de correlación se somete a la dicotomía clásica que vengo apreciando que existe en las técnicas de análisis e investigación científica: la dimensión cualitativa y la dimensión cuantitativa. El caso es que yo sugiero a mis alumnas y alumnos que interpreten el coeficiente de correlación tanto en su dimensión cualitativa como en su dimensión cuantitativa. Por dimensión cualitativa me refiero al significado, sentido, signo o dirección de la relación entre las dos variables implicadas. En este sentido, un coeficiente de correlación² puede ser negativo, igual a cero o positivo. Cuando el coeficiente de correlación es positivo decimos que existe una relación lineal directamente proporcional entre dos variables. Ésto es, cuando una variable aumenta la otra también lo hace mientras que cuando una de las variables disminuye la

²Al menos en los que vamos a tratar en este capítulo. No obstante, es cierto que existen otros estadísticos de correlación que se interpretan de modo diferente. De modo genérico todas estas explicaciones son apropiadas para el coeficiente de correlación de Pearson aunque pueden extenderse a otros índices de correlación.

Capítulo 8 - Correlación y regresión lineal

otra también lo hace. Un ejemplo de este tipo de asociación entre variables lo podríamos encontrar al estudiar la relación que existe entre el número de horas que trabaja un artista y el número de obras de arte que genera dado que, en un mundo ideal, cuantas más horas se trabajen mayor cantidad de elementos se producirán. Por su parte, una correlación de signo negativo indica una relación lineal inversamente proporcional entre variables. En este caso, cuando una variable aumenta la otra disminuye o cuando una variable disminuye la otra aumenta. Por ejemplo, la relación que existe entre el número de trabajadores y el tiempo que se tarda en realizar cierta tarea sería un ejemplo de correlación negativa. Así, cuantos más trabajadores se dispusiesen a realizar una tarea (por ejemplo, construir un barco) menor sería el tiempo que tardarían en finalizar. Por último, cuando una correlación tiene un valor de cero decimos que no existe relación lineal entre las variables. Valga, como ejemplo de correlación nula, la relación que podría existir entre la motivación laboral de los trabajadores de una cadena de montaje y la luminosidad del color del bolígrafo que se utiliza para firmar sus nóminas.

Por su parte, la interpretación cuantitativa del coeficiente de correlación está referida a la magnitud de relación o a la fuerza de asociación que se establece entre las variables. En este sentido, cuanto más cercano a uno sea el valor del coeficiente de correlación (en valor absoluto) mayor será la fuerza de asociación entre las variables. Esto quiere decir que, independientemente del signo, cuanto más cercano sea un coeficiente de correlación a sus extremos posibles (el -1 y el +1) mayor será la magnitud de relación entre las variables.

En términos gráficos podemos identificar la correlación en la medida que su nube de puntos se aproxima a una recta en un gráfico de dispersión. En un gráfico de dispersión se representan los puntos (o nube de puntos) que corresponden a cada par de valores para cada una de las variables. Cuanto más se aproxima la nube de puntos a una línea recta mayor es la correlación entre las variables. En el caso extremo y poco probable en que el coeficiente de correlación valga uno o menos uno se dice que existe una relación perfecta entre las variables y la nube de puntos sería una línea recta. En la Figura 8.1 aparecen algunos ejemplos de gráficos de dispersión donde las nubes de puntos implican diferentes niveles de relación entre las variables x e y .

8.1 - Correlación

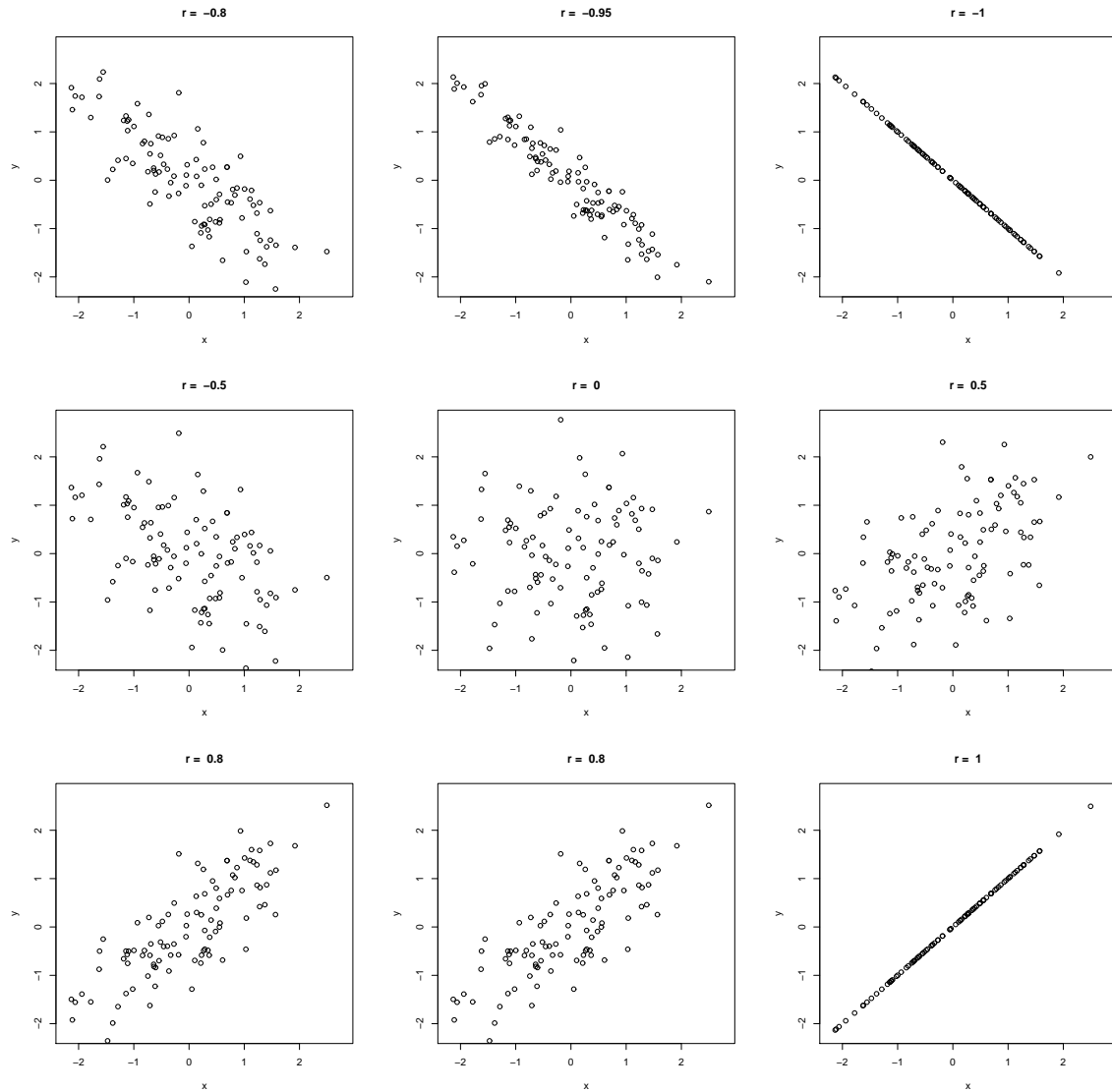


Figura 8.1: Ejemplos de gráficos de dispersión. En la fila superior aparecen correlaciones negativas que aumentan en fuerza de relación de izquierda a derecha. En la fila inferior aparecen diagramas de dispersión que representan correlaciones positivas que van aumentando en magnitud de asociación de izquierda a derecha. En la fila central aparece un gráfico de dispersión donde no se aprecia relación lineal entre las variables (centro), mientras que a los lados aparecen una correlación del punto medio de la «subescala» negativa (izquierda) y una correlación del punto medio de la «subescala» positiva (derecha). Los gráficos han sido creados con el *plug-in* TeachingDemos diseñado para el paquete [Rcmdr](#), motivo por el cual el delimitador decimal es un punto.

8.1.1. Coeficiente de Pearson

El coeficiente de correlación de Pearson (también denominado *coeficiente de correlación lineal producto-momento de Pearson* y simbolizado r o r_{xy}) es un parámetro adimensional que representa la relación que se establece entre dos variables de tipo cuantitativo. Este índice es la razón entre la covarianza y el producto de las desviaciones típicas de ambas variables (ecuación 8.1).

$$r_{xy} = \frac{COV(x, y)}{s_x \times s_y} \quad (8.1)$$

La covarianza (ecuación 8.2) es una medida de la asociación lineal entre dos variables cuantitativas y tiene el valor 0 cuando no existe relación lineal entre las variables, es positiva cuando la relación es directamente proporcional y es negativa cuando la relación es inversamente proporcional. En \mathbb{R} podemos calcular la covarianza entre dos variables utilizando la función `cov()` incluyendo las variables objeto de análisis separadas por una coma. El problema de la covarianza es que no está acotada lo que hace dificultosa su interpretación como índice de relación entre dos variables por ello, el coeficiente de Pearson es el estadístico más apropiado para estimar la relación lineal entre dos variables independientemente del nivel de medida de las mismas.

$$COV(x, y) = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})(y_i - \bar{y})}{n} \quad (8.2)$$

La función que estima el coeficiente de correlación de Pearson en \mathbb{R} es `cor()` cuyos argumentos serán los nombres de las variables separadas por una coma. Para estimar la correlación entre variables utilizando la interfaz de `Rcmdr` accedemos al menú *Estadísticos* \rightarrow *Resúmenes* \rightarrow *Matriz de correlaciones...* En el cuadro de diálogo que nos aparece (Figura 8.2) tenemos un listado de las variables numéricas en la parte superior. Tendremos que seleccionar dos o más y elegir un método de estimación, por defecto aparece el de Pearson. También podemos pedir un contraste de hipótesis para el valor del coeficiente de correlación donde se testa la hipótesis nula de que el coeficiente de correlación es igual a cero, esto es

$$H_0 : r_{xy} = 0$$

$$H_1 : r_{xy} \neq 0.$$

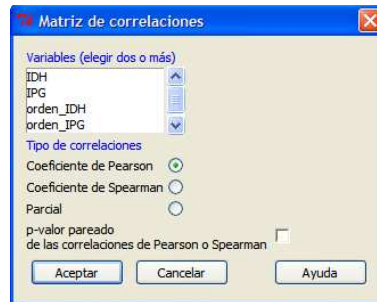


Figura 8.2: Matriz de correlaciones en Rcmdr.

Por ejemplo, si queremos estimar la correlación que existe entre las variables cuantitativas continuas de nuestro fichero (*IDH*, *IPG*, *PIB* y *SWL*) solicitando un test de hipótesis para cada correlación obtendríamos el siguiente resultado:

1		IDH	IPG	PIB	SWL
2	IDH	1.00	-0.54	0.22	0.59
3	IPG	-0.54	1.00	-0.02	-0.51
4	PIB	0.22	-0.02	1.00	0.17
5	SWL	0.59	-0.51	0.17	1.00
6					
7	n=	136			
8					
9	P				
10		IDH	IPG	PIB	SWL
11	IDH		0.0000	0.0094	0.0000
12	IPG	0.0000		0.8408	0.0000
13	PIB	0.0094	0.8408		0.0524
14	SWL	0.0000	0.0000	0.0524	
15					
16	Adjusted p-values (Holm's method)				
17		IDH	IPG	PIB	SWL
18	IDH		0.0000	0.0281	0.0000
19	IPG	0.0000		0.8408	0.0000
20	PIB	0.0281	0.8408		0.1049
21	SWL	0.0000	0.0000	0.1049	

En primer lugar, aparece la matriz de correlaciones donde aparecen los coeficientes de correlación para cada comparación de pares de variables. En la línea 7

Capítulo 8 - Correlación y regresión lineal

nos informa del tamaño de la muestra sobre la que se han ejecutado los cálculos tras eliminar los casos en los que existe algún valor perdido. Entre las líneas 10 y 14 aparecen los p -valores para cada correlación mientras que entre las líneas 16 y 21 tenemos p -valores corregidos con el método de Holm que controla las comparaciones múltiples.

Como se puede observar, el índice de desarrollo humano correlaciona positivamente con la satisfacción vital y con el producto interior bruto mientras que lo hace negativamente con el índice de paz global. Resulta curioso observar que la relación entre el PIB y el índice de paz global es prácticamente cero en la muestra utilizada y que la relación entre el PIB y satisfacción con la vida es sólo marginalmente significativa mientras que su asociación desaparece cuando se controlan comparaciones múltiples.

Cuando partimos de hipótesis relativas al sentido de la correlación entre dos variables podemos realizar un test unilateral sobre el valor de la correlación utilizando la función `cor.test()` e indicando como uno de sus parámetros si el valor de la hipótesis alternativa (`alternative`) es en un sentido o en otro. Imaginemos que queremos testar la hipótesis nula de ausencia de correlación entre las variables *IDH* e *IPG* frente a la hipótesis alternativa de que el coeficiente de correlación es inferior a cero. Para ejecutar este análisis usando la interfaz gráfica accedemos al menú *Estadísticos* → *Resúmenes* → *Test de correlación*. Como podrás apreciar (Figura 8.3) hay que elegir únicamente un par de variables, el método de estimación y un formato para la hipótesis alternativa.

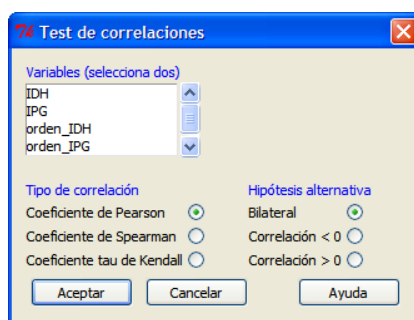


Figura 8.3: Test de correlación en Rcmdr.

Al ejecutar el test obtenemos una salida análoga a esta:

```
1 Pearson's product-moment correlation
2
```



```

3 data:  ecopaz$IDH and ecopaz$IPG
4 t = -7.734, df = 136, p-value = 1.045e-12
5 alternative hypothesis: true correlation is less than 0
6 95 percent confidence interval:
7  -1.0000000 -0.4467886
8 sample estimates:
9      cor
10 -0.5526907

```

En la línea 4 se nos proporciona el estadístico de contraste (que en este caso es una t de Student) con sus grados de libertad y su nivel de significación estimado. En la línea 7 se nos proporcionan los dos valores que forman el intervalo de confianza del parámetro al 95 %. Finalmente, en la línea 10, tenemos la estimación muestral del coeficiente de correlación.

8.1.2. ρ de Spearman y τ de Kendall

Cuando tenemos variables que son consideradas como variables medidas en una escala ordinal no sería recomendable utilizar el coeficiente de correlación de Pearson. En esta sección se tratan dos índices que proporciona **R** para estimar la relación que se establece entre variables de tipo ordinal.

ρ de Spearman

El coeficiente ρ (*rho*) de Spearman (también simbolizado como r_s o conocido como *coeficiente de correlación por rangos*) no es ni más ni menos que el coeficiente de correlación de Pearson aplicado sobre variables de tipo ordinal. Esto es, cada variable objeto de ser analizada con este parámetro es recodificada a una nueva variable cuya puntuación representa la posición o el número de orden que ocupaba el antiguo valor en la variable original. Una vez hecho esto se aplica el coeficiente de correlación de Spearman cuya ecuación, en su modo más sencillo, es

$$r_s = 1 - \frac{6 \sum_{i=1}^{i=n} d_i^2}{n(n^2 - 1)}, \quad (8.3)$$

donde d_i se refiere a la diferencia de rangos entre los valores de las variables implicadas en el análisis.

Capítulo 8 - Correlación y regresión lineal

Para obtener el coeficiente de correlación de Spearman tendríamos que acceder al mismo cuadro de diálogo en el que solicitamos el coeficiente de correlación de Pearson (Figura 8.2) utilizando la ruta *Estadísticos* → *Resúmenes* → *Matriz de correlaciones...* del menú. En este caso tendríamos que seleccionar la opción correspondiente al coeficiente de correlación de Spearman y marcar las variables sobre las que queremos ejecutar el análisis. Por ejemplo, podríamos analizar las correlaciones que se establecen entre las variables ordinales (*orden_IDH*, *orden_IPG*, *orden_PIB*, *orden_SWL*) que tenemos en nuestra base de datos y que representan las ordenaciones de las variables cuantitativas que hemos utilizado previamente. Si ejecutamos ese análisis obtenemos la salida:

1		<i>orden_IDH</i>	<i>orden_IPG</i>	<i>orden_PIB</i>	<i>orden_SWL</i>
2	<i>orden_IDH</i>	1.00	0.59	0.56	0.60
3	<i>orden_IPG</i>	0.59	1.00	0.16	0.48
4	<i>orden_PIB</i>	0.56	0.16	1.00	0.36
5	<i>orden_SWL</i>	0.60	0.48	0.36	1.00
6					
7	n=	136			
8					
9	P				
10		<i>orden_IDH</i>	<i>orden_IPG</i>	<i>orden_PIB</i>	<i>orden_SWL</i>
11	<i>orden_IDH</i>		0.0000	0.0000	0.0000
12	<i>orden_IPG</i>	0.0000		0.0561	0.0000
13	<i>orden_PIB</i>	0.0000	0.0561		0.0000
14	<i>orden_SWL</i>	0.0000	0.0000	0.0000	
15					
16	Adjusted p-values (Holm's method)				
17		<i>orden_IDH</i>	<i>orden_IPG</i>	<i>orden_PIB</i>	<i>orden_SWL</i>
18	<i>orden_IDH</i>		0.0000	0.0000	0.0000
19	<i>orden_IPG</i>	0.0000		0.0561	0.0000
20	<i>orden_PIB</i>	0.0000	0.0561		0.0000
21	<i>orden_SWL</i>	0.0000	0.0000	0.0000	

Los resultados se estructuran de manera análoga a como se ha comentado para el caso del coeficiente de correlación de Pearson. En primer lugar aparece la matriz de correlaciones, seguidamente tenemos la matriz de significaciones o *p*-valores y, por último, aparecen valores ajustados para las significaciones de cada parámetro teniendo en cuenta comparaciones múltiples. Por otro lado, si queremos ejecutar contrastes de hipótesis unilaterales semejantes a los que hemos realizado con el coeficiente de Pearson podemos hacerlo accediendo al comando *Estadísticos* → *Resúmenes* → *Test de correlación* de la interfaz gráfica.

τ de Kendall

El coeficiente τ (*tau*) de Kendall es otro índice de correlación adecuado para evaluar la relación que se establece entre variables de tipo ordinal pero con el que se evalúa la concordancia y la discordancia entre ordenaciones de pares de observaciones. Dado que los datos sobre los que se aplica el análisis consisten en dos variables (x e y) y ya que cada observación consta de una pareja de datos (por ejemplo, x_i, y_i) podríamos definir formalmente una *concordancia* o coincidencia cuando

$$(x_i < x_j) \cap (y_i < y_j) \cup (x_i > x_j) \cap (y_i > y_j),$$

mientras que diríamos que existen *discordancias* o desacuerdos si

$$(x_i < x_j) \cap (y_i > y_j) \cup (x_i > x_j) \cap (y_i < y_j).$$

Hay tres versiones del coeficiente de correlación τ de Kendall. `R` calcula el coeficiente τ_b cuando la tabla de contingencia que se genera con las variables objeto de estudio es cuadrada dado que el conocido como τ_a es considerado como un estadístico sesgado (Solanas et al., 2005). En caso de tablas de contingencia rectangulares es recomendable usar la versión τ_c del coeficiente.

Como te habrás cerciorado, para obtener los coeficientes de correlación τ de Kendall entre las variables ordinales del archivo únicamente tendríamos que seleccionar la opción correspondiente en el cuadro de diálogo (8.3) que ya hemos utilizado para obtener los contrastes de hipótesis para los índices de Perason y de Spearman. Por ejemplo, tras calcular el coeficiente de correlación τ de Kendall entre las variables *orden_IDH* y *orden_IPG*, utilizando un contraste de hipótesis bilateral, obtenemos la siguiente salida:

```

1 Kendall's rank correlation tau
2
3 data:  ecopaz$orden_IDH and ecopaz$orden_IPG
4 z = 7.3739, p-value = 1.656e-13
5 alternative hypothesis: true tau is not equal to 0
6 sample estimates:
7     tau
8 0.4239378

```

De manera análoga a como hemos visto con el coeficiente de correlación de Pearson, en la línea 4 tenemos el estadístico de contraste (que en este caso es una z) y el correspondiente nivel de significación. En la línea 8 tenemos la estimación muestral del parámetro cuyo valor indica que cuanto más arriba está un país en el ranking del *IDH* también lo estará en su ordenación del *IPG*.

8.2. Introducción a la regresión lineal

Para terminar con este capítulo quería dar algunas pinceladas sobre la técnica del análisis de regresión lineal y de su implementación en \mathbb{R} por medio del uso de \mathbb{R}_m . En primer lugar, creo que tendríamos que intentar conceptualizar el término *regresión*. Según la vigésima edición del Diccionario de la Real Academia de la Lengua (www.rae.es), el vocablo «regresión» proviene del latín *REGRESSIO*, *-ŌNIS*, y viene a referirse a una «retrocesión o acción de volver hacia atrás». Sin embargo, aunque podríamos darle vueltas al asunto para encontrar una relación con el sentido matemático del término, lo cierto es que en el contexto científico *regresión* suele ser considerado como sinónimo de *predicción* (Silva y Barroso, 2004).

Más concretamente, el análisis de regresión lineal (como su nombre indica) pretende predecir el valor de una variable (denominada *resultado*, *dependiente*, *explicada* o *predicha*) a partir de otra variable o variables (llamadas *predictoras*, *independientes*, *explicativas* o *indicadores*). Para ello utiliza como modelo subyacente la ecuación de la línea recta que queda definida por la ecuación $y = a + b \times x$; donde y es la variable predicha, x es la predictora, a es el origen (intercepto) o punto de corte con el eje de ordenadas (y), y b es la pendiente de la recta. En definitiva, lo que conseguimos cuando ejecutamos un análisis de regresión lineal es una ecuación de la recta que nos servirá para predecir los valores de nuestra variable resultado a partir del valor o valores de nuestra variable/s predictora o predictoras.

Por ejemplo, cuando estimamos la matriz de correlaciones de Pearson vimos que la mayor correlación observada entre los pares de variables era la que estimaba la relación entre los índices *IDH* y *SWL*. Como consecuencia, podríamos preguntarnos si el índice de desarrollo humano en un país es un buen predictor de la satisfacción vital que experimentan sus ciudadanos. Es decir, podríamos estimar un modelo de regresión lineal que implicase a éstas variables y, de este

8.2 - Introducción a la regresión lineal

modo, evaluar su grado de idoneidad o verosimilitud. Para estimar el modelo de regresión lineal que predice la satisfacción con la vida en función del índice de desarrollo humano habría que ejecutar la siguiente sintaxis:

```
1 RegModel.1 <- lm(SWL ~ IDH, data=ecopaz)
2 summary(RegModel.1)
```

o, alternativamente, acceder al comando *Estadísticos* → *Ajuste de modelos* → *Regresión lineal...* de la interfaz y seleccionar las variables explicada y explicativa en sus correspondientes listas del cuadro de diálogo que aparece (Figura 8.4).



Figura 8.4: Regresión lineal en Rcmdr.

Como resultado aparecerá una salida análoga a esta:

```
1 Call:
2 lm(formula = SWL ~ IDH, data = ecopaz)
3
4 Residuals:
5      Min       1Q   Median       3Q      Max
6 -90.526 -17.289   6.625  22.921  70.633
7
8 Coefficients:
9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept)  112.522     9.319  12.075  <2e-16 ***
11 IDH          134.436    13.630   9.863  <2e-16 ***
12 ---
13 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
14
15 Residual standard error: 30.94 on 172 degrees of freedom
16 Multiple R-squared:  0.3613, Adjusted R-squared:  0.3576
17 F-statistic: 97.29 on 1 and 172 DF, p-value: < 2.2e-16
```

Como se puede ver, lo primero que aparece (línea 2) en la salida es una especificación del modelo estimado y del conjunto de datos que se ha utilizado. En las líneas 5 y 6 aparecen algunos estadísticos (el mínimo, el máximo y los cuartiles)

Capítulo 8 - Correlación y regresión lineal

sobre los residuos (diferencias entre los valores reales de y y los estimados por la regresión lineal) del modelo. Entre la línea 8 y la 11 aparece la tabla de coeficientes del modelo. Dado que sólo se ha incluido una variable predictora en el modelo tendremos dos coeficientes: uno para el intercepto y otro para la pendiente asociada a la variable predictora. Para cada parámetro tendremos un estadístico t de contraste que testa la hipótesis nula de que el parámetro del modelo al que se asocia sea cero. En la última columna de esa tabla aparece la significación del parámetro. En este caso se aprecia que tanto el intercepto como el coeficiente asociado a la variable IDH son diferentes de cero. Como consecuencia, la ecuación de la recta estimada que predice la satisfacción con la vida en función del índice de desarrollo humano quedaría de la siguiente forma

$$\widehat{SWL} = 112,522 + 134,436 \times IDH.$$

En la línea 16 aparecen dos versiones del estadístico R^2 que se utilizan para evaluar la bondad de ajuste global del modelo de regresión lineal mientras que en la línea 17 aparece el estadístico F de Snedecor que testa la hipótesis nula de que el modelo que contiene la variable predictora IDH predice mejor a la variable SWL que usar la media de ésta última variable para hacer las predicciones. Si tuviésemos que hacer una descripción verbal del modelo generado podríamos decir que por cada unidad que aumenta el IDH la SLW aumenta en 134,436 unidades. Por otro lado, cuando el IDH es cero la satisfacción con la vida tiene un valor de 112,522 puntos.

Una vez que hemos estimado un modelo podemos, como hicimos con el caso del análisis de la varianza unifactorial, realizar una serie de tests diagnósticos sobre el modelo o, incluso, comparar diferentes modelos en su habilidad para predecir la variable de respuesta. Consideremos ahora la posibilidad de estimar un modelo de regresión lineal múltiple. Sin preocuparnos ahora mismo por los supuestos o requisitos técnicos necesarios (como he comentado al principio del capítulo, recomiendo que se acceda a manuales especializados en estos temas para aclarar estas ideas) para que el análisis goce de calidad estadística (como, por ejemplo, en lo relativo al problema de la multicolinealidad), podríamos tratar de estimar la ecuación que predice la satisfacción con la vida en función del IDH , del PIV y del IPG . Si solicitamos este análisis en el cuadro de diálogo que aparece en la Figura 8.4 obtendríamos este resultado:

8.2 - Introducción a la regresión lineal

```
1 Call:
2 lm(formula = SWL ~ IDH + IPG + PIB, data = ecopaz)
3
4 Residuals:
5     Min       1Q   Median       3Q      Max
6 -80.350 -18.633   5.595  20.767  61.678
7
8 Coefficients:
9             Estimate Std. Error t value Pr(>|t|)
10 (Intercept) 184.246932  23.577523   7.815 1.53e-12 ***
11 IDH          94.117991  18.577467   5.066 1.34e-06 ***
12 IPG         -25.063382   7.192607  -3.485 0.000669 ***
13 PIB           0.001621   0.001644   0.986 0.325982
14 ---
15 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
16
17 Residual standard error: 30.41 on 132 degrees of freedom
18 (38 observations deleted due to missingness)
19 Multiple R-squared:  0.3998, Adjusted R-squared:  0.3861
20 F-statistic: 29.3 on 3 and 132 DF,  p-value: 1.381e-14
```

Como se puede apreciar en la salida que ha generado el programa el *PIB* no contribuye significativamente al modelo de regresión lineal múltiple que serviría para predecir la satisfacción con la vida ($p = 0,326$) mientras que el parámetro asociado al *IPG* sí alcanza a ser estadísticamente diferente de cero. En cualquier caso, como aparece en las líneas 19 y 20, el modelo sigue teniendo una bondad de ajuste aceptable y la ecuación que representa la relación entre las variables sería la siguiente

$$\widehat{SWL} = 184,25 + 94,12 \times IDH - 25,06 \times IPG + 0,002 \times PIB.$$

□ EJERCICIOS □

1. Estima el modelo de regresión lineal que predice el Producto Interior Bruto (*PIB*) de un país en función de la satisfacción con la vida que experimentan sus ciudadanos (*SWL*). ¿Qué conclusiones extraes?
2. Estima el modelo de regresión lineal que predice el *PIB* en función del *IPG*, del *IDH* y del *SWL*. Escribe la ecuación de la recta y reflexiona sobre los resultados que obtienes.

9

Creación y manipulación de gráficas

Para muchas personas, como es mi caso, uno de los mayores atractivos que presenta \mathbb{R} se concreta en las opciones y potencialidades gráficas que ofrece. Aunque enfrentarse a la creación de gráficos con \mathbb{R} puede atemorizarnos en un primer momento por las líneas de código que tenemos que manejar, lo cierto es que los resultados que se pueden conseguir son tremendamente llamativos y espectaculares.

La gestión y creación de gráficos con \mathbb{R} es un mundo. Existen multitud de posibilidades y opciones que se pueden personalizar en los gráficos que generamos con este software. Es más, podemos generar nuestros propios tipos de gráficos personalizados. Por ello, aquí sólo se dedicarán unas pocas páginas a describir ligeramente algunas de las características generales sobre la creación de gráficos con \mathbb{R} y \mathbb{R}_{char} para que el lector interesado pueda continuar avanzando en su autoaprendizaje sobre este tema. Recomiendo encarecidamente que se acceda a la introducción a \mathbb{R} que se puede encontrar en Venables et al. (2011) donde aparece una sección dedicada específicamente a la creación y gestión de gráficos.

Hay un par de cosas interesantes que creo conveniente comentar cuando tra-

Capítulo 9 - Creación y manipulación de gráficas






tamos el tema de la generación de gráficos con . En primer lugar, hay que destacar que cuando creamos un gráfico con  o  se abre una nueva ventana donde se proyectarán los gráficos que vayamos generando. Este *visor gráfico* tiene un nuevo menú (Figura 9.1) desde donde se puede guardar el gráfico en diferentes formatos o modificar sus dimensiones.



Figura 9.1: Menú del visor gráfico en R.

Por otro lado, es conveniente tener en mente que existen tres tipos de comandos que pueden ser utilizados para producir los gráficos en . En primer lugar, los *comandos de alto nivel* son aquellos que crean un gráfico totalmente nuevo sobre el visor de gráficos. Por su parte, los *comandos de bajo nivel* añaden información a los gráficos previamente creados mientras que los *comandos de interacción* sirven para añadir o extraer información interactivamente del gráfico que está proyectado sobre el visor. En este capítulo se introducirán brevemente algunos de estos tipos de comandos y se comentarán algunas de las funcionalidades que proporciona  para generar gráficos.

9.1. Comandos de alto nivel

Los comandos de alto nivel, como se ha comentado anteriormente, generan un gráfico totalmente nuevo y reemplazan (si es que tenemos alguno) el gráfico existente en el visor de gráficos.

Un ejemplo de comando de alto nivel es la función `plot()` que se comporta de manera diferente dependiendo del tipo de vectores o variables que contenga como argumentos. Por ejemplo, si los argumentos de la función son dos vectores numéricos se genera un gráfico de dispersión de las variables. Por ejemplo, la sintaxis:

```
1 x <- -100:100
2 y <- x^2
3 plot(x,y)
```

9.1 - Comandos de alto nivel

generará un gráfico de dispersión donde se representa la función $y = x^2$. No obstante, si introducimos una variable tipo factor como argumento de la función se creará un gráfico de barras. Por ejemplo, imagina que estamos haciendo un seguimiento de los errores que comente un niño al escribir las letras A, V y R. Si creamos un vector (línea 1 del código que aparece más abajo) donde cada letra significa que el niño ha cometido un error de escritura podemos crear un gráfico de barras utilizando la función `plot()`¹ de esta manera:

```
1 letras <- c("R","V","V","V","V","V","V","V","V","R","R","R","R","A","A","A")
2 f <- as.factor(letras)
3 plot(f)
```

Otro ejemplo básico de comando de alto nivel es la función `hist()`. Esta función crea un histograma de la variable que tiene como argumento. Por ejemplo, la siguiente sintaxis genera un objeto (línea 1) que es un vector de 1000 valores aleatorios que siguen una distribución normal (con media 0 y desviación típica 1) y luego genera un histograma con ese vector

```
1 x <- rnorm(1000)
2 hist(x)
```

La función `boxplot()`, por su parte, genera un diagrama de caja sobre la variable que toma como argumento. Si ejecutas esta función sobre el objeto `x` que hemos creado en la función anterior obtendrás un gráfico similar al que aparece en la Figura 9.2².

Una característica interesante de la función `plot()` es que si el primero de los argumentos es una variable de tipo factor se crearán diagramas de caja por cada nivel del factor. Por ejemplo, si ejecutamos la siguiente sintaxis que carga el conjunto de datos `iris` (línea 1) que está contenido en \mathbb{R} , generaremos un gráfico de caja por cada tipo de flor (setosa, versicolor y virgínica) para la longitud del sépalo:

```
1 data(iris)
2 attach(iris)
3 plot(Species, Sepal.Length)
```

¹La función `barplot()` también genera un gráfico de barras pero a partir de vectores numéricos que indican la altura de las barras.

²No será exactamente el mismo dado que el vector de números aleatorios no será diferente en ambos casos pero, a grandes rasgos, será muy parecido.

Capítulo 9 - Creación y manipulación de gráficas

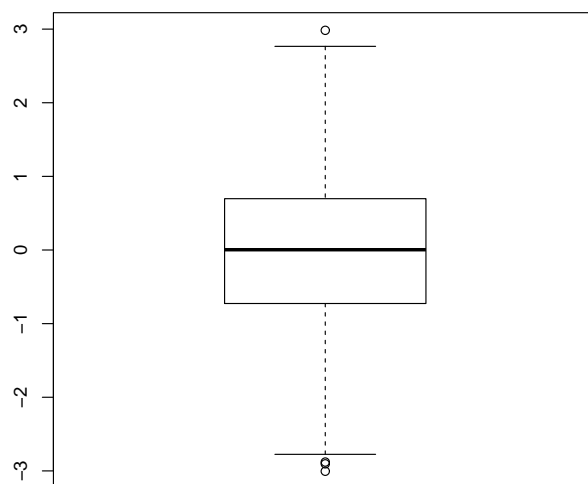


Figura 9.2: Ejemplo de diagrama de caja en R.

Los comandos de alto nivel pueden ser modificados añadiendo diferentes parámetros que controlan las propiedades de cada tipo de gráfico. Por ejemplo, la siguiente función modifica el gráfico que hemos generado anteriormente de la función $y = x^2$ sustituyendo los puntos por una línea:

```
1 x <- -100:100
2 y <- x^2
3 plot(x,y,type="l")
```

Los parámetros gráficos que controlan los títulos de los ejes de coordenadas x e y son `xlab` e `ylab` respectivamente. También podemos cambiar el título principal con el parámetro `main` y podemos personalizar el color de fondo con el parámetro `bg`. Por ejemplo, a continuación aparece una sintaxis donde se obtiene un histograma como el que se ha generado previamente, pero con algunos parámetros que personalizan la apariencia del gráfico. Los parámetros están comentados para aclarar el aspecto del gráfico que controlan:

```
1 x <- rnorm(1000)
2 hist(x,
3 main= "Histograma personalizado", # Título principal
```

```

4 xlab="Variable aleatoria normal (n=1000)", # Título del eje x
5 ylab="Frecuencia", # Título del eje y
6 col="blue", # Color de las barras
7 border ="green" # Color del borde de las barras
8 )

```

Cada comando gráfico de alto nivel tendrá parámetros propios aunque algunos son comunes a muchos de ellos. Se recomienda acceder a la documentación específica de cada función para controlar el aspecto que queremos dar a los gráficos que se quieran generar.

9.2. Comandos de bajo nivel

En numerosas ocasiones los gráficos que se generan con los comandos de alto nivel no satisfacen nuestras necesidades, incluso aunque hayamos modificado parámetros de la función gráfica. Por ello, \mathbb{R} proporciona funciones que permiten añadir elementos a nuestros gráficos con el objetivo de que los personalizemos a nuestro antojo.

Por ejemplo, podemos añadir textos, símbolos o líneas a los gráficos que generamos. En el código que aparece a continuación se crea un gráfico donde se representa la función $y = \frac{1}{x}$ y donde se añaden ciertos elementos que aparecen comentados:

```


1 x <- -100:100
2 y <- 1/x
3 plot(x,y,type="l")
4 points(50,1) # Añade un punto en la coordenada (50,1)
5 legend(-80,0.75,legend="Función y = 1/x") # Añade una leyenda donde se
   especifica la función representada en la coordenada (-80,0,75)
6 title(main="Gráfico de Ejemplo", sub="Función matemática") # Añade un título y
   un subtítulo al gráfico
7 abline(0,0.005) # Añade una línea con intercepto 0 y con pendiente 0,005

```

9.3. Personalización de parámetros gráficos

Por lo general, nuestras necesidades o nuestras preferencias estéticas hacen que los gráficos que han sido generados por defecto con \mathbb{R} tengan que ser modificados

Capítulo 9 - Creación y manipulación de gráficas


para hacerlos más ajustados a nuestros deseos. La personalización de los gráficos se realiza utilizando lo que denominamos como *parámetros gráficos*.  permite la manipulación de un gran número de parámetros que permitan ajustar la figura del visor gráfico a nuestras necesidades. Parámetros como el estilo de las líneas, el color, el tipo de letra o la justificación de los textos son características que pueden ser manipuladas en cada gráfico. Cuando creamos un gráfico, nos aparecerá en el visor de gráficos con unas características por defecto que se pueden cambiar temporalmente (afectando sólo al gráfico concreto que hemos generado) o permanentemente (afectando a todos los gráficos que creamos en una sesión determinada).


La función `par()` se utiliza para introducir cambios permanentes en los gráficos que mandamos al visor gráfico. Así, si queremos generar gráficos que tengan un aspecto similar podemos utilizar esta función que nos garantizará que todos los gráficos generados serán similares. Por ejemplo, si utilizamos la siguiente sintaxis (demasiado esperpéntica, por cierto) podremos obligar a que todos los gráficos generados tengan las propiedades que definen los siguientes parámetros:

```
1 par(bg="violet", # Define el color de fondo del gráfico
2 col.lab="red", # Define el color de las etiquetas de los ejes
3 font.axis=6, # Define el tipo de letra de los ejes
4 font.main=3, # Define el tipo de letra del título principal
5 col.main="yellow", # Define el color del título principal
6 font.lab=11) # Define el tipo de letra de las etiquetas de los ejes
```

Puedes probar estos parámetros, variando sus valores si te apetece, para ver cómo afectan al gráfico. También puedes probar con diferentes tipos de gráficos dado que dependiendo del gráfico que generes tendrás uno u otro resultado.

9.4. Facilidades que proporciona R Commander

 permite generar gráficos a través del uso de cuadros de diálogo que facilitan su creación. Una vez generados, se pueden modificar sus parámetros o añadir más a la sintaxis que crea el programa.

Todos los tipos de gráficos que podemos generar con  los encontramos en el menú *Gráficas* (aunque también se pueden generar, por ejemplo, gráficos básicos de diagnóstico de la bondad de ajuste de los modelos desde el menú *Modelos*).

9.4 - Facilidades que proporciona R Commander

Así, desde este menú, podemos generar gráficos de series temporales, histogramas, de sectores, de barras, de medias, de caja, de dispersión o gráficos tridimensionales que podemos rotar e inspeccionar desde diferentes perspectivas. Para cada tipo de gráfico tendremos un cuadro de diálogo en el que podremos definir sus características.

Capítulo 9 - Creación y manipulación de gráficas

Referencias

- Ajzen, I., y Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood, NJ: Prentice-Hall.
- Ajzen, I., y Fishbein, M. (2005). The influence of attitudes on behavior. En D. Albarracín, B. T. Hohnson, y M. P. Zanna (Eds.), *The hadnbook of attitudes* (pp. 173–221). Mahwah, NJ: Erlbaum.
- Allport, G. W. (1935). Attitudes. En C. Murchison (Ed.), *Handbook of social psychology* (pp. 798–844). Worcester, MA: Clark University Press.
- Arriaza, A. J., Fernández, F., López, M. A., Muñoz, M., Pérez, S., y Sánchez, A. (2008). *Estadística básica con R y R-Commander*. Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.
- Bachrach, A. J. (1994). *Cómo investigar en psicología* (4^a ed.). Madrid: Morata. (Trabajo original publicado en 1966)
- Bardin, J. (2012, Marzo 22). Making connections. Is a project to map the brain's ful communications network worth the money? *Nature*, *483*, 394–396. doi: 10.1038/483394a.
- Bond, M. (2009, Octubre 28). Decision-making: risk school. *Nature*, *461*, 1189–1192. doi: 10.1038/4611189a.
- Carlson, N. R. (2000). *Fisiología de la conducta* (3^a ed.). Barcelona: Ariel. (Trabajo original publicado en 1993)
- Computer Music. (1999). Software pirata realidad y mito. *Computer Music*, *4*, 55–61.
- Cook, D. A., y Beckman, T. J. (2006). Current concepts in validity and reliability for psychometrics instruments: theory and application. *The American Journal of Medicine*, *119*, 166e7–166e16. doi: 10.1016/j.amjmed.2005.10.036.
- De la Fuente, E. I. (1998). Presentación. En E. I. De la Fuente y J. García

Referencias

- (Eds.), *Análisis de datos en psicología: ejercicios de estadística descriptiva* (pp. 5–6). Granada: Urbano Delgado, J. C.
- de Leeuw, J., y Mair, P. (2007). An introduction to the special volume on “psychometrics in R”. *Journal of Statistical Software*, 20, 1–5.
- Elosua, P. (2009). ¿Existe vida más allá de SPSS? Descubre R. *Psicothema*, 21, 652–655.
- Elosua, P. (2011). *Introducción al entorno R*. Bilbao: Euskal Herriko Unibertsitateko Argitalpen Zerbitzua / Servicio Editorial de la Universidad del País Vasco.
- Elosua, P., y Etxeberria, J. (2012). *R Commander. Gestión y análisis de datos*. Madrid: La Muralla.
- Field, A. (2009). *Discovering statistics* (3^a ed.). Londres: SAGE.
- García, J., De la Fuente, L., y Martín, E. (1998). Transformaciones en los datos de investigación. En E. I. De la Fuente y J. García (Eds.), *Análisis de datos en psicología: ejercicios de estadística descriptiva* (pp. 56–63). Granada: Urbano Delgado, J. C.
- Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.
- Hair, J. F., Anderson, R. E., Tatham, R. L., y Black, W. C. (1998). *Multivariate data analysis*. Englewood Cliffs, NY: Prentice Hall.
- Hothersall, D. (1997). *Historia de la psicología* (3^a ed.). México: McGraw-Hill. (Trabajo original publicado en 1995)
- Jovel, A. J. (1995). *Análisis de regresión logística*. Madrid: Centro de Investigaciones Sociológicas.
- León, O. G., y Montero, I. (2003). *Métodos de investigación en psicología y educación* (3^a ed.). Madrid: McGraw-Hill.
- López, J. (2009). Modelos predictivos en actitudes emprendedoras. Análisis comparativo de las condiciones de ejecución de las redes bayesianas y la regresión logística (Tesis Doctoral, Facultad de Psicología). *Repositorio Institucional de la Universidad de Almería*. URI: <http://hdl.handle.net/10835/356>.
- López, J. (2012). Evolución de la reflexión cognitiva en la universidad. *Boletín de la Titulación de Matemáticas de la UAL*, 5, 17–18.
- Mair, P., y Hatzinger, R. (2007). Psychometrics task view. *R News*, 7, 38–40.
- Pabón, J. M. (1997). *Diccionario manual Griego-Español* (9^a ed.). Barcelona: Vox.

- Pagano, R. R. (1999). *Estadística para las ciencias del comportamiento* (5ª ed.). Madrid: Thomson. (Trabajo original publicado en 1998)
- Pardo, A., Ruiz, M. A., y San Martín, R. (2007). Cómo ajustar e interpretar modelos multinivel con SPSS. *Psicothema*, *19*, 308–321.
- Pinel, J. P. J. (2011). *Biopsychology* (8ª ed.). Boston, MA: Allyn & Bacon.
- R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. (<http://www.R-project.org>. ISBN: 3-900051-07-0)
- Scheines, R., Spirtes, P., Glymour, C., Meek, C., y Richardson, T. (2005). TETRAD 3: tools for causal modeling. User's manual. Descargado el 14 de Febrero de 2005, desde <http://www.phil.cmu.edu/projects/tetrad/>.
- Sáez, J. A. (2010). Métodos estadísticos con R y R commander. Descargado el 15 de Diciembre de 2011, desde <http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf>.
- Silva, L. C., y Barroso, I. M. (2004). *Regresión logística*. Madrid: La Muralla / Hespérides.
- Solanas, A., Salafranca, L., Fauquet, J., y Núñez, M. I. (2005). *Estadística descriptiva en ciencias del comportamiento*. Madrid: Thomson.
- Spirtes, P., Glymour, C., y Scheines, R. (2000). *Causation, prediction and search* (2ª ed.). Cambridge, MA: MIT Press.
- Spirtes, P., Scheines, R., Glymour, C., Richardson, T., y Meek, C. (2004). Causal inference. En D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 447–477). Thousand Oaks, CA: Sage Publications.
- Stevens, S. S. (1946, Junio 7). Theory of scales of measurement. *Science*, *103*, 677–680.
- Thompson, S. C. G., y Barton, M. A. (1994). Ecocentric and anthropocentric attitudes toward the environment. *Journal of Environmental Psychology*, *14*, 149–157. doi: 10.1016/S0272-4944(05)80168-9.
- Valero-Mora, P., y Ledesma, R. (2012, Junio). *Graphical user interfaces for R: a summary of the state of the art*. Comunicación presentada en el V European Congress of Methodology. Santiago de Compostela.
- Venables, W., Smith, D. M., y the R Development Core Team. (2011). *An introduction to R. Notes on R: a programming environment for data*

Referencias

analysis and graphics. Versión 2.15.0: Descargado desde: <http://cran.r-project.org/doc/manuals/R-intro.pdf>.