Estadística

Básica

con

R y R–Commander

Estadística Básica con R y R–Commander

Autores:
A. J. Arriaza Gómez
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
S. Pérez Plaza
A. Sánchez Navas

UCA Universidad de Cádiz
Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN:

Depósito legal:

Índice general

Prólogo

1. Introducción

La Universidad de Cádiz es pionera en España en la búsqueda de soluciones de conocimiento abierto, consciente de que es la forma más eficiente de lograr sus objetivos institucionales relacionados con la docencia y la investigación. En concreto, el Punto 1 del Artículo 2 de sus Estatutos, que describe los fines esenciales de la institución, establece como objetivo fundamental: “La creación, desarrollo, transmisión y crítica de la ciencia, la técnica y la cultura y su integración en el patrimonio intelectual heredado”. Mientras que en el Punto 6 del mismo artículo dice: “Acoger, defender y promover los valores sociales e individuales que le son propios, tales como la libertad, el pluralismo, el respeto de las ideas y el espíritu crítico, así como la búsqueda de la verdad”.

La creación de la Oficina de Software Libre (OSLUCA) el 15 de marzo de 2004, la aprobación de la Normativa para el intercambio de información institucional el 27 de septiembre de 2004 y la utilización de herramientas de formato abierto en las aplicaciones de comunicación y gestión de la Universidad, son actuaciones que ponen de manifiesto el decidido apoyo del Equipo de Gobierno de la UCA a las soluciones basadas en formatos abiertos.

Desde un plano mucho más modesto, bajo el auspicio del Vicerrectorado de Tecnologías de la Información e Innovación Docente y a través de la `Oficina de Software Libre de la Universidad de`

`Cádiz (OSLUCA)`, nace el Proyecto R UCA. Dicho proyecto, cuyas líneas principales de actuación pueden consultarse en la página web del proyecto `http://knuth.uca.es/R`, contempla, entre otras acciones, la elaboración de material para la docencia y la investigación, siendo en el primero de estos aspectos, el docente, en el que se enmarca este manual.

En la misma línea que nuestros órganos de gobierno, pensamos que una institución como la Universidad debe preocuparse por proveer a sus miembros de las mejores herramientas para desarrollar su tarea, en aras de la mejora global del conocimiento. Pero la creación de conocimiento se verá muy mermada si se emplean soluciones tecnológicas que se ofrecen como cajas negras, es decir que no pueden ser analizadas ni modificadas, y que además limita fuertemente el uso que se haga de los resultados que se consigan a partir de ellas.

El uso de software propietario en áreas como la Estadística, donde existen alternativas con igual o mejor calidad con licencia libre, no sólo tiene consecuencias negativas desde un punto de vista económico, sino que supone un auténtico "harakiri" intelectual, porque limita el ejercicio de uno de los aspectos que mejor caracterizan a nuestra institución: su espíritu analítico y crítico, ¿cómo se va a fomentar ese espíritu con el uso de herramientas absolutamente herméticas?, y si alguien consiguiera descifrarlas y manipularlas se convertiría formalmente en un delincuente.

Centrándonos en los aspectos intrínsecos de la cuestión, cuando nos planteamos confeccionar este manual, tuvimos claro que no queríamos enseñar a manejar un programa, sino a hacer análisis estadísticos con el apoyo de una herramienta que facilitara el cálculo y la aplicación de los procedimientos. De ahí el nombre del libro: "Estadística básica con R y Rcmdr".

La decisión de elegir **R** fue fácil, ningún otro programa en la actualidad reúne las condiciones de madurez, cantidad de recursos y manejabilidad que posee **R**, además de ser el que tiene una mayor implantación en la comunidad científica. El incorporar la interfaz gráfica de usuario (GUI) **Rcmdr** pretende, en primera instancia, facilitar el manejo de **R** y, en segundo lugar, servir como generador de instrucciones **R**. Es posible que muchos de nuestros alumnos no necesiten otro nivel de uso que el que proporciona **Rcmdr**, pero unos pocos y la mayoría del personal investigador, una vez superado el respeto inicial a la herramienta, se decantarán por manejarse directamente con la consola de **R**, creando y editando instrucciones con una evidente economía de recursos y, lo que es más importante, con un control total sobre los procedimientos que en cada momento se van a aplicar.

Respecto a los contenidos, el libro pretende abarcar las necesidades prácticas de un programa básico de estadística, y así, salvo el primer capítulo, donde se presenta de forma muy sucinta el software, el resto está dedicado a los tópicos habituales de un curso introductorio: Análisis Exploratorio en una y dos Dimensiones, Distribuciones de Probabilidad, Inferencia Paramétrica y no Paramétrica y Análisis de la Varianza de un Factor. El esquema de presentación de los temas incluye una breve descripción de los conceptos, la resolución de una serie de ejemplos con la ayuda de **R** y la propuesta de ejercicios para evaluar los conocimientos adquiridos.

Al objeto de facilitar el uso del software, los primeros capítulos están soportados básicamente sobre la interfaz **Rcmdr**. A partir del capítulo 5 aumenta el uso de funciones construidas directamente en el indicador de mandatos, en parte por necesidad y en parte por motivos estratégicos, puesto que para entonces consideramos que nuestros alumnos están bien familiarizados con la sintaxis de las funciones de **R**.

Esperamos que este manual sea de utilidad y, en cualquier caso y con más motivos, dado que se trata de la primera versión, ponemos nuestro trabajo a disposición de la comunidad científica para que se hagan las mejoras, ampliaciones y adaptaciones que se deseen.

Los autores,

2. History (Histórico)

Este libro surge como material de apoyo a un curso de estadística básica con R. La génesis está en la creación del proyecto R UCA en mayo del 2007 y su primera versión ve la luz en enero de ese mismo año. Los autores en orden alfabético inverso son *Antonio Sánchez Navas*, *Sonia Pérez Plaza*, *Manuel Muñoz Márquez*, *María Auxiliadora López Sánchez*, *Fernando Fernández Palacín* y *Antonio Jesús Arriaza Gómez*.

Una versión electrónica de este documento se encuentra en:
`http://knuth.uca.es/ebrcmdr`

3. Licencia de Documentación Libre de GNU

This is an unofficial translation of the GNU Free Documentation License (Version 1.2, Noviembre 2002) into Spanish. It was not published by the Free Software Foundation, and does not legally state the distribution terms for documentation that uses the GNU FDL – only the original English text of the GNU FDL does that. However, we hope that this translation will help Spanish speakers understand the GNU FDL better.

Ésta es una traducción no oficial de la GNU Free Document License (Versión 1.2, Noviembre 2002) a Español (Castellano). No ha sido publicada por la Free Software Foundation y no establece legalmente los términos de distribución para trabajos que usen la GFDL (sólo el texto de la versión original en Inglés de la GFDL lo hace). Sin embargo, esperamos que esta traducción ayude los hispanohablantes a entender mejor la GFDL. La versión original de la GFDL esta disponible en la Free Software Foundation. `http://www.gnu.org/copyleft/fdl.html` Esta traducción está basada en una de la versión 1.1 de Igor Támara y Pablo Reyes. Sin embargo la responsabilidad de su interpretación es de Joaquín Seoane.

Copyright (C) 2000, 2001, 2002 Free Software Foundation, Inc. 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA. Se permite la copia y distribución de copias literales de este documento de licencia, pero no se permiten cambios[1].

Preámbulo

El propósito de esta Licencia es permitir que un manual, libro de texto, u otro documento escrito sea **"libre"** en el sentido de libertad: asegurar a todo el mundo la libertad efectiva de copiarlo y redistribuirlo, con o sin modificaciones, de manera comercial o no. En segundo término, esta Licencia proporciona al autor y al editor[2] una manera de obtener reconocimiento por su trabajo, sin que se le considere responsable de las modificaciones realizadas por otros.

Esta Licencia es de tipo **"copyleft"**, lo que significa que los trabajos derivados del documento deben a su vez ser libres en el mismo sentido. Complementa la Licencia Pública General de GNU, que es una licencia tipo copyleft diseñada para el software libre.

[1]Ésta es la traducción del Copyright de la Licencia, no es el Copyright de esta traducción no autorizada.

[2]La licencia original dice **"publisher"**, que es, estrictamente, quien publica, diferente de editor, que es más bien quien prepara un texto para publicar. En castellano editor se usa para ambas cosas.

Hemos diseñado esta Licencia para usarla en manuales de software libre, ya que el software libre necesita documentación libre: un programa libre debe venir con manuales que ofrezcan la mismas libertades que el software. Pero esta licencia no se limita a manuales de software; puede usarse para cualquier texto, sin tener en cuenta su temática o si se publica como libro impreso o no.

Recomendamos esta licencia principalmente para trabajos cuyo fin sea instructivo o de referencia.

1. Aplicabilidad y definiciones

Esta Licencia se aplica a cualquier manual u otro trabajo, en cualquier soporte, que contenga una nota del propietario de los derechos de autor que indique que puede ser distribuido bajo los términos de esta Licencia. Tal nota garantiza en cualquier lugar del mundo, sin pago de derechos y sin límite de tiempo, el uso de dicho trabajo según las condiciones aquí estipuladas. En adelante la palabra **"Documento"** se referirá a cualquiera de dichos manuales o trabajos. Cualquier persona es un licenciatario y será referido como **"Usted"**. Usted acepta la licencia si copia. modifica o distribuye el trabajo de cualquier modo que requiera permiso según la ley de propiedad intelectual.

Una **"Versión Modificada"** del Documento significa cualquier trabajo que contenga el Documento o una porción del mismo, ya sea una copia literal o con modificaciones y/o traducciones a otro idioma.

Una **"Sección Secundaria"** es un apéndice con título o una sección preliminar del Documento que trata exclusivamente de la relación entre los autores o editores y el tema general del Documento (o temas relacionados) pero que no contiene nada que entre directamente en dicho tema general (por ejemplo, si el Documento es en parte un texto de matemáticas, una Sección Secundaria puede no explicar nada de matemáticas). La relación puede ser una conexión histórica con el tema o temas relacionados, o una opinión legal, comercial, filosófica, ética o política acerca de ellos.

Las **"Secciones Invariantes"** son ciertas Secciones Secundarias cuyos títulos son designados como Secciones Invariantes en la nota que indica que el documento es liberado bajo esta Licencia. Si una sección no entra en la definición de Secundaria, no puede designarse como Invariante. El documento puede no tener Secciones Invariantes. Si el Documento no identifica las Secciones Invariantes, es que no las tiene.

Los **"Textos de Cubierta"** son ciertos pasajes cortos de texto que se listan como Textos de Cubierta Delantera o Textos de Cubierta Trasera en la nota que indica que el documento es liberado bajo esta Licencia. Un Texto de

Cubierta Delantera puede tener como mucho 5 palabras, y uno de Cubierta Trasera puede tener hasta 25 palabras.

Una copia **“Transparente”** del Documento, significa una copia para lectura en máquina, representada en un formato cuya especificación está disponible al público en general, apto para que los contenidos puedan ser vistos y editados directamente con editores de texto genéricos o (para imágenes compuestas por puntos) con programas genéricos de manipulación de imágenes o (para dibujos) con algún editor de dibujos ampliamente disponible, y que sea adecuado como entrada para formateadores de texto o para su traducción automática a formatos adecuados para formateadores de texto. Una copia hecha en un formato definido como Transparente, pero cuyo marcaje o ausencia de él haya sido diseñado para impedir o dificultar modificaciones posteriores por parte de los lectores no es Transparente. Un formato de imagen no es Transparente si se usa para una cantidad de texto sustancial. Una copia que no es **“Transparente”** se denomina **“Opaca”**.

Como ejemplos de formatos adecuados para copias Transparentes están ASCII puro sin marcaje, formato de entrada de Texinfo, formato de entrada de LaTeX, SGML o XML usando una DTD disponible públicamente, y HTML, PostScript o PDF simples, que sigan los estándares y diseñados para que los modifiquen personas. Ejemplos de formatos de imagen transparentes son PNG, XCF y JPG. Los formatos Opacos incluyen formatos propietarios que pueden ser leídos y editados únicamente en procesadores de palabras propietarios, SGML o XML para los cuáles las DTD y/o herramientas de procesamiento no estén ampliamente disponibles, y HTML, PostScript o PDF generados por algunos procesadores de palabras sólo como salida.

La **“Portada”** significa, en un libro impreso, la página de título, más las páginas siguientes que sean necesarias para mantener legiblemente el material que esta Licencia requiere en la portada. Para trabajos en formatos que no tienen página de portada como tal, **“Portada”** significa el texto cercano a la aparición más prominente del título del trabajo, precediendo el comienzo del cuerpo del texto.

Una sección **“Titulada XYZ”** significa una parte del Documento cuyo título es precisamente XYZ o contiene XYZ entre paréntesis, a continuación de texto que traduce XYZ a otro idioma (aquí XYZ se refiere a nombres de sección específicos mencionados más abajo, como **“Agradecimientos”**, **“Dedicatorias”**, **“Aprobaciones”** o **“Historia”**. **“Conservar el Título”** de tal sección cuando se modifica el Documento significa que permanece una sección **“Titulada XYZ”** según esta definición[3].

[3] En sentido estricto esta licencia parece exigir que los títulos sean exactamente **“Acknowledgements”**, **“Dedications”**, **“Endorsements”** e **“History”**, en

El Documento puede incluir Limitaciones de Garantía cercanas a la nota donde se declara que al Documento se le aplica esta Licencia. Se considera que estas Limitaciones de Garantía están incluidas, por referencia, en la Licencia, pero sólo en cuanto a limitaciones de garantía: cualquier otra implicación que estas Limitaciones de Garantía puedan tener es nula y no tiene efecto en el significado de esta Licencia.

2. Copia literal

Usted puede copiar y distribuir el Documento en cualquier soporte, sea en forma comercial o no, siempre y cuando esta Licencia, las notas de copyright y la nota que indica que esta Licencia se aplica al Documento se reproduzcan en todas las copias y que usted no añada ninguna otra condición a las expuestas en esta Licencia. Usted no puede usar medidas técnicas para obstruir o controlar la lectura o copia posterior de las copias que usted haga o distribuya. Sin embargo, usted puede aceptar compensación a cambio de las copias. Si distribuye un número suficientemente grande de copias también deberá seguir las condiciones de la sección 3.

Usted también puede prestar copias, bajo las mismas condiciones establecidas anteriormente, y puede exhibir copias públicamente.

3. Copiado en cantidad

Si publica copias impresas del Documento (o copias en soportes que tengan normalmente cubiertas impresas) que sobrepasen las 100, y la nota de licencia del Documento exige Textos de Cubierta, debe incluir las copias con cubiertas que lleven en forma clara y legible todos esos Textos de Cubierta: Textos de Cubierta Delantera en la cubierta delantera y Textos de Cubierta Trasera en la cubierta trasera. Ambas cubiertas deben identificarlo a Usted clara y legiblemente como editor de tales copias. La cubierta debe mostrar el título completo con todas las palabras igualmente prominentes y visibles. Además puede añadir otro material en las cubiertas. Las copias con cambios limitados a las cubiertas, siempre que conserven el título del Documento y satisfagan estas condiciones, pueden considerarse como copias literales.

Si los textos requeridos para la cubierta son muy voluminosos para que ajusten legiblemente, debe colocar los primeros (tantos como sea razonable colocar) en la verdadera cubierta y situar el resto en páginas adyacentes.

Si Usted publica o distribuye copias Opacas del Documento cuya cantidad exceda las 100, debe incluir una copia Transparente, que pueda ser leída

---

inglés.

por una máquina, con cada copia Opaca, o bien mostrar, en cada copia Opaca, una dirección de red donde cualquier usuario de la misma tenga acceso por medio de protocolos públicos y estandarizados a una copia Transparente del Documento completa, sin material adicional. Si usted hace uso de la última opción, deberá tomar las medidas necesarias, cuando comience la distribución de las copias Opacas en cantidad, para asegurar que esta copia Transparente permanecerá accesible en el sitio establecido por lo menos un año después de la última vez que distribuya una copia Opaca de esa edición al público (directamente o a través de sus agentes o distribuidores).

Se solicita, aunque no es requisito, que se ponga en contacto con los autores del Documento antes de redistribuir gran número de copias, para darles la oportunidad de que le proporcionen una versión actualizada del Documento.

4. Modificaciones

Puede copiar y distribuir una Versión Modificada del Documento bajo las condiciones de las secciones 2 y 3 anteriores, siempre que usted libere la Versión Modificada bajo esta misma Licencia, con la Versión Modificada haciendo el rol del Documento, por lo tanto dando licencia de distribución y modificación de la Versión Modificada a quienquiera posea una copia de la misma. Además, debe hacer lo siguiente en la Versión Modificada:

A. Usar en la Portada (y en las cubiertas, si hay alguna) un título distinto al del Documento y de sus versiones anteriores (que deberán, si hay alguna, estar listadas en la sección de Historia del Documento). Puede usar el mismo título de versiones anteriores al original siempre y cuando quien las publicó originalmente otorgue permiso.

B. Listar en la Portada, como autores, una o más personas o entidades responsables de la autoría de las modificaciones de la Versión Modificada, junto con por lo menos cinco de los autores principales del Documento (todos sus autores principales, si hay menos de cinco), a menos que le eximan de tal requisito.

C. Mostrar en la Portada como editor el nombre del editor de la Versión Modificada.

D. Conservar todas las notas de copyright del Documento.

E. Añadir una nota de copyright apropiada a sus modificaciones, adyacente a las otras notas de copyright.

F. Incluir, inmediatamente después de las notas de copyright, una nota de licencia dando el permiso para usar la Versión Modificada bajo los términos de esta Licencia, como se muestra en la Adenda al final de este documento.

G. Conservar en esa nota de licencia el listado completo de las Secciones Invariantes y de los Textos de Cubierta que sean requeridos en la nota de Licencia del Documento original.

H. Incluir una copia sin modificación de esta Licencia.

I. Conservar la sección Titulada **"Historia"**, conservar su Título y añadirle un elemento que declare al menos el título, el año, los nuevos autores y el editor de la Versión Modificada, tal como figuran en la Portada. Si no hay una sección Titulada **"Historia"** en el Documento, crear una estableciendo el título, el año, los autores y el editor del Documento, tal como figuran en su Portada, añadiendo además un elemento describiendo la Versión Modificada, como se estableció en la oración anterior.

J. Conservar la dirección en red, si la hay, dada en el Documento para el acceso público a una copia Transparente del mismo, así como las otras direcciones de red dadas en el Documento para versiones anteriores en las que estuviese basado. Pueden ubicarse en la sección **"Historia"**. Se puede omitir la ubicación en red de un trabajo que haya sido publicado por lo menos cuatro años antes que el Documento mismo, o si el editor original de dicha versión da permiso.

K. En cualquier sección Titulada **"Agradecimientos"** o **"Dedicatorias"**, Conservar el Título de la sección y conservar en ella toda la sustancia y el tono de los agradecimientos y/o dedicatorias incluidas por cada contribuyente.

L. Conservar todas las Secciones Invariantes del Documento, sin alterar su texto ni sus títulos. Números de sección o el equivalente no son considerados parte de los títulos de la sección.

M. Borrar cualquier sección titulada **"Aprobaciones"**. Tales secciones no pueden estar incluidas en las Versiones Modificadas.

N. No cambiar el título de ninguna sección existente a **"Aprobaciones"** ni a uno que entre en conflicto con el de alguna Sección Invariante.

O. Conservar todas las Limitaciones de Garantía.

Si la Versión Modificada incluye secciones o apéndices nuevos que califiquen como Secciones Secundarias y contienen material no copiado del Documento, puede opcionalmente designar algunas o todas esas secciones como invariantes. Para hacerlo, añada sus títulos a la lista de Secciones Invariantes en la nota de licencia de la Versión Modificada. Tales títulos deben ser distintos de cualquier otro título de sección.

Puede añadir una sección titulada **"Aprobaciones"**, siempre que contenga únicamente aprobaciones de su Versión Modificada por otras fuentes –por ejemplo, observaciones de peritos o que el texto ha sido aprobado por una organización como la definición oficial de un estándar.

Puede añadir un pasaje de hasta cinco palabras como Texto de Cubierta Delantera y un pasaje de hasta 25 palabras como Texto de Cubierta Trasera en la Versión Modificada. Una entidad solo puede añadir (o hacer que se añada) un pasaje al Texto de Cubierta Delantera y uno al de Cubierta Trasera. Si el Documento ya incluye textos de cubiertas añadidos previamente por usted o por la misma entidad que usted representa, usted no puede añadir otro; pero puede reemplazar el anterior, con permiso explícito del editor que agregó el texto anterior.

Con esta Licencia ni los autores ni los editores del Documento dan permiso para usar sus nombres para publicidad ni para asegurar o implicar aprobación de cualquier Versión Modificada.

5. Combinación de documentos

Usted puede combinar el Documento con otros documentos liberados bajo esta Licencia, bajo los términos definidos en la sección 4 anterior para versiones modificadas, siempre que incluya en la combinación todas las Secciones Invariantes de todos los documentos originales, sin modificar, listadas todas como Secciones Invariantes del trabajo combinado en su nota de licencia. Así mismo debe incluir la Limitación de Garantía.

El trabajo combinado necesita contener solamente una copia de esta Licencia, y puede reemplazar varias Secciones Invariantes idénticas por una sola copia. Si hay varias Secciones Invariantes con el mismo nombre pero con contenidos diferentes, haga el título de cada una de estas secciones único añadiéndole al final del mismo, entre paréntesis, el nombre del autor o editor original de esa sección, si es conocido, o si no, un número único. Haga el mismo ajuste a los títulos de sección en la lista de Secciones Invariantes de la nota de licencia del trabajo combinado.

En la combinación, debe combinar cualquier sección Titulada **"Historia"** de los documentos originales, formando una sección Titulada **"Historia"**; de la misma forma combine cualquier sección Titulada **"Agradecimientos"**, y cualquier sección Titulada **"Dedicatorias"**. Debe borrar todas las secciones tituladas **"Aprobaciones"**.

6. Colecciones de documentos

Puede hacer una colección que conste del Documento y de otros documentos liberados bajo esta Licencia, y reemplazar las copias individuales de esta Licencia en todos los documentos por una sola copia que esté incluida en la colección, siempre que siga las reglas de esta Licencia para cada copia literal de cada uno de los documentos en cualquiera de los demás aspectos.

Puede extraer un solo documento de una de tales colecciones y distribuirlo individualmente bajo esta Licencia, siempre que inserte una copia de esta Licencia en el documento extraído, y siga esta Licencia en todos los demás aspectos relativos a la copia literal de dicho documento.

7. Agregación con trabajos independientes

Una recopilación que conste del Documento o sus derivados y de otros documentos o trabajos separados e independientes, en cualquier soporte de almacenamiento o distribución, se denomina un **“agregado”** si el copyright resultante de la compilación no se usa para limitar los derechos de los usuarios de la misma más allá de lo que los de los trabajos individuales permiten. Cuando el Documento se incluye en un agregado, esta Licencia no se aplica a otros trabajos del agregado que no sean en sí mismos derivados del Documento.

Si el requisito de la sección 3 sobre el Texto de Cubierta es aplicable a estas copias del Documento y el Documento es menor que la mitad del agregado entero, los Textos de Cubierta del Documento pueden colocarse en cubiertas que enmarquen solamente el Documento dentro del agregado, o el equivalente electrónico de las cubiertas si el documento está en forma electrónica. En caso contrario deben aparecer en cubiertas impresas enmarcando todo el agregado.

8. Traducción

La Traducción es considerada como un tipo de modificación, por lo que usted puede distribuir traducciones del Documento bajo los términos de la sección 4. El reemplazo de las Secciones Invariantes con traducciones requiere permiso especial de los dueños de derecho de autor, pero usted puede añadir traducciones de algunas o todas las Secciones Invariantes a las versiones originales de las mismas. Puede incluir una traducción de esta Licencia, de todas las notas de licencia del documento, así como de las Limitaciones de Garantía, siempre que incluya también la versión en Inglés de esta Licencia y las versiones originales de las notas de licencia y Limitaciones de Garantía. En caso de desacuerdo entre la traducción y la versión original en Inglés de esta Licencia, la nota de licencia o la limitación de garantía, la versión original en Inglés prevalecerá.

Si una sección del Documento está Titulada **"Agradecimientos"**, **"Dedicatorias"** o **"Historia"** el requisito (sección 4) de Conservar su Título (Sección 1) requerirá, típicamente, cambiar su título.

9. Terminación

Usted no puede copiar, modificar, sublicenciar o distribuir el Documento salvo por lo permitido expresamente por esta Licencia. Cualquier otro intento de copia, modificación, sublicenciamiento o distribución del Documento es nulo, y dará por terminados automáticamente sus derechos bajo esa Licencia. Sin embargo, los terceros que hayan recibido copias, o derechos, de usted bajo esta Licencia no verán terminadas sus licencias, siempre que permanezcan en total conformidad con ella.

10. Revisiones futuras de esta licencia

De vez en cuando la Free Software Foundation puede publicar versiones nuevas y revisadas de la Licencia de Documentación Libre GNU. Tales versiones nuevas serán similares en espíritu a la presente versión, pero pueden diferir en detalles para solucionar nuevos problemas o intereses. Vea

`http://www.gnu.org/copyleft/`.

Cada versión de la Licencia tiene un número de versión que la distingue. Si el Documento especifica que se aplica una versión numerada en particular de esta licencia o **"cualquier versión posterior"**, usted tiene la opción de seguir los términos y codiciones de la versión especificada o cualquiera posterior que haya sido publicada (no como borrador) por la Free Software Foundation. Si el Documento no especifica un número de versión de esta Licencia, puede escoger cualquier versión que haya sido publicada (no como borrador) por la Free Software Foundation.

ADENDA: Cómo usar esta Licencia en sus documentos

Para usar esta licencia en un documento que usted haya escrito, incluya una copia de la Licencia en el documento y ponga el siguiente copyright y nota de licencia justo después de la página de título:

> Copyright (c) AÑO SU NOMBRE. Se concede permiso para copiar, distribuir y/o modificar este documento bajo los términos de la Licencia de Documentación Libre de GNU, Versión 1.2 o cualquier otra versión posterior publicada por la Free Software Foundation; sin Secciones Invariantes ni Textos de Cubierta Delantera ni Textos de Cubierta Trasera. Una copia de la licencia está incluida en la sección titulada GNU Free Documentation License.

Si tiene Secciones Invariantes, Textos de Cubierta Delantera y Textos de Cubierta Trasera, reemplace la frase **"sin ... Trasera"** por esto:

> siendo las Secciones Invariantes LISTE SUS TÍTULOS, siendo los Textos de Cubierta Delantera LISTAR, y siendo sus Textos de Cubierta Trasera LISTAR.

Si tiene Secciones Invariantes sin Textos de Cubierta o cualquier otra combinación de los tres, mezcle ambas alternativas para adaptarse a la situación.

Si su documento contiene ejemplos de código de programa no triviales, recomendamos liberar estos ejemplos en paralelo bajo la licencia de software libre que usted elija, como la Licencia Pública General de GNU (**"GNU General Public License"**), para permitir su uso en software libre.

4. GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free"in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The **"Document"**, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as **"you"**. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A **"Modified Version"** of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A **"Secondary Section"** is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The **"Invariant Sections"** are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The **"Cover Texts"** are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A **"Transparent"** copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called **"Opaque"**.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The **"Title Page"** means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page"means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section **"Entitled XYZ"** means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as **"Acknowledgements"**, **"Dedications"**, **"Endorsements"**, or **"History"**.) To **"Preserve the Title"** of such a section when you modify the Document means that it remains a section "Entitled XYZ.ªccording to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History"in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History"section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements.ºr "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled “Endorsements”. Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled “Endorsements.ºr to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version’s license notice. These titles must be distinct from any other section titles.

You may add a section Entitled “Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version.ªpplies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright ©YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Capítulo 1

Comenzando con R

1. Introducción

El que un libro que pretende incidir sobre los aspectos prácticos de la Estadística, comience con un capítulo dedicado al software, no debería sorprender, aun cuando en el Prólogo se haya dejado claro que no es un objetivo fundamental enseñar a manejar un programa informático. De hecho, este manual seguiría teniendo utilidad aun cuando se usara otra interfaz gráfica distinta a la que se propone o, incluso, otro software; bastaría en ese caso con acomodar los menús y/o la sintaxis. No obstante, el que existan varias soluciones informáticas, no quiere decir que optar por una de ellas no tenga un interés determinante y, por tanto, deben emplearse para su elección criterios objetivos de eficiencia, no solo de carácter estadístico, sino que atiendan también a su facilidad de uso.

Para la elección de **R** se han evaluado pues distintos aspectos, siendo especialmente destacables sus bondades en lo que se refiere a calidad, a la cantidad de técnicas y funciones implementadas, a que es libre y a la gran comunidad científica que lo usa como estándar para el análisis de datos. Dicha comunidad ha desarrollado y desarrolla herramientas integradas en paquetes–en la actualidad más de 800–, que dan solución a una gran variedad de problemas estadísticos.

R es un lenguaje de programación y un entorno para análisis estadístico y la realización de gráficos. Debido a su naturaleza es fácilmente adaptable a una gran variedad de tareas. Fue inicialmente escrito por Robert Gentleman y Ross Ihaka del *Departamento de Estadística* de la *Universidad de Auckland* en Nueva Zelanda. **R** actualmente es el resultado de un esfuerzo de colaboración de personas del todo el mundo. Desde mediados de 1997 se formó lo que se conoce como núcleo de desarrollo de **R**, que actualmente es el que tiene la posibilidad de modificación directa del código fuente. Por otra parte, **R** es un proyecto GNU similar a S, desarrollado éste por los *Laboratorios Bell*. Las diferencias entre **R** y S son importantes, pero la mayoría del código escrito para S corre bajo **R** sin modificaciones.

R abarca una amplia gama de técnicas estadísticas que van desde los modelos lineales a las más modernas técnicas de clasificación pasando por los test clásicos y el análisis de series temporales. Proporciona una amplia gama de gráficos que además son fácilmente adaptables y extensibles. La calidad de los gráficos producidos y la posibilidad de incluir en ellos símbolos y fórmulas matemáticas, posibilitan su inclusión en publicaciones que suelen requerir gráficos de alta calidad.

El código de **R** está disponible como software libre bajo las condiciones de la licencia GNU-GPL. Además está disponible precompilado para una multitud de plataformas. La página principal del proyecto es `http://www.r-project.org`.

Una diferencia importante entre **R**, y también `S`, con el resto del software estadístico es el uso del `objeto` como entidad básica. Cualquier expresión evaluada por **R** tiene como resultado un objeto. Cada objeto pertenece a una clase, de forma que las funciones pueden tener comportamientos diferentes en función de la clase a la que pertenece su objeto argumento. Por ejemplo, el resultado de la función `print` evaluada sobre un vector da como resultado la impresión de todos los elementos del vector mientras que la misma función evaluada sobre una función muestra información sobre ella. De la misma manera, la función `plot` no se comporta igual cuando su argumento es un vector que cuando es un fichero de datos o una función.

A continuación se dan unas breves instrucciones que permitirán comenzar a usar **R** y su interfaz gráfica **R-Commander**, que se denotará abreviadamente como **Rcmdr**. Instrucciones más detalladas y actualizadas pueden encontrarse en `http://knuth.uca.es/R` en la sección `R Wiki`. Por último, existen multitud de documentos que ilustran sobre el manejo de **R**, algunos de ellos pueden descargarse desde `http://knuth.uca.es/R` en la sección `Documentación`. Los autores de este manual han redactado un somero documento técnico sobre el uso de **R**, a cuyo repositorio puede accederse en la dirección `http://knuth.uca.es/R-basico`.

2. Instalación de R y R–Commander

2.1. Instalación en GNU/Linux

Para la instalación, distribuciones derivadas de debian (Ubuntu, Guadalinex,...), en una consola se introduce en una sola línea:
`sudo apt-get install r-base-html r-cran-rcmdr r-cran-rodbc r-doc-html r-recommended`

Otra opción es utilizar el gestor de paquetes de la propia distribución e instalar los paquetes `r-base-html`, `r-cran-rcmdr`, `r-cran-rodbc`, `r-doc-html` y `r-recommended`.

2.2. Instalación en Windows

La descarga de **R** en el equipo se efectua desde:
`http://cran.es.r-project.org/bin/windows/base/release.htm`

Luego se procede con la ejecución, siguiendo las instrucciones. Para la instalación de **Rcmdr**, se arranca **R** desde `Inicio→Todos los programas→` **R**. A continuación, `Paquetes→Instalar Paquete(s)` y elegido el mirror desde el cual se quiere instalar el paquete, por ejemplo *Spain (Madrid)*, se selecciona **Rcmdr**.

R–Nota 1.1

*Harán falta más paquetes para la instalación completa de **Rcmdr**, pero se instalarán automáticamente la primera vez que se ejecute.*

3. Ejecución de Rcmdr

En ambos sistemas operativos, la carga de la librería se efectuará mediante la instrucción `library("Rcmdr")`.

R–Nota 1.2

*Si se cierra **Rcmdr** (sin cerrar **R**), para volver a cargarlo se debe ejecutar la instrucción `Commander()`.*

Capítulo 2

Análisis Exploratorio de Datos Unidimensional

En este módulo, a través de una serie de medidas, gráficos y modelos descriptivos, se caracterizará a un conjunto de individuos, intentando descubrir regularidades y singularidades de los mismos y, si procede, comparar los resultados con los de otros grupos, patrones o con estudios previos. Se podría considerar que este estudio es una primera entrega de un estudio más completo o, por contra, tener un carácter finalista; en cualquier caso, se trata de un análisis calificable como de *exploratorio*, y de ahí el nombre del capítulo.

Las conclusiones obtenidas serán aplicables exclusivamente a los individuos considerados explícitamente en el estudio, sin que puedan hacerse extrapolaciones con validez científica fuera de ese contexto. Los resultados del Análisis Exploratorio de Datos (AED) sí que podrían emplearse para establecer hipótesis sobre individuos no considerados explícitamente en dicho análisis, que deberían ser posteriormente contrastadas.

Formalmente, se podría definir el AED como un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas; aunque esta primera entrega se centrará en un análisis de tipo unidimensional.

1. La organización de la información

Al conjunto de individuos físicos considerados en un análisis se le denominará *Colectivo* o *Población*, aunque también se utilizarán esos mismos términos para referirse a la(s) característica(s) de esos individuos que son objeto de estudio. De hecho, desde un punto de vista estadístico, los individuos sólo interesan como portadores de rasgos que son susceptibles de marcar diferencias entre ellos. La obtención y materialización en formato analógico o digital de las características consideradas constituirá el conjunto de datos que será estadísticamente analizado.

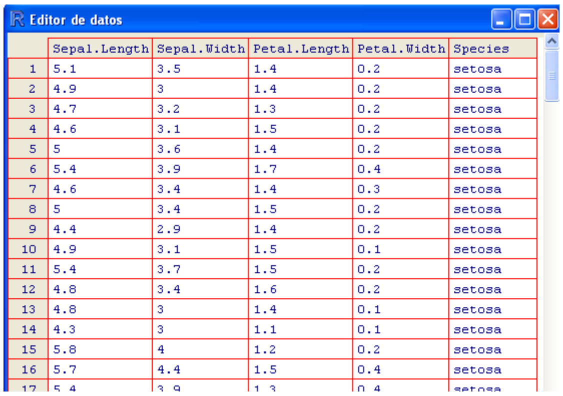
Los datos constituyen pues la materia prima de la Estadística, pudiéndose establecer distintas clasificaciones en función de la forma en que éstos vengan dados. Se obtienen datos al realizar cualquier tipo de prueba, experimento, valoración, medición, observación, ..., dependiendo de la naturaleza de los mismos y del método empleado para su obtención. Una vez obtenidos los datos por los procedimientos que se consideren pertinentes, pueden generarse nuevos datos mediante transformación y/o combinación de las variables originales. Al conjunto de datos convenientemente organizados se le llamará *modelo de datos*.

1.1. La matriz de datos

En una primera instancia se supondrá que, sobre un conjunto de n individuos físicos, se obtienen una serie de k caracteres u observaciones de igual o distinta naturaleza. Es importante tener en cuenta, ya desde este momento, que la calidad del análisis que se realice, va a depender de la habilidad que se tenga a la hora de seleccionar los caracteres que se obtendrán del conjunto de individuos seleccionados.

Los datos obtenidos se organizarán en una matriz $n \times k$, donde cada fila representa a un individuo o registro y las columnas a las características observadas. Las columnas tendrán naturaleza homogénea, pudiendo tratarse de caracteres nominales, dicotómicos o politómicos, presencias–ausencias, ordenaciones, conteos, escalas de intervalo, razones,...; también se podrían tener variables compuestas como ratios, densidades,... En ocasiones se añade una columna que se suele colocar en primer lugar y que asigna un nombre a cada individuo; dicha columna recibe el nombre de *variable etiqueta*.

Físicamente, la estructura de una matriz de datos se corresponde con el esquema de una base de datos o una hoja de cálculo. Al igual que pasa con los editores de los programas de tratamiento de datos, las dos dimensiones de una pantalla se acomodan perfectamente al tanden individuo–variable. Si se consideran los individuos identificados por los términos $I_1, I_2, \ldots, I_n$ y los caracteres por $C_1, C_2, \ldots, C_k$, la casilla x_{ij} representa el comportamiento del individuo I_i respecto al carácter C_j. En la figura se muestra la matriz de datos del fichero `Iris` del paquete `datasets` de **R**.



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa
14	4.3	3	1.1	0.1	setosa
15	5.8	4	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa

R se refiere a este tipo de estructura de datos como *data.frame*. Este es el formato que requiere el programa para aplicar la mayoría de los procedimientos estadísticos.

1.1.1. Anomalías de la matriz de datos

Hay veces en que por distintos motivos la matriz de datos presenta casillas vacías, ello se debe a que no se ha podido medir un dato o a que se ha perdido la observación. En otras ocasiones un dato presente en la matriz ha sido *depurado* por presentar algún tipo de anomalía, como haber sido mal medido, mal transcrito a la matriz de datos, pertenecer a un colectivo distinto del que se está analizando, etc... La identificación de estos elementos anómalos se realiza mediante un proceso de detección de inconsistencias o de evaluación de valores extremos, muy grandes o muy pequeños, que determinará si razonablemente pueden pertenecer al colectivo bajo estudio. A veces se sustituye el valor depurado de un individuo por uno que sea congruente con el resto de caracteres del mismo, mediante técnicas que se conocen como de *imputación*. Los huecos que definitivamente queden en la matriz se referirán como *valores omitidos* o, más comunmente, como *valores missing*. En **R** estos valores se representan con `NA` (Not Available). En función del tipo de análisis que se esté realizando, el procedimiento desestimará sólo el dato o todo el registro completo.

En este módulo se analizarán –salvo excepciones que se indicarán con antelación– de forma independiente cada uno de los caracteres de la matriz de datos, de forma que cada carácter describirá parcialmente al conjunto de individuos. La integración de todos los análisis deberá dar una cierta visión general de la población. En cualquier caso, este enfoque está muy lejos de ser eficiente, entre otras cosas porque habitualmente las variables individuales comparten información y dicha redundancia distorsionaría las conclusiones del estudio, siendo en general preferible decantarse por un análisis global en vez del secuencial. Por tanto, la pretensión de este capítulo es tratar algunos conceptos básicos y adquirir destreza en el manejo de medidas estadísticas que serán empleadas masivamente cuando se aborden, más adelante, modelos más sofisticados.

2. Naturaleza de los caracteres: Atributos y Variables

Respecto a la *cantidad de información* que porta cada tipo de carácter, se puede considerar que los caracteres nominales son los más "pobres", puesto que ni siquiera poseen orden, mientras que los más ricos serían las escalas de intervalos y las razones, que tienen orden, son cuantitativas y en el caso de las razones el cero lo es en términos absolutos, es decir, el 0 representa la ausencia de la característica. En posiciones intermedias se situarían el resto en el orden en que se han introducido en la figura 2.1.

Ejemplo 2.1

El caso más evidente para apreciar las diferencias entre las escalas de intervalo y las razones o escalas de cociente, lo ofrece el termómetro. Un termómetro genera una variable de escala de intervalo, porque la

Figura 2.1: Esquema de cantidad de información

diferencia real entre 2 y 3 grados es la misma que entre 40 y 41 grados, pero no se puede decir que cuando el termómetro marca 30 grados hace el doble de calor que cuando marca 15.

Por otra parte, muchas magnitudes físicas, como el peso, la longitud o la intensidad de corriente, son razones porque, por ejemplo en el caso del peso, un objeto de 20 kilogramos pesa el doble que otro de 10 kilogramos. Es decir existe el cero absoluto.

Como ya se ha comentado, la naturaleza del carácter condicionará su tratamiento, aunque en ningún caso hay que confundir la cantidad de información que porta con su valor intrínseco para analizar a los individuos del colectivo.

En una primera instancia, se distinguirá entre los caracteres que no están ordenados y los que sí lo están, los primeros jugarán en general un rol de *atributos* mientras que los segundos habitualmente actuarán como *variables*. Los atributos tendrán la misión de establecer clases, dividiendo el colectivo global en subgrupos o categorías; por su parte, las variables caracterizarán a dichos subgrupos e intentarán establecer diferencias entre unos y otros, para lo que necesariamente se debe considerar algun tipo de métrica. Pero ello es una regla general que tiene muchas excepciones y así, en ocasiones, un carácter llamado a adoptar el papel de variable podría, mediante una operación de *punto de corte*, actuar como atributo, mientras que es factible definir una medida de asociación sobre caracteres intrínsecamente de clase que permita caracterizar a los individuos del colectivo en base a una serie de atributos.

Ejemplo 2.2

Es habitual que la edad, que es intrínsecamente una variable –medida en un soporte temporal– se emplee para dividir la población en clases dando cortes en el intervalo de tiempo, obteniéndose por ejemplo grupos de alevines, adultos y maduros de una comunidad de peces y adoptando por tanto la variable un rol de atributo.

En el extremo opuesto, hay investigaciones médicas que relacionan el tipo de patología con el sexo del paciente y con el desenlace de la enfermedad, caracteres todos ellos intrínsecamente atributos.

Las variables pueden clasificarse según su conjunto soporte. El soporte de una variable es el conjunto de todos los posibles valores que toma. Cuando el conjunto soporte es finito o numerable se habla de variable discreta. Por el contrario, cuando el conjunto soporte es no numerable, se habla de variable continua. Si la variable continua no toma valores en puntos aislados se dice absolutamente continua. Esta diferencia tendrá relevancia cuando se planteen, más adelante, estructuras de probabilidad para modelizar la población bajo estudio.

Ejemplo 2.3

El número de lunares en la piel de pacientes aquejados de una cierta patología, el número de hijos de las familias de una comunidad o el número de meteoritos que surcan una cierta región estelar en periodos de tiempo determinados son variables discretas. La distancia por carretera entre las capitales de provincia peninsulares españolas, el tiempo de reacción de los corredores de una carrera de 100 metros o las longitudes de los cabellos de una persona son variables continuas.

Una vez identificadas, recolectadas y organizadas, las variables serán tratadas estadísticamente combinando un análisis numérico, a través de una serie de medidas estadísticas, con representaciones gráficas. El software estadístico **R** ofrece una amplia gama de ambos elementos: numéricos y gráficos, aunque conviene ser selectivos y tomar aquellos

Figura 2.2: Ventana de selección de datos en paquetes adjuntos

que verdaderamente aportan información relevante. A tal efecto, se proponen las siguientes opciones:

Escala de Medida	**Medidas centrales**	**Medidas de dispersión**	**Representaciones gráficas**
Atributo	Moda Porcentajes		Diagrama de sectores
Ordenación	Mediana Percentiles	Recorrido Intercuartílico	Diagrama de barras
Recuento	Media	Desviación típica	Diagramas de barras
Intervalo	Media	Desviación típica	Histograma
Razón	Media geométrica	Coeficiente de variación	Histograma Diagrama de dispersión Diagrama de cajas

Tabla 2.1: Medidas y gráficos según tipo de variable

En última instancia corresponde al investigador el tomar las decisiones correctas en cada momento, de forma que sin transgredir los principios básicos, den como resultado un análisis eficiente de los datos.

3. Análisis de atributos

Los atributos son susceptibles de ser tratados de forma individual o en grupo, para obtener los porcentajes de cada subgrupo en el colectivo global. De hecho, cada carácter o conjunto de ellos establece una partición o catálogo de la población bajo estudio. Por otra parte, el

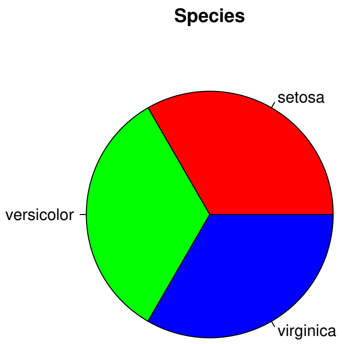


Figura 2.3: Diagrama de sectores del fichero `iris`

tratamiento gráfico más usual que se le daría a un atributo individual sería a través de un *diagrama de sectores* o *diagrama de tarta*.

Ejemplo 2.4

Se consideran ahora los datos del ejemplo `iris` *del paquete* `datasets` *de* **R** *que se describe en el apéndice A. Se carga el fichero en* **Rcmdr** *mediante la selección de las opciones del menú* `Datos`→ `Datos en paquetes`→`Leer datos desde paquete adjunto...`, *en el cuadro de diálogo se elige el paquete* `datasets` *y dentro de éste el juego de datos* `iris`,*figura 2.2. Del conjunto de variables de la matriz se considera la denominada* `Species`, *que es un atributo con los tres tipos de flores de Iris:* `Setosa, Virginica y Versicolor`.

Análisis numérico: *Se selecciona* `Estadísticos`→`Resúmenes`→ `Distribuciones de frecuencias...` *y en el cuadro de diálogo se elige el único atributo,* `Species`. *Se observa que los 150 individuos se reparten a partes iguales entre las tres variedades de flores, 50 para cada una, y que por tanto los porcentajes son iguales a* $33,33$. *No tiene sentido hablar de moda, puesto que las tres clases lo son.*

```
> .Table <- table(iris$Species)
> .Table # counts for Species
setosa     versicolor     virginica
50             50             50
> 100*.Table/sum(.Table) # percentages for Species
setosa     versicolor     virginica
33.33333    33.33333       33.33333
```

Análisis gráfico: *A continuación se selecciona el diagrama de sectores mediante* `Gráficas→Gráfica de sectores...`

Si el fichero de datos activo tiene más de una variable de clase se permite seleccionar la que se quiera. En este caso, la única variable elegible es `Species`*, que el programa da por defecto. Si se pulsa el botón* `Aceptar` *el programa dibuja el gráfico de sectores que se muestra en la figura 2.3. Como era de esperar, la tarta se divide en tres trozos exactamente iguales.*

4. Análisis de variables ordenadas

Las diferencias que se establecen entre variables de clase pura y ordenada se concretan desde el punto de vista del análisis numérico en que el grupo de medidas recomendables son las de posición, es decir los cuantiles en sus distintas versiones. Como medidas de representación, pensando que en general se dispondrá de pocas clases, se recurrirá a los cuartiles y como medida de dispersión al recorrido intercuartílico. En cuanto al análisis gráfico, se recomienda el uso del diagrama de barras.

Este tipo de variables ordenadas suele venir dada en forma de tabla de frecuencias. Por ello, en el ejemplo que ilustra el tratamiento de este tipo de variables, se comenzará explicando como transformar una tabla de frecuencias en una matriz de datos, al objeto de que puedan ser tratadas por **R** como un `data.frame`.

Ejemplo 2.5

Un caso de variable ordenada es la correspondiente a un estudio estadístico sobre el nivel académico de la población gaditana en el año 2001 (Fuente: Instituto Estadístico de Andalucía).

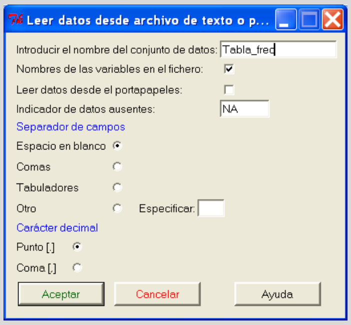
Los valores que toma la variable son: `Sin estudios`, `Elementales` *(primaria),* `Medios` *(secundaria, bachillerato y fp grado medio) y* `Superiores` *(fp superior, diplomatura, licenciatura y doctorado).*

Los datos se recogen en la tabla:

	NIVEL DE ESTUDIOS			
SEXO	*Sin estudios*	*Elementales*	*Medios*	*Superiores*
Hombre	*79309*	*107156*	*183488*	*70594*
Mujer	*108051*	*109591*	*174961*	*64858*

Debido al gran número de individuos que forman esta muestra puede ser útil almacenar la variable estudiada a partir de su tabla de frecuencias, transformándola en base de datos en el momento de realizar los análisis. El fichero en cuestión se ha guardado bajo el nombre de `tabla_freq_niv_estudios.dat`*, conteniendo tres variables:* `sexo`*,* `nivel` *y* `frec`*. En total consta de 8 filas que se correponden con los cruces de las clases sexo y nivel.*

Para cargar en **Rcmdr** *la tabla de frecuencias se selecciona* `Datos`→ `Importar datos desde archivo de texto o portapapeles...`*, en este ejemplo se ha elegido el nombre* `Tabla_frec` *para denominar al fichero que contendrá los datos de la tabla de frecuencias, como se muestra en la ventana de diálogo. A continuación se elige el archivo* `tabla_freq_niv_estudios.dat`*.*



Ahora se tendrá que transformar esta tabla de frecuencias en un conjunto de datos, `data.frame`*, con el que* **R** *pueda trabajar. Para conseguir esto se procede de la siguiente manera:*

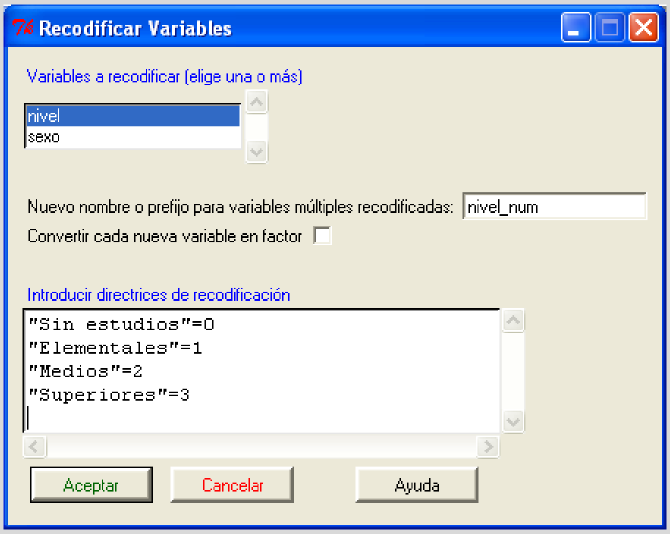
```
>nivel<-rep(Tabla_frec$nivel,Tabla_frec$frec)
>sexo<-rep(Tabla_frec$sexo,Tabla_frec$frec)
>niv_estudios_cadiz< -data.frame(nivel,sexo)
```

Es decir, se crean las variables `nivel` *y* `sexo` *a partir de la repetición de cada una de las clases de las respectivas variables, tantas veces como indique su frecuencia. A partir de ahí, se construye el* `data.frame` `niv_estudios_cadiz` *con las dos variables creadas.*

Este `data.frame` *se encuentra entre los datos que se facilitan en este libro y se puede cargar directamente sin realizar las operaciones anteriores. Para ello, basta con seleccionar* `Datos`→`Importar datos`→ `desde archivo de texto o portapapeles...`*, eligiendo ahora el ar-*

chivo `niv_estudios_cadiz.dat`.

Análisis numérico: *En variables de tipo ordenado es aconsejable utilizar, como medida de posición, los cuartiles.*



Para realizar este análisis la variable `nivel` *debe ser codificada numéricamente. Se creará una nueva variable en la base de datos, que se llamará* `nivel_num` *y que representará los valores numéricos de la variable* `nivel`. *Los valores* `Sin estudios`, `Elementales`, `Medios` *y* `Superiores` *han sido codificados mediante los valores 0, 1, 2 y 3, respectivamente. En* **Rcmdr** *esto se realizará seleccionando* `Datos`→`Modificar variables de los datos activos`→ `Recodificar variables...` *, desmarcando la pestaña* `Convertir cada nueva variable en factor`.

Para realizar el análisis numérico de la variable `nivel_num` *se selecciona:* `Estadísticos`→`Resúmenes`→`Resúmenes numéricos...`, *eligiendo en la ventana emergente la variable* `nivel_num` *y marcando la opción de* **cuantiles**. *Se puede observar entre los cuartiles que la mediana recae sobre el valor 2.*

```
> numSummary(Niv_estudios[,''niv_num''],
statistics=c(''quantiles''))
0%    25%    50%    75%    100%
0      1      2      2      3
```

Desde **Rcmdr** *existe otra forma de realizar el análisis numérico de una variable ordenada.*

Para ello, se reordenan los niveles de la variable factor usando las opciones del menú `Datos`→`Modificar variables del conjunto de datos activo`→`Reordenar niveles de factor...`*, almacenando la variable* `nivel` *como factor de tipo ordenado. A la nueva variable se le ha llamado* `nivel_ord`*. A continuación se almacena ésta como variable de tipo numérico, escribiendo en la ventana de instrucciones:*

```
Datos$nivel_num< −as.numeric(Datos$nivel_ord)
```

siendo ya posible calcular los cuantiles, para la variable numérica `Datos$nivel_num`*.*

Como medida de dispersión se ha recomendado el recorrido intercuartílico relativo, definido como el cociente entre la diferencia de los cuartiles tercero y primero, y la mediana. **Rcmdr** *no proporciona directamente este estadístico, pero se puede implementar fácilmente en la ventana de instrucciones, mediante las órdenes siguientes:*

```
>Q1<-quantile(niv_estudios_cadiz$nivel_num, 0.25)
>Q2<-quantile(niv_estudios_cadiz$nivel_num, 0.5)
>Q3<-quantile(niv_estudios_cadiz$nivel_num, 0.75)
>RIR<-as.numeric((Q3-Q1)/Q2)
>RIR
[1] 0.5
```

Análisis gráfico: *Para realizar el análisis gráfico de la variable se utiliza el diagrama de barras. En* **Rcmdr** *se selecciona:* `Gráficas`→ `Gráfica de barras...` *y se elige en la ventana de diálogo, la variable* `nivel_ord`*.*

En **R** *existe una gran variedad de opciones que ayudan a mejorar el aspecto de los gráficos. Se puede acceder a ellas escribiéndolas en la ventana de instrucciones. En este ejemplo se ha optado por modificar el*

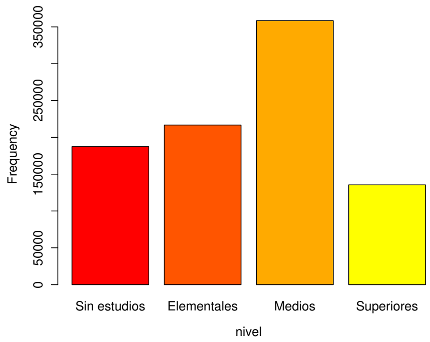


Figura 2.4: Diagrama de barras de la variable nivel de estudios

color, siguiendo una escala de colores cálidos. Esto se consigue agregando `col=heat.colors(5)` *a las opciones de* `barGraph` *(figura 2.4).*

5. Análisis de variables de escala

Ejemplo 2.6

Se estudiará ahora el tratamiento de una variable continua. Para ello se considera la base de datos `chickwts`*, del paquete* `datasets` *de* **R**. *En ella se recogen los pesos finales, en gramos, de 71 polluelos, según el tipo de dieta seguida durante un periodo de 6 semanas.*

Análisis numérico: *Para la variable que da el peso de los polluelos las medidas básicas recomendadas son la media y la desviación típica. Estas medidas se calculan desde* Estadísticos→Resúmenes→ Resúmenes numéricos..., *seleccionando para la variable* `weight` *las opciones deseadas.*

```
> numSummary(chickwts[,''weight''], statistics=c(''mean'',
''sd''))
mean      sd      n
261.3099 78.0737 71
```

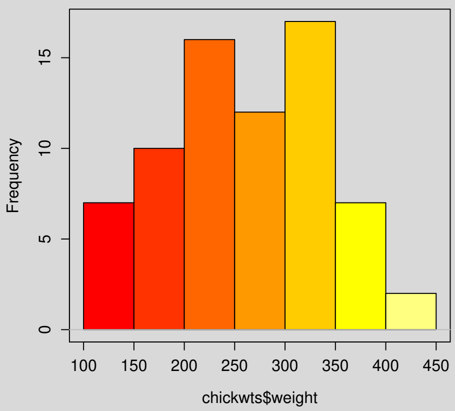
Aunque se está hablando de la desviación típica, la función `sd` *calcula en realidad la cuasidesviación típica. Cabe la posibilidad de que* *se necesiten otro tipo de medidas que completen el estudio, como la simetría, el apuntamiento, ... Para ello, en el apéndice B, se incluye una tabla de medidas estadísticas. Por ejemplo, si se deseara calcular la simetría y la curtosis de la variable* `weight`*, habría en primer lugar que instalar y cargar en* ***R****, si no lo está ya, el paquete* `fBasics`*. Y a continuación:*

```
> kurtosis(chickwts$weight)
 -0.9651994
attr(,‘‘method’’)
‘‘excess’’
```

```
> skewness(chickwts$weight)
 -0.01136593
attr(,‘‘method’’)
‘‘moment’’
```

Ambos coeficientes están calculados a partir de los momentos y, en el caso de la curtosis, se le ha restado 3. Se podría concluir que la distribución es bastante simétrica y algo aplastada.

Análisis gráfico: *Para analizar gráficamente la variable peso se comienza con la realización del histograma que se muestra al margen mediante las instrucciones* `Gráficas→Histograma...` *En el histograma se observa un comportamiento bastante simétrico y la posibilidad de que existan dos modas.*



A continuación, se construye el diagrama de caja (figura 2.5). Se puede observar en el gráfico que la variable no posee valores atípicos, es simétrica y está relativamente dispersa.

El `data.frame` *que se está utilizando incluye un factor,* `Feed`*, que se corresponde con las diferentes dietas sumimistradas a los pollos. Ello permite la realización de un análisis por grupo, tanto numérico como gráfico, que permita evaluar las diferencias de peso en función del tipo de alimentación seguida. Los valores que toma la variable* `Feed` *son:*

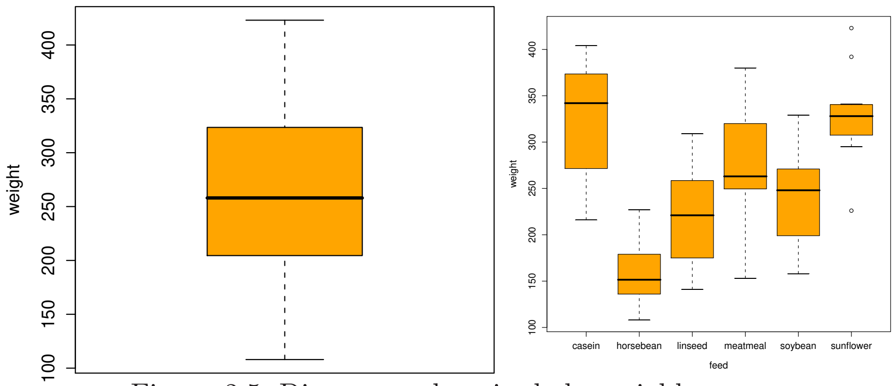


Figura 2.5: Diagramas de caja de la variable peso

`horsebean` (habas), `linseed` (linaza), `soybean` (soja), `sunflower` (girasoles), `meatmeal` (carne) y `casein` (caseína).

Es interesante la representación del diagrama de caja de la variable peso, según el tipo de alimentación (figura 2.5). Se observa que los valores de la variable peso están más concentrados para la dieta `sunflower`. También éste es el único grupo en el que se dan valores atípicos. Por contra la mayor dispersión de los datos se produce con la dieta `casein`. Una evaluación inicial, parece indicar que la dieta que produce pollos de mayor peso es `sunflower`, ya que los pesos que consigue están más concentrados en torno a uno de los valores más altos.

El análisis numérico ofrece los siguientes resultados:

```
> numSummary(chickwts[,‘‘weight’’], groups=chickwts$feed,
statistics=c(‘‘mean’’))


          mean      sd        n
casein    323.5833  64.43384  12
horsebeen 160.2000  38.62584  10
lindseed  218.7500  52.23570  12
meatmeal  276.9091  64.90062  11
soybean   246.4286  54.12907  14
sunflower 328.9167  48.83638  12
```

6. Ejercicios

2.1 Al comenzar el curso se pasó una encuesta a los alumnos del primer curso de un colegio, preguntándoles, entre otras cuestiones, por el número de hermanos que tenían. Se obtuvieron los siguientes resultados:

3, 3, 2, 2, 8, 5, 2, 4, 3, 1, 4, 5, 3, 3, 3, 3, 3, 2, 5
1, 3, 3, 2, 2, 4, 3, 3, 2, 2, 4, 4, 3, 6, 3, 3, 2, 2, 4
3, 4, 3, 2, 2, 4, 4, 3, 3, 4, 2, 5, 4, 1, 2, 8, 2 ,3, 3, 4

a) Represente este conjunto de datos con un diagrama de barras.

b) Calcule media, moda y mediana.

c) Estudie la dispersión de los datos.

d) Analice la simetría de la distribución.

2.2 Los pesos de un colectivo de niños son:

60, 56, 54, 48, 99, 65, 58, 55, 74, 52, 53, 58, 67, 62, 65
76, 85, 92, 66, 62, 73, 66, 59, 57, 54, 53, 58, 57, 55, 60
65, 65, 74, 55, 73, 97, 82, 80, 64, 70, 101, 72, 96, 73, 55
59, 67, 49, 90, 58, 63, 96, 100, 70, 53, 67, 60, 54

Obtenga:

a) La distribución de frecuencias agrupando por intervalos.

b) La mediana de la distribución.

c) La media de la distribución, indicando su nivel de representatividad.

d) Utilizando la agrupación en intervalos, el porcentaje de alumnos que tienen un peso menor de 65 kg y el número de alumnos con un peso mayor de 60 kg dentro del grupo de los que pesan menos de 80 kg.

2.3 En el Consejo de Apuestas del Estado se han ido anotando, durante una temporada, el número de premiados de quinielas según la cantidad de aciertos. Los resultados se recogen en la siguiente tabla:

Nº de aciertos	11	12	13	14	15
Nº de personas (miles)	52	820	572	215	41

Calcule:

a) La mediana, la moda y los cuartiles de la distribución.

b) La simetría de la distribución.

2.4 En un puerto se controla diariamente la entrada de pesqueros según su tonelaje, resultando para un cierto día los siguientes datos:

Peso(Tm.)	0-25	25-50	50-70	70-100	100-500
N^o de barcos	5	17	30	25	3

Se pide:

a) El peso medio de los barcos que entran en el puerto diariamente, indicando la representatividad de dicha medida.

b) El intervalo donde se encuentra el 60 % central de la distribución.

c) El grado de apuntamiento.

d) El tonelaje más frecuente en este puerto.

Capítulo 3

Análisis Exploratorio de Datos multidimensional

Una vez estudiados los distintos caracteres de la matriz de datos de forma individual, resulta muy interesante realizar análisis conjuntos de grupos de ellos, de hecho, la mayoría de los análisis estadísticos tienen carácter multivariable. Los motivos para adoptar este enfoque son variados, aunque de nuevo la cuestión de la naturaleza de los caracteres y los objetivos del estudio serán determinantes a la hora de fijar las técnicas que se emplearán.

Aunque en posteriores entregas se tratarán técnicas multivariables muy potentes, los objetivos en este capítulo son mucho más modestos y se limitarán a un primer acercamiento de naturaleza descriptiva; empleándose para ello tanto medidas de relación entre caracteres como representaciones gráficas. En la mayoría de las ocasiones sólo se contemplarán dos caracteres de forma conjunta, realizándose, por tanto, un análisis bidimensional.

En este capítulo también se hará una primera incursión en el tema de la modelización. Un modelo estadístico relaciona mediante una o varias expresiones matemáticas a un grupo de caracteres, que ocasionalmente deben cumplir algunos requisitos. En este caso, se abordará un *modelo de ajuste bidimensional*, en el que se tratará de explicar el comportamiento de una variable *causa* a partir de otra que se denomina *efecto.*

Siempre existe un cierto grado de tolerancia para asimilar caracteres de menor nivel de información a los de nivel superior, aunque existe una marca que no se debe transgredir, que es la de la ordenación. Así, podría justificarse el tratar una variable contada como variable de escala, pero nunca se podría asimilar un atributo a una variable ordenada.

1. Tipos de relaciones entre caracteres

En principio se podrían establecer tantos tipos de relación como los que resultarían de cruzar los diferentes caracteres definidos en el capítulo anterior. No obstante, el número de cruces sería demasiado elevado y muchos de ellos no tendrían interés práctico, por lo que se limitará el estudio a aquellos que habitualmente se encuentran en la práctica, que básicamente se corresponden con los que relacionan caracteres de la misma naturaleza. Se expondrán previamente algunas matizaciones y precauciones que conviene tener presente.

- En general funcionan mejor los cruces entre caracteres de la misma naturaleza. Ello se debe a que para realizar el análisis se debe especificar algún tipo de *disimilaridad* que establezca la diferencia, en función de los caracteres considerados, que existe entre cada par de individuos de la matriz de datos. Así, la disimilaridad entre dos individuos sobre los que se han medido dos variables de escala es habitualmente la distancia euclídea, que como se sabe posee buenas propiedades, mientras que si un carácter es de clase y el otro una variable de escala la disimilaridad que se elija tendrá, con toda seguridad, propiedades mucho más débiles.

- Como consecuencia de lo anterior cuando se incluyan en el mismo análisis caracteres de distinta naturaleza conviene, siempre que sea posible, asignarles roles distintos.

- La asignación de roles a variables de la misma naturaleza en ningún caso se soportará por motivos estadísticos, sino que dependerá exclusivamente del criterio del investigador.

A, B	B_1	$\cdots$	B_j	$\cdots$	B_s	
A_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$\mathbf{n_{1\cdot}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$\mathbf{n_{i\cdot}}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$\mathbf{n_{r\cdot}}$
	$\mathbf{n_{\cdot 1}}$	$\cdots$	$\mathbf{n_{\cdot j}}$	$\cdots$	$\mathbf{n_{\cdot s}}$	$\mathbf{n}$

Tabla 3.1: Distribuciones conjuntas y marginales de (A, B)

- La investigación combinatoria, es decir aquella que considera todos los grupos posibles de variables, está fuertemente desaconsejada, aunque se trate, como es el caso, de un análisis de carácter exploratorio. La violación de este principio puede llevar a aceptar como válidas asociaciones meramente espúreas.

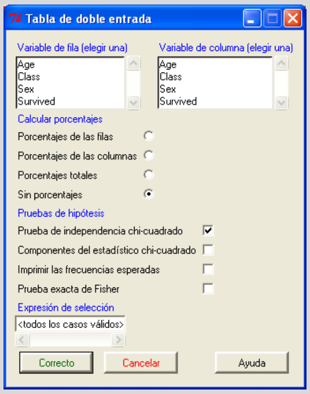
2. Análisis de relaciones entre dos atributos

Para relacionar dos atributos, tanto dicotómicos como politómicos, se construirá la tabla de frecuencias conjunta o *tabla de doble entrada*. Así, si se considera que el atributo A está conformado por las clases $A_1, A_2, \ldots, A_r$ y el atributo B por las clases $B_1, B_2, \ldots, B_s$, la información a tratar quedaría conformada por la tabla 3.1; donde n_{ij} representa la frecuencia absoluta del par (A_i, B_j), es decir el número de individuos que presentan de forma conjunta la clase A_i de A y la B_j de B. La última columna y la última fila de la tabla 3.1 representan las *distribuciones marginales* de A y B, respectivamente.

Cuando se consideran dos atributos dicotómicos se tendrá una tabla 2×2, que en ocasiones necesitará un tratamiento diferenciado. Mención aparte merece el caso en que uno o los dos atributos son del tipo *presencia-ausencia* de una cualidad.

Ejemplo 3.1

Como caso práctico para analizar la relación entre atributos se ha elegido el archivo de datos `titanic.dat`*, en el que aparecen las variables* `Class`*,* `Sex`*,* `Age` *y* `Survived`*, que aportan información, respectivamente, sobre la clase que ocupaba el pasajero, su sexo, edad y si sobrevivió o no al naufragio del famoso trasatlántico. En concreto, se intentará establecer una posible asociación entre la supervivencia y la clase en la que viajaban los pasajeros del Titanic.*



En primer lugar se construirá la tabla de doble entrada con las variables seleccionadas. Con **Rcmdr** *esto se consigue desde* `Estadísticos`→`Tablas de contingencia`→`Tabla de doble entrada...`*, con lo que se abre la ventana de diálogo mostrada arriba, en la que se seleccionan los correspondientes atributos fila (*`Survived`*) y columna (*`Class`*), además se eligen* `Porcentajes totales` *y se deja marcada la opción* `Prueba de independencia chi-cuadrado`*. Los resultados son:*

```
> .Table < −xtabs(∼Survived+Class, data=Datos)

> .Table
Class

Survived    1st    2nd    3rd    Crew
No          122    167    528    673
Yes         203    118    178    212

> totPercents(. Table) # Percentage of Total

       1st      2nd      3rd      Crew     Total
No      5.5     7.6      24.0     30.6     67.7
Yes     9.2     5.4      8.1      9.6      32.3
Total  14.8    12.9      32.1     40.2     100.0

> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
Pearson's Chi-squared test
data: .Table
X-squared=190.4011 ,df=3, p-value < 2.2e-16
```

R *además de proporcionar las tablas de valores absolutos y de porcentajes sobre el total, da información sobre el grado de relación entre los atributos, a través del coeficiente χ^2. De momento se considera sólo el valor del estadístico $\chi^2 = 190{,}4$. Este estadístico indica el grado de relación entre la clase que ocupaba el pasajero y si sobrevivió o no al naufragio; si $\chi^2 = 0$ indicaría una ausencia de relación y a medida que χ^2 crece la relación va en aumento.*

El estadístico no está acotado en un rango de valores que permita interpretar la intensidad de la relación, por lo que se debe recurrir a algún coeficiente derivado que esté acotado. Los más usuales son el coeficiente de contingencia y el coeficiente de Cramer, ambos acotados en el intervalo $[0, 1)$. Se empleará en este caso el primero que viene dado por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

donde n es el tamaño muestral. En nuestro caso el coeficiente de contingencia vale $0,28$, lo que indica una cierta relación entre ambos atributos. Si se observa la tabla de doble entrada se ve que porcentualmente se salvaron más pasajeros de primera clase, mientras que los de tercera clase y la tripulación fueron los que más sufrieron las consecuencias del naufragio. Más adelante, se verá que se puede ser más contundente a la hora de concluir la existencia de relación utilizando los Contrastes de Hipótesis.

Para poder visualizar la relación entre las variables puede ser muy útil la realización de un diagrama de barras de la variable supervivencia según la clase de los pasajeros. Para ello, se almacena en primer lugar la tabla de contingencia de las variables `Survived` *frente a* `Class`*, a la que se ha llamado* `Tabla`*, ejecutando en la ventana de instrucciones:*

```
>Tabla <-xtabs(~ Survived+Class, data=Datos)
```

A continuación se obtiene el diagrama de barras mediante las órdenes **R***:*

```
>barplot(Tabla, xlab=''Clase'', ylab=''Frecuencia'',
legend.text=c(''No superviviente'', ''Superviviente''),
beside=TRUE,col=cm.colors(2))
```

Observando el diagrama de barras de valores absolutos (figura 3.1), se aprecia que éste ofrece una visión que podría llevar a confusión, aparentando, por ejemplo, que el número de supervivientes de primera clase

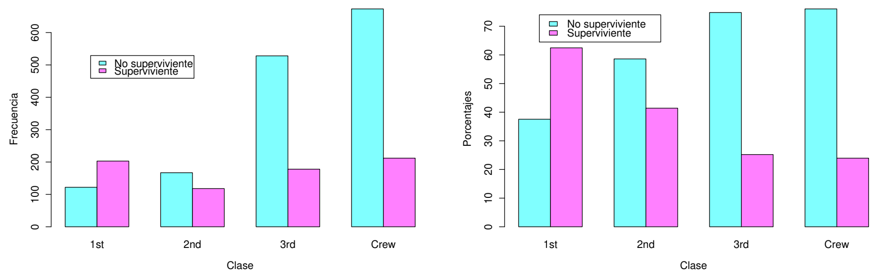


Figura 3.1: Diagramas de barras de la supervivencia

es prácticamente igual al número de supervientes de la tripulación. Ello se debe a que se han comparado las frecuencias absolutas de estos dos grupos, y mientras que en primera clase viajaban 325 individuos, los miembros de la tripulación eran 885. Una alternativa para apreciar la relación existente entre los dos atributos es construir el diagrama de barras de las frecuencias relativas, o porcentajes de supervivencia respecto a cada clase, en lugar de usar las frecuencias absolutas. Igual que antes, se debe almacenar previamente la tabla de porcentajes, lo que se consigue con las siguientes instrucciones **R***:*

```
>Tabaux <-colPercents(Tabla)
>Tablarel <-Tabaux[1:2][1:4]
```

`Tabaux` *contiene la tabla de porcentajes, los porcentajes totales y las frecuencias marginales. Para representar el diagrama de barras no son necesarias las dos últimas filas, por lo que se ha construido una nueva tabla denominada* `Tablarel` *con la información que interesa.*

Ahora se está en condiciones de construir el diagrama de barras; para ello se sustituye, en la secuencia de instrucciones usada para el diagrama de barras de valores absolutos, `Tabla` *por* `Tablarel` *(figura 3.1).*

Por último, se construirá un gráfico de mosaico, figura 3.2, con todos los atributos del fichero `Titanic`*. Para ello, se ejecuta la instrucción:*

```
>mosaicplot(Titanic, main=‘‘Supervivientes del Titanic’’,
color=c(‘‘red’’,‘‘green’’))
```

Se han seleccionado los colores verde para los supervivientes y rojo para los no supervivientes.

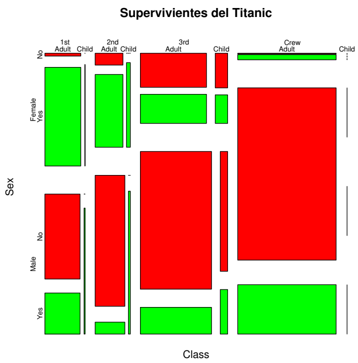


Figura 3.2: Gráfico de mosaico de los datos Titanic

R–Nota 3.1

*Éste puede ser un buen momento para analizar someramente la sintaxis de las instrucciones **R**, dado que en ocasiones, como ha ocurrido en este ejemplo, se necesita crear o editar una instrucción. Como el lector habrá podido comprobar, cada vez que se ha utilizado un procedimiento de **Rcmdr**, éste ha generado una o varias instrucciones **R**; en realidad, **Rcmdr** no es otra cosa que lo que se conoce como un* frontend *de **R**, es decir un forma más amigable de acceder a los recursos de **R**.*

*Las instrucciones de **R** pueden ser una expresión o una asignación. Una expresión se evalúa, se muestra su resultado y se descarta. Una asignación se evalúa obteniendo un nuevo objeto que se almacena con el nombre especificado.*

Concretamente, si se analiza la estructura de la instrucción:

```
>Tabla <-xtabs(∼ Survived+Class, data=Datos)
```

se observa que se crea el objeto `Tabla`*, al que se le asigna* $(<-)$ *el resultado de la evaluación de la función* `xtabs`*, que genera una tabla de doble entrada con las variables* `Survived` *y* `Class` *del* `data.frame` *con nombre* `Datos`*. Si ahora se fija la atención en la instrucción:*

```
>barplot(Tabla, xlab=‘‘Clase’’, ylab=‘‘Frecuencia’’,
legend.text=c(‘‘No superviviente’’, ‘‘Superviviente’’),
beside=TRUE,col=cm.colors(2))
```

*Ésta le indica a **R** que cree un gráfico de barras,* `barplot`*, de la* *tabla de doble entrada `Tabla`, siendo las etiquetas de los ejes, `xlab e ylab, Clase y Frecuencia`, que la leyenda de las clases, `legend.text`, sea `No superviviente` y `Superviviente`, que el tipo de barras sea pegada, `beside=TRUE`, y que utilice la gama de colores `col=cm.colors(2)`.*

R–Nota 3.2

En los diagramas de barras anteriores se usa el argumento `legend.text` para incluir una leyenda de los datos, pero de esta forma la leyenda se dibuja en ocasiones sobre las barras. Para mejorar los resultados gráficos se pueden utilizar las siguientes instrucciones:

1. *Escribir la orden del gráfico de barras sin `legend.text`:*

   ```
   >barplot(Tablarel, xlab=‘‘Clase’’, ylab=‘‘Porcentajes’’,
   beside=TRUE,col=cm.colors(2))
   ```

2. *Para localizar las coordenadas del gráfico en las que se desea insertar la leyenda se emplea la orden `locator(n)`, donde `n` es el número de puntos de los que se quiere averiguar las coordenadas, en nuestro caso `n`= 1.*

3. *Una vez ejecutada la orden, se pincha en la gráfica anterior con el botón izquierdo del ratón en el lugar donde se desee insertar la leyenda y automáticamente aparecerán las coordenadas `(x,y)` del punto elegido.*

4. *Por último, se incluirá la leyenda en la posición elegida con la orden:*

   ```
   legend(x,y,c(‘‘No superviviente’’,‘‘Superviviente’’),
   fill=cm.colors(2))
   ```

   El argumento `fill` sirve para indicarle los colores de las barras.

3. Análisis de relaciones entre dos variables

Una vez analizada la relación entre dos atributos, se aborda el estudio de la relación entre dos variables medidas. Este estudio se hará a través de la construcción de una *función de ajuste*, que expresa matemáticamente cómo una de las variables denominada *causa* explica el comportamiento de la otra variable llamada *efecto*. A la variable causa se le conoce también con los nombres de *independiente*, *explicativa*, *exógena*, ..., mientras que la variable efecto es llamada también *dependiente*, *explicada*, *endógena*, ... Desde el punto de vista de la investigación que se esté realizando es fundamental la selección de las variables que entrarán en el análisis y la asignación de roles, causa-efecto, para cada una de ellas.

Es muy habitual confundir los conceptos de *ajuste* y de *regresión*, y aunque no es objeto de este manual entrar en temas teóricos en profundidad, si habría que aclarar que la idea de ajuste implica la selección de un modelo matemático que aproxime lo mejor posible la relación entre las variables, mientras que el concepto de regresión hace referencia a la idea de predecir mediante alguna regla, un valor de la variable dependiente para cada valor de la independiente. Dicho lo cual, y como suele ocurrir en muchos textos estadísticos, a partir de ahora se admitirá, y usará, de forma indistinta ambos conceptos.

Por otra parte, en la mayoría de las ocasiones la matriz de datos contiene varias variables numéricas y el investigador desea estudiar cómo se explica el comportamiento de una de ellas sobre la que tiene un especial interés (dependiente) a partir del conocimiento de un conjunto del resto de variables (independientes). En esta situación, el análisis dos a dos, en el que se consideraría la variable dependiente con cada una de las independientes es claramente ineficiente, siendo necesario la construcción de un modelo de ajuste múltiple que relacione de forma conjunta la variable dependiente con el conjunto de las independientes. La explicación para plantear este enfoque es que las variables independientes suelen estar relacionadas también entre ellas, es decir comparten información de los individuos que se están estudiando, de forma que si se hiciera el análisis dos a dos se estaría utilizando la misma información de forma reiterada.

En lo sucesivo, se consideran sólo dos variables, la independiente (X) y la dependiente (Y), dando lugar a n parejas de valores $(x_i,\ y_i)$. Desde un punto de vista gráfico estos valores se pueden representar en un plano, siendo el conjunto de puntos la denominada *nube de puntos* o *diagrama de dispersión*. El objeto del ajuste es la obtención de una función que se adapte lo mejor posible a la nube de puntos.

$$Y^* = f(X)$$

El conocimiento previo que se puede tener de la relación Y/X junto con el análisis de la nube de puntos debe ofrecer las claves para la selección de la función f. En realidad seleccionar f es elegir una clase funcional que dependerá de unos parámetros que habrá que estimar. Es decir, se elige una recta $Y = a+bX$, una parábola $Y = a+bX+cX^2$, una función exponencial $Y = ab^X$, una función potencial $Y = aX^b$, una hipérbola $Y = a + \frac{b}{X}$, ... Se puede apreciar que mediante alguna transformación muchas de estas funciones se convierten en rectas.

Ejemplo 3.2

- *La clase funcional exponencial $Y = ab^X$ aplicando una transformación logarítmica se linealiza, $logY = loga + Xlogb$.*
- *La clase funcional hiperbólica $Y = a + \frac{b}{X}$ también se convierte en una recta transformando $X' = \frac{1}{X}$.*

Cuando antes se ha escrito «la selección de un modelo matemático que aproxime lo "mejor posible" la relación entre las variables» o la «obtención de una curva que se adapte lo "mejor posible" a la nube de puntos», en realidad se estaba indicando la necesidad de establecer un criterio de ajuste que minimice las diferencias entre la curva de ajuste y la nube de puntos. El criterio más generalizado es el de los *mínimos cuadrados*, que establece que la suma de las distancias al cuadrado entre los valores observados de la variable Y, es decir los y_i, y las predicciones que se obtienen de ésta a partir de la función de ajuste, $y_i^* = f(x_i) \forall i$, sea mínima. La aplicación de este criterio permite la estimación de los parámetros del modelo y la determinación de forma unívoca de la función de ajuste.

La figura 3.3 ilustra lo dicho para el caso lineal $Y = a + bX$, donde a representa el punto de corte de la recta con el eje Y y b el incremento–decremento de Y para un incremento unitario de X.

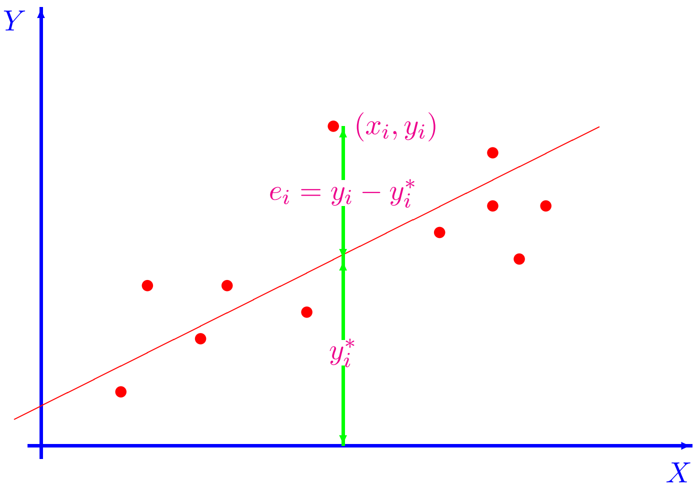


Figura 3.3: Recta de ajuste

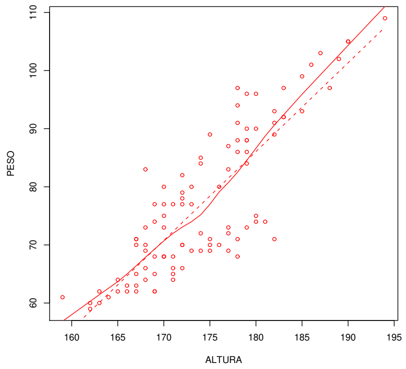
- **Predicciones.** Una de las utilidades más importantes del ajuste es la de realizar predicciones de la variable explicada para distintos valores de la variable explicativa. En realidad, se trata de sustituir en el ajuste los valores de X para obtener los correspondientes valores de Y. Cuando se sustituyen los valores de X que se han empleado para calcular la función de ajuste, $x_1, x_2, \ldots, x_n$ se obtienen los correspondientes valores ajustados por el modelo, $y_1^*, y_2^*, \ldots, y_n^*$, mientras que si se asigna a X cualquier valor factible para esta variable, el valor que se obtiene para Y es una predicción. Obsérvese que la diferencia entre los valores observados de Y, y_i, y sus correspondientes valores ajustados, y_i^*, son los errores del ajuste $e_i = y_i - y_i^*$. Los puntos ajustados (x_i, y_i^*) pertenecen a la recta de ajuste y los y_i^* tienen menos varianza que los y_i, de hecho, se puede demostrar para una gran cantidad de modelos, en particular para el lineal, que la varianza de Y es igual a la de Y^* más la varianza del error, $S_Y^2 = S_{Y^*}^2 + S_e^2$.

Las predicciones para valores de X distintos a los empleados en el ajuste se denominan interpolaciones cuando dichos valores se encuentran dentro del rango de valores de ajuste para X, y extrapolaciones cuando se encuentran fuera de dicho rango. La validez estadística de las interpolaciones es mayor que las de las extrapolaciones, de hecho la calidad de la predicción decrece cuando aumenta la distancia al centro de gravedad de la nube de puntos, $(\bar{x}, \bar{y})$.

- **Análisis de bondad del ajuste.** El ajuste no estaría totalmente resuelto si no viniera acompañado de una medida de su bondad, es decir, de un valor, a ser posible acotado en un intervalo, que exprese en qué porcentaje la variable dependiente se explica por la independiente a través del ajuste realizado. Si el ajuste fuera perfecto todos los valores observados se situarían sobre la nube de puntos y los residuos y su varianza se anularían, mientras que en el extremo contrario sería la variable ajustada la que tendría varianza nula.

  La medida que sintetiza lo expresado en el párrafo anterior es el *coeficiente de determinación*, $R^2 = \frac{S_{Y^*}}{S_Y^2}$ que, como puede verse, toma valores en $[0, 1]$; interpretándose que la variable Y se explica en un $100 * R^2\,\%$ por la variable X, mientras que el resto, es decir el $100 * (1 - R^2)\,\%$, se explicaría por una parte a través de una mejora de la función de ajuste, por otra incorporando, si es factible, información nueva (otras variables, con lo que se tendría un modelo de regresión múltiple) y por la variabilidad intrínseca de los datos.

  Para el caso de ajuste lineal existe un coeficiente específico de bondad de ajuste denominado *coeficiente de correlación lineal* r, que toma valores en el intervalo $[-1, 1]$ y que además de medir la intensidad de la relación indica si ésta es de tipo directo, cuando X crece Y crece, o inverso, cuando X crece Y decrece. Se verifica que $r^2 = R^2$.



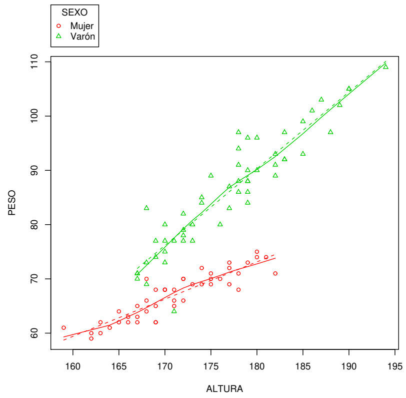


Figura 3.4: Diagramas de dispersión peso-altura

- **Análisis de residuos del modelo.** Conviene examinar, tanto desde un punto de vista numérico como sobre todo gráfico, los residuos que genera el ajuste, es decir las diferencias entre los valores observados, Y, y los ajustados por la función de ajuste, Y^*. En particular, resulta de especial interés el análisis de los residuos extremos y de las gráficas de los residuos frente a valores de X, indexados o frente a las predicciones. También es interesante el análisis de puntos influyentes, entendiendo esto como aquellos puntos que tienen un sobrepeso en la construcción de la función de ajuste. Estos puntos van a estar localizados en los extremos de la nube de puntos, ver ejemplo 3.3.

- **Mejora del modelo.** Para terminar, conviene indicar que reemplazar una función de ajuste por otra más sofisticada, con más parámetros y más compleja, sólo se justifica si la mejora en términos de R^2 es alta, pues en otro caso se complica la interpretación del modelo sin apenas recompensa.

Ejemplo 3.3

Para ilustrar los conceptos sobre el ajuste lineal se procederá a analizar la relación entre `peso` *y* `altura` *del fichero de datos* `peso_altura.dat`*, en*

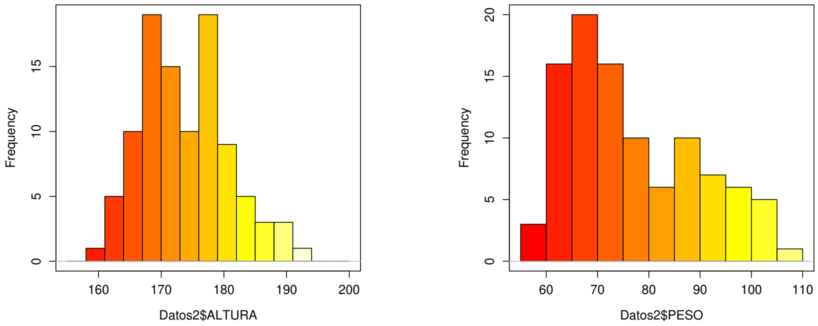


Figura 3.5: Histogramas de peso y altura

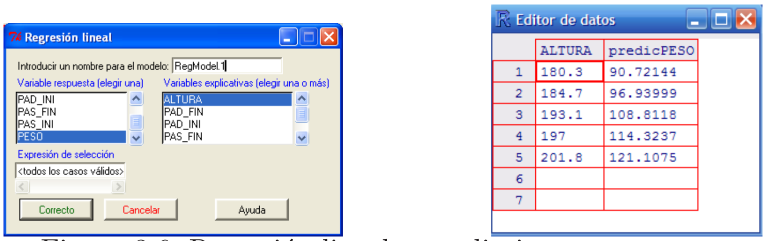


Figura 3.6: Regresión lineal y predicciones

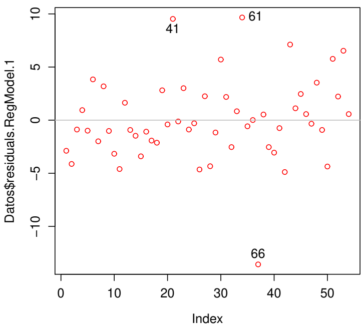


Figura 3.7: Residuos indexados

el que aparecen, entre otras variables, el sexo, peso y altura de un grupo de personas. Como se ha indicado anteriormente es necesario establecer qué variable será la explicada y cuál la explicativa. Dado que se trata de un ejemplo y que no se cuenta con elementos adicionales para avalar la decisión, se decide explicar el `peso` *en función de la* `altura`*.*

1. *Histogramas. Antes de abordar el análisis bidimensional propiamente dicho, se representarán los histogramas de las variables* `peso` *y* `altura`*, operando para ello tal y como se indicó en el capítulo anterior. Al objeto de fijar el número de clases de los histogramas y los colores, se retocan las instrucciones* **R** *que genera* **Rcmdr**, *cambiando en ambos casos las opciones del número de intervalos (breaks) y los colores (col) y se vuelven a ejecutar, con lo que se obtiene las figuras en 3.5. Las instrucciones retocadas son respectivamente:*

   ```
   >Hist(Datos$ALTURA, scale=‘‘frequency’’, breaks=seq(155,200,3),
   col=heat.colors(13))
   >Hist(Datos$PESO, scale=‘‘frequency’’, breaks=seq(55,110,5),
   col=heat.colors(12))
   ```

   Una primera visión de los histogramas permite detectar una bimodalidad tanto en la variable `peso` *como en la* `altura`*, aunque ello es un indicio claro de mezcla de poblaciones, se continuará con los siguientes pasos del ajuste con todos los datos, en un ejercicio básicamente didáctico, en busca de establecer la relación que justifique el* `peso` *en función de la* `altura`*.*

2. *Diagrama de dispersión. Al objeto de decidir el tipo de función de ajuste que se utilizará, se representa el diagrama de dispersión. En* **Rcmdr** *se seleccionan las opciones* `Gráficas`→ `Diagrama de dispersión...`*, para las variables mencionadas. Por defecto aparece marcada la opción* `línea suavizada`*, que ofrece una regresión a los puntos y que da una idea de la clase funcional más eficiente bajo el criterio de mínimos cuadrados.*

   A la vista de la figura 3.4 se observa la existencia de relación entre las dos variables. La línea de regresión suavizada y la línea discontinua de ajuste lineal, sugieren que los ajustes más eficientes son tipo lineal y posiblemente parabólico o potencial. No obstante, la escala de representación de las variables podría ser un factor *distorsionador que podría llevar a pensar, erróneamente, que las variables mantienen un grado de relación lineal mayor del que realmente existe. Para confirmar la existencia de una alta correlación se calculará el* coeficiente de correlación lineal de Pearson.

3. *Análisis de la correlación. Se selecciona la secuencia de opciones* `Estadísticos→Resúmenes→Test de correlación`, *eligiéndose en el cuadro de diálogo las variables que interesan. La salida que ofrece* **Rcmdr** *es:*

```
> cor.test(Datos$ALTURA, Datos$PESO, alternative=''two.sided'',
method=''pearson'')
    Pearson's product-moment correlation
data: Datos$ALTURA and Datos$PESO
t = 15.8396, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7818060 0.8952982
sample estimates:
cor
0.8480039
```

El coeficiente de correlación es positivo y relativamente alto, $r = 0,848$, *lo que indica que existe relación directa entre las variables. En cuanto a la intensidad, el coeficiente de determinación* $R^2 = r^2 = 0,719$ *implica que un* $28\,\%$ *de la variación de* Y *no se explica por* X *a través de la recta de ajuste.*

En este momento, y si no se hubiera detectado la bimodalidad en el histograma, habría que plantearse la posibilidad de mejorar la función de ajuste utilizando una clase funcional que se adaptara mejor a la nube de puntos; en el diagrama de dispersión se ha visto que la regresión suavizada sugería la posibilidad de un crecimiento de tipo parabólico o potencial. Pero como ya se ha comentado antes, la bimodalidad del histograma parece indicar la confusión de dos poblaciones. En efecto, se están considerando conjuntamente los dos sexos, hombre y mujer, cuando los patrones de relación peso–altura no tienen porqué coincidir y de hecho no lo hacen. Si se observa atentamente el diagrama de dispersión se puede entrever la existencia de dos poblaciones, para confirmarlo se representará el diagrama de dispersión pero diferenciando los individuos de ambos sexos.

4. *Análisis por grupo. En* **Rcmdr** *se eligen las opciones* `Gráficas→ Diagrama de dispersión...`, *seleccionando en la ventana de diálogo la opción* `Gráfica por grupos...` *la variable* `sexo`. *La visualización del gráfico 3.4 es muy elocuente, las dos líneas de ajuste se acomodan mucho mejor a sus respectivos grupos y la regresión suavizada, al contrario de lo que ocurría antes, no presenta desviaciones claras de la linealidad. Por lo que procede ajustar de forma diferenciada las variables* `peso-altura` *para cada sexo.*

   Para dividir el conjunto de datos según la variable `SEXO`, *se procede en* **Rcmdr** *desde* `Datos→Datos activos→ Filtrar los datos activos...` *tomando como expresión de selección* `SEXO==''Mujer''` *para la muestra femenina y* `SEXO==''Varón''` *para la masculina.* **R** *crea nuevos conjuntos de datos con los nombres que se le hayan indicado en el correspondiente apartado de la opción de filtrado. En este caso se han denominado* `Peso_Altura_Mujer` *y* `Peso_Altura_Varon`, *respectivamente.*

   Para analizar cada grupo de sexo, se elige como juego de datos activos el que interese y se calcula su coeficiente de correlación de Pearson. Se observa como la correlación para las mujeres es de $0,897$, *mientras que para los hombres llega hasta* $0,928$, *con* R^2 *iguales, respectivamente a* $0,804$ *y* $0,861$, *mucho más altas que las que se tenían para el ajuste conjunto.*

```
> cor.test(Peso_Altura_Mujer$ALTURA, Peso_Altura_Mujer$PESO,
alternative=''two.sided'', method=''pearson'')
    Pearson's product-moment correlation
data: Peso_Altura_Mujer$ALTURA and Peso_Altura_Mujer$PESO
t = 13.4879, df = 44, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8208994 0.9422066
sample estimates:
cor
0.8973532
```

```
> cor.test(Peso_Altura_Varon$ALTURA, Peso_Altura_Varon$PESO,
alternative=''two.sided'', method=''pearson'')
    Pearson's product-moment correlation
data: Peso_Altura_Varon$ALTURA and Peso_Altura_Varon$PESO
t = 13.0335, df = 52, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8793910 0.9580797
sample estimates:
cor
0.9285171
```

5. *Recta de ajuste. Se obtendrá ahora una de las dos rectas de ajuste del peso en función de la altura, concretamente se ha elegido el subgrupo de los hombres. Una vez elegido el conjunto de datos activo correspondiente a los hombres, se selecciona* `Estadísticos`→ `Ajuste de modelos`→`Regresión lineal...`, *y en la ventana de la figura 3.6, se elige* `PESO` *como variable explicada y* `ALTURA` *como variable explicativa.*

```
> RegModel.1 <- lm(PESO~ ALTURA, data=Peso_Altura_Varon)
> summary(RegModel.1)
Call:
lm(formula = PESO ~ ALTURA, data = Peso_Altura_Varon)
Residuals:

Min        1Q        Median        3Q        Max
-13.578    -2.091    -0.491        2.213     9.662

Coefficients:

            Estimate        Std. Error      t value      Pr(> |t|)
(Intercept)    -164.09760     13.89222      -11.81     2.43e-16 ***
ALTURA         1.41331        0.07837       18.03      < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.937 on 52 degrees of freedom
Multiple R-Squared: 0.8621, Adjusted R-squared: 0.8595
F-statistic: 325.2 on 1 and 52 DF, p-value: < 2,2e-16
```

A la vista de los resultados se sabe que la recta de regresión es `Y=-164,09760 +1,41331X`*. Si sólo se quisieran obtener los coeficientes de la recta éstos se pueden obtener con las órdenes:*

```
> RegModel.1 <- lm(PESO~ ALTURA, data=Peso_Altura_Varon)
> coef(RegModel.1)
(Intercept) ALTURA
-164.097600 1.413306
```

6. *Valores ajustados y predicciones. Para obtener los valores ajustados por el modelo se selecciona* `Modelos→ Añadir las estadísticas de las observaciones a los datos...` *y se marcan las opciones deseadas, en este caso* `Valores ajustados` *y* `residuos`*.* **R** *añade al conjunto de datos activos dos nuevas columnas llamadas* `fitted`*.RegModel.1 y* `residuals`*.RegModel.1 con los correspondientes valores ajustados y residuos del modelo activo.*

   Al realizar las estadísticas descriptivas de Y, Y^* y e, seleccionando las opciones media y desviación típica en resúmenes numéricos, se tiene:

```
> numSummary(Hombres[,c(''fitted.RegModel.1'', ''PESO'',
''residuals.RegModel.1'')], statistics=c(''mean'', ''sd''))


                       mean            sd          n
fitted.RegModel.1      8.624074e+01    9.753284    54
PESO                   8.624074e+01    10.504150   54
residuals.RegModel.1   -3.781456e-17   3.900081    54
```

   y efectivamente se comprueba que $S_Y^2 = S_{Y^*}^2 + S_e^2$, ya que $10,504^2 = 9,753^2 + 3,9^2$; pudiéndose calcular el coeficiente de determinación como $R^2 = \frac{9,753^2}{10,504^2} = 0,8621$.

   Para realizar predicciones para cualquier valor de X, se necesita crear previamente un nuevo conjunto de datos, que en este caso se ha llamado `pred` *y que contendrá una variable cuyo nombre se hace coincidir con el nombre de la variable independiente del modelo:*

```
>pred<-data.frame(ALTURA=c(180.3,184.7,193.1,197.0,201.8))
```

   Se incluyen en el fichero `pred` *los valores 180.3, 184.7, 193.1, 197.0 y 201.8 cms. Seguidamente se asigna a la variable* `predicPESO` *las predicciones que genera el modelo con la orden* `predict` *para los valores de la variable* `ALTURA` *del* `data.frame pred`*:*

```
> predicPESO <-predict(nombreModelo,pred)
```

   Por último se añade la variable `predicPESO` *al conjunto de datos* `pred`*:*

```
>pred<-data.frame(pred,predicPESO)
```

El nuevo conjunto de datos se puede ver en la figura 3.6. Puesto que el rango de valores de la altura es $(167, 194)$, se estarían realizando tres interpolaciones y dos extrapolaciones para los valores 197,0 y $201, 8$; además, puesto que $\bar{x} = 177, 1$, la predicción más fiable corresponde al valor 180,3 y la menos al valor 201,8.

7. *Análisis de Residuos. Para obtener los residuos, tanto absolutos como estudentizados, se selecciona de nuevo* `Modelos→ Añadir las estadísticas de las observaciones a los datos...` *y se marcan las opciones correspondientes, generándose por parte de* **R** *dos nuevas columnas en el fichero de datos activos, denominadas* `residuals.(RegModel.1)` *y* `rstudent.(RegModel.1)`*, donde RegModel.1 hace referencia al modelo usado.*

   Aunque en este capítulo se está abordando la regresión desde un punto de vista descriptivo y por tanto no se exigen condiciones a los datos, resulta interesante hacer una diagnosis de los residuos que detecte básicamente problemas de mala elección del modelo, existencia de otras variables relevantes, presencia de valores atípicos,... Para ello se suelen utilizar algunas representaciones gráficas, entre las que destacan la de `Residuos indexados` *y la de* `Residuos frente a ajustados`*. De su observación se pueden extraer valiosas conclusiones.*

- ***Residuos indexados***. *Detecta sobre todo problemas relacionados con la influencia que valores previos de la variable X ejercen sobre los posteriores. Ocurre sobre todo cuando la variable independiente es el tiempo, desde el punto de vista estadístico se dice que existe un problema de autocorrelación y la solución pasa por enfocar el tema desde la óptica de las series* temporales. *El gráfico de los residuos indexados se obtiene desde* `Gráficas→Gráfica secuencial...` *seleccionando la variable* `residuals.RegModel.1`*, la opción* `Identificar puntos con el ratón` *y por último elegir la representación por* `puntos`*. En este caso, la figura 3.7 presenta una distribución de residuos* *sin ninguna relación y no se obtiene mayor anormalidad que la existencia de los candidatos a valores atípicos.*

- ***Residuos estudentizados frente a valores ajustados.*** *Es probablemente el gráfico que proporciona más información sobre la calidad del ajuste realizado, informando sobre la falta de linealidad de la relación, la presencia de valores atípicos, la existencia de terceras variables que aportarían información relevante sobre Y, etc.*

  Usando las opciones `Gráficas→Diagrama de dispersión...`*, tomando* `fitted.RegModel.1` *como variable explicativa y* `rstudent.RegModel.1` *como explicada, se obtiene la figura 3.8. En el que, al igual que en el gráfico de residuos indexados, sólo destaca la presencia de los candidatos a valores atípicos.*

- ***Obtención de valores influyentes.*** *Se buscan ahora valores especialmente determinantes a la hora de estimar los parámetros del modelo. Normalmente estos valores van a coincidir con valores extremos para una de las dos variables. Uno de los criterios para detectar estos valores influyentes se basa en el cálculo de la distancia de* Cook. *La distancia de Cook para la observación i-ésima calcula la diferencia entre los parámetros del modelo que se obtiene incluyendo la observación i-ésima y sin incluirla. En general se deben tener en cuenta aquellas observaciones cuya distancia de Cook sea mayor que 1. La figura 3.8, se genera a través de* `Gráficas→ Gráfica secuencial...` *y se puede apreciar que los valores más influyentes coinciden con las observaciones 41, 61 y 66.*

  Otra forma de ver la influencia de una observación es a través de su potencial, que estima el peso de cada observación a la hora de realizar predicciones. Los potenciales se obtienen como los elementos de la diagonal principal de la matriz de Hat, $H = X(X'X)^{-1}X'$. En la figura 3.9 se tienen la representación indexada de los potenciales Hat, realizada a partir de la misma opción gráfica anterior. Los puntos influyentes serían aquellos que superaran el doble del cociente entre el número de variables regresoras más uno y el número de observaciones. En este caso el valor de referencia es $0,074$ y los

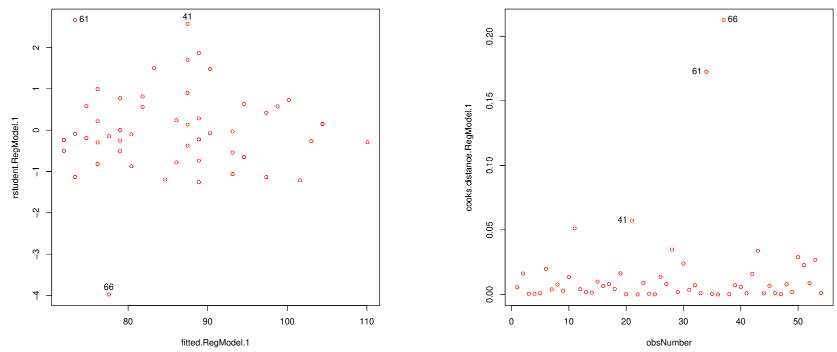


Figura 3.8: Residuos estundentizados frente a Y^* y distancias de Cook

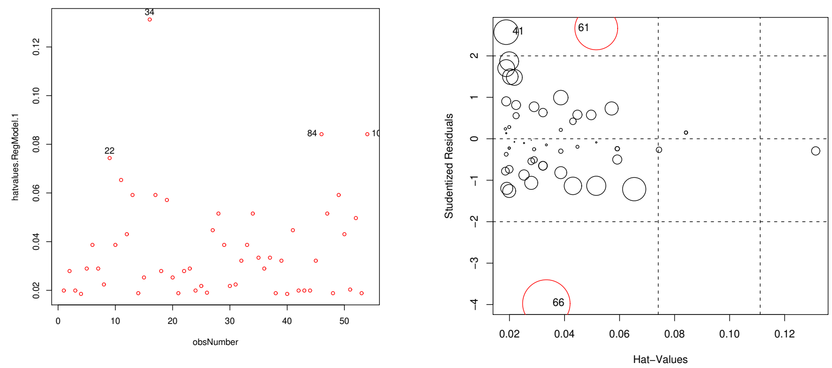


Figura 3.9: Potenciales Hat y puntos influyentes

puntos que superan esta cota son el 32, el 34, el 84 y el 100.

Por último, la gráfica de potenciales hat frente a residuos estudentizados, donde cada observación está identificada por un círculo cuyo diámetro es proporcional a su distancia de cook, sintetiza toda la información a tener en cuenta a la hora de identificar los puntos influyentes. La gráfica ha sido creada desde `Modelos`→`Gráficas`→ `Gráfica de influencia` *y refleja de nuevo que los valores a considerar son el 61 y el 66, ver figura 3.9.*

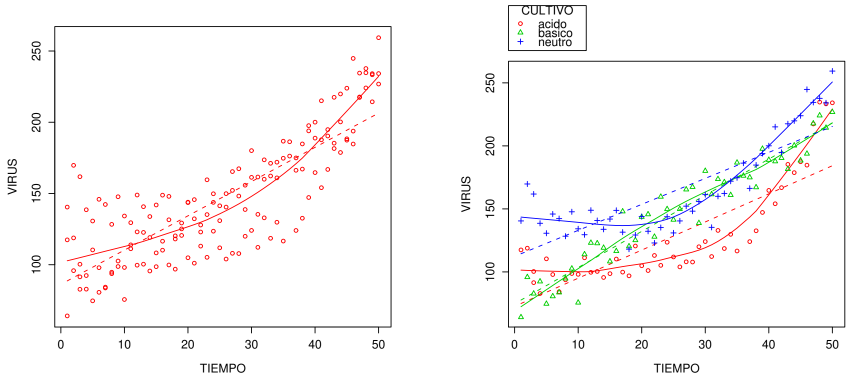


Figura 3.10: Dispersión y dispersión según cultivo

R–Nota 3.3

Supóngase un conjunto de datos del cual se desea obtener un modelo para un subconjunto de estos datos. Por ejemplo en los datos `peso_altura` *se quiere hacer un modelo para los datos femeninos, se selecciona* `Estadísticos`→`Ajuste de modelos`→`Regresión lineal...` *y en la ventana de diálogo aparecerá la opción* `Expresión de selección` *donde se puede elegir el subconjunto deseado, en este caso* `SEXO=='Mujer'`. *El problema surge si se quiere añadir, por ejemplo, la columna de valores ajustados seleccionando* `Modelos`→ `Añadir estadísticas de las observaciones a los datos...`, *esto se debe a que el conjunto de datos activos no se corresponde con el modelo activo, para solucionar esto, sólo se debe hacer en primer lugar el filtrado de los datos para el subconjunto y seguidamente aplicar el modelo.*

Ejemplo 3.4

Para ilustrar la realización de un ajuste de tipo polinomial, se consideran los datos del fichero `reproduccion_vir.dat` *en el que se muestran el* `número de virus` *reproducidos en función del* `tiempo (minutos)` *y de la* `temperatura (grados)`, *según el* `tipo de cultivo (ácido,`

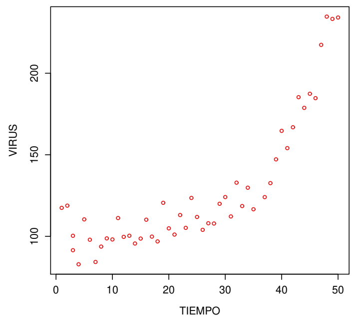


Figura 3.11: Diagrama de dispersión del cultivo ácido

`básico o neutro)`. *Se está interesado en ver como influye el tiempo en el número de virus.*

Se realiza en primer lugar el diagrama de dispersión de la variable `número de virus` *frente al* `tiempo`. *La observación de la figura 3.10 revela para el conjunto de datos una disposición no lineal, aunque la evidente variabilidad presente en cualquier rango de valores del tiempo hace presuponer que el factor* `tipo de cultivo` *debería tenerse en cuenta (figura 3.10).*

Si se rehace el gráfico para cada uno de los subgrupos que determina la variable cultivo, se observa que los cultivos de `tipo básico` *tienen un comportamiento aproximadamente lineal, mientras los de* `tipo neutro` *y* `ácido` *no lo tienen.*

El estudio se centrará en el cultivo ácido, para ello se filtran los datos (se almacenan como `reproduccion_vir_acido`*) y se representan de nuevo. El diagrama de dispersión, figura 3.11, parece sugerir un comportamiento de tipo parabólico.*

Para realizar el ajuste parabólico se selecciona `Estadísticos`→ `Ajuste de modelos`→`Modelo lineal...`, *tomando como fórmula del modelo VIRUS∼ 1+ TIEMPO+ I(TIEMPOˆ2) (figura 3.12). Los resultados obtenidos son:*

```
> LinearModel.3 <- lm(VIRUS ~ 1 + TIEMPO +I( TIEMPO^2),
data=acido)
summary(LinearModel.1)
 Call:
lm(formula = VIRUS ~ 1 + TIEMPO + I(TIEMPO^2), data = acido)

Residuals:

Min       1Q       Median     3Q     Max
-23.295   -6.140   1.510      6.491  24.271

Coefficients:
Estimate Std. Error t value Pr(> |t|)
(Intercept) 115.552345 4.917038 23.500 < 2e-16 ***
TIEMPO -2.901809 0.455127 -6.376 7.25e-08 ***
I(TIEMPO^2) 0.101647 0.008731 11.642 1.89e-15 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 11.73 on 47 degrees of freedom
Multiple R-Squared: 0.9179, Adjusted R-squared: 0.9144
F-statistic: 262.8 on 2 and 47 DF, p-value: < 2.2e-16
```

Se concluye que el tiempo explica casi el 92 % del número de virus a través del ajuste parabólico estimado.

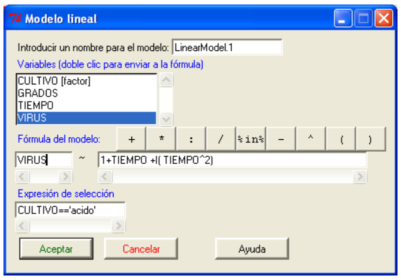
Después de representar el gráfico de dispersión de la variable `VIRUS` *frente al* `TIEMPO` *(de los datos* `reproduccion_vir_acido`*) (figura 3.11) es posible representar en la misma ventana la parábola del modelo (figura 3.12) mediante las instrucciones:*

```
> x<- seq(0,50)
> y<- 115,552345 – 2,901809*x + 0,101647*x^2
> lines(x,y,col=''green'')
```

Llegados a este punto, se podría plantear si los datos se ajustarían mejor a un polinomio de grado tres. Aunque no existen evidencias en el gráfico de dispersión, se procederá a realizar este ajuste por motivos básicamente pedagógicos.

Al ser un modelo más general que el parabólico se producirá una mejora del ajuste, aunque la cuestión es si esta mejora es lo suficientemente importante para justificar la mayor complejidad del modelo.

Para realizar el ajuste de grado tres, se selecciona `Estadísticos→ Ajuste de modelos→Modelo lineal...`*, tomando como fórmula del modelo VIRUS~ 1+ TIEMPO+ I(TIEMPO^2)+I(TIEMPO^3) (figura 3.13).*



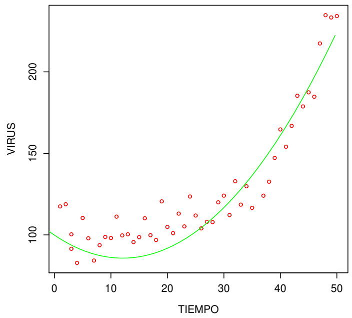


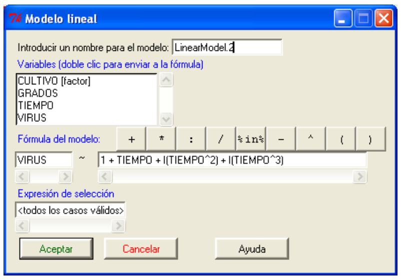
Figura 3.12: Opciones y representación del modelo parabólico

```
> summary(LinearModel.2)
 Call:
lm(formula = VIRUS ~ 1 + TIEMPO + I(TIEMPO^2) + I(TIEMPO^3),
data = Virus_acido)
Residuals:

Min        1Q       Median     3Q      Max
-21.1995   -5.1259  -0.1860    7.1273  21.0148

Coefficients:

                 Estimate       Std. Error       t value       Pr(> |t|)
(Intercept)      98.1018701       5.6855078        17.255       < 2e-16 ***
TIEMPO           1.1938655        0.9905237        1.205        0.2343
I(TIEMPO^2)      -0.1006612       0.0457034        -2.202       0.0327 *
I(TIEMPO^3)      0.0026659        0.0005944        4.485        4.83e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 9.892 on 46 degrees of freedom
Multiple R-Squared: 0.9429, Adjusted R-squared: 0.9392
F-statistic: 253.2 on 3 and 46 DF, p-value: < 2.2e-16
```

El coeficiente de determinación es igual a 0,9429, con una mejora de un 2 %, lo que no parece justificar la adopción de este modelo más complejo. Igual que antes es posible representar el ajuste cúbico como puede observarse en la figura 3.13.

R–Nota 3.4

Para realizar un ajuste polinomial con ***Rcmdr*** *se selecciona la opción*



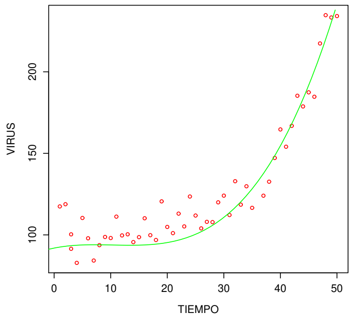


Figura 3.13: Opciones y representación del modelo cúbico

`Estadísticos→Ajustes de modelos→Modelo lineal...` *y en la ventana de diálogo se escribe la expresión del modelo deseado:*

- *Para indicar un modelo lineal con término independiente se escriben cualquiera de las dos fórmulas siguientes:*

$$Y \sim X$$

$$Y \sim 1 + X$$

- *Si se desea omitir el término independiente en un modelo lineal se utiliza una de las fórmulas siguientes:*

$$Y \sim -1 + X$$

$$Y \sim 0 + X$$

- *En general para un modelo polinomial con término independiente se escribe:*

$$Y \sim X + I(X^2) + I(X^3) + \cdots + I(X^n) \text{ o bien}$$

$$Y \sim 1 + X + I(X^2) + I(X^3) + \cdots + I(X^n)$$

  y con un -1 *ó* 0 *para un modelo sin término independiente.*

Si se quiere observar la notación que utiliza ***R*** *para formular estos modelos, véase el apéndice C.*

4. Ejercicios

3.1 Para los datos del fichero `peso_altura.dat`, analice el comportamiento del peso en función de la altura para el grupo de las mujeres.

3.2 La tabla 3.2 muestra una serie histórica sobre el olivar español que recoge la superficie, rendimiento y producción, durante el periodo 1965-1979, donde:
X = Superficie en miles de Ha.
Y = Rendimiento en Qm/Ha.
Z = Producción en miles de Tm.

Se pide:

a) El diagrama de dispersión de las variables X e Y.

b) Las medidas más representativas para cada una de las variables, indicando su representatividad.

c) El estudio de la relación entre las variables XY, XZ e YZ.

3.3 La siguiente tabla muestra la relación existente entre la lluvia caída, en l/m^2, en el periodo octubre–mayo y la producción obtenida en kilogramos por olivo.

X	300	400	500	600	700
Y	13	26	40	57	64
Y	24	21	31	45	69
Y	17	17	38	51	57
Y	11	26	34	58	76
Y	20	30	27	44	74

donde X representa la lluvia e Y la producción.

a) Represente el diagrama de dispersión.

b) Indique si existe alguna tendencia.

c) Cuantifique y comente la relación existente entre las dos variables.

Año	X	Y	Z
1965	73,6	69,8	8,5
1966	98,1	62,5	6
1967	99,8	98,5	8,7
1968	107,7	102,5	6
1969	107,7	97,4	3,7
1970	122	113,8	8,9
1971	127	118	7,9
1972	138,1	128,1	10,1
1973	152,1	145,8	6,8
1974	144,8	139,8	5
1975	160,7	152,9	11,1
1976	150,2	143,4	9,8
1977	152,1	146	9,5
1978	167,3	162,1	10,8
1979	165	160,2	10

Tabla 3.2: Datos ejercicio 3.2

3.4 Dada la siguiente tabla de doble entrada con valores porcentuales:

$Y \setminus X$	2	3	4
0	$0,22$	$0,13$	$0,04$
1	$0,16$	$0,11$	$0,05$
2	$0,08$	$0,16$	$0,05$

a) Obtenga la distribución marginal de X. Calcule su media, moda y mediana.

b) Calcule la media de Y cuando X toma el valor 3.

c) Estudie la dependencia de las variables X e Y.

3.5 A un grupo de estudiantes se les preguntó por el tiempo que tardan en llegar desde su hogar hasta la facultad, X (minutos), el tiempo que le dedican diariamente al estudio, Y (horas), y las calificaciones obtenidas en la asignatura de Estadística, Z, obteniéndose las siguientes respuestas:

(40, 4, 4), (45, 3, 3), (30, 4, 5), (40, 4, 5), (80, 2, 5), (20, 3, 5)
(10, 1,5, 6), (10, 4, 6), (20, 4, 6), (45, 3, 3), (20, 4, 4), (30, 4, 7)
(30, 3, 7), (20, 4, 6), (30, 1, 6), (10, 5, 5), (15, 5, 5), (20, 6, 5)
(20, 3, 7), (20, 4, 5), (20, 5, 6), (60, 2, 3), (60, 5, 5)

a) Obtenga el diagrama de dispersión correspondiente al tiempo dedicado al estudio y las calificaciones obtenidas en Estadística.

b) ¿Se aprecia alguna tendencia?

c) Estudie las relaciones existentes entre XY, XZ e YZ.

3.6 Al mismo grupo del ejercicio anterior se le ha pedido que escriba un dígito al azar entre 0 y 9 así como el número de hermanos que tiene, obteniéndose los siguientes pares de valores:

(7, 4), (0, 1), (2, 1), (2, 0), (9, 4), (7, 4), (6, 3), (8, 5)
(7, 3), (3, 2), (7, 3), (2, 1), (7, 4),(7, 3), (8, 4), (8, 5)
(5, 3), (3, 1), (4, 2), (4, 2), (5, 3), (2, 0), (4, 2)

¿Existe alguna relación entre las variables?, ¿de qué tipo?

3.7 Se examinan 300 alumnos de una asignatura y durante el examen se les pregunta por el tiempo que han dedicado a su preparación (menos de una hora, entre una hora y tres, más de tres), obteniéndose la siguiente tabla de calificaciones según el tiempo de estudio:

Nota \ Horas Estudio	< 1	$1 - 3$	> 3
Suspenso	43	32	10
Aprobado	31	48	81
Notable	7	13	20
Sobresaliente	3	4	8

¿Están relacionadas las calificaciones con las horas de estudio?

3.8 Dada la distribución:

X	1	$1,5$	2	$2,5$	3	$3,75$	$4,5$	5
Y	1	$1,5$	$2,95$	$5,65$	$8,8$	15	25	32

a) Elija la mejor clase funcional para ajustar Y/X y estime sus parámetros.

b) Establezca la bondad del ajuste.

c) Calcule la previsión para Y cuando $X = 7$. Analice dicha previsión.

3.9 Dada la distribución:

X	$2,5$	$3,75$	5	$7,5$	10	$12,5$	20
Y	8	14	$23,75$	40	62	90	165

a) Utilice una ecuación del tipo aX^b para ajustar Y/X.

b) Dé una medida de la bondad del ajuste.

3.10 Dada la distribución:

X	1	$1,5$	2	3	4	5	6	7
Y	1	$1,75$	$2,65$	$4,7$	7	$9,5$	12	15

a) Ajuste Y/X utilizando una función del tipo aX^b.

b) Analice la bondad del ajuste.

3.11 Dada la distribución:

X	5	6	8	10	13	18	20
Y	$1,5$	$1,25$	$0,93$	$0,7$	$0,46$	$0,23$	$0,15$

a) Estime los parámetros de la clase funcional $ab^{-0,2X}$ para ajustar Y/X.

b) Estudie la bondad del ajuste.

Capítulo 4

Distribuciones de Probabilidad

La existencia de fenómenos o experimentos no determinísticos, donde el conocimiento de las condiciones en las que éstos se desarrollan no determinan los resultados, hace imprescindible el uso de una función que asigne niveles de certidumbre a cada uno de los desenlaces del fenómeno y ahí es donde aparece la *teoría de la probabilidad*. Los experimentos o fenómenos que poseen la característica anterior se denominan aleatorios. Intuitivamente, la concreción numérica del fenómeno mediante la asignación de valores con un cierto criterio, da origen a la *variable aleatoria*. Una correcta proyección de estos conceptos es lo que va a permitir estudiar grandes colectivos a partir de pequeñas partes de ellos, llamadas muestras, dando lugar a lo que se conoce como *inferencia estadística*.

La teoría de la probabilidad y la variable aleatoria van a permitir establecer un amplio catálogo de modelos teóricos, tanto discretos como continuos, a los cuales se van a poder asimilar muchas de las situaciones de la vida real. El estudio de los modelos teóricos, incluyendo la caracterización a través de sus parámetros, el cálculo de probabilidades en sus distintos formatos y la generación de números aleatorios, van a facilitar enormemente el análisis de estas situaciones reales. Ese será el objetivo del capítulo.

Antes de entrar en materia se describirán una serie de fenómenos que se podrán asimilar a las distribuciones de probabilidad que se describirán en este capítulo.

Ejemplo 4.1

- *Si se contesta al azar un examen tipo test de 10 preguntas, donde cada una de ellas tiene 4 posibilidades siendo sólo una de ellas cierta, ¿qué número de aciertos es más probable?*

- *Cuando alguien pregunta por el número que salió en el sorteo de la ONCE, la respuesta suele ser la unidad de dicho número: el 7, el 5,... ¿cómo se distribuyen las unidades de los premios en el sorteo de la ONCE?*

- *En las oposiciones es frecuente que se realice un sorteo público extrayendo una serie de bolas o papeletas de una urna o bolsa. Imagínese un opositor que se ha preparado 60 temas de 100, de los que se seleccionan al azar dos de ellos, ¿qué probabilidad tiene el opositor de que sea elegido al menos uno de los temas que lleva preparado?*

- *Sabemos que el servicio de autobuses entre Cádiz y San Fernando tiene salidas cada media hora entre las 6 am y las 12 pm, una persona que se ha olvidado el reloj en casa llega a la estación de autobuses en Cádiz ¿cuál es la probabilidad de que espere menos de 10 minutos para coger el autobús?*

- *Se sabe que las bombillas de bajo consumo de 14 w tienen una vida media útil de 10000 horas, mientras que las bombillas clásicas por incandescencia de 60 w tienen una vida media útil de 1000 horas. Si cada día se encienden unas 4 horas ¿cuál es la probabilidad de que después de un año estén funcionando las dos?, ¿y ninguna de ellas?, ¿y al menos una de ellas?, ¿y como mucho una de ellas?*

- *Si se controlan el peso, la edad, la estatura, la talla de pantalón, las horas de estudio, la nota de selectividad, ... de los 350 alumnos que están matriculados en 1º de Empresariales y Económicas en el campus de Cádiz y Jerez, ¿qué estructura tiene su distribución?*

Cada una de las situaciones anteriores conlleva la realización de un experimento aleatorio: “elegir una de las cuatro posibles respuestas en cada una de las preguntas”, “extraer la bola del número de las unidades entre las 10 posibles”, “sacar 2 temas entre 100”, ..., que proporcionan resultados de distinta naturaleza. Así, el número de aciertos que se puede obtener al responder las 10 preguntas “variará” entre 0 y 10, o sea, tiene un número finito de posibles valores, mientras que el tiempo de espera para coger el autobús puede tomar infinitos valores dentro del intervalo $(0, 30)$, sólo condicionado por la precisión de los aparatos de medición. Esto lleva a una primera gran clasificación entre modelos de probabilidad discretos y continuos. El primer problema a resolver será la elección del modelo teórico apropiado para cada caso en estudio.

Para tener un buen manejo matemático de las distintas situaciones que se puedan plantear dada la distinta naturaleza y la diversidad de los resultados que proporcionan los experimentos, se necesita realizar una abstracción cuantificada del experimento. Para ello se asignará a cada uno de los posibles resultados del experimento aleatorio (suceso elemental) un número real. A esta aplicación se le llamará *variable aleatoria* y se designará por X, $X : \Omega \to R$. Así en el primer ejemplo, la variable aleatoria consistiría en asignar al suceso “responder correctamente siete preguntas” el número 7. Esta asignación no es única, se le podría haber asignado otro número, por ejemplo 17, lo que proporcionaría otra variable aleatoria, pero en este caso los valores no serían fácilmente identificables en términos del experimento de partida. Como norma, se intentará que la asignación se realice de la forma más natural posible.

Además, por abuso de lenguaje, se tiende a confundir la aplicación X con los valores del conjunto imagen y se traslada la probabilidad de ocurrencia de un suceso al valor correspondiente de la variable aleatoria; por lo tanto, se puede hablar de la probabilidad de que la variable aleatoria tome un determinado valor. Las probabilidades asociadas a cada uno de los valores de la variable aleatoria pueden ser organizadas como una distribución de probabilidad, expresándose mediante una tabla, una gráfica o una fórmula, denominándose en este último caso, a la regla de correspondencia valores–probabilidades, *función de probabilidad*.

DISCRETAS		
Distribución	**Parámetros**	**En Rcmdr**
Binomial	$n = size; p = prob$	binom
Binomial negativa	$n = size; p = prob$	nbinom
Geométrica	$p = prob$	geom
Hipergeométrica	$(N, K, n) = (m, n, k)$	hyper
Poisson	$\lambda = lambda$	pois

Tabla 4.1: Tabla de distribuciones discretas

Como se ha indicado, según la naturaleza de la variable aleatoria pueden considerarse distribuciones de probabilidad *discretas* o *continuas*. Las principales distribuciones de probabilidad de variables discretas son: *Binomial*, *Binomial Negativa*, *Geométrica*, *Hipergeométrica* y de *Poisson*. Entre los modelos de variable continua destacan las distribuciones: *Normal*, *T-Student*, *Chi-Cuadrado*, *F-Snedecor*, *Exponencial*, *Uniforme*, *Beta*, *Cauchy*, *Logística*, *Lognormal*, *Gamma*, *Weibull* y *Gumbel*. Todas estas distribuciones están recogidas en **Rcmdr**. Se puede acceder a ellas en: `Distribuciones`→`Distribuciones continuas`, o en `Distribuciones`→`Distribuciones discretas`, o también escribiendo directamente en la ventana de instrucciones el nombre de la distribución, poniendo delante una `d`, si se quiere *la función de densidad*, una `p` para la *función de distribución*, una `q` para los *cuantiles* y una `r` para generar una *muestra aleatoria* de la distribución; además, por supuesto, de los argumentos necesarios en cada caso.

1. Distribuciones discretas

En la tabla 4.1 están resumidas todas las distribuciones contenidas en la versión actual de **Rcmdr**, sus parámetros (el nombre teórico y el usado en el programa) y las instrucciones correspondientes. Para cada una de las distribuciones discretas están disponibles las siguientes opciones:

- **Cuantiles:** Permite calcular el valor de la variable que deja a derecha o a izquierda (según se seleccione) una determinada probabilidad.

- **Probabilidades:** Determina la probabilidad de que la variable tome un valor dado.

- **Gráfica de la distribución:** Genera la gráfica de la función de cuantía o de distribución.

- **Muestra de la distribución:** Genera muestras aleatorias extraídas de la distribución.

- **Probabilidades Acumuladas:** Calcula bien el valor de $P(X \leq x)$ (cola de la izquierda), o bien, $P(X > x)$ (cola de la derecha) para cada valor x.

Con el fin de familiarse con las distribuciones y su uso desde **Rcmdr**, se verán ahora algunos ejemplos representativos de las distribuciones más usuales.

1.1. Distribución Binomial

Ejemplo 4.2

Si un estudiante responde al azar a un examen de 8 preguntas de verdadero o falso.

a) ¿Cuál es la probabilidad de que acierte 4?

La variable X=“número de aciertos” sigue una distribución Binomial de parámetros $n = 8$ y $p = 1/2$. Para calcular las probabilidades en ***Rcmdr*** *se selecciona:* `Distribuciones`→`Distribuciones discretas`→`Distribución binomial`→`Probabilidades binomiales...`

En este caso se introduce `Ensayos binomiales= 8` y `Probabilidad de éxito= 0.5` y se puede ver que $P(X = 4) = 0{,}2734375$.

```
>.Table <- data.frame(Pr=dbinom(0:8, size= 8, prob= 0.5))
>rownames(.Table) <- 0:8
>.Table


  Pr
0 0.00390625
1 0.03125000
2 0.10937500
3 0.21875000
4 0.27343750
5 0.21875000
6 0.10937500
7 0.03125000
8 0.00390625
```

b) ¿Cuál es la probabilidad de que acierte 2 o menos?
Se calculan ahora las probabilidades acumuladas: `Distribuciones→ Distribuciones discretas→Distribución binomial→ Probabilidades binomiales acumuladas...` *Para calcular la probabilidad de que acierte 2 preguntas o menos, en la ventana que aparece, se debe indicar* `Valor de la variable= 2` *y* `Ensayos binomiales= 8`*, dejando marcada la opción* `Cola izquierda`*.*

```
>pbinom(c(2), size= 8, prob= 0.5, lower.tail=TRUE)
[1] 0.1445313
```

c) ¿Cuál es la probabilidad de que acierte 5 o más?
Para determinar la probabilidad de que acierte 5 o más preguntas se realiza el mismo procedimiento, pero señalando en la ventana emergente `Valor de la variable= 4`*, y* `Ensayos binomiales= 8`*, tomándose la opción* `Cola Derecha`*.*

```
>pbinom(c(4), size=8, prob=0.5, lower.tail=FALSE)
[1] 0.3632813
```

1.2. Distribución de Poisson

Ejemplo 4.3

Una cierta área de Estados Unidos es afectada, en promedio, por 6 hura-

canes al año. Encuentre la probabilidad de que en un determinado año esta área sea afectada por:

a) Menos de 4 huracanes.

Se define la variable $X =$“número de huracanes por año” y se sabe que ésta se distribuye mediante una Poisson, porque describe el número de éxitos por unidad de tiempo y porque son independientes del tiempo desde el último evento. Se calcularán ahora las probabilidades:

Como en el caso anterior se señala `Probabilidades binomiales acumuladas...` *tomando ahora en la ventana emergente* `Valor(es) de la variable= 4`*, y* `Media= 6`*, para la opción* `Cola izquierda`*.*

```
>ppois(c(3), lambda = 6, lower.tail=TRUE)
[1] 0.1512039
```

b) Entre 6 y 8 huracanes.

Para calcular la probabilidad de que ocurran entre 6 y 8 huracanes, se pueden sumar las probabilidades $P(X = 6) + P(X = 7) + P(X = 8)$ o restar las probabilidades acumuladas, con la opción `Cola izquierda`*, $P(X \leq 8) - P(X \leq 5)$. Como antes se realizan en primer lugar las probabilidades acumuladas y se restan los resultados obtenidos:*

```
>a <- ppois(c(8), lambda = 6, lower.tail=TRUE)
>b <- ppois(c(5),lambda = 6, lower.tail=TRUE)
>a-b
[1] 0.4015579
```

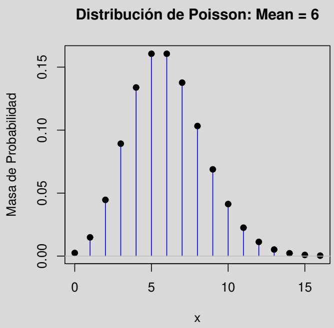


Fig. 4.1: Distribución de Poisson

c) Represente la función de probabilidad de la variable aleatoria que mide el número de huracanes por año. La gráfica se realiza en `Distribuciones→ Distribuciones discretas →Distribución de Poisson→Gráfica de la distribución de Poisson...`*(figura 4.1).*

1.3. Distribución Hipergeométrica

Ejemplo 4.4

En un juego se disponen 15 globos llenos de agua, de los que 4 tienen premio. Los participantes en el juego, con los ojos vendados, golpean los globos con un palo por orden hasta que cada uno consigue romper 2.

a) ¿Cuál es la probabilidad de que el primer participante consiga un premio?

Para el primer participante la variable X="número de premios conseguidos entre 2 posibles" sigue una distribución Hipergeométrica de parámetros $m = 11, n = 4, K = 2$. *Para obtener respuesta a las cuestiones en* ***Rcmdr*** *se selecciona:* `Distribuciones`→`Distribuciones discretas`→`Distribución hipergeométrica...`
Para calcular la probabilidad de que consiga un sólo premio se elige la opción ***`probabilidades hipergeométricas...`****, con* *`m(número de bolas blancas en la urna)= 11`*, *`n(número de bolas negras en la urna)= 4`* *y* *`k(número de extracciones)= 2`*, *resultando* $P(X = 1) = 0{,}41904762$.

```
>.Table <- data.frame(Pr=dhyper(0:2, m=11, n=4, k=2))

>rownames(.Table) <- 0:2

>.Table

        Pr
0 0.05714286
1 0.41904762
2 0.52380952
```

b) Construya la gráfica de la función de distribución.

Ésta se obtiene en `Distribuciones`→`Distribuciones discretas`→ *`Distribución hipergeométrica`*→*`Gráfica de la distribución hipergeométrica...`*, *marcando la opción* ***`gráfica de la función de distribución`*** *(figura 4.2).*

c) Si el primer participante ha conseguido sólo un premio, ¿cuál es la probabilidad de que el segundo participante consiga otro?

Para el segundo participante la variable seguirá una hipergeométrica de parámetros *`m= 10`*, *`n= 3`* *y* *`k= 2`*, *resultando* $P(X = 1) = 0{,}38461538$.

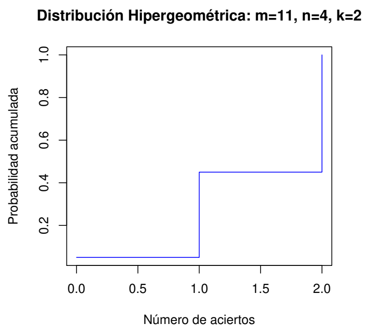


Figura 4.2: Distribución hipergeométrica

1.4. Distribución Geométrica. Distribución Binomial Negativa

Ejemplo 4.5

Un vendedor de alarmas de hogar tiene éxito en una casa de cada diez que visita. Calcula:

a) La probabilidad de que en un día determinado consiga vender la primera alarma en la sexta casa que visita.

Se define la variable X="número de casas que visita antes de conseguir vender la primera alarma", que sigue una distribución Geométrica con `Probabilidad de éxito= 0.1`*. Se selecciona en* ***Rcmdr*** `Distribuciones`→`Distribuciones discretas`→ `Distribución geométrica`→`Probabilidades geométricas...`.

Habrá que calcular la probabilidad de que tenga 5 fracasos antes del primer éxito, obteniendo de la tabla la probabilidad $P(X = 5) = 5{,}904900e{-}02$.

b) La probabilidad de que no venda ninguna después de siete viviendas visitadas.

La variable X="número de alarmas vendidas en 7 viviendas" sigue una distribución Binomial con `Ensayos binomiales= 8` *y* `Probabilidad de éxito= 0.1`*, luego en nuestro caso se tiene* $P(X = 0) = 0{,}4782969$.

c) Si se plantea vender tres alarmas, ¿cuál es la probabilidad de que consiga su objetivo en la octava vivienda que visita?

CONTINUAS		
Distribución	**Parámetros**	**En Rcmdr**
Normal	$\mu = mean; \sigma = sd$	norm
T-Student	$n = df$	t
Chi-Cuadrado	$n = df$	chisq
F-Snedecor	$n = df1; m = df2$	f
Exponencial	$\lambda = rate$	exp
Uniforme	$(a, b) = (min, max)$	unif
Beta	$p = shape1; q = shape2$	beta
Cauchy	$t = location; s = scale$	cauchy
Logística	$t = location; s = scale$	logis
Lognormal	$\mu = meanlog;\sigma = sdlog$	lnorm
Gamma	$p = shape; \alpha = scale$	gamma
Weibull	$p = shape; \alpha = scale$	weibull
Gumbel	$p = shape; \alpha = scale$	gumbel

Tabla 4.2: Tabla de distribuciones continuas

Para abordar esta cuestión, se define la variable Y= "número de casas que visita antes de conseguir vender la tercera alarma". Esta variable sigue una distribución Binomial Negativa de parámetros `Número de éxitos= 3`, `Probabilidad de éxito= 0.1`. *En* **Rcmdr** *se selecciona* `Distribuciones`→ `Distribuciones discretas`→`Distribución binomial negativa`→ `Probabilidades binomiales negativas...`, *de donde:* $P(Y = 5) = 1{,}240029e{-}02$.

2. Distribuciones continuas

En la tabla 4.2 están resumidas todas las distribuciones continuas contenidas en la versión actual de **Rcmdr**, sus parámetros (el nombre teórico y el usado en el programa) y las correspondientes instrucciones.

Para cada una de las distribuciones continuas están disponibles las siguientes opciones:

- **Cuantiles:** Permite calcular el valor de la variable que deja a derecha o a izquierda (según seleccionemos) una determinada probabilidad.
- **Probabilidades:** Determina la probabilidad que queda acumulada a izquierda (o a derecha) de un valor dado.
- **Gráfica de la distribución:** Genera la gráfica de la función de densidad o de distribución.
- **Muestra de la distribución:** Genera muestras aleatorias extraídas de la distribución.

2.1. Distribución Normal

Trabajando directamente en **R**, para calcular los cuantiles normales se usaría `qnorm`, agregando a ésta los argumentos necesarios. En concreto, para hallar el valor que, en una $N(0,1)$, deja en la cola izquierda una probabilidad de 0,25:

```
qnorm(c(.25), mean = 0, sd = 1, lower.tail = TRUE)
```

R–Nota 4.1

$lower.tail = TRUE$ usa la cola de la izquierda, mientras que $lower.tail = FALSE$ usa la derecha. Los parámetros $lower.tail = TRUE$, $mean = 0$ y $sd = 1$ pueden ser omitidos, pues son los valores por defecto en esta función.

Ejemplo 4.6

Una empresa está buscando personal para su departamento de marketing. El perfil solicitado es el de sujetos extrovertidos y creativos. Se han presentado 50 candidatos y la empresa ha establecido como criterio *de selección el que los candidatos superen el percentil 80 en creatividad y extroversión. Sabiendo que la variable extroversión* (X) *se distribuye según una Normal de media 5 y desviación típica 1, que la variable creatividad* (Y) *sigue una t-Student de 10 grados de libertad y que las puntuaciones de creatividad y extroversión son independientes:*

a) ¿Cuántos candidatos serán seleccionados?

Al ser X *e* Y *independientes, la probabilidad* $P(X \geq P_{80} \cap Y \geq P_{80}) = P(X \geq P_{80}) \cdot P(Y \geq P_{80}) = 0{,}20 \cdot 0{,}20 = 0{,}04$*. Como se han presentado* 50 *aspirantes, serán seleccionadas* $0{,}04 \cdot 50 = 2$ *personas.*

b) ¿Qué puntuaciones debe superar un aspirante en creatividad y extroversión para ser admitido?

Según el criterio de selección se debe superar el percentil 80*, en ambas variables, para ser admitido. Se calculará pues el percentil* P_{80} *de la variable* X *e* Y*, utilizando los cuantiles normales para la variable* X*:*

```
> qnorm(c(.8), mean=5, sd=1, lower.tail=TRUE)
[1] 5.841621
```

y los t-cuantiles para la variable Y*:*

```
> qt(c(.8), df=10, lower.tail=TRUE)
[1] 0.8790578
```

c) Si se extraen al azar 16 *candidatos, ¿cuál es la probabilidad de que su media aritmética en extroversión sea mayor que* 4,5*?*

Se sabe que al extraer una muestra de una población normal de tamaño n*, la media de la muestra,* $\bar{X}$*, sigue otra distribución normal de media igual que la poblacional y desviación típica* $\frac{\sigma}{\sqrt{n}}$*. Por lo que en este caso* $\bar{X} \sim N(5, \frac{1}{4})$*. Como se desea calcular* $P(\bar{X} \geq 4{,}5)$*, se selecciona* `Cola derecha` *en la entrada de* `Probabilidades normales...`

```
> pnorm(c(4.5),mean=5,sd=0.25,lower.tail=FALSE)
[1] 0.9772499
```

d) Dibuje las gráficas de densidad de las variables `Extroversión` *y* `Creatividad`*.*

Para ello se selecciona la función de densidad de ambas variables en `Distribuciones→Distribuciones Continuas...`*, obteniéndose las figuras 4.3 y 4.4.*

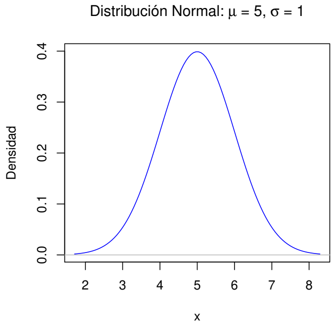


Figura 4.3: Función de densidad de la variable extroversión (normal)

2.2. Distribución Uniforme Continua

Ejemplo 4.7

Una persona informal hace esperar a su pareja aleatoriamente entre 0 *y* 90 *minutos. Harto de esta situación, la persona que sufre la espera se plantea un ultimátum; si al día siguiente su pareja tarda menos de* 15 *minutos mantiene la relación, si la espera está entre* 15 *y* 55 *minutos, decide en la siguiente cita con los mismos criterios, mientras que si tarda más de* 55 *minutos la relación termina en ese momento.*

a) Represente gráficamente la función de densidad de la variable que modeliza esta situación.

Se define la variable X=“tiempo de espera”, que sigue una distribución uniforme continua definida en el intervalo $(0, 90)$*. En* ***Rcmdr*** *se selecciona* `Distribuciones`→`Distribuciones continuas`→ `Distribución uniforme...` *Se elige* `Gráfica de la distribución uniforme...`*, marcando* `Función de densidad` *(figura 4.5).*

b) Calcule la probabilidad de que la relación continúe hasta la siguiente cita.

En `Probabilidades uniformes...` *se indica el valor de la variable y los límites del intervalo, dejando la opción* `Cola Izquierda`*.*

```
> punif(c(55), min=0, max=90, lower.tail=TRUE)
[1] 0.6111111
```

c) Calcule la probabilidad de que la relación termine en la segunda

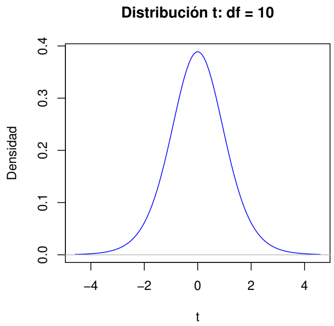


Figura 4.4: Función de densidad de la variable creatividad (t-student)

cita.

b) En `Probabilidades uniformes...` *se indica el valor de la variable y los límites del intervalo, dejando la opción* `Cola Izquierda`*.*

```
> punif(c(55), min=0, max=90, lower.tail=TRUE)
[1] 0.6111111
```

c) Suponiendo que el tiempo de espera en una cita es independiente respecto de otras citas, se calcula la probabilidad $P(15 < X < 55) = P(X < 55) - P(X \leq 15) = 0{,}6111 - 0{,}1666 = 0{,}4445$*, que es la probabilidad de que aplace la decisión para la segunda cita y, en la segunda cita, la probabilidad de que lo deje definitivamente es* $P(X > 55) = 0{,}3888$*, luego multiplicando ambas probabilidades se obtiene el valor pedido* $0{,}1728$*.*

2.3. Distribución Exponencial

Ejemplo 4.8

La duración media de un modelo de marcapasos es de 7 años.

a) ¿Cuál es la probabilidad de que dure al menos 5 años? ¿y menos de 3?

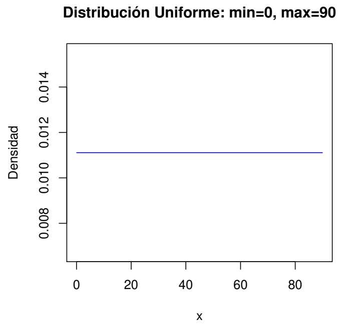


Figura 4.5: Función de densidad

La variable X="tiempo de funcionamiento del marcapasos" sigue una distribución exponencial con parámetro $\lambda = 1/7$. Utilizando la opción `Distribuciones→Distribuciones continuas→ Distribución exponencial→Probabilidades exponenciales...` *se obtiene $P(X \geq 5)$*

```
> pexp(c(5), rate=0.1428, lower.tail=FALSE)
[1] 0.4896815
```

y de igual forma $P(X < 3)$:

```
> pexp(c(3), rate=0.1428, lower.tail=TRUE)
[1] 0.3484493
```

b) Si han transcurrido ya 4 años desde su implantación, ¿cuál es la probabilidad de que dure otros 4?
Teniendo en cuenta que $1-F(x) = e^{-\lambda \cdot x}$, se tiene que $1-F(8) = e^{-8 \cdot \lambda} = (e^{-4 \cdot \lambda})^2 = (1 - F(4))^2$, con lo que $P(X \geq 8/X \geq 4) = (1 - F(8))/(1 - F(4)) = 1 - F(4) = 0{,}5647182$.

c) ¿Cuánto tiempo debería funcionar un marcapasos para estar entre el 10 % de los más duran? Hay que calcular el percentil 90 seleccionando:

`Distribuciones→Distribuciones Continuas→ Distribución exponencial→Cuantiles exponenciales...,` *con las opciones* `Probabilidades= 0.9`*,* `Parámetro de la exponencial= 0.14285` *y* `Cola Izquierda`*, o de forma similar,* `Probabilidades= 0.1`*,* `Parámetro de la exponencial= 0.14285` *y* `Cola Derecha`*,*

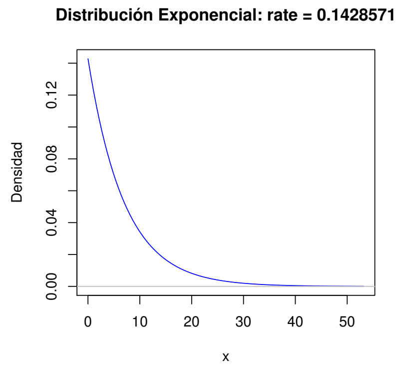


Figura 4.6: Gráfica de la función de densidad de una Exp(0.14285≈ 1/7)

resultando 16,12 *años.*

d) Calcular el valor que deben tener a y b para que $P(X < a) = 0{,}5$ *y* $P(X > b) = 0{,}32$*, De forma análoga al apartado anterior, en el primer caso habría que calcular la mediana,* $a = 4{,}852$*, y en el segundo, el percentil* 68, $b = 7{,}97$. *e) Represente la función de densidad de la variable aleatoria asociada. Figura 4.6.*

2.4. Distribución t-Student

Ejemplo 4.9

Una variable X sigue una distribución t-Student con 16 *grados de libertad.*

a) Calcular la mediana y el percentil 85.

Habría que calcular Me de forma que $P(t_{16} \geq Me) = 0{,}5$*, para ello se selecciona* `Distribuciones→Distribuciones Continuas→ Distribución t→Cuantiles t...`*, con las opciones* `Probabilidades= 0.5`*,* `Grados de libertad= 16` *y* `Cola Izquierda` *o, de forma similar,* `Probabilidades= 0.5`*,* `Grados de libertad= 16` *y* `Cola Derecha`*, resulta que el valor de la mediana es* 0.

```
> qt(c(0.5), df=16, lower.tail=TRUE)
[1] 0
```

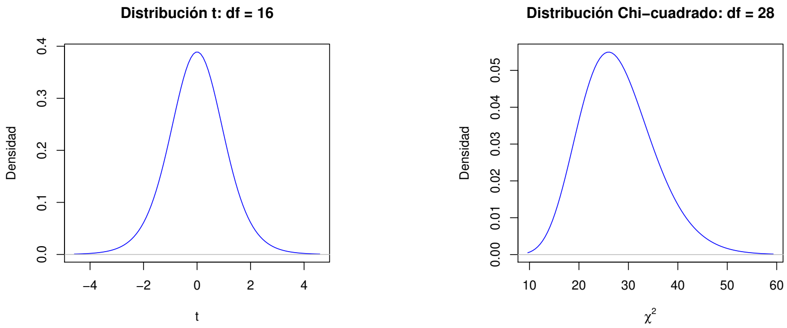


Figura 4.7: Gráfica de la función de densidad t_{16} y χ_{28}

El percentil 85 se calcula de forma parecida:

```
> qt(c(0.85), df=16, lower.tail=TRUE)
[1] 1.071137
```

b) Encontrar el valor de a de forma que $P(-1 < X < a) = 0{,}7$. Para calcular a, se descompone la probabilidad $P(-1 < X < a) = P(X < a) - P(X \leq -1)$, se calcula $P(X \leq -1)$ utilizando la opción `Probabilidades t...`

```
> pt(c(-1), df=16, lower.tail=TRUE)
[1] 0.1660975
```

y, se despeja $P(X < a)$, resultando ser $P(X < a) = 0{,}7 + 0{,}166 = 0{,}866$. Se selecciona ahora la opción `Cuantiles t...`*,*

```
> qt(c(0.866), df=16, lower.tail=TRUE)
[1] 1.147611
```

resultando el valor de `a`$=1{,}147611$.

c) Obtener la gráfica de su función de densidad. ¿Qué similitud tiene con la normal $N(0,1)$?
Como se puede observar en la figura 4.7 su estructura es similar a la $N(0;1)$ con la particularidad de que en la zona central la t_{16} se encuentra por debajo de la normal, consecuencia de tener una varianza mayor.

2.5. Distribución Chi-cuadrado. Distribución F-Snedecor

Ejemplo 4.10

La variable X sigue una distribución Chi-cuadrado con 28 grados de libertad.

a) Calcule la probabilidad de que X sea mayor de 7,5.

La probabilidad pedida $P(\chi_{28} > 7{,}5)$, se obtiene en `Distribuciones→ Distribuciones Continuas→Distribución Chi-cuadrado→ Probabilidades Chi-cuadrado...`, *con las opciones* ***`Valor(es) de la variable= 7.5`***, ***`Grados de libertad= 28`*** *y* ***`Cola derecha`***. *Su valor es 0,9999611.*

```
> pchisq(c(7.5), df=28, lower.tail=FALSE)
[1] 0.9999611
```

b) Obtenga la función de densidad, ¿qué características se observan?. Otra variable Y sigue una distribución F de Snedecor con $n_1 = 8$ y $n_2 = 14$ grados de libertad, si se representa su función de densidad. Como se puede observar en la figura 4.7 sólo toma valores positivos y es asimétrica con forma campaniforme, salvo para $n \leq 2$. c) ¿Qué similitudes hay entre las gráficas?

Como se aprecia en 4.8, en general, sus características son muy similares a la función de densidad de la χ^2.

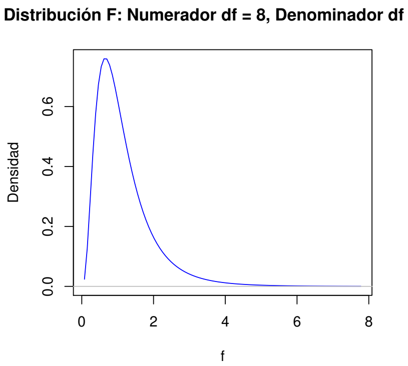


Figura 4.8: Función de densidad $F_{8,14}$

3. Generación de valores aleatorios

Hay situaciones donde es necesario generar valores aleatorios que sigan un determinado patrón y que permitan estudiar el comportamiento de determinados modelos, simular situaciones de laboratorio, generar la distribución de una combinación de variables, comparar valores muestrales con los extraídos de la verdadera población en estudio, ... En **Rcmdr**, para cada una de las distribuciones de probabilidad que tiene implementadas, se puede seleccionar la opción `Muestra de una distribución...`. Así, para generar una muestra de tamaño 15 de una distribución uniforme en el intervalo $[0, 1]$, se selecciona en `Distribuciones`→`Distribuciones continuas`→`Distribución uniforme`→`Muestra de una distribución uniforme...`, y se introducen los parámetros, en este caso, para obtener los datos en formato de columna, `Mínimo= 0`, `Máximo= 1`, `Número de muestras (filas)= 15` y `Número de observaciones (columnas)= 1`.

```
> Muestras_uniformes <- as.data.frame(matrix(runif(15*1,
min=0, max=1), ncol=1))
> rownames(Muestras_uniformes) <- paste(''sample'', 1:15,
sep='''')
> colnames(Muestras_uniformes) <- ''obs''
```

Para mostrarlos en pantalla se escribe en la ventana de instrucciones el nombre que se le haya asignado a la muestra:

```
> Muestras_uniformes
        obs
sample1 0.22597988
sample2 0.65997127
sample3 0.07038248
sample4 0.52902704
sample5 0.04517561
sample6 0.73990437
sample7 0.90452613
sample8 0.60055627
sample9 0.99432508
sample10 0.70652675
sample11 0.97110556
sample12 0.24558711
sample13 0.68375576
sample14 0.95487024
sample15 0.80651304
```

O también se puede pulsar el botón `Visualizar conjunto de datos` en **Rcmdr**. De la misma forma se podrían generar muestras aleatorias para el resto de las distribuciones de probabilidad.

4. Ejercicios

4.1 Se responde al azar un examen tipo test de 10 preguntas donde en cada una de ellas se plantean 4 posibilidades siendo sólo una de ellas cierta. Si se responden todas las preguntas y, las preguntas con respuestas correcta suman un punto mientras que las contestadas incorrectamente restan un cuarto de punto, se pide:

a) La variable aleatoria asociada.

b) Las gráficas de la función de cuantía y distribución y coméntelas.

c) La probabilidad de obtener 3 aciertos.

d) La probabilidad de aprobar.

e) ¿Qué número de aciertos es más probable?

f) ¿Cuántos aciertos debe tener para quedar por encima de la mitad de la clase?

g) ¿Y por encima de un tercio de la clase?

4.2 Dada la distribución $B(10; 0{,}4)$, calcule las siguientes probabilidades:

a) $P(X \leq 8)$

b) $P(2 < X \leq 5)$

c) $P(X \geq 7)$

4.3 Un conocido fumador gorrón ha explotado tanto a sus compañeros que por término medio cada uno de ellos le da un cigarrillo de cada diez veces que éste les pide.

a) ¿Cuál es la probabilidad de que consiga 1 cigarrillo en menos de 5 intentos?

b) Si pretende hacer acopio de cigarrillos para el fin de semana, ¿cuántas veces, en promedio, tendrá que pedir tabaco para conseguir 20 unidades?

4.4 En las oposiciones es frecuente que se realice un sorteo público extrayendo una serie de bolas o papeletas de una urna o bolsa. Imagínese que un opositor se ha preparado 60 temas entre 100, de los que se seleccionan al azar dos temas. Se pide:

a) La variable aleatoria asociada.

b) Las gráficas de la función de cuantía y distribución y coméntelas.

c) La probabilidad de que le salga uno de los temas que lleva preparado.

d) La probabilidad de que le salgan dos de los temas que lleva preparado.

e) ¿Qué ocurre con la probabilidad anterior si aumenta el número de temas preparados a 80?

4.5 A un establecimiento de apuestas deportivas llega 1 cliente cada 3 minutos por término medio.

a) ¿Cuál es la probabilidad de que en un periodo de 5 minutos lleguen más de 5 clientes?

b) ¿Cuál es el número más probable de llegadas en media hora?

4.6 Las compañías aéreas acostumbran a reservar más plazas de las existentes en sus vuelos, dado el porcentaje de anulaciones que se produce. Si el porcentaje medio de anulaciones es del 5 %, ¿cuántas reservas deberá hacer una compañía para un vuelo con 200 plazas, si quiere con una probabilidad del 97 % que todos sus clientes tengan cabida en dicho vuelo?

4.7 El servicio de reclamaciones de una asociación de consumidores recibe por término medio 3 quejas a la hora.

a) Calcule la probabilidad de que en 1 hora no reciba ninguna reclamación.

b) Calcule la probabilidad de que en 2 horas reciba entre 2 y 6 reclamaciones.

4.8 En una pecera hay 10 peces machos y 8 hembras, si se extraen aleatoriamente 5 peces, calcule la probabilidad de que 3 sean machos y 2 hembras.

4.9 Un jugador apuesta 5€ por tirada a un número de los 37 que componen la ruleta, si acierta, gana 180€. Calcule los beneficios esperados al cabo de 100 jugadas.

4.10 El servicio de autobuses entre Cádiz y San Fernando tiene salidas cada media hora entre las 6 am y las 12 pm, una persona que se ha olvidado el reloj en casa llega a la estación de autobuses en Cádiz, se pide:

a) La variable aleatoria adecuada para esta situación.

b) Las gráficas de la función de densidad y distribución y coméntelas.

c) ¿Cuál es su media? ¿y su mediana? ¿y su moda?

d) La probabilidad de que espere menos de 10 minutos.

e) La probabilidad de que espere más de 15 minutos, pero menos de 20.

f) ¿Cuál es la probabilidad de que espere exactamente 11 minutos y medio?

4.11 Se sabe que las bombillas de bajo consumo de 14 w tienen una vida útil media de 10000 horas, mientras que las bombillas clásicas por incandescencia de 60 w tienen una vida útil media de 1000 horas.Si cada día se encienden unas 4 horas, en esta situación

a) Defina la variable aleatoria asociada.

b) Obtenga las gráficas de la función de densidad y distribución y coméntelas.

c) ¿Cuál es su media? ¿y su mediana?

d) ¿Cuál es la probabilidad de que después de un año estén funcionando?

4.12 ¿Cuál es la probabilidad de que de 10 personas elegidas al azar al menos 2 cumplan años en el mes de Enero?

4.13 Durante la Segunda Guerra Mundial los alemanes bombardearon repetidas veces Londres. Los expertos demostraron que se trataba de bombardeos indiscriminados y que caían en cada acción y por término medio 2 bombas por cada cuadrícula de 100 metros de lado. En vista a lo anterior, calcule la probabilidad de que en una cierta cuadrícula de 50 metros de lado no haya caído ninguna bomba durante un bombardeo.

4.14 Dada una distribución normal de media 3 y varianza 9, calcule las siguientes probabilidades:

a) $P(2 \leq X \leq 5)$
b) $P(X \geq 3)$
c) $P(X \leq -2)$

4.15 La centralita de un programa de televisión que premia aquellos concursantes que llaman dando la respuesta correcta de un concurso, atiende 1 de cada 10 llamadas que se realizan.

a) ¿Qué número medio de llamadas se tendrán que realizar para ser atendido?

b) ¿Cuál es la probabilidad de ser atendido a la primera?

4.16 Calcule en los siguientes casos el valor de a, sabiendo que $X \sim N(1, 5)$.

a) $P(0 \leq X \leq a) = 0{,}28$
b) $P(1 - a \leq X < 1 + a) = 0{,}65$

4.17 Se sabe que la alarma de un reloj saltará en cualquier momento entre las siete y las ocho de la mañana. Si el propietario del reloj se despierta al oír dicha alarma y necesita, como mínimo, veinticinco minutos para arreglarse y llegar al trabajo,

a) ¿Cuál es la probabilidad de que llegue antes de las ocho?

b) Si el dueño del reloj sigue programando el reloj de la misma manera durante 10 días, calcule el número más probable de días en que llegará después de las ocho.

4.18 Si se controlan el peso, la edad, la estatura, talla de pantalón, horas de estudio, nota de selectividad, ... de los 350 alumnos que están matriculados en 1º de Empresariales y Económicas en el campus de Cadiz y Jerez. ¿Qué estructura tiene su distribución?

4.19 De una tribu indígena se sabe que los hombres tienen una estatura que se distribuye según una ley normal con media 1,70 y desviación típica σ. Si a través de estudios realizados se conoce que la probabilidad de que su estatura sea mayor a 1,80 es 0,12, calcule la probabilidad de que un individuo elegido al azar mida entre 1,65 y 1,75.

4.20 Calcule la probabilidad de obtener más de 200 seises en 1200 lanzamientos de un dado no trucado.

4.21 Genere muestras de tamaño 10, 100, 500 y 1000 de una población que sigue una distribución normal de media 3,5 y desviación típica 2. Estudie el comportamiento de la media y desviación típica en las cuatro muestras.

4.22 Obtenga una muestra aleatoria de tamaño 50 para una característica poblacional que sigue una distribución binomial de parámetros $n = 12$ y $p = 0{,}7$. Calcule su media y desviación típica comparándolas con los respectivos valores poblacionales. Además, represente los datos mediante un diagrama de barras y compare los resultados con los observados en la gráfica de la función de cuantía de la distribución binomial. ¿Qué ocurre si se aumenta el tamaño de la muestra a 500?

Capítulo 5

Inferencia clásica en poblaciones Normales

1. Conceptos fundamentales

Hasta ahora los objetivos planteados se han limitado a explorar un conjunto de datos describiendo sus características principales o las relaciones entre distintos caracteres. La intención de este capítulo es hacer una primera incursión en lo que se conoce como *análisis inferencial*, en el que a partir del estudio de una muestra pequeña y representativa de miembros de un gran colectivo, se extraen conclusiones que afectan a todos los elementos del mismo. Interesa, por ejemplo, conocer aproximadamente las principales características del colectivo, como pueden ser la media, la desviación típica, su estructura probabilística,...

El enfoque que se le va a dar a este tema se conoce como *clásico*. En él, las características poblacionales a estudiar se consideran parámetros (constantes desconocidas), mientras que los elementos de la muestra se consideran variables aleatorias. La alternativa a este enfoque vendría dada por la *teoría bayesiana*, en el que los parámetros son variables aleatorias, mientras que los datos que se poseen de la población son considerados constantes.

Desde un punto de vista intuitivo, parece razonable que si efectivamente la muestra representa bien al colectivo, los parámetros muestrales sean muy parecidos a los poblacionales y aunque ciertamente este enfoque de *estimación puntual* es básicamente correcto, adolece de ciertas carencias que lo convierten sólo en una parte del proceso inferencial.

Interesa dar una mayor consistencia al análisis inferencial y ello se consigue desde dos puntos de vista, que en muchas ocasiones son complementarios: la construcción de *intervalos de confianza* y la realización de *contrastes de hipótesis*. Tanto uno como otro tienen en cuenta el margen de error derivado de cierta pérdida de información, que se produce al intentar explicar el comportamiento de una población a partir del conocimiento de una parte muy pequeña de sus miembros. Para ilustrar lo dicho se introduce el siguiente ejemplo:

Ejemplo 5.1

Una máquina está preparada para fabricar piezas de 7 cms de longitud. En una inspección se toman 1000 piezas fabricadas por dicha máquina, comprobándose que la media de éstas es de 7,0037 cms. Si se tomaran decisiones sólo a partir de esta estimación puntual habría que concluir que la máquina se ha desajustado y actuar en consecuencia. Pero se está desaprovechando información importante, como si la varianza de los datos es alta o pequeña, o si, como parece, la distribución de las longitudes es normal. La utilización de dicha información va a permitir construir un intervalo de confianza para la media de la población o confirmar directamente si ésta se puede considerar igual a 7 cms. En todo caso se estará asumiendo un margen de error derivado del proceso de extracción aleatorio de la muestra, ya que si se eligieran otras 1000 piezas la media sería distinta a la anterior.

En el caso de los intervalos de confianza, el objetivo es dar una cierta "garantía" de la presencia del parámetro dentro de un intervalo construido a partir de la muestra, mientras que para el caso de los contrastes, la pretensión es dar respuesta a si el valor del parámetro se encuentra, a la luz de la evidencia muestral, dentro de un conjunto de valores especificados en lo que se conoce como *hipótesis nula* (H_0) o, por el contrario, se haya dentro de su alternativo especificado por la *hipótesis alternativa* (H_1).

Se llama *nivel de confianza*, $1-\alpha$, de un intervalo a la probabilidad (a priori) de que el intervalo contenga el valor del parámetro a estimar. La interpretación habitual del nivel de confianza es la probabilidad de que el intervalo de confianza, ya obtenido, contenga el valor del parámetro. Esta interpretación es incorrecta pues una vez obtenido el intervalo el valor del parámetro está o no está y no tiene sentido hablar de la probabilidad de que esto ocurra. $1-\alpha$ debe interpretarse como la proporción teórica de intervalos (ya construidos) que contiene al valor del parámetro.

Para el caso de los contrastes, α es la probabilidad de rechazar la hipótesis nula cuando ésta es cierta y se conoce también como probabilidad de *error de tipo I*, $1-\alpha$ también se llama aquí nivel de confianza. En el caso de los contrastes, existe un error asociado al α que se conoce como β y que indica la probabilidad de no rechazar la hipótesis nula cuando es falsa, conocido también como probabilidad de *error de tipo II*, $1-\beta$ se conoce como *potencia del test.* Ambos errores son contrapuestos y fijado un tamaño muestral cuando uno de los dos crece el otro decrece. El cuadro que sigue recoge las distintas situaciones que pueden darse a la hora de realizar un contraste en término de los errores y aciertos.

		Decisión estadística	
		No rechazar H_0	Rechazar H_0
Estado Real	H_0 cierta	Correcta	Error tipo I
de la cuestión	H_0 falsa	Error tipo II	Correcta

En el peor de los casos, a la hora de realizar un estudio inferencial se cuenta con la información muestral, mientras que en las ocasiones más favorables, se tiene un conocimiento bastante aproximado de la estructura de probabilidad de la población analizada. Cuando se hace uso de la distribución de probabilidad de la población estudiada se dice que la inferencia realizada es *paramétrica*, mientras que si sólo se hace uso de la muestra, la inferencia es *no paramétrica*. El objetivo en los contrastes paramétricos es intentar obtener información sobre los parámetros desconocidos de la distribución de la población bajo estudio. En el caso de los contrastes no paramétricos, su objetivo es intentar determinar alguna característica de la población o de la muestra bajo estudio.

Puesto que los contrastes paramétricos utilizan más información que los no paramétricos, ofrecen mejores resultados. Por ello, siempre que sea posible se debe recurrir a los primeros.

Dependiendo de la estructura de sus hipótesis, se distingue entre los siguientes tipos de contrastes:

1. *Contrastes bilaterales*: en ellos se propone un valor puntual para el parámetro bajo estudio, de forma que se rechazará bien porque la evidencia muestral lleve a decidir que el valor es mayor que el propuesto o bien que es menor. Formalmente:

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

2. *Contrastes unilaterales*: en ellos se propone que el valor del parámetro se encuentre por debajo (o por encima) de un cierto valor. Las dos situaciones se plantearían de la siguiente forma:

$$\begin{aligned} H_0 &: \theta \geq \theta_0 \qquad & H_0 &: \theta \leq \theta_0 \\ H_1 &: \theta < \theta_0 \qquad & H_1 &: \theta > \theta_0 \end{aligned}$$

Se puede observar que en todos los casos el signo igual está incluido en la hipótesis nula, el motivo de ello se encuentra en el procedimiento que se va a utilizar para realizar el contraste.

Las distribuciones asociadas al proceso de muestreo son la `normal` y la `t de student` para el estudio de medias, la `Chi-cuadrado` para la varianza y la `F de Snedecor` para la comparación de varianzas; todas ellas estudiadas en el anterior capítulo. En general, interesa analizar el comportamiento de la media, aunque el mismo va a depender del conocimiento o no que se tenga de su varianza o si, para el caso de dos poblaciones sus varianzas coinciden. No hay que olvidar que la varianza determina la escala de la variable y siempre es más fácil comparar aquellas poblaciones con el mismo factor de escala.

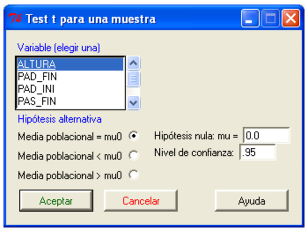


Figura 5.1: Ventana de diálogo para el test t

Es muy importante entender que en el *contraste de hipótesis* los roles que juegan las hipótesis nula y alternativa no son equiparables y mucho menos intercambiables. En todo caso, hay que ver este enfoque como una regla de confirmación sobre una cuestión que el investigador cree razonablemente que es cierta, siendo la función del contraste la de validarla o, por el contrario, si la evidencia muestral en contra es muy fuerte, la de rechazarla.

En este capítulo se estudiarán problemas que involucran a una o dos poblaciones, mientras que en el capítulo 7 se generalizarán los resultados a más de dos poblaciones. Se aceptará, a expensas de poder comprobarlo en el próximo capítulo, que las poblaciones siguen distribuciones normales; caso de que esto no fuera cierto, habría que replantear el análisis desde una perspectiva no paramétrica. Además, se supondrá que las muestras extraídas son aleatorias y que no existen valores anómalos. Igual que para la normalidad, en el próximo capítulo se comprobarán estos supuestos.

2. Inferencias sobre una población

En esta sección se abordará el estudio de la media de una población, de la que se dispone de una muestra aleatoria simple de tamaño n. Aunque en el caso, poco frecuente, de que se conozca la varianza de la población se podría utilizar la distribución Normal, y que cuando el tamaño de la muestra sea grande ($n \geq 50$) la distribución t de student se puede reemplazar por la $N(0,1)$, en general se empleará la propia t de student.

Ejemplo 5.2

Se considera que el fichero de datos `peso_altura.dat` *es una muestra aleatoria simple de la población adulta de un municipio andaluz. Dicha muestra se utilizará para estudiar los valores medios del* `peso` *y la* `altura` *de la población.*

- *Las características muestrales se obtienen como siempre en* `Estadísticos→Resúmenes→Resúmenes numéricos...`*, seleccionando las correspondientes variables e indicando que se haga en función del sexo:*

```
> numSummary(Datos[,c(''ALTURA'', ''PESO'')],
groups=Datos$SEXO, statistics=c(''mean'', ''sd'',
''quantiles''))
 Variable: ALTURA

          mean       sd        0%    25%    50%   75%   100%   n
Mujer    171.0000  5.676462   159   167.00  170.5  175   182   46
Varón    177.1296  6.901043   167   171.25  178.0  182   194   54

Variable: PESO

          mean       sd        0%    25%    50%   75%   100%   n
Mujer    66.95652  4.340796   59    63.00   68.0   70    75    46
Varón    86.24074  10.504150  64    77.25   86.5   93    109   54
```

- **Intervalos de confianza.** *A continuación se obtendrán los intervalos de confianza del 95 % para la altura de los hombres. Para ello se filtra la base de datos por la variable* `sexo`*. A continuación se marca* `Estadísticos→Medias→Test t para una muestra`*, seleccionando en la ventana de diálogo la variable que interesa, en este caso la* `altura`*, y comprobando que el* `nivel de confianza` *está fijado en el* 0,95*(fig 5.1). Las instrucciones que se generan son:*

```
> t.test(Hombres$ALTURA, alternative='two.sided', mu=0.0,
conf.level=.95)
 One Sample t-test
data: Hombres$ALTURA
t = 188.6138, df = 53, p-value <2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
175.2460 179.0133
sample estimates:
mean of x
177.1296
```

De la salida interesa la parte que hace referencia al intervalo de confianza, la media de altura de la población de hombres se encuentra dentro del intervalo $(175{,}24; 179{,}01)$ *con una confianza, que no una probabilidad, del* 95 %.

- **Contraste bilateral.** *Como se puede observar en las instrucciones de* ***R*** *generadas por* ***Rcmdr****, además de la variable y el nivel de confianza, el procedimiento* `t.test` *incluye dos opciones más. La primera de ellas es* `alternative` *y admite tres posibilidades: contraste bilateral* `two.sided`*, contraste unilateral* $H_1 : \mu < \mu_0$ `less` *y contraste unilateral* $H_1 : \mu > \mu_0$ `greater`*. La segunda opción permite fijar un valor para la hipótesis nula* `mu=0.0`*. Para realizar los distintos contrastes se va a retocar la línea de instrucciones. En primer lugar se desea realizar el contraste:*

  $H_0 : \mu = 175$

  $H_1 : \mu \neq 175$

  con un nivel de significación $\alpha = 0{,}01$*. Editando la línea de instrucciones y ejecutando se tiene:*

```
> t.test(Hombres$ALTURA, alternative='two.sided', mu=175.0,
conf.level=.99)
 One Sample t-test
data: Hombres$ALTURA
t = 2.2677, df = 53, p-value = 0.02745
alternative hypothesis: true mean is not equal to 175
99 percent confidence interval:
174.6205 179.6388
sample estimates:
mean of x
177.1296
```

  Se puede observar que, respecto a la salida anterior al aumentar el nivel de confianza ha aumentado la amplitud del intervalo y que el resto es prácticamente igual. Respecto al contraste se concluye que puesto que el `p-value`$= 0{,}027$*, es mayor que el nivel de significación,* $\alpha = 0{,}01$*, no hay evidencias para rechazar la hipótesis nula. Se puede ver que en este caso el valor que* H_0 *propone para la media se encuentra dentro del intervalo de confianza. Esto no ocurría en la salida anterior donde se había fijado el nivel de confianza en* $0,95$*, pues en ese caso* 175 *estaba fuera del intervalo.*

- **Contraste unilateral.** *Se plantea ahora la realización del contraste:*

  $H_0 : \mu \geq 180$

  $H_1 : \mu < 180$

  con un nivel de significación $\alpha = 0{,}1$. *Se edita de nuevo la línea de instrucciones y se ejecuta:*

```
> t.test(Hombres$ALTURA, alternative='less', mu=180.0,
conf.level=.90)
 One Sample t-test
data: Hombres$ALTURA
t = -3.0565, df = 53, p-value = 0.001752
alternative hypothesis: true mean is less than 180
90 percent confidence interval:
-Inf 178.3483
sample estimates:
mean of x
177.1296
```

  En este caso el `p-valor`*=0,0017 es mucho menor que el nivel de significación y por tanto se rechaza la hipótesis nula. Igualmente se puede comprobar que 180 no pertenece al intervalo de confianza.*

3. Inferencias sobre dos poblaciones

Para el caso de comparar las medias de dos poblaciones, además de comprobar las hipótesis sobre normalidad y aleatoriedad, que como ya se ha comentado se verán en el próximo capítulo, se plantean distintas situaciones. En primer lugar habrá que determinar si se tienen muestras independientes o pareadas (relacionadas). La diferencia entre uno y otro caso es que en el segundo, se dan dos mediciones de la misma o similar característica para cada individuo o para dos individuos de idénticas, respecto de los restantes, características relevantes de la muestra.

Si se miden el peso de 50 alevines de truchas antes y después de una cierta dieta alimenticia, ambas observaciones están relacionadas. La aplicación de dos pomadas en diferentes zonas de la piel de un individuo y la observación de ambas respuestas conduce a observaciones pareadas. A veces la dependencia no resulta tan evidente. La longitud de la cola de trabajo de dos impresoras pueden parecer dos observaciones independientes, sin embargo, si ambas impresoras presentan idénticas características tanto en prestaciones como en accesibilidad, la elección del usuario dependerá de las longitudes de las colas existentes, introduciendo dependencia entre ambas longitudes.

Otra cuestión a tener en cuenta, para el caso de muestras independientes, es si las varianzas de las poblaciones se pueden considerar iguales o no.

3.1. Muestras independientes

Ejemplo 5.3

Para el caso de muestras independientes se usará el fichero `parque_eolico.dat`*, que contiene datos de la velocidad del viento, registrados durante 730 horas de forma simultánea, en dos localizaciones alternativas (*`Parque1` *y* `Parque2`*). Se tratará de establecer la localización más aconsejable para la instalación de un parque de producción de energía eólica.*

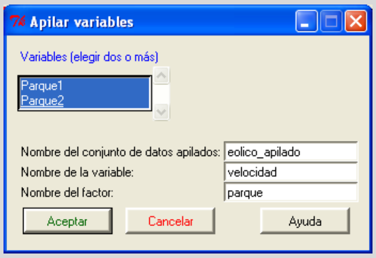


Fig. 5.2: Ventana para apilar `parque_eolico.dat`

Hay que tener en cuenta, al importar este conjunto de datos, que el carácter decimal viene dado en este fichero mediante una coma. Por otra parte, la estructura de la base de datos es de dos columnas, conteniendo cada una de ellas las mediciones en cada localización. Aunque ***R*** *puede trabajar con esta estructura de datos, resulta más manejable para* ***Rcmdr*** *si es transformada en dos variables, una continua que contenga las mediciones de viento y otra factor que indique la localización. Esto se realiza desde el menú* `Datos→Conjunto de datos activo→ Apilar variables del conjunto de datos activo...` *En la ventana de diálogo (fig. 5.2) se pide el nombre de la nueva base de datos que se ha venido a llamar* `eolico_apilado`*, el nombre de la variable apilada,* `velocidad`*, y el nombre de la nueva variable factor,* `parque`*, cuyas clases se han denominado* `Parque1` *y* `Parque2`*.*

Como se ha dicho es conveniente saber si las varianzas se pueden considerar iguales o no a la hora de comparar las dos poblaciones. Una primera idea sobre la igualdad de varianzas es mediante la representación simultánea de los diagramas de caja de las muestras. Desde `Gráficas→Diagrama de caja...`, *se selecciona la variable* `velocidad` *y el grupo* `parque`, *obteniéndose la figura 5.3.*

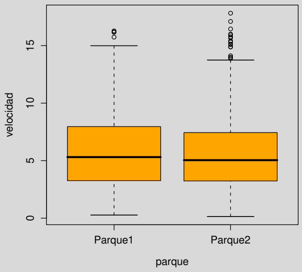


Fig. 5.3: Velocidad según tipo de parque

La comparación de los diagramas sugiere la igualdad de varianzas. El `test F` *permite constrastar dicha hipótesis, desde* `Estadísticos→Varianzas→Test F para dos varianzas...` *seleccionando en este caso como factor la variable* `parque` *y como explicada la variable* `velocidad`.

```
> tapply(eolico_apilado$velocidad, eolico_apilado$parque, var,
na.rm=TRUE)
 Parque1 Parque2
10.50574 10.59477
> var.test(velocidad ~ parque, alternative='two.sided',
conf.level=.95, data=eolico_apilado)
F test to compare two variances
data: velocidad by parque
F = 0.9916, num df = 729, denom df = 729, p-value = 0.9093
alternative hypothesis: true ratio of variances is not equal to
1
95 percent confidence interval:
0.8574994 1.1466647
sample estimates:
ratio of variances
0.9915968
```

Como `p-valor`$= 0,9093 > 0,05$ *no hay motivos para rechazar la igualdad de varianzas. Siendo así, como se supone que los datos están distribuidos normalmente y las varianzas son iguales, los dos parques eólicos serán igualmente productivos cuando la diferencia de sus medias no se separe significativamente de* 0. *Para realizar este contraste se selecciona* `Estadísticos→Medias→Test t para muestras independientes...` *y en la ventana de diálogo emergente se selecciona como grupo la variable* `parque` *y como variable explicada la* `velocidad`, *marcando la opción* `bilateral` *con el* $95\,\%$ *de nivel de confianza y suponiendo las varianzas iguales.*

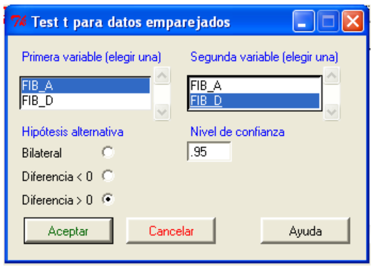


Figura 5.4: Contraste unilateral de fenofibrato

```
> t.test(velocidad~parque, alternative='two.sided',
conf.level=.95, var.equal=TRUE, data=eolico_apilado)
Two Sample t-test
data: velocidad by parque
t = 0.9937, df = 1458, p-value = 0.3205
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
-0.1645533 0.5024437
sample estimates:
mean in group Parque1 mean in group Parque2
5.801795 5.632849
```

Al ser el `p-valor`$= 0,32 > 0,05$ *no se rechaza que la diferencia de las medias sea cercana a cero.*

3.2. Muestras pareadas

Ejemplo 5.4

Para el caso de muestras pareadas se tomará el conjunto de datos `fenofibrato.dat` *en el que se quiere analizar si el tratamiento durante un año con fenofibrato reduce el fibrinógeno, contando para ello con una muestra de 32 individuos. Se efectúa el* `Test t` *en* `Estadísticos`→`Medias`→`Test t para datos relacionados...`*, realizando un contraste unilateral (figura 5.4).*

```
> t.test(Datos$FIB_A, Datos$FIB_D, alternative='greater',
conf.level=.95,paired=TRUE)
Paired t-test
data: Datos$FIB_A and Datos$FIB_D
t = 7.5391, df = 31, p-value = 8.48e-09
alternative hypothesis: true difference in means is greater than
0
95 percent confidence interval:
57.8178 Inf
sample estimates:
mean of the differences
74.59375
```

Al ser el $p - valor < 0,001$ *se rechaza la hipótesis nula, con lo que se acepta que la diferencia, entre los niveles iniciales y finales, es positiva. Con ello se puede deducir que el tratamiento anual con fenofibrato reduce los niveles de fibrinógeno en el organismo y existen así evidencias acerca de su efectividad. Si se deseara confirmar que el tratamiento produce un descenso de más de 50 puntos en el nivel de fenofibrato, se debería tocar ligeramente la instrucción* ***R*** *incluyendo ese dato:*

```
> t.test(Datos$FIB_A, Datos$FIB_D, alternative='greater',
conf.level=.95, paired=TRUE, mu=50)
 Paired t-test
data: Datos$FIB_A and Datos$FIB_D
t = 2.4857, df = 31, p-value = 0.009265
alternative hypothesis: true difference in means is greater than
50
95 percent confidence interval:
57.8178 Inf
sample estimates:
mean of the differences
74.59375
```

De nuevo dado que $p < 0,001$ *se rechaza la hipótesis de que* $\mu_A \leq \mu_D + 50$ *y se concluye que el medicamento produce una disminución de más de 50 puntos en el nivel de fenofibrato.*

4. Ejercicios

5.1 Utilizando el fichero de datos `peso_altura.dat` realice los siguientes ejercicios:

a) Obtenga el intervalo de confianza del 90 % para la altura de las mujeres.

b) Obtenga los intervalos del 95 % para el peso de hombres y mujeres.

c) Para un nivel de confianza del 99 % contraste si la media de la altura de las mujeres es mayor o igual a 173 cms y la de los hombres menor o igual a 175 cms. ¿Puede indicar la razón de este aparente contrasentido?

5.2 Para estudiar la diferencia de estaturas medias, medidas en centímetros, de estudiantes varones en las facultades de ciencias de Cádiz y Málaga, se toma una muestra aleatoria de 15 estudiantes en cada facultad, obteniéndose:

Cádiz	182	170	175	167	171	174	181	169
	174	174	170	176	168	178	180	
Málaga	181	173	177	170	170	175	169	169
	171	173	177	182	179	165	174	

Obtenga el intervalo de confianza al 99 % para la diferencia de estaturas medias entre ambos colectivos de estudiantes. Se supone que las estaturas siguen una distribución normal.

5.3 Se está realizando un estudio sobre la evolución del nivel de colesterol de las personas, para lo cual se seleccionan 10 individuos al azar y se les somete a una nueva dieta alimenticia durante seis meses, tras la cual se les volvió a medir el nivel de colesterol en mg/dl. Suponiendo normalidad, obtenga un intervalo de confianza al 90 % para la diferencia de medias.

Antes	200	156	178	241	240	256	245	220	235	200
Después	190	145	160	240	240	255	230	200	210	195

5.4 Una fábrica produce barras de hierro cuya longitud sigue una distribución Normal. A partir de la muestra:

$$\begin{array}{ccccc} 100,9 & 101,2 & 100,2 & 100,4 & 99,8 \\ 100,1 & 101,5 & 100,4 & 101,7 & 99,5. \end{array}$$

a) Encuentre un intervalo de confianza para la longitud media.

b) Tras revisar la maquinaria, se obtuvo una nueva muestra:

$$\begin{array}{ccccc} 99,7 & 100,7 & 97,8 & 98,8 & 101,4 \\ 100,3 & 98,7 & 101,1 & 99,4 & 99,5. \end{array}$$

Estudie si se produjo algún cambio en la longitud media de la barras.

5.5 Una empresa de transporte de mercancías tiene dos oficinas en una determinada ciudad. Al objeto de asignar un nuevo trabajador a una de las dos oficinas, la dirección de la empresa decide analizar la productividad de cada una de ellas, contabilizándose las facturaciones en los últimos doce meses (miles de euros).

Ofic. 1	13,7	12,1	12,3	8,9	9,7	10,1	12,7	11,0	13,2	9,7	10,1	9,9
Ofic. 2	9,8	9,9	10,0	10,3	9,5	9,3	11,1	13,9	9,8	9,5	7,3	7,9

Suponiendo la normalidad de ambas poblaciones, ¿existen diferencias de facturación entre las dos oficinas?

5.6 Una empresa le propone al director de una fábrica un nuevo método que, supuestamente, reduce el tiempo empleado en el montaje de uno de sus productos. Con el propósito de comparar tal método con el empleado habitualmente, seleccionó aleatoriamente a siete de sus empleados para que llevasen a cabo el montaje con los dos sistemas y anotó los tiempos empleados en el montaje, obteniendo los siguientes resultados:

Trabajador	1	2	3	4	5	6	7
Método habitual	38	32	41	35	42	32	45
Método nuevo	30	32	34	37	35	26	38

Supuesto que el tiempo de montaje sigue una distribución Normal, ¿se puede afirmar que efectivamente el nuevo método reduce el tiempo en más de dos minutos?

Capítulo 6

Inferencia no paramétrica. Diagnosis del modelo

En este capítulo se aborda en primer lugar la realización de contrastes sobre la calidad de la muestra, a continuación se estudian test de bondad de ajuste, haciendo especial énfasis en los de normalidad y, por último, se dan alternativas no paramétricas para el caso de que las poblaciones no sean normales.

1. Pruebas de aleatoriedad

En esta sección se abordará el estudio de la calidad de la muestra extraída de la población, y aunque el procedimiento de obtención debería garantizar unos niveles mínimos de calidad, lo cierto es que en ocasiones los datos vienen impuestos sin que el investigador haya podido supervisar el procedimiento de extracción. No obstante y como en todo contraste, debe tenerse en cuenta que el test sólo desestimará la hipótesis si la evidencia muestral en su contra es muy fuerte.

En ocasiones, los elementos de la muestra se han obtenido en un marco territorial o temporal. Imagine por ejemplo mediciones de una cierta magnitud económica a lo largo de un periodo de tiempo o niveles de un determinado elemento químico en estudios de contaminación, bien en aire, agua o tierra. En estas situaciones es de esperar que las mediciones tomadas en un cierto entorno tengan ciertas analogías o presenten tendencias. Para estudiar este tipo de situaciones se debe acudir a modelos específicos, como son las series temporales o los modelos geoespaciales, en ambos casos existe un elemento que sirve de variable de referencia o longitudinal: la fecha o el posicionamiento gps. Sin embargo, en otras situaciones donde no se contempla esa variable de referencia, las personas encargadas de realizar el muestreo, por comodidad o descuido, no adoptan las medidas para garantizar la independencia de las mediciones.

Ejemplo 6.1

Para analizar si existe autocorrelación entre los elementos de una muestra, se consideran los datos del PIB en billones de euros durante los últimos diez años: 13, 14, 18, 21, 22, 19, 20, 23, 27 y 30. Parece que debería existir influencia del PIB de años precedentes sobre los posteriores. Para comprobarlo se aplicará el test de autocorrelación de Ljung-Box, *contemplando autocorrelaciones de primer y segundo orden. Para la de primer orden, se fija la opción* `lag=1`.

```
> x<- c(13, 14, 18, 21, 22, 19, 20, 23, 27, 30)
> Box.test(x, lag = 1, type = c(''Ljung-Box''))
Box-Ljung test
data: x
X-squared = 4.2281, df = 1, p-value = 0.03976
```

Lo que indica, dado que $p = 0,03976$*, que para un* $\alpha = 0,05$ *se rechazaría la hipótesis de indepedencia lineal de primer orden, por lo que el valor del PIB del año* T *influye sobre la del año* $T+1$*. Si se analiza la correlación de segundo orden,* `lag=2`*, se tiene:*

```
> Box.test(x, lag = 2, type = c(''Ljung-Box''))
Box-Ljung test
data: x
X-squared = 4.4046, df = 2, p-value = 0.1105
```

En esta ocasión y puesto que $p > 0,05$ *no se rechaza la hipótesis de independencia y se descarta la autocorrelación de segundo orden.*

Otra perspectiva desde la que analizar la aleatoriedad de la muestra, si ésta viene dada en forma de variable binaria, es comprobar si existen muy pocas o muchas rachas, entendiendo por racha al grupo de valores consecutivos iguales interrumpido por uno de signo distinto. Si la variable no es de tipo binario, se la puede transformar para que lo sea asignando las clases de la dicotomía en función de que el elemento muestral esté por encima o por debajo de un determinado valor, típicamente la mediana.

Ejemplo 6.2

*Para analizar la independencia de los mismos datos del PIB del ejemplo anterior se aplicará ahora el test de rachas. Previamente habrá que cargar el paquete `tseries` de series temporales, bien desde el menú o con la instrucción `library(‘‘tseries’’)`. En este caso se realizará un contraste bilateral, rechazándose la hipótesis nula tanto si existen muchas rachas como si hay muy pocas, aunque las opciones de la función de **R** admitirían que se especificaran contrastes de carácter unilateral.*

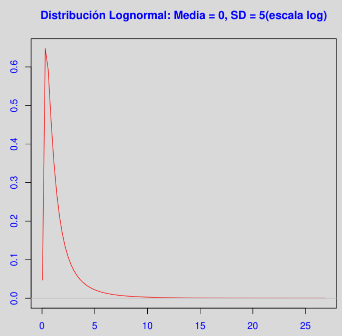
```
> runs.test(as.factor(x>median(x)))
Runs Test
data: as.factor(x > median(x))
Standard Normal = -1.3416, p-value = 0.1797
alternative hypothesis: two.sided
```

Con la orden `as.factor(x>median(x))` se convierte a la variable `x` en dicotómica, dando códigos distintos en función de que el valor esté por debajo o por encima de la mediana (20,5). La salida del procedimiento indica, puesto que $p > 0,05$, que no hay evidencias para considerar los datos no aleatorios.

2. Pruebas de bondad de ajuste

En este epígrafe se contrastará si la estructura de la población analizada se ajusta a una determinada distribución. En principio el procedimiento de obtención de la información deberá ofrecer pautas para decidir si la población tiene una u otra estructura probabilística. Así, en el caso que más nos interesa, si la variable se genera a partir de la medición objetiva de alguna característica, ésta será en general normal; la excepción se dará cuando se haya considerado un conjunto de individuos no homogéneos, mezclando grupos de edad, sexos, . . . Si realmente se han mezclado grupos de individuos, un análisis exploratorio arrojará una estructura probabilística multimodal, mientras que si, por el contrario, la población física es homogénea, la distribución presentará, si acaso, problemas de simetría; en algunas ocasiones estos problemas se pueden solucionar mediante transformaciones de los datos. También puede darse la circunstancia de que distribuciones que converjan a la normal en situaciones ideales y para muestras grandes, como es el caso de la binomial o la Poisson, necesiten alguna transformación para mejorar la simetría. Se analizará esta cuestión en el capítulo de Análisis de la Varianza. Por último, hay que indicar que en muchas ocasiones hay que realizar una operación de truncamiento para adaptar la distribución teórica al rango de valores de los datos en estudio.

Ejemplo 6.3



En problemas ecológicos es muy habitual que la abundancia de una especie tenga una distribución de tipo lognormal respecto a los parámetros ambientales, por tanto una transformación logarítmica convertiría a la abundancia en una variable normal. Como se puede ver, no se trata de una medición de una característica de los individuos, sino de una medida de su abundancia respecto a una variable ambiental.

A continuación se presentará un contraste específico de normalidad, como es el test de Shapiro-Wilk, y un par de test genéricos para evaluar la bondad del ajuste, uno para cuando los datos son continuos, el de Kolmogorov-Smirnov, y otro para variables categóricas, el test de la χ^2. En el caso de contrastes de normalidad, se recomienda el uso del test de Shapiro-Wilk para muestras pequeñas $n \leq 50$, mientras que si las muestras son grandes es preferible utilizar el test de Kolmogorov-Smirnov, salvo que los datos vengan dados en una distribución de frecuencias por intervalos donde se empleará la χ^2.

Ejemplo 6.4

El archivo de datos que se utilizará en este ejemplo es el `caracoles.dat` *que incluye las mediciones de dos variables, diámetro de las conchas (mm) y separación entre las espirales (μm), para un conjunto de 20 individuos adultos de una especie de caracoles. Dado el tamaño de la muestra, se contrastará la hipótesis de normalidad mediante el test de Shapiro-Wilk. Utilizando en este caso* **Rcmdr** *y marcando las opciones* `Estadísticos`→`Resúmenes`→ `Test de normalidad de Shapiro-Wilk...` *se obtiene el cuadro de diálogo, donde se selecciona la variable diámetro (*`Diam`*).*

En la ventana de resultados de **Rcmdr** *se tiene tanto la instrucción de* **R** *como la salida del procedimiento. En este caso el* `p-valor`$= 0,6869$ *viene a indicar que los datos se pueden considerar normales.*


```
>shapiro.test(Datos$Diam)
 Shapiro-Wilk normality test
data: Datos$Diam
W = 0.9668, p-value = 0.6869
```

Ejemplo 6.5

Se estudiará la normalidad de la variable peso del fichero `peso_altura.dat`*. Dado que el número de individuos es grande, $n = 100$, se utilizará el test de Kolmogorov-Smirnov. En primer lugar, con* **Rcmdr** *se calcula la media y la desviación típica del conjunto de datos, resultando $\bar{x} = 73,37$ y $\sigma = 12,69$. A continuación se computarán las diferencias entre la función de distribución empírica muestral y la distribución teórica $N(73,37; 12,69)$. Para ello se empleará el procedimiento* `ks.test`*.*

```
> ks.test(Datos$PESO,pnorm,73.37,12.69)
 One-sample Kolmogorov-Smirnov test
data: Datos$PESO
D = 0.136, p-value = 0.04939
alternative hypothesis: two-sided
```

En este caso y para un $\alpha = 0,05$ se rechaza la hipótesis de que los pesos sigan una distribución normal.

El test de Kolmogorov-Smirnov también se puede utilizar para comparar las distribuciones empíricas de dos conjuntos de datos, para ello en la instrucción se sustituiría la distribución a ajustar por la segunda variable.

Ejemplo 6.6

*Se generan mediante instrucciones de **R** dos muestras aleatorias de 100 y 150 elementos procedentes de distribuciones exponenciales de parámetros 1 y 1,5, respectivamente, mediante las instrucciones:*

```
x<-rexp(100,1); y<-rexp(150,1.5)
```

Aplicando de nuevo el test de Kolmogorov-Smirnov para comparar las funciones de distribución empírica de ambas muestras, se tendría:

```
>ks.test(x,y)
 Two-sample Kolmogorov-Smirnov test
data: x and y
D = 0.2833, p-value = 0.0001310
alternative hypothesis: two-sided
```

Se puede comprobar que el test rechaza la hipótesis de igualdad de funciones de distribución empíricas con un `p-valor`$= 0,00013$.

El análisis de la bondad de ajuste de una serie de datos a una distribución de probabilidad se estudia mediante el test de la chi-cuadrado de Pearson. Básicamente, el estadístico χ^2 evalúa las diferencias entre los valores observados y los valores ajustados por la ley de probabilidad. Se verán a continuación distintas situaciones y cómo se resuelven con **R**.

Ejemplo 6.7

Para contrastar si un dado no está trucado se lanza 60 veces, obteniéndose los siguientes resultados:

x_i	1	2	3	4	5	6
n_i	7	12	10	11	8	12

La hipótesis a contrastar es que $p_i = 1/6,\ \forall i$, *con lo que se tiene que* $E_i = 60(1/6) = 10,\ \forall i$.

Para resolver el contraste con ***R*** *basta introducir el vector de frecuencias,* $n = (7, 12, 10, 11, 8, 12),$ *y escribir las instrucciones de* ***R***.

```
> n< -c(7,12,10,11,8,12)
>chisq.test(n)
 Chi-squared test for given probabilities
data: n
X-squared = 2.2, df = 5, p-value = 0.8208
```

A la vista del p-valor no se rechaza que el dado no está trucado.

El test Chi-cuadrado permite contrastar la hipótesis de independencia entre dos atributos organizados en tabla de contingencia.

Ejemplo 6.8

Se desea analizar la relación entre el nivel de estudios del padre y la orientación del alumno hacia las ciencias en un determinado instituto de bachillerato. Se cuenta para ello con la información obtenida en el centro.

	Estudios padre			
Orientación	*Ninguno*	*Básico*	*Medio*	*Superior*
Orientado	*23*	*12*	*34*	*32*
No orientado	*18*	*42*	*16*	*27*

Para contrastar esta relación se introduce la matriz de datos en ***Rcmdr*** *como se describe en el ejemplo 3.1, obteniéndose los siguientes resultados:*

```
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
Pearson's Chi-squared test
data: .Table
X-squared = 24.1629, df= 3, p-value = 2.31e-05
```

Lo que indica que se rechaza la hipótesis de independencia y existe una relación entre los estudios de los padres y la orientación hacia las ciencias de sus hijos.

Para el caso de tablas 2×2 se aplica el *test exacto de Fisher*, aunque existe la alternativa de aplicar el test Chi-cuadrado con la corrección de Yates. Para aplicar esta corrección bastaría especificar,`correct=TRUE`, en la instrucción de dicho test.

Ejemplo 6.9

En el conservatorio de música de una ciudad se pretende estudiar la relación existente entre el sexo del alumnado y su afición por los instrumentos de viento. Para ello, observados los 482 estudiantes se tiene:

	Hombre	*Mujer*
Aficionado	*150*	*97*
No aficionado	*123*	*112*

Se introduce la matriz de datos de la misma forma que en el ejemplo 3.1 seleccionando la opción de `Prueba exacta de Fisher`

```
>fisher.test(.Table)
Fisher's Exact Test for Count Data
data: .Table
p-value = 0.06655
alternative hypothesis: true odds ratio is not equal to 1
```

Por lo que para un nivel de significación $\alpha = 0,05$ no se rechaza, aunque con poca evidencia, la hipótesis de independencia entre el sexo y la afición a los instrumentos de viento.

Se analizará ahora la bondad de ajuste de unos datos a una distribución teórica no uniforme.

Ejemplo 6.10

Durante la Segunda Guerra Mundial los alemanes bombardearon en diversas ocasiones Londres. Al objeto de analizar si los bombardeos eran indiscriminados o se hacían con intención, se procedió a dividir la ciudad en cuadrículas y a contar el número de impactos en cada una de ellas. Los resultados se recogen en la siguiente tabla

Impactos	0	1	2	3	4	5
Número cuadrículas	229	211	93	35	7	1

Las hipótesis podrían ser expresadas, en términos probabilísticos, de la siguiente manera

$$H_0: \quad X \sim P(\lambda)$$
$$H_1: \quad X \nsim P(\lambda)$$

puesto que si las bombas caen indiscriminadamente, lo hacen de forma independiente en un soporte continuo. Lo que, de ser cierto, indicaría que la variable que mide el número de impactos por cuadrículas debe ser Poisson.

En primer lugar, se estimará el parámetro de la Poisson a partir de la media muestral, resultando que $\hat{\lambda} = 0,929$. A continuación se calcularán las probabilidades $P(X = i)$, con $i = 0, 1, 2, 3, 4$ y $P(X \geq 5)$ mediante ***Rcmdr***.

Las probabilidades discretas se obtienen en: `Distribuciones`→`Distribuciones discretas`→`Distribución de Poisson`→`Probabilidades de Poisson...` *tomando* `media`= 0,929.

```
>.Table
    Pr
0 0.3949
1 0.3669
2 0.1704
3 0.0528
4 0.0123
5 0.0023
6 0.0004
7 0.0000
```

La probabilidad $P(X \geq 5)$ se obtiene desde: `Distribuciones`→ `Distribuciones discretas`→`Distribución de Poisson`→

`Probabilidades de Poisson acumuladas...`, *tomando* `valor(es) de la variable`= 4 *ya que* ***Rcmdr*** *realiza* $P(X > 4)$=$P(X \geq 5)$, *para la cola de la derecha y* `media`= 0,929, *resulta:*

```
> ppois(c(4), lambda=0.929, lower.tail=FALSE)

 [1] 0.002682857
```

Con objeto de comprobar si se verifica la restricción de que todos los valores esperados deben ser mayores a tres, se calcula $n \cdot P[X \geq 5] = 576 \cdot 0,0027 = 1{,}5552 < 3$, *por lo que debe procederse a una agrupación de clases y considerar ahora* $P(X \geq 4)$. *Se obtiene que* $n \cdot P[X \geq 4] = 576 \cdot 0{,}015 = 8{,}64 > 3$.

Se almacenan ahora estas probabilidades en un vector `p`, *las frecuencias de los valores que toma la variable en otro vector* `x` *y se aplica el test chi-cuadrado resultando:*

```
>p< -c(0.3949,0.3669,0.1704,0.0528,0.0150)
>x< -c(229,211,93,35,8)
>chisq.test(x,p=p,rescale.p=TRUE)
 Chi-squared test for given probabilities
data: x
X-squared = 1.0205, df = 4, p-value = 0.9067
```

Por lo que se puede afirmar de forma contundente, dado el valor de p, que los bombardeos alemanes fueron indiscriminados.

3. Contrastes de localización y escala

Si se desestima la hipótesis de normalidad de los datos, no son aplicables los test vistos en el capítulo anterior basados en dicha distribución, siendo necesario utilizar contrastes no paramétricos. Este tipo de test se basan en el análisis de la situación de los elementos de la muestra respecto a determinadas medidas de posición, muy en especial respecto a la mediana. De esta forma, se estudia si los datos muestrales están por encima o por debajo de la mediana, es decir, se analiza el signo de su diferencia con la mediana; o bien, se estudia la distancia ordenada a la que se encuentra de la mediana, es decir, se considera el rango o la posición que ocupa dicho elemento en la secuencia ordenada de las diferencias.

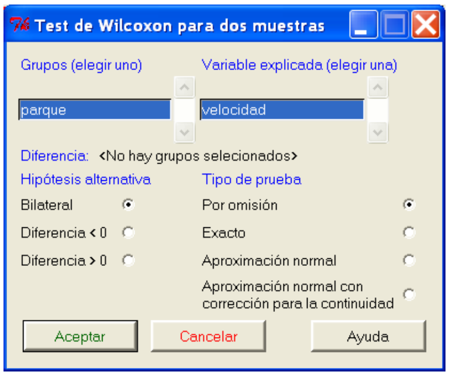


Figura 6.1: Test de Wilcoxon

En todo caso, las situaciones a analizar son las mismas del capítulo anterior: una muestra, dos muestras independientes y dos muestras apareadas, a las que se intentará dar respuesta con los ejemplos que siguen.

3.1. Dos muestras independientes

Ejemplo 6.11

Se estudiará mediante el test de Wilcoxon para muestras independientes si las dos ubicaciones del parque eólico, cuya información se encuentra en el archivo `eolico_apilado.dat`*, tienen la misma potencialidad eólica. Para ello, en el menú de* ***Rcmdr*** *se seleccionan las opciones de menú,* `Estadísticos`→`Test no paramétricos`→ `Test de Wilcoxon para dos muestras...`*, con lo que abre la ventana de diálogo 6.1.*

Seleccionados los únicos elementos de la base de datos, variable y factor, los resultados del análisis son:

```
> wilcox.test(velocidad~parque, alternative="two.sided",
data=Datos)
 Wilcoxon rank sum test with continuity correction
data: velocidad by parque
W = 276269.5, p-value = 0.2228
alternative hypothesis: true location shift is not equal to 0
```

Lo que implica el no rechazo de la hipótesis nula de igualdad de *medianas, siendo indistinta, desde esta óptica, la ubicación del parque eólico.*

3.2. Una muestra

Ejemplo 6.12

Se desea contrastar la hipótesis nula, con $\alpha = 0,05$, de que la separación mediana entre las espirales (variable `Separ`*) de los caracoles del fichero* `caracoles.dat` *es menor o igual a 110 μm. Se supondrá que los datos son aleatorios pero no normales y se utilizará por tanto el test de Wilcoxon para una muestra. Trabajando directamente con **R** se tiene:*

```
> wilcox.test(Datos$Separ,alternative=c("greater"),mu=110)
 Wilcoxon signed rank test with continuity correction
data: Datos$Separ
V = 157, p-value = 0.006617
alternative hypothesis: true location is greater than 110
```

Por lo que se rechaza la hipótesis nula y se concluye que la separación mediana es superior a 110 μm.

3.3. Dos muestras pareadas

Ejemplo 6.13

Para documentar el caso de muestras pareadas se considera el mismo ejemplo que se usó en el capítulo anterior, la eficacia del tratamiento con fenofibrato, suponiendo ahora que la distribución de la diferencia de medias no es normal. En este caso se quiere probar la afirmación del fabricante de que el tratamiento durante un año con fenofibrato reduce el fibrinógeno en al menos 50 puntos. Se aplicará pues el `test de Wilcoxon para muestras pareadas`*. Para acceder al test, se ejecuta la secuencia de **Rcmdr**:*

`Estadísticos→Test no paramétricos→Test de Wilcoxon para muestras pareadas...`

Aunque las opciones de la ventana no admiten que se especifiquen diferencias, bastará con retocar mínimamente la instrucción añadiendo al final de la línea la opción `mu=50`.

```
> wilcox.test(Datos$FIB_A, Datos$FIB_D, alternative='greater',
paired=TRUE, mu=50)
 Wilcoxon signed rank test with continuity correction
data: Datos$FIB_A and Datos$FIB_D
V = 354, p-value = 0.01934
alternative hypothesis: true location shift is greater than 50
```

Así para $\alpha = 0,05$ *se rechaza la hipótesis de que* $\text{med}_A - \text{med}_D \leq 50$ *y se concluye que el medicamento produce una disminución de más de 50 puntos en el nivel de fenofibrato.*

4. Ejercicios

6.1 Contraste la normalidad de la variable separación entre las espirales (`Separ`) del fichero `caracoles.dat`.

6.2 Mediante el test de Kolmogorov-Smirnov, compruebe la hipótesis de igualdad de las funciones de distribución empírica de dos muestras de tamaño 200, procedentes de poblaciones N(0;1) y N(0;1,3) previamente generadas.

6.3 Compruebe la hipótesis de normalidad de la velocidad para cada una de las ubicaciones en el fichero `parque_eolico.dat`.

6.4 Contraste la hipótesis de que los datos siguientes, generados aleatoriamente mediante ordenador, procedan de una distribución Uniforme en el intervalo $[0, 1]$ con un nivel de significación $\alpha = 0{,}05$.

$$\begin{array}{cccccc} 0,582 & 0,501 & 0,497 & 0,026 & 0,132 & 0,561 \\ 0,642 & 0,994 & 0,948 & 0,081 & 0,179 & 0,619 \end{array}$$

6.5 En un grupo de 100 personas se estudian los atributos color del cabello (moreno, rubio y castaño) y color de los ojos (negro, marrón, azul y verde), obteniéndose la siguiente tabla de contingencia:

	Cabello		
Ojos	Moreno	Rubio	Castaño
Negros	20	8	4
Marrones	16	2	11
Azules	5	8	8
Verdes	10	5	3

¿Están relacionados dichos atributos?

6.6 Contraste si los datos de la siguiente muestra organizada como distribución de frecuencias proceden de una Normal.

$(L_{i-1}, L_i]$	n_i
$(0, 1]$	1
$(1, 2]$	3
$(2, 3]$	7
$(3, 4]$	12
$(4, 5]$	6
$(5, 6]$	2
$(6, 7]$	1

6.7 Estudie, utilizando el contraste χ^2 de bondad de ajuste, si la siguiente muestra de tamaño 30 procede de una Normal.

107	96	91	80	103	88	101	106	112	106
93	88	101	109	102	99	93	86	100	99
104	116	87	93	106	102	89	96	104	90

6.8 Con el fin de estudiar el tiempo de vida, en horas, de las baterías de 7 voltios, se extrae aleatoriamente un muestra de 10 de ellas, obteniéndose los siguientes resultados:

28.9	15.2	28.7	72.5	48.6
52.4	37.6	49.5	62.1	54.5

Proponga un modelo de distribución de probabilidad y estudie su ajuste.

6.9 Para medir la introversión se aplica a 12 individuos un test de personalidad en sus dos variantes, 1 y 2, que se supone la miden por igual. A partir de los datos de la siguiente tabla, compruebe mediante el test de rangos de Wilcoxon, con un nivel de significación del 5 %, si es cierto que las formas 1 y 2 miden por igual la introversión.

Individuo	1	2	3	4	5	6	7	8	9	10	11	12
Forma 1	12	18	21	10	15	27	31	6	15	13	8	10
Forma 2	10	17	20	5	21	24	29	7	9	13	8	11

6.10 Para estudiar cuál de los dos tratamientos contra la artrosis es más eficaz se eligen aleatoriamente dos muestras de 10 y 22 pacientes a los cuales se les somete a los tratamientos 1 y 2, respectivamente. Pasados tres meses se valoran ambos tratamientos de manera que el que tenga mayor puntuación será más eficaz. La tabla siguiente refleja los resultados obtenidos.

Tratamiento 1	12	15	21	17	38	42	10	23	35	28	
Tratamiento 2	21	18	42	25	14	52	65	40	43	35	18
	56	29	32	44	15	68	41	37	43	58	42

Utilice el test de Wilcoxon para evaluar si existen diferencias entre los dos tratamientos.

Capítulo 7

Introducción al Análisis de la Varianza

1. Conceptos básicos

Aunque en origen el *Análisis de la Varianza* (ANOVA) fue introducido por Fisher para evaluar los efectos de los distintos niveles de un factor sobre una variable respuesta continua, desde un punto de vista puramente abstracto el ANOVA va a permitir generalizar el contraste de igualdad de medias de dos a k poblaciones. Y esa es la perspectiva en la que se va a centrar este último capítulo. No se propondrá pues ningún modelo teórico, sino que el objetivo se limitará a usar la técnica para contrastar la hipótesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$. Eso sí, al igual que se ha hecho para una y dos poblaciones, se evaluarán las hipótesis previas relativas a la calidad de la muestra, a la estructura de probabilidad, normal o no, de la población y a si las distintas poblaciones tienen varianzas iguales o distintas, propiedad esta última conocida como *homocedasticidad*.

El ANOVA en su versión paramétrica del *test de la F*, como todos los procedimientos estadísticos, tiene un cierto grado de *robustez* frente a un relativo incumplimiento de alguna(s) de sus hipótesis. En concreto, el test de la F soporta mejor las deficiencias respecto a la normalidad que las relacionadas con la homocedasticidad. En todo caso, los test son menos sensibles a las desviaciones de las hipótesis exigidas cuando el número de observaciones de las muestras es aproximadamente el mismo.

Como libro de ruta se propone que, cuando se verifiquen todas las hipótesis exigidas la alternativa preferida sea el test de la F. Cuando se dé la normalidad pero no la homocedasticidad, se recomienda el uso del *test de Welch* o el *test de Kruskal Wallis*. Si falla, aunque no de forma drástica la normalidad, con valores de p entre $0,01$ y $0,05$, la robustez del test de la F le hace seguir siendo una buena opción. Por último, si fallara fuertemente la normalidad, se recomienda el uso del test de Kruskal Wallis.

Si la conclusión del test aplicado fuera el rechazo de la hipótesis nula, no ocurriría como en el caso de dos poblaciones en el que claramente una de ellas tendría media superior a la otra, sino que habría que evaluar las relaciones entre las k poblaciones, bien dos a dos o a través de combinaciones entre ellas, mediante los denominados *test de comparaciones múltiples*. El resultado final de estas comparaciones desembocará en un mapa de relaciones que, debido a la naturaleza intrínseca de los test, no verificará en general el principio de transitividad.

Existe una gran cantidad de test que realizan las comparaciones múltiples, tratando cada uno de ellos de adaptarse mejor a determinadas circunstancias. Cabe destacar, por ser de uso más extendido, los contrastes de Duncan, Newman-Keuls, Bonferroni, Scheffé y HSD de Tukey. Dependiendo de que las comparaciones sean entre parejas de medias o más generales, combinaciones de las mismas, será más aconsejable el test de Tukey o el de Scheffé. En el caso de comparaciones de parejas de medias, puesto que el de Tukey proporciona intervalos de confianza de menor longitud, se preferirá al de Scheffé.

2. Diagnosis del modelo

Como se ha puesto de manifiesto, los primeros pasos a dar son los de comprobar si las muestras son aleatorias y las poblaciones normales a través de los test descritos en el capítulo anterior. A continuación, si la muestra no está contaminada y no hay desviaciones importantes de normalidad, se comprobará la hipótesis de homocedasticidad y a la vista de ambas pruebas se elegirá el contraste adecuado. Puesto que ya se han visto los test de aleatoriedad y de normalidad, se dedicará este epígrafe a validar la hipótesis de homocedasticidad. Para ello, se empleará el *test de homogeneidad de varianzas de Barlett*.

Ejemplo 7.1

El archivo `cebada.dat` *contiene información sobre la producción de cuatro variedades de cebada. Utilizando el test de Barlett se estudiará la homocedasticidad de los datos. En* ***Rcmdr****, una vez cargados los datos, se selecciona:* `Estadísticos`→`Varianzas`→`Test de Barlett`*, tomando en la ventana de diálogo, en* `Grupos`*, el factor tipo de cebada,* `tipo`*, y en la* `variable explicada` *la producción de la misma,* `prod`*.*

```
> bartlett.test(prod~tipo, data=Datos)
Bartlett test of homogeneity of variances
data: prod by tipo
Bartlett's K-squared = 5.9371, df = 3, p-value = 0.1147
```

Dado que `p-valor`$= 0,1147$ *no se rechaza la hipótesis de igualdad de varianzas para los cuatro tipos del factor.*

En muchas ocasiones las muestras que se emplean son de tamaño muy pequeño, menores de 10 elementos, y dado que los test son en general muy conservativos, van a tender a no rechazar la hipótesis nula debido a la escasez de información. Por ello, en este tipo de situaciones, además de la aplicación del contraste para validar la hipótesis, es bueno analizar la naturaleza de los datos. En particular, cuando se trata de validar la normalidad de los datos, si éstos no se han obtenido por un procedimiento de medición sino por observación o conteo, los datos no van a ser intrínsecamente normales aunque pasen el test de normalidad. Para mitigar el problema se recomienda realizar una transformación de los datos. Entre las transformaciones más importantes destacan la raíz cuadrada y la arco seno. La transformación raíz cuadrada se emplea cuando los datos se obtienen a partir de un conteo de elementos, pues en ese caso la distribución de los mismos suele ser de tipo Poisson. Por otra parte, cuando se tienen los datos en forma de tanto por uno, p, es decir que proceden de una binomial, se aconseja la transformación $\mathrm{arcsen}\sqrt{p}$.

3. Test de la F

En este epígrafe se estudiará el contraste de igualdad de medias suponiendo que los datos son normales y homocedásticos. El test que se utilizará será el de la F, que no es sino la generalización del test de la t de student a k poblaciones.

Ejemplo 7.2

Para evaluar el índice de alfabetización de cuatro municipios de una determinada comarca, se ha pasado un test a varios habitantes de cada una de ellas con los siguientes resultados.

Pueblo 1	*Pueblo 2*	*Pueblo 3*	*Pueblo 4*
78	52	82	57
85	48	91	61
90	60	85	45
77	35	74	46
69	51	70	
	47		

Los datos se han recogido en el fichero `alfabeto.dat`*. Suponiendo que los datos son normales y que las varianzas son iguales se aplicará el test de la F. En* ***Rcmdr****, una vez cargados los datos, se selecciona* `Estadísticos→Medias→ANOVA de un factor...`*, lo que da acceso a la ventana de diálogo del procedimiento donde se indicarán las variables a tratar, obteniendo en* ***Rcmdr*** *la siguiente salida:*

```
> .Anova <- lm(Ind~Pueblo, data=Datos)
> anova(.Anova)
 Analysis of Variance Table
Response: Ind
             Df   Sum Sq   Mean Sq   F value      Pr(> F)
   Pueblo     3   4499.0    1499.7    22.433    5.632e-06 ***
 Residuals   16   1069.6     66.8
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> tapply(Datos$Ind, Datos$Pueblo, mean, na.rm=TRUE) # means
    P1          P2          P3          P4
 79.80000   48.83333   80.40000   52.25000
```

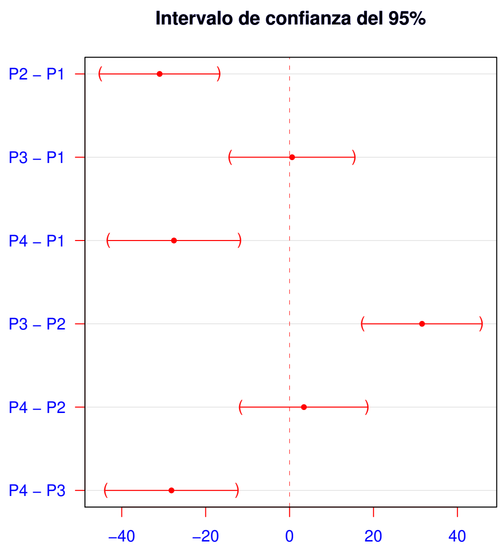


Figura 7.1: Intervalos de confianza de Tukey

```
> tapply(Datos$Ind, Datos$Pueblo, sd, na.rm=TRUE) # std.
deviations
      P1         P2         P3         P4
  8.043631   8.183316   8.443933   7.973916
> tapply(Datos$Ind, Datos$Pueblo, function(x) sum(!is.na(x))) #
counts
 P1   P2   P3   P4
  5    6    5    4
> remove(.Anova)
```

Puesto que el p-valor $< 0,001$ *se rechaza la hipótesis de igualdad de medias en el índice de alfabetización de los cuatro municipios.*

3.1. Comparaciones múltiples

Bajo las mismas hipótesis del test de la F, si se rechaza la hipótesis nula de igualdad de medias se debe proceder a la realización de contrastes de medias dos a dos.

Ejemplo 7.3

Con los datos del ejemplo anterior y puesto que se ha rechazado la hipótesis de igualdad global se realizarán las comparaciones de medias dos a dos. Se accede mediante la misma secuencia de menú, `Estadísticos→Medias→ANOVA de un factor...`, *a la ventana de introducción de datos y opciones, marcando ahora* `Comparaciones dos a dos de las medias`.

Además de la salida anterior ***Rcmdr*** *crea dos bloques de instrucciones, una que genera la salida numérica de intervalos para las diferencias de medias y otra que construye el gráfico de dichos intervalos.*

Análisis numérico:

El siguiente grupo de instrucciones crea la salida numérica.

```
> .Pairs <- glht(.Anova, linfct = mcp(Pueblo = ''Tukey''))
> confint(.Pairs)
Simultaneous Confidence Intervals for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = Ind~Pueblo, data = Datos)
Estimated Quantile = 2.8607
Linear Hypotheses:
                 Estimate     lwr        upr
 P2 - P1 == 0   -30.9667   -45.1295   -16.8038
 P3 - P1 == 0     0.6000   -14.1926    15.3926
 P4 - P1 == 0   -27.5500   -43.2399   -11.8601
 P3 - P2 == 0    31.5667    17.4038    45.7295
 P4 - P2 == 0     3.4167   -11.6810    18.5143
 P4 - P3 == 0   -28.1500   -43.8399   -12.4601
95% family-wise confidence level
```

El análisis de la salida lleva a que P_1 es igual a P_3 y mayor que P_2 y P_4, que P_2 es igual a P_4 y menor que P_3 y que P_3 es mayor que P_4.

Análisis gráfico:

Por otra parte, el siguiente grupo de instrucciones crea el gráfico de intervalos de confianza para la diferencia de medias (figura 7.1).

```
> old.oma <- par(oma=c(0,5,0,0))
> plot(confint(.Pairs), col= ''red'', main=''Intervalo
de confianza del 95%'',col.main=''blue'', xlab='' '',
col.axis=''blue'')
> par(old.oma)
> remove(.Pairs)
```

4. Alternativa no paramétrica. Test de Kruskal Wallis

Como se ha indicado, si fallan las hipótesis de normalidad y/o homocedasticidad se debe recurrir a una alternativa no paramétrica para realizar el test de igualdad de medias. La solución más extendida la proporciona el test de Kruskal Wallis. Dicho test es una prueba basada en rangos con signos y es una generalización del test de Wilcoxon al caso de k muestras.

Ejemplo 7.4

Suponga que se desea comparar el rendimiento de 5 tipos de neumáticos, A, B, C, D y E, para lo que decide probarlos en distintos coches de similares características. Sus vidas medias en rodaje, medidas en miles de kilómetros, vienen dadas en la siguiente tabla:

Llantas	*Vidas medias*				
A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Para contrastar que no hay diferencias entre los cinco tipos de neumáticos se elige el test de Kruskal Wallis. Los datos han sido almacenados en el fichero `neumaticos.dat` *dentro del repositorio de datos. En* ***Rcmdr*** *se activa la secuencia de menú* `Estadísticos`→`Test no paramétricos`→`Test de Kruskal Wallis`*, abriéndose la correspondiente ventana de diálogo donde se seleccionan variable y factor, en este caso* `Km` *y* `Neum`*.* ***Rcmdr*** *proporciona en primer lugar las medianas de cada grupo y seguidamente el estadístico de Kruskal Wallis junto con su p-valor.*

```
> tapply(Datos$Km, Datos$Neum, median, na.rm=TRUE)
  A   B   C   D   E
 68  53  72  57  65
```

```
> kruskal.test(Km~Neum, data=Datos)

Kruskal-Wallis rank sum test
data: Km by Neum
Kruskal-Wallis chi-squared = 6.4949, df = 4, p-value = 0.1651
```

A la vista de los resultados, p-valor $= 0,1651$*, se concluye que no hay diferencias significativas entre los rendimientos de los cinco tipos de neumáticos.*

5. Ejercicios

7.1 Estudie, a partir de la tabla de datos porcentuales que se da, si las medias de los tres niveles de un determinado factor son iguales.

Nivel I	Nivel II	Nivel III
$8,1$	$8,6$	12
$9,2$	$8,9$	$13,2$
$9,5$	$7,4$	$13,1$

7.2 Una empresa tiene en un establecimiento cuatro vendedores y pretende asignar primas en función de las ventas. A la vista de la tabla de ventas en los últimos cinco meses (miles de euros), indique si los cuatro vendedores son igualmente eficaces. De no ser así elabore el ranking en razón de las ventas.

Vend. 1	Vend. 2	Vend. 3	Vend. 4
$6,46$	$5,79$	$8,37$	$4,94$
$4,83$	$5,13$	$7,57$	$4,11$
$5,89$	$6,17$	$8,69$	$5,45$
$5,30$	$4,72$	$8,06$	$5,21$
$6,33$	$5,60$	$7,23$	$5,00$

7.3 A partir de la cuenta de resultados que presentaban 13 entidades financieras englobadas en los ámbitos europeo, nacional y regional se ha calculado el porcentaje destinado a la generación bruta de fondos, con los siguientes resultados:

Ámbito	Generación bruta de fondos				
Europeo	$0,4$	$3,8$	$2,5$	$2,9$	
Tipo II	$4,7$	$2,0$	$1,8$	$2,8$	
Tipo III	$0,9$	$3,7$	$3,1$	$6,2$	$2,7$

¿Puede considerarse que la proporción de fondos es igual independientemente del ámbito de actuación?

7.4 Una cierta planta ha sido cultivada con cinco fertilizantes distintos. Se desea estudiar si el tipo de fertilizante influye en la longitud de la planta, para lo cual se han medido las longitudes de cinco series de 10 plantas, obteniéndose para cada serie los resultados que aparecen en el fichero plantas.dat. ¿Hay evidencia estadística suficiente para afirmar que las medias son diferentes? De ser así, ¿existen tipos de fertilizante que no se diferencien entre sí?

7.5 Un fabricante está interesado en la resistencia a la tensión de una fibra sintética. Se sospecha que la resistencia está relacionada con el porcentaje de algodón en la fibra. Suponer que la distribución para cada porcentaje son aproximadamente normales y se da la homogeneidad de las varianzas. Para ello, se emplean cinco niveles de porcentaje de algodón. De 5 réplicas aleatorias se obtienen los siguientes datos:

Porcentaje de algodón	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

¿Puede considerarse que la resistencia de las prendas es la misma independiente del porcentaje de algodón presente en sus fibras?

Apéndice A

Ficheros de datos

Puede accederse a los ficheros documentados en esta sección en la dirección `http://knuth.uca.es/ebrcmdr`.

`caracoles.dat` Conjunto de datos que recoge las medidas del diámetro y la separación entre espirales (μm) de las conchas de 20 caracoles adultos.

`cebada.dat` Contiene información sobre la producción de cuatro variedades de cebada A, B, C y D.

`chickwts` Datos contenidos en el paquete "datasets" de R. Peso de 71 pollos sometidos a distintos tipos de alimentación Contiene dos variables, una numérica weight: peso y un factor feed: tipo de alimentación, con 6 niveles.

`eolico_apilado.dat` Los datos del fichero `parque_eolico.dat` apilados según las variables velocidad y parque. Estos datos permiten trabajar más cómodamente en Rcmdr.

`fenofibrato.dat` Niveles de fibrinógeno de 32 pacientes, antes y después de ser tratados durante un año con fenofibrato.

`iris` Datos contenidos en el paquete "datasets" de R. Provienen del famoso estudio realizado por el estadístico y genetista Sir Ronald A.

Fisher. sobre la clasificación de 3 especies de iris (setosa, versicolor y virginica). Las variables de estudio son la longitud y el ancho del sépalo y, la longitud y el ancho del pétalo de las 3 especies mencionadas.

`neumaticos.dat` Vidas medias en rodaje de 5 tipos de neumáticos A, B, C, D y E, medidas en miles de kilómetros, probados en distintos coches de similares características.

`niv_estudios_cadiz.dat` Nivel académico de la población gaditana. Fuente: Instituto Estadístico de Andalucía.

`peso_altura.dat` Fichero en el que se proporcionan peso, altura y presión arterial inicial y final de un grupo de 100 pacientes sometidos a cierto fármaco (Ca Antagonista + diurético, IECA o placebo).

`reproduccion_vir.dat` Número de virus reproducidos en función del tiempo (minutos) y de la temperatura (grados), según el tipo de cultivo (ácido, básico o neutro).

`titanic.dat` Recoge información sobre el naufragio del buque Titanic (estatus económico, sexo, edad y supervivientes). Éste es el fichero incluido en el paquete "datasets" de R y está modificado para que se cargue correctamente en Rcmdr.

`parque_eolico.dat` Mediciones de la velocidad del viento (m/s) en dos localizaciones alternativas (Parque1 y Parque2) registradas de forma simultánea durante 730 horas.

Apéndice B

Tabla de medidas estadísticas

En la siguiente tabla se ofrece un resumen de las medidas más usadas en estadística descriptiva con sus correspondientes instrucciones en **R**

Medidas de posición	**Instrucciones en R**
Cuantiles	> `quantile(datos,p)` con `p` vector de cuantiles deseados. > `quantile(datos)` obtenemos todos los cuartiles.
Medidas de centralización	
Media	> `mean(datos)`
Mediana	> `median(datos)`
Medidas de dispersión	
Cuasivarianza	> `var(datos)`
Cuasidesviación típica	> `sd(datos)`
Varianza	> `var(datos)*` `(length(datos)-1)/length(datos)`

Medidas de dispersión	**Instrucciones en R**
Desviación típica	`>sqrt(var(datos)*` `(length(datos)-1)/length(datos))`
Rango muestral	`>max(datos)-min(datos)`
Rango intercuartílico	`>quantile(datos,.75)` `-quantile(datos,.25)`
Coeficiente de variación	`>sd(datos)/abs(mean(datos))`
Medidas de forma	En el paquete `fBasics`
Coeficiente de curtosis	`>kurtosis(datos)`
Coeficiente de asimetría	`>skewness(datos)`

Apéndice C

Tabla de modelos

Modelo	**Instrucción**	**Ecuación**
Lineal	$>$`lm(`$Y \sim X$`, data=Datos)`	$Y = a + b \cdot X$
Lineal sin término independiente	$>$`lm(`$Y \sim 0 + X$`, data=Datos)`	$Y = a \cdot X$
Polinomial	$>$`lm(`$Y \sim X + I(X^2)+$ $+I(X^3) + \cdots + I(X^n)$`,` `data=Datos)`	$Y = a_0 + a_1 \cdot X+$ $+ \cdots + a_n \cdot X^n$
Polinomial sin término independiente	$>$`lm(`$Y \sim 0 + X + I(X^2)+$ $+I(X^3) + \cdots + I(X^n)$`,` `data=Datos)`	$Y = a_1 \cdot X+$ $+ \cdots + a_n \cdot X^n$
Potencial	$>$`lm(`$\log(Y) \sim \log(X)$`,` `data=Datos)`	$Y = a' \cdot X^b$, (1)
Exponencial	$>$`lm(`$\log(Y) \sim X$`, data=Datos)`	$Y = e^{a+b \cdot X}$
Logarítmico	$>$`lm(`$Y \sim \log(X)$`, data=Datos)`	$Y = a + b \cdot \log(X)$
Hiperbólico	$>$`lm(`$Y \sim I(1/X)$`, data=Datos)`	$Y = a + \frac{b}{X}$
Doble inverso	$>$`lm(`$I(1/Y) \sim I(1/X)$`,` `data=Datos)`	$Y = \frac{1}{a+\frac{b}{X}}$
Lineal generalizado	$>$`glm(fórmula, family=` `=familia(link), data=Datos)`	(2)

(1) Los coeficientes a y b obtenidos en **Rcmdr** corresponden a la ecuación $\log(Y) = a + b \cdot \log(X)$, con lo que el modelo potencial sería $Y = e^a \cdot X^b$.

(2) `familia` puede tomar los valores `gaussian, binomial, poisson, Gamma, inverse.gaussian, quasibinomial y quasipoisson`. La función de enlace (`link`) puede tomar distintos valores según la familia seleccionada. Podemos ver las distintas opciones consultando en la ayuda de **R** la función `family (help(family) o ?family)`.