Resultados de la encuesta realizada en el proyecto

PI1_12_057

"Desarrollo de material didáctico para la docencia de estadística utilizando el software estadístico, R con R-commander"

Marzo 2012

Universidad de Cádiz

1. Preámbulo

Para conocer las experiencias y opiniones de los alumnos en este proyecto, se ha realizado una encuesta, cuyos resultados recoge el presente informe.

La encuesta ha sido enviada por correo electrónico a los alumnos de aquellas asignaturas que se recogían en este proyecto, cuyo listado se consiguió a través del campus virtual de dichas asignaturas que incluía 139 registros.

Una vez cumplimentadas las encuestas, éstas se añadían, de forma anónima, en sendos registros de una base de datos para su posterior tratamiento. A aquellas personas que no cumplimentaron la encuesta, en una primera instancia, se les remitió hasta un máximo de dos recordatorios en las dos semanas siguientes del envío inicial.

En todo el proceso se han garantizado los preceptos recogidos en la Ley Orgánica de Protección de Datos (LOPD).

2. Descripción del cuestionario

El cuestionario se divide en varios bloques, donde las preguntas han sido codificadas en una escala de Likert $1-5$, correspondiéndose el valor 1 con un nivel mínimo y el 5 con un nivel máximo de satisfacción.

El cuestionario es el siguiente:

Aspectos generales de organización	1	2	3	4	5
Instalaciones donde se desarrollan las clases	O	O	O	O	O
Duración de las clases	O	O	O	O	O
Calendario y Horario de la asignatura	O	O	O	O	O
Docentes de la asignatura	1	2	3	4	5
Dominio de la materia	O	O	O	O	O
Claridad en la comunicación	O	O	O	O	O
Aspectos globales	1	2	3	4	5
Guiones de trabajo	O	O	O	O	O
Medios y recursos disponibles	O	O	O	O	O
Datos y problemas analizados relacionados con la titulación	O	O	O	O	O
Utilidad de la asignatura para formación académica	O	O	O	O	O
Utilidad asignatura para la formación investigadora	O	O	O	O	O
Valoración general de la asignatura	O	O	O	O	O
Desarrollo de la asignatura	1	2	3	4	5
Método expositivo	O	O	O	O	O
Estudio de casos	O	O	O	O	O
Resolución de ejercicios	O	O	O	O	O
Aprendizaje basado en problemas	O	O	O	O	O
Proyectos tutorizados	O	O	O	O	O
Participación del grupo	1	2	3	4	5
Constancia en la asignatura	O	O	O	O	O
Implicación y motivación de los participantes	O	O	O	O	O

En relación a las aplicaciones informáticas, todas las herramientas empleadas tienen licencia libre (GNU-GPL):

- La plataforma usada ha sido apache sobre ubuntu-linux
- La edición y elaboración de informes se ha realizado con LaTeX
- La herramienta elegida para la administración automatizada del cuestionario ha sido `limesurvey`
- El análisis de datos se ha hecho con el software estadístico `R`

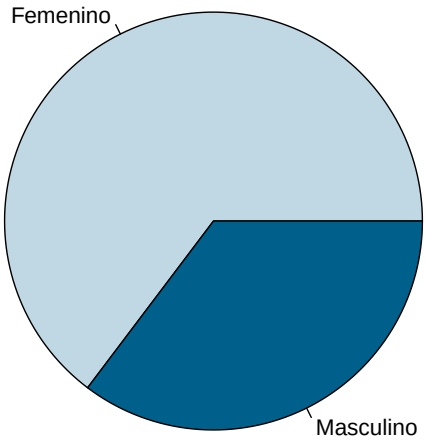
La encuesta ha sido cumplimentada en su totalidad por un total de 34 personas, lo que representa un 24.46 % del total de 139 posibles participantes. Un total de 4 invitados declinaron explícitamente participar en el estudio.

3. Caracterización de perfiles de los que han contestado la encuesta

En este apartado se caracteriza a los encuestados en función del sexo y de la titulación a la que pertenecen.

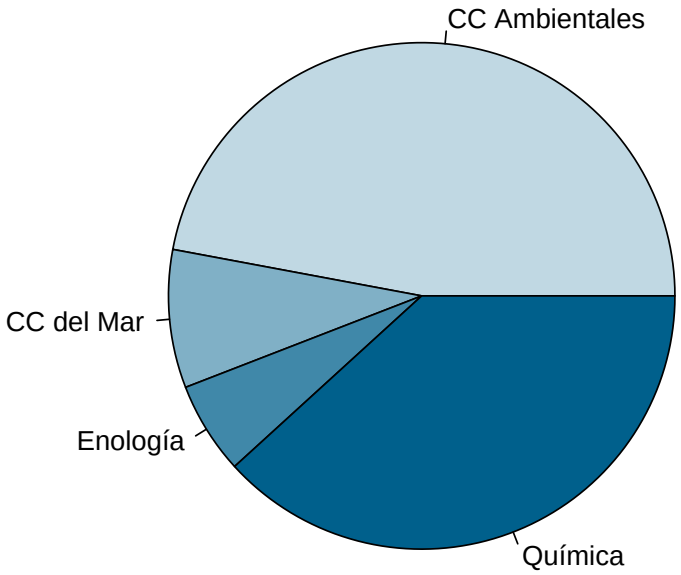
Sexo

	n	%
Femenino	22	64.7
Masculino	12	35.3



Titulación

	n	%
Grado en Ciencias Ambientales	16	47.1
Grado en Ciencias del Mar	3	8.8
Grado en Enología	2	5.9
Grado en Quimica	13	38.2



3.1. Resultados del cuestionario

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	34	3.82	0.97
Duración de las clases	34	3.56	1.11
Calendario y Horario de la asignatura	34	3.24	1.23
Docentes de la asignatura			
Dominio de la materia	34	4.03	1.09
Claridad en la comunicación	34	3.15	1.37
Aspectos globales			
Guiones de trabajo	34	2.85	1.31
Medios y recursos disponibles	34	3.24	1.13
Datos y problemas analizados relacionados con la titulación	34	3.26	1.16
Utilidad de la asignatura para formación académica	34	3.76	1.07
Utilidad asignatura para la formación investigadora	33	3.67	1.08
Valoración general de la asignatura	32	3.00	1.24
Desarrollo de la asignatura			
Método expositivo	34	3.26	1.29
Estudio de casos	34	3.15	1.26
Resolución de ejercicios	34	2.79	1.23
Aprendizaje basado en problemas	34	2.85	1.28
Proyectos tutorizados	34	2.91	1.19
Participación del grupo			
Constancia en la asignatura	33	4.06	0.93
Implicación y motivación de los participantes	33	3.30	1.13

4. Resultados según titulación

4.1. Grado en Química

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	13	3.92	1.04
Duración de las clases	13	4.15	0.80
Calendario y Horario de la asignatura	13	3.85	0.99
Docentes de la asignatura			
Dominio de la materia	13	4.23	1.01
Claridad en la comunicación	13	3.38	1.12
Aspectos globales			
Guiones de trabajo	13	3.38	0.96
Medios y recursos disponibles	13	3.62	0.87
Datos y problemas analizados relacionados con la titulación	13	3.54	0.97
Utilidad de la asignatura para formación académica	13	3.31	1.18
Utilidad asignatura para la formación investigadora	12	3.33	1.07
Valoración general de la asignatura	12	3.58	0.90
Desarrollo de la asignatura			
Método expositivo	13	3.69	1.03
Estudio de casos	13	3.38	1.04
Resolución de ejercicios	13	3.00	0.91
Aprendizaje basado en problemas	13	2.77	1.17
Proyectos tutorizados	13	3.15	0.90
Participación del grupo			
Constancia en la asignatura	13	4.00	1.00
Implicación y motivación de los participantes	13	3.08	1.04

4.2. Grado en Enología

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	2	4.50	0.71
Duración de las clases	2	4.00	1.41
Calendario y Horario de la asignatura	2	4.50	0.71
Docentes de la asignatura			
Dominio de la materia	2	5.00	0.00
Claridad en la comunicación	2	5.00	0.00
Aspectos globales			
Guiones de trabajo	2	5.00	0.00
Medios y recursos disponibles	2	4.50	0.71
Datos y problemas analizados relacionados con la titulación	2	3.00	1.41
Utilidad de la asignatura para formación académica	2	5.00	0.00
Utilidad asignatura para la formación investigadora	2	4.50	0.71
Valoración general de la asignatura	2	4.50	0.71
Desarrollo de la asignatura			
Método expositivo	2	5.00	0.00
Estudio de casos	2	5.00	0.00
Resolución de ejercicios	2	4.50	0.71
Aprendizaje basado en problemas	2	4.50	0.71
Proyectos tutorizados	2	5.00	0.00
Participación del grupo			
Constancia en la asignatura	2	5.00	0.00
Implicación y motivación de los participantes	2	5.00	0.00

4.3. Grado en Ciencias del Mar

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	3	4.00	1.00
Duración de las clases	3	3.67	1.53
Calendario y Horario de la asignatura	3	3.00	1.00
Docentes de la asignatura			
Dominio de la materia	3	4.67	0.58
Claridad en la comunicación	3	4.33	1.15
Aspectos globales			
Guiones de trabajo	3	3.33	0.58
Medios y recursos disponibles	3	3.67	0.58
Datos y problemas analizados relacionados con la titulación	3	4.33	0.58
Utilidad de la asignatura para formación académica	3	4.00	1.00
Utilidad asignatura para la formación investigadora	3	4.00	1.00
Valoración general de la asignatura	3	3.67	0.58
Desarrollo de la asignatura			
Método expositivo	3	4.00	1.00
Estudio de casos	3	3.67	0.58
Resolución de ejercicios	3	3.00	1.00
Aprendizaje basado en problemas	3	3.00	1.00
Proyectos tutorizados	3	2.33	0.58
Participación del grupo			
Constancia en la asignatura	3	4.67	0.58
Implicación y motivación de los participantes	3	4.33	0.58

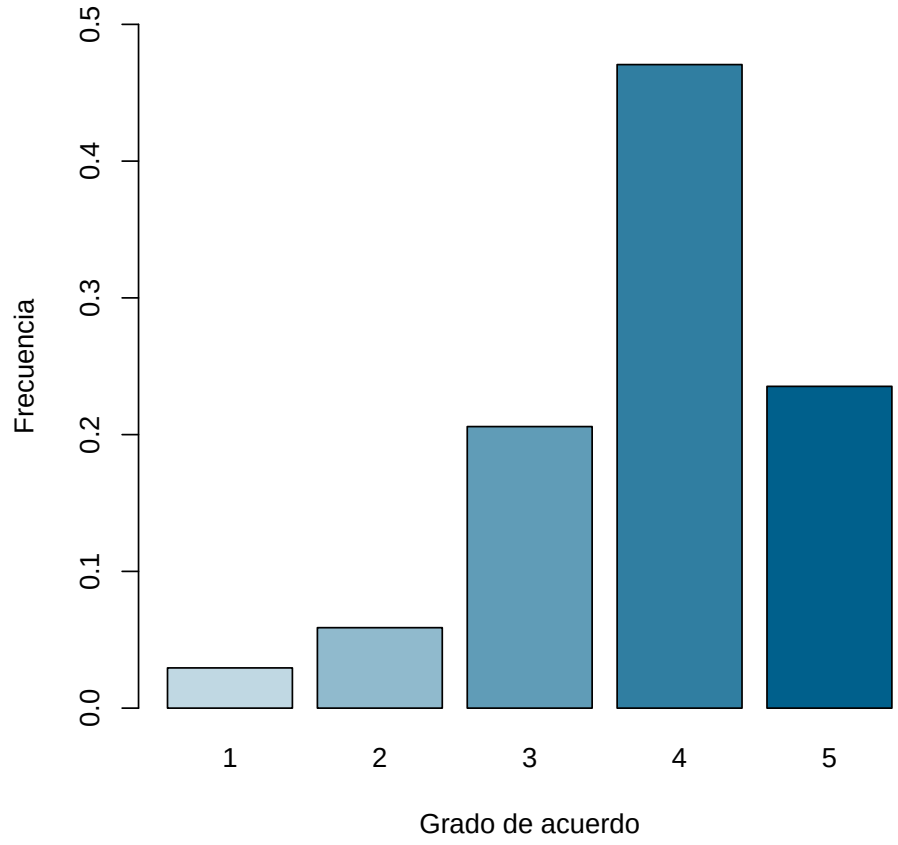
4.4. Grado en Ciencias Ambientales

	n	Media	D. típica
Aspectos generales de organización			
Instalaciones donde se desarrollan las clases	16	3.62	0.96
Duración de las clases	16	3.00	1.03
Calendario y Horario de la asignatura	16	2.62	1.20
Docentes de la asignatura			
Dominio de la materia	16	3.62	1.15
Claridad en la comunicación	16	2.50	1.32
Aspectos globales			
Guiones de trabajo	16	2.06	1.18
Medios y recursos disponibles	16	2.69	1.20
Datos y problemas analizados relacionados con la titulación	16	2.88	1.26
Utilidad de la asignatura para formación académica	16	3.94	0.93
Utilidad asignatura para la formación investigadora	16	3.75	1.13
Valoración general de la asignatura	15	2.20	1.15
Desarrollo de la asignatura			
Método expositivo	16	2.56	1.21
Estudio de casos	16	2.62	1.31
Resolución de ejercicios	16	2.38	1.36
Aprendizaje basado en problemas	16	2.69	1.40
Proyectos tutorizados	16	2.56	1.26
Participación del grupo			
Constancia en la asignatura	15	3.87	0.92
Implicación y motivación de los participantes	15	3.07	1.10

5. Aspectos generales de organización

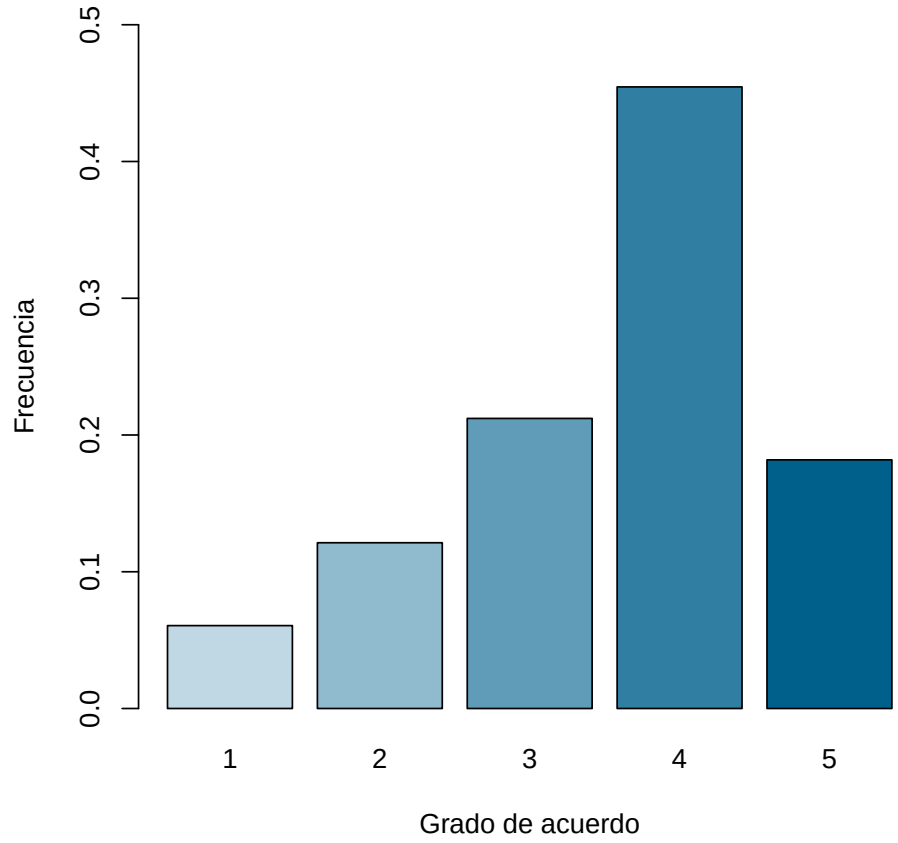
Instalaciones donde se desarrollan las clases

Cuenta	34.00
Mínimo	1.00
Media	3.82
Mediana	4.00
Máximo	5.00
Des. Típica	0.97



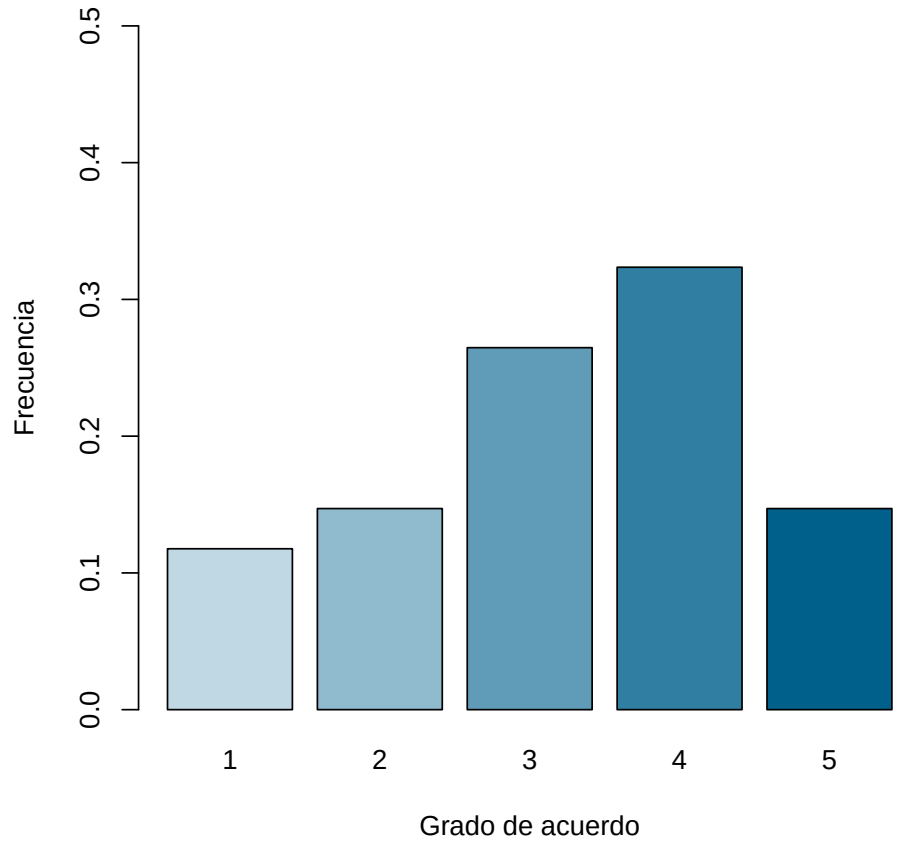
Duración de las clases

Cuenta	34.00
Mínimo	1.00
Media	3.56
Mediana	4.00
Máximo	5.00
Des. Típica	1.11



Calendario y Horario de la asignatura

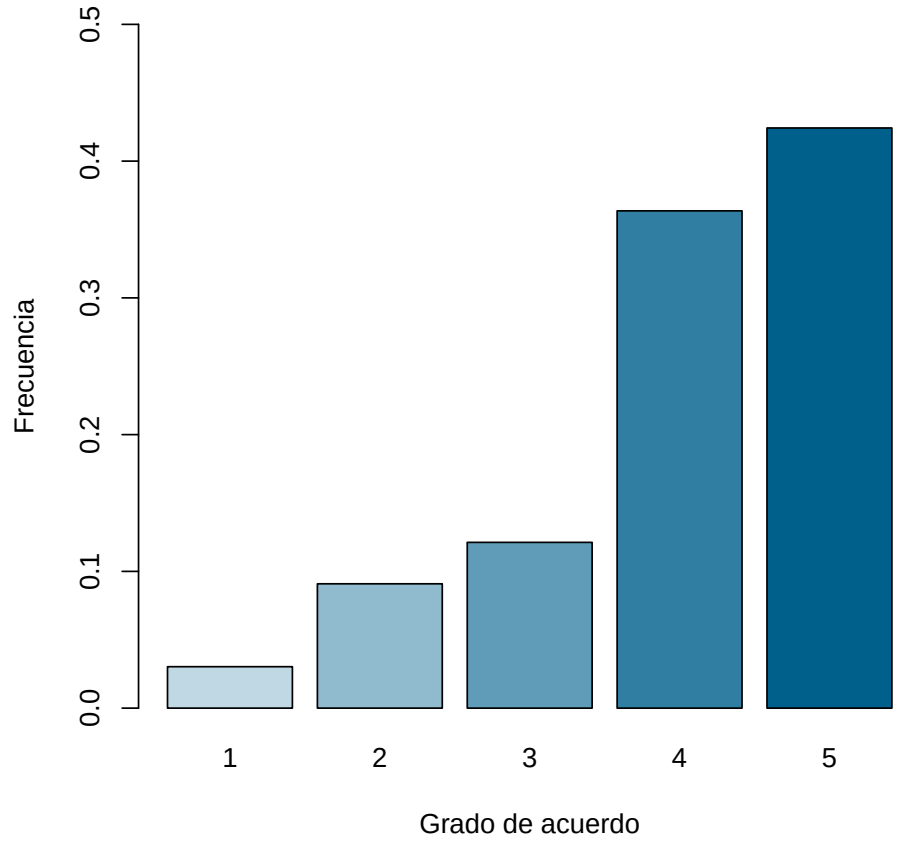
Cuenta	34.00
Mínimo	1.00
Media	3.24
Mediana	3.00
Máximo	5.00
Des. Típica	1.23



6. Docentes de la asignatura

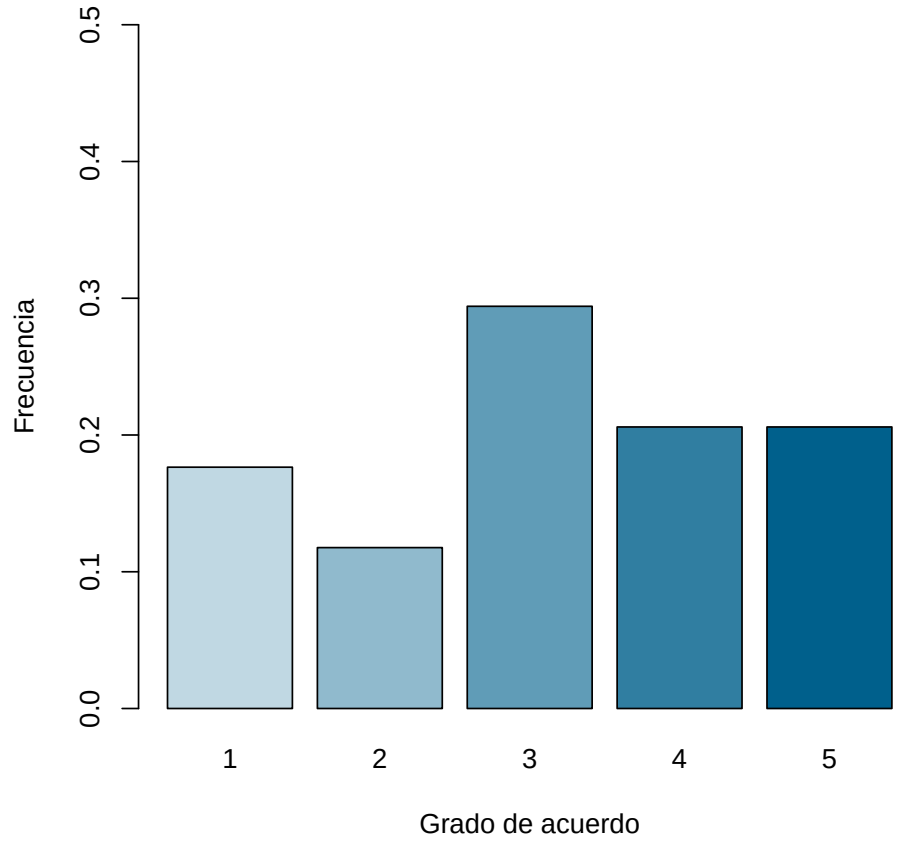
Dominio de la materia

Cuenta	34.00
Mínimo	1.00
Media	4.03
Mediana	4.00
Máximo	5.00
Des. Típica	1.09



Claridad en la comunicación

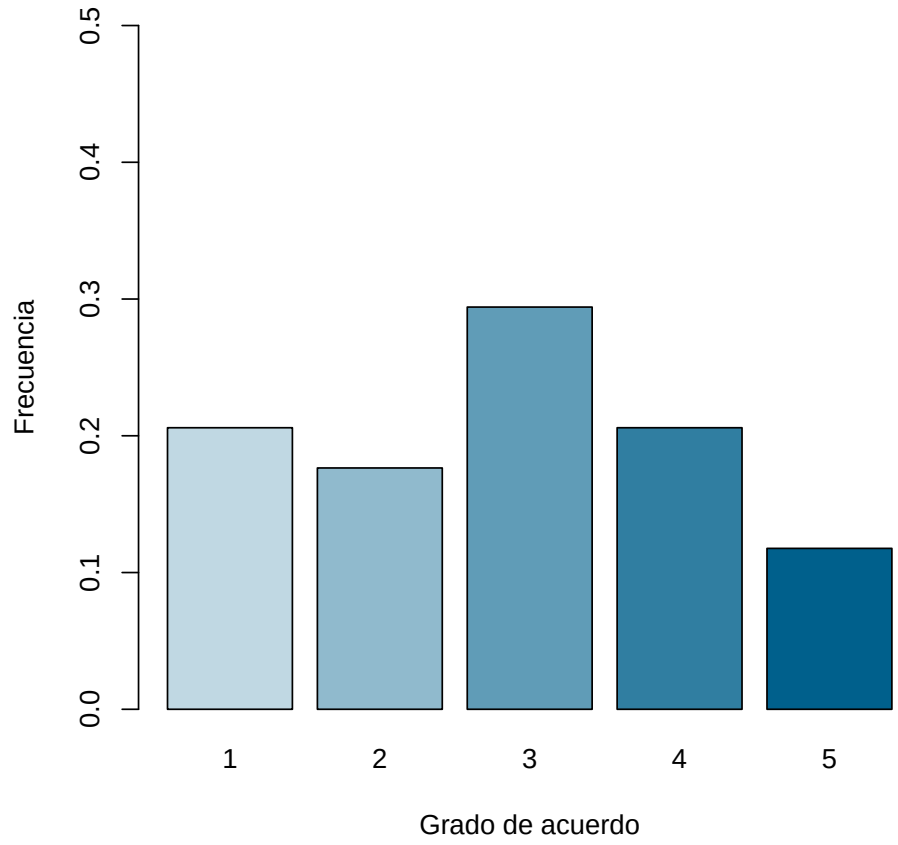
Cuenta	34.00
Mínimo	1.00
Media	3.15
Mediana	3.00
Máximo	5.00
Des. Típica	1.37



7. Aspectos globales

Guiones de trabajo

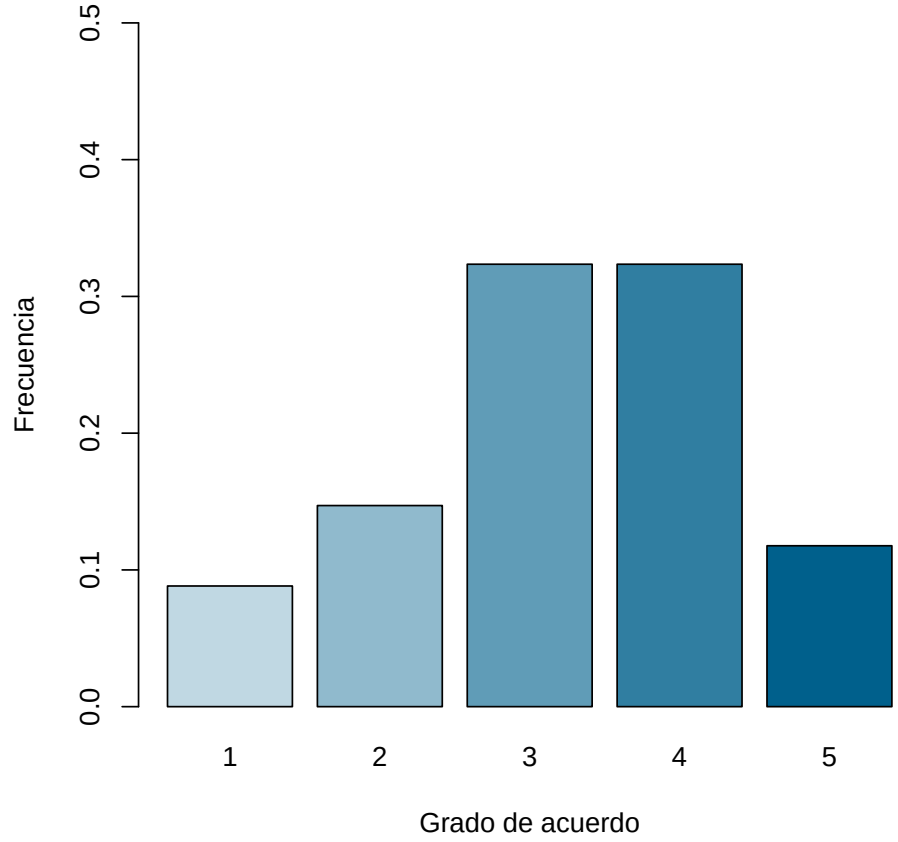
Cuenta	34.00
Mínimo	1.00
Media	2.85
Mediana	3.00
Máximo	5.00
Des. Típica	1.31



Se puede observar que existe un número elevado de los encuestados que valoran positivamente los guiones de trabajo en el desarrollo metodológico de las distintas materias, ya que un 60 % han elegido una puntuación entre 3 y 5 (incluidos) para dicha cuestión.

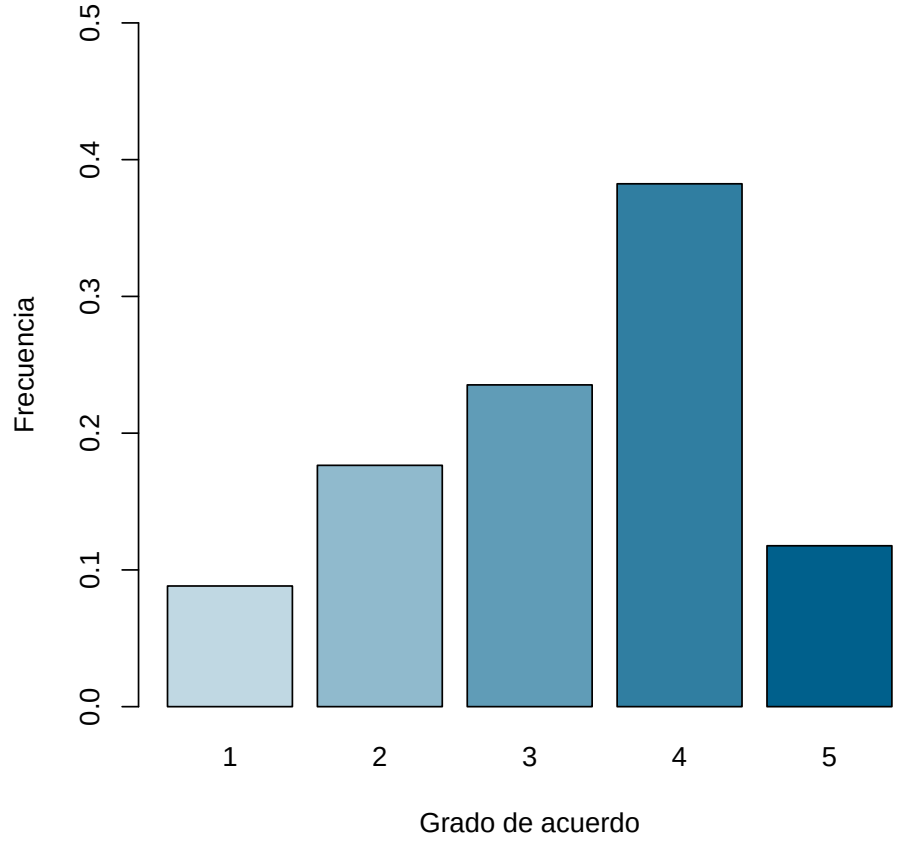
Medios y recursos disponibles

Cuenta	34.00
Mínimo	1.00
Media	3.24
Mediana	3.00
Máximo	5.00
Des. Típica	1.13



Datos y problemas analizados relacionados con la titulación

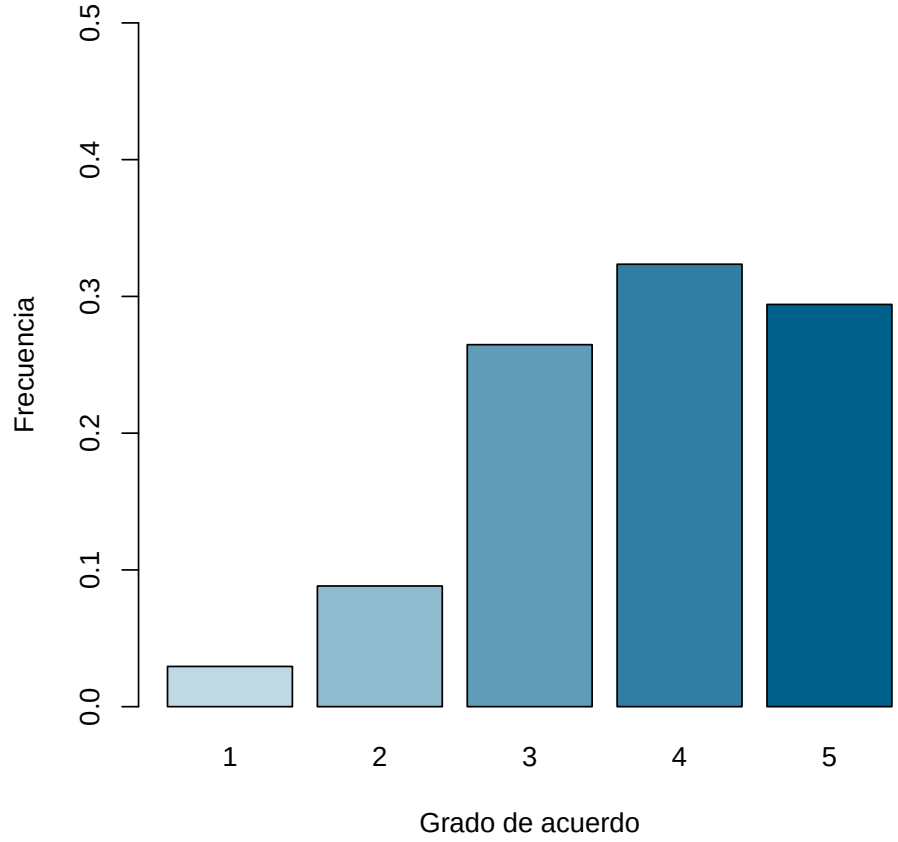
Cuenta	34.00
Mínimo	1.00
Media	3.26
Mediana	3.50
Máximo	5.00
Des. Típica	1.16



Se puede observar que existe un número elevado de los encuestados que valoran con alta estima la relación con cada titulación mantenida por los distintos conjuntos de datos y problemas analizados. Sólo un 26 %, tomó una valoración de 1 o 2, que pudiese interpretarse como no favorable.

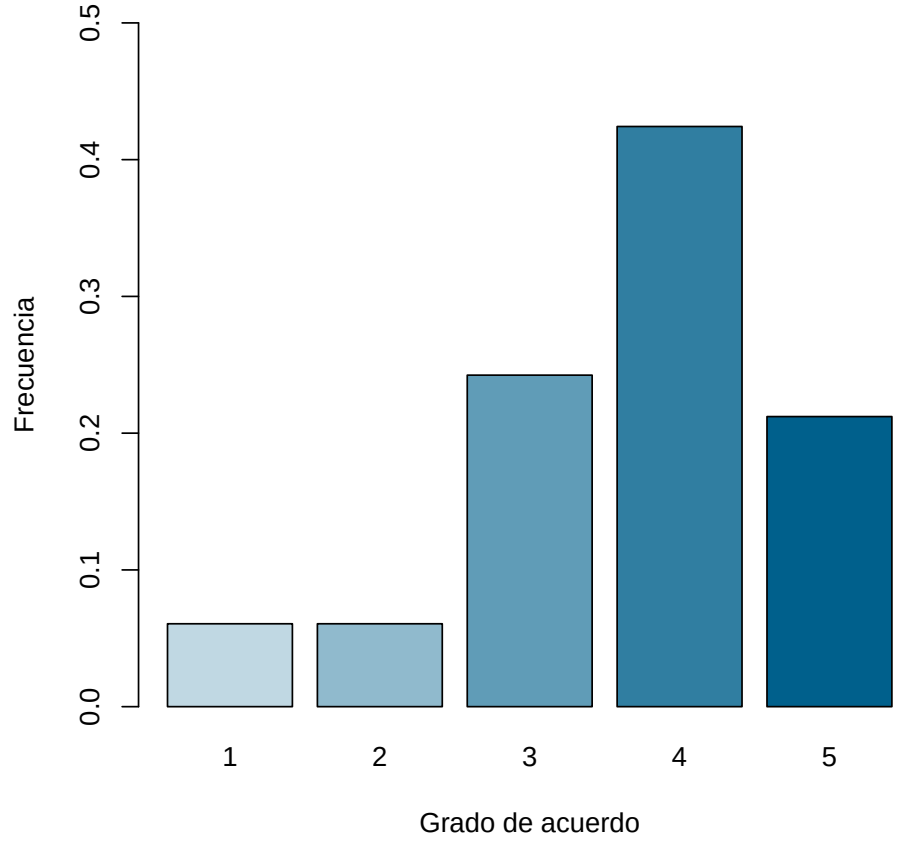
Utilidad de la asignatura para formación académica

Cuenta	34.00
Mínimo	1.00
Media	3.76
Mediana	4.00
Máximo	5.00
Des. Típica	1.07



Utilidad de la asignatura para la formación investigadora

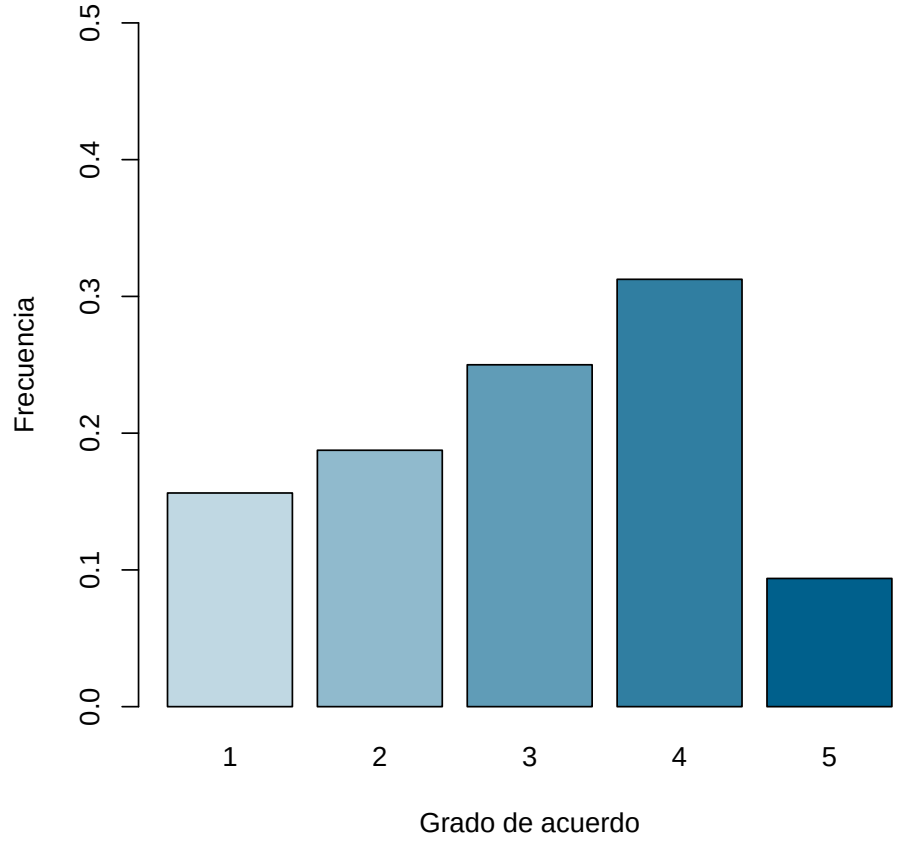
Cuenta	33.00
Mínimo	1.00
Media	3.67
Mediana	4.00
Máximo	5.00
Des. Típica	1.08



Existe un número elevado de los encuestados que valoran positivamente la utilidad de la asignatura para su formación académica e investigadora. En el primer caso, el 86 % de los encuestados, han elegido valores de 3 a 5. Para el aspecto correspondiente a la utilidad de la asignatura para cubrir su formación investigadora, el 60 % de los encuestados han elegido un valor de 4 o 5, el 23 % han elegido el valor de 3, y sólo un 12 % han valorado esta cuestión con las dos clases más bajas.

Valoración general de la asignatura

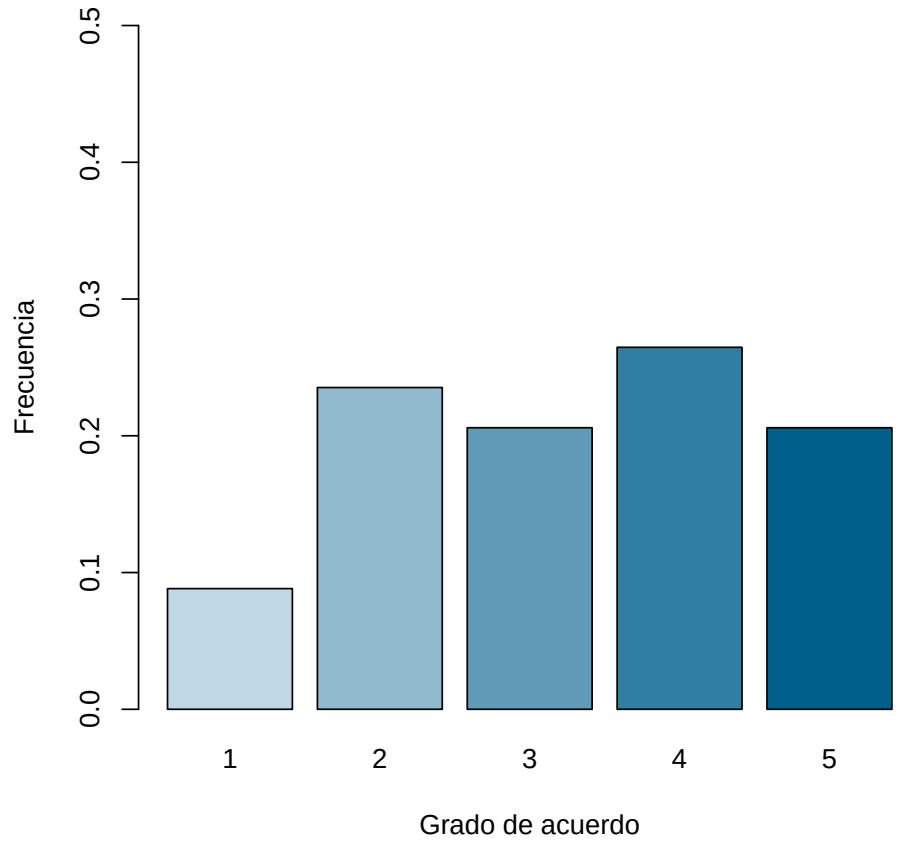
Cuenta	32.00
Mínimo	1.00
Media	3.00
Mediana	3.00
Máximo	5.00
Des. Típica	1.24



8. Desarrollo de la asignatura

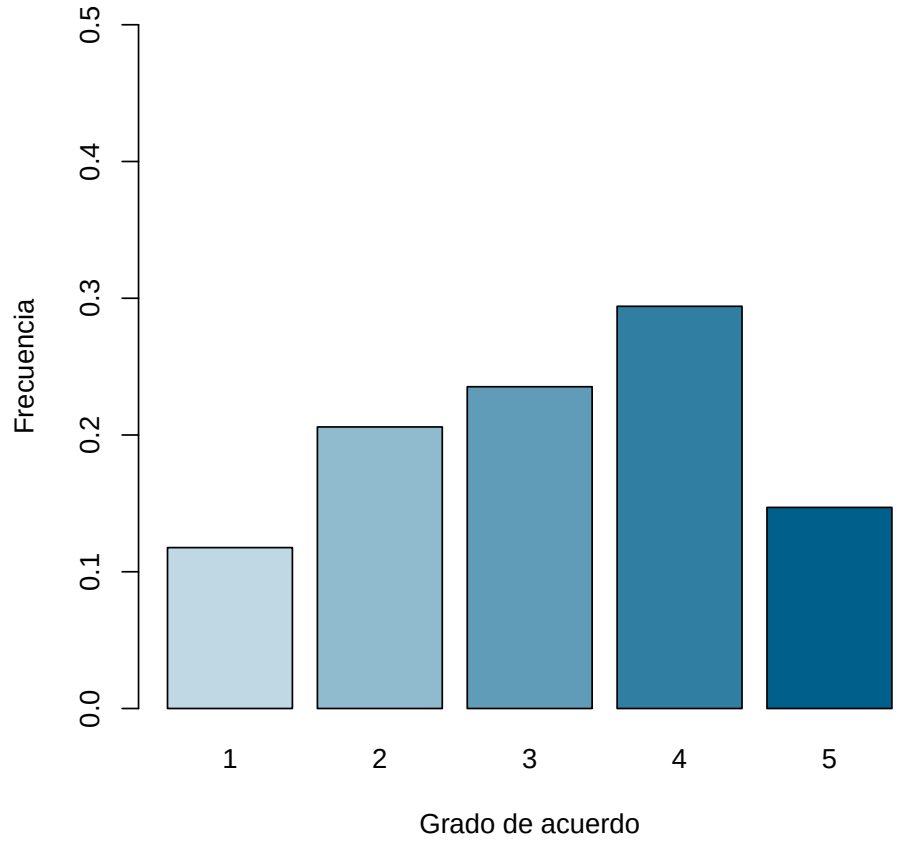
Método expositivo

Cuenta	34.00
Mínimo	1.00
Media	3.26
Mediana	3.00
Máximo	5.00
Des. Típica	1.29



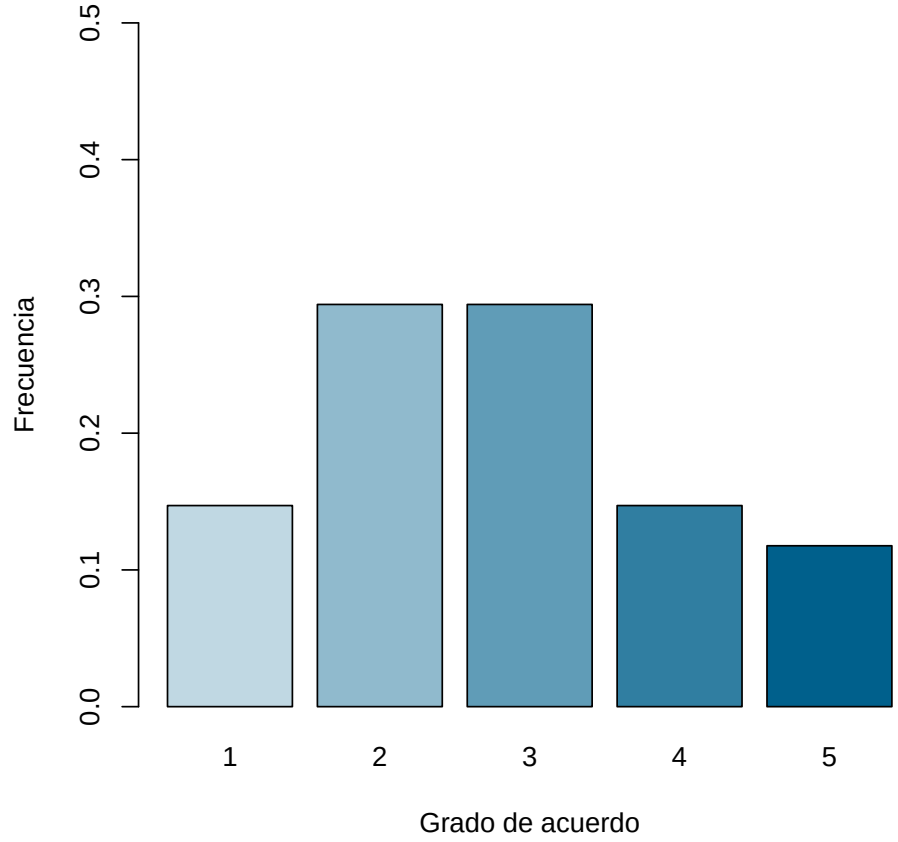
Estudio de casos

Cuenta	34.00
Mínimo	1.00
Media	3.15
Mediana	3.00
Máximo	5.00
Des. Típica	1.26



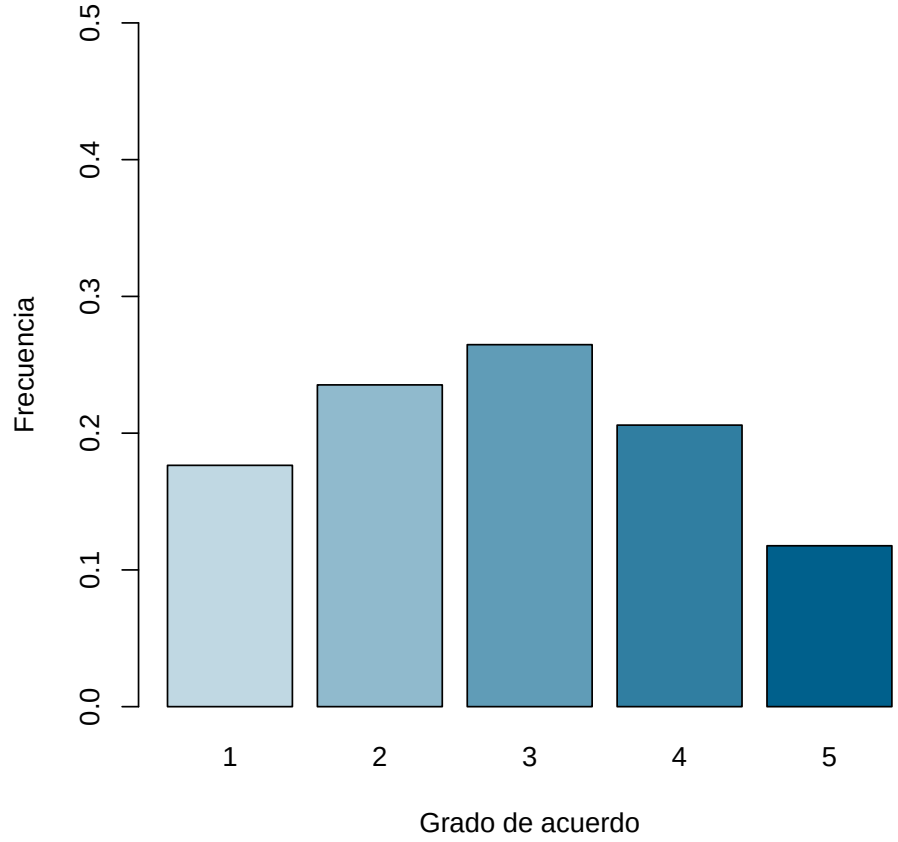
Resolución de ejercicios

Cuenta	34.00
Mínimo	1.00
Media	2.79
Mediana	3.00
Máximo	5.00
Des. Típica	1.23



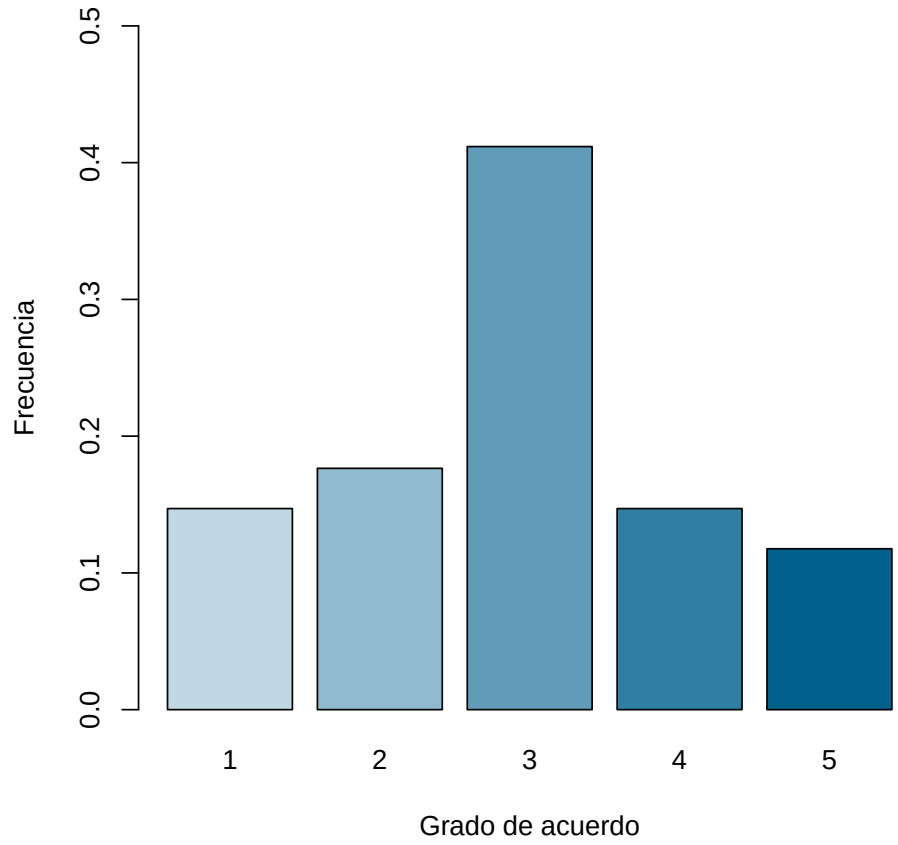
Aprendizaje basado en problemas

Cuenta	34.00
Mínimo	1.00
Media	2.85
Mediana	3.00
Máximo	5.00
Des. Típica	1.28



Proyectos tutorizados

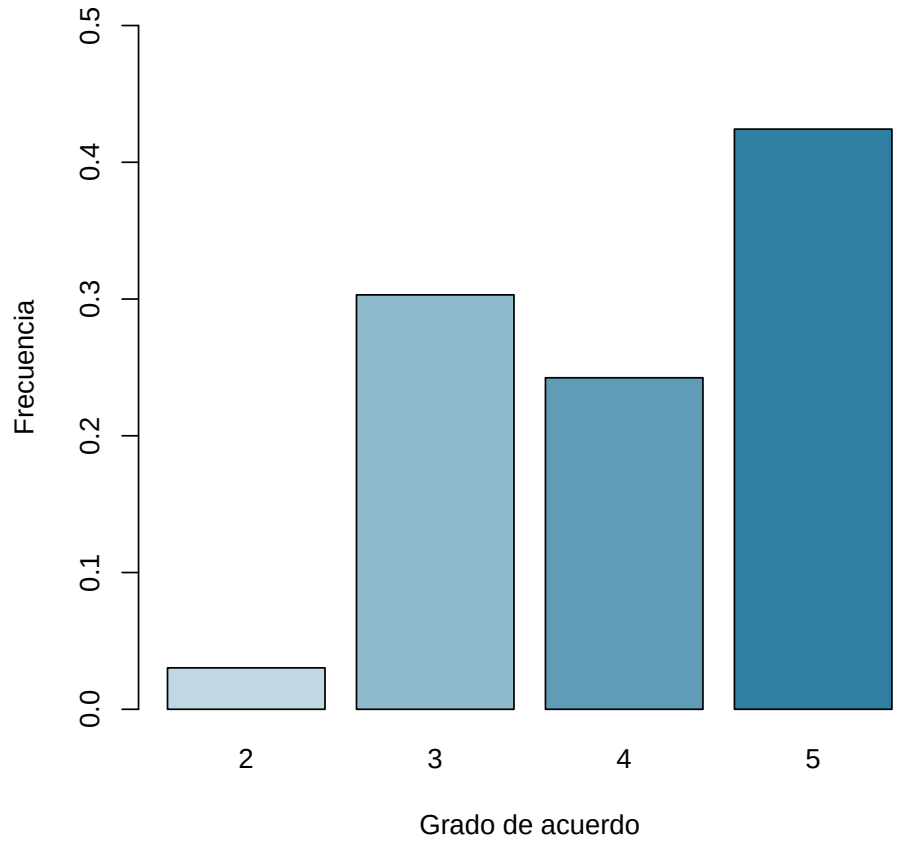
Cuenta	34.00
Mínimo	1.00
Media	2.91
Mediana	3.00
Máximo	5.00
Des. Típica	1.19



9. Participación del grupo

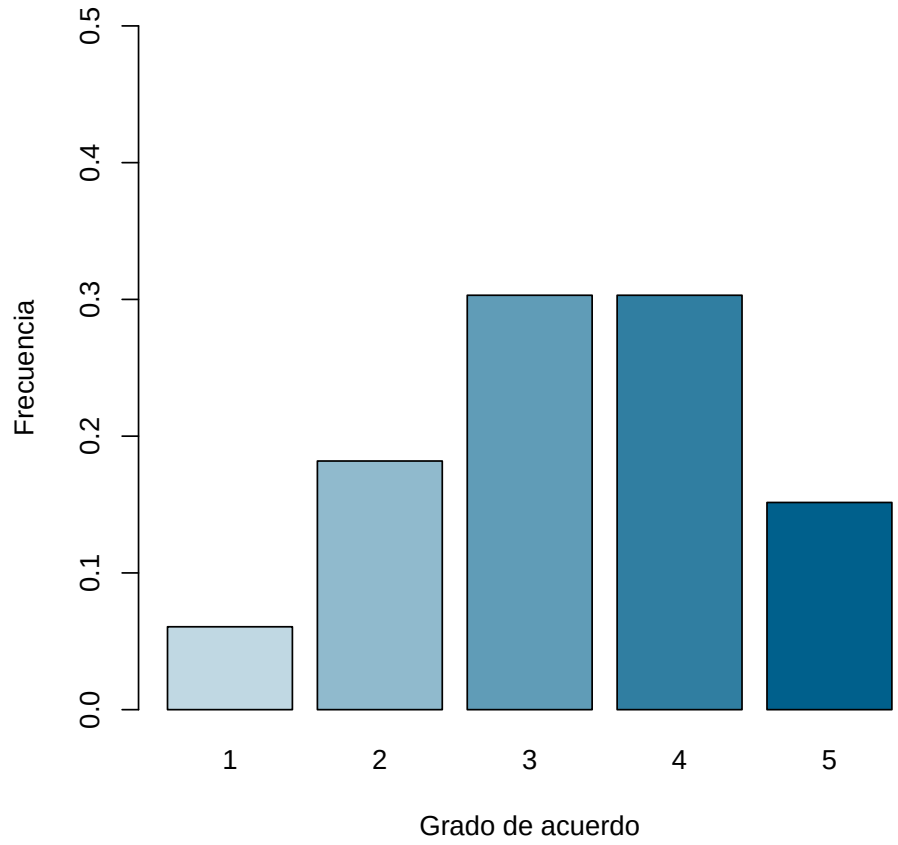
Constancia en la asignatura

Cuenta	33.00
Mínimo	2.00
Media	4.06
Mediana	4.00
Máximo	5.00
Des. Típica	0.93



Implicación y motivación de los participantes

Cuenta	33.00
Mínimo	1.00
Media	3.30
Mediana	3.00
Máximo	5.00
Des. Típica	1.13



Proyecto de Innovación Docente
Universidad de Cádiz

Cuadernillo de Prácticas

Estadística usando R y R-Commander

Curso 2011-2012

Autores:

Antonio Jesús Arriaza Gómez antoniojesus.arriaza@uca.es
Alfonso José Bello Espina alfonsojose.bello@uca.es
Fernando Fernández Palacín fernando.fernandez@uca.es
M. Auxiliadora López Sánchez auxiliadora.lopez@uca.es
Sonia María Pérez Plaza sonia.perez@uca.es
Antonio Sánchez Navas antonio.navas@uca.es

Proyecto de Innovación Docente:

Código PI1_12_057

Versión 1.0

Licencia de Documentación Libre de GNU

El formato usado para el presente documento ha sido una modificación de otro formato con licencia libre, extraído del proyecto de innovación IE-26.

Índice general

Parte I

Instalación y primeros pasos

1 Primeros Pasos

Contenidos

En esta práctica se van a dar los primeros pasos para empezar a trabajar con R además de familiarizarse con la interfaz R-commander. Se aprenderá a organizar los datos, cargarlos en **Rcmdr** y editarlos al objeto de adaptarlos a diferentes necesidades. Se verá como introducir datos en un editor, cargarlos e importarlos y exportarlos desde o hacia distintos formatos. En la dirección http://knuth.uca.es/R/ se pueden encontrar todos los recursos públicos generados a lo largo del proyecto R-UCA, incluyendo documentación, wiki, foros, estadísticas, etc.

1.1 Objetivos

- Instalar R y R-commander
- Manejar los distintos conjuntos de datos con R

1.2 Introducción a R

R es un paquete estadístico de última generación, al mismo tiempo que un lenguaje de programación, lo cual lo hace muy versátil. El 28 de marzo de 2012 se encontraban disponibles más de 3700 paquetes para R, que cubren multitud de campos, por citar algunos: estadística bayesiana, aplicaciones financieras, dibujado de mapas, series cronológicas, wavelets, análisis de datos espaciales, localización, computación en clúster, generación automática de informes, integración con servidores web, etc.

R es software libre y funciona bajo múltiples plataformas y sistemas operativos. En particular, funciona bajo una amplia variedad de UNIX, Windows y MacOs. Se puede obtener gratuitamente desde The R Project for Statistical Computing.

Para facilitar a los usuarios la instalación de R, se ha elaborado el Paquete R-UCA, que instala en un sólo paso todo lo necesario para el curso. Así mismo está disponible el Foro de discusión y soporte para usuarios de R. Ambas iniciativas se integran dentro del Proyecto R-UCA.

Ventajas de usar R-UCA

- Se instala en un único paso R, R-Commander y los otros paquetes recomendados
- Permite instalar R en un ordenador sin conexión a internet
- Se configura R para que inicie automáticamente R-Commander al iniciar R
- Permite comprobar fácilmente la existencia de nuevas versiones
- En caso de desinstalación se borran todos los ficheros

Inicio de R

En primer lugar será necesario iniciar R. Si utiliza Windows o MacOS pulse el icono de R que tendrá en su menú de aplicaciones y si utiliza un sistema tipo linux abra una consola y teclee: 'R' o bien 'R -g Tk'.

1.3 R-Commander

R-Commander es una interface gráfica para R desarrollada por John Fox.
La página web de R-Commander es
http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/.
Algunas ventajas de R-Commander son:

- Es sencilla de usar
- Está disponible en español
- Permite el acceso a las funciones y gráficos estadísticos más comunes
- Facilita el aprendizaje de R y la realización de tareas más complejas
- Es multisistema y multiplataforma al estar basada en la librería Tcl/Tk
- Es fácilmente extensible y personalizable

¿Cómo iniciar R-Commander?

La forma de iniciar de R-Commander dependerá del sistema operativo en uso. En general, se iniciará primero R y después se procederá a cargar el paquete Rcmdr.

El paquete R-UCA para Windows configura R para que cargue automáticamente Rcmdr al iniciarse R por lo que es suficiente con pulsar sobre el icono de R.

En otros sistemas operativos, como por ejemplo ubuntu, se añade al menú un icono para la ejecución directa de R-Commander.

Si al iniciar R no se inicia automáticamente R-Commander. Puede iniciarlo introduciendo la instrucción: 'library(Rcmdr)' en la consola de R.

Rcmdr

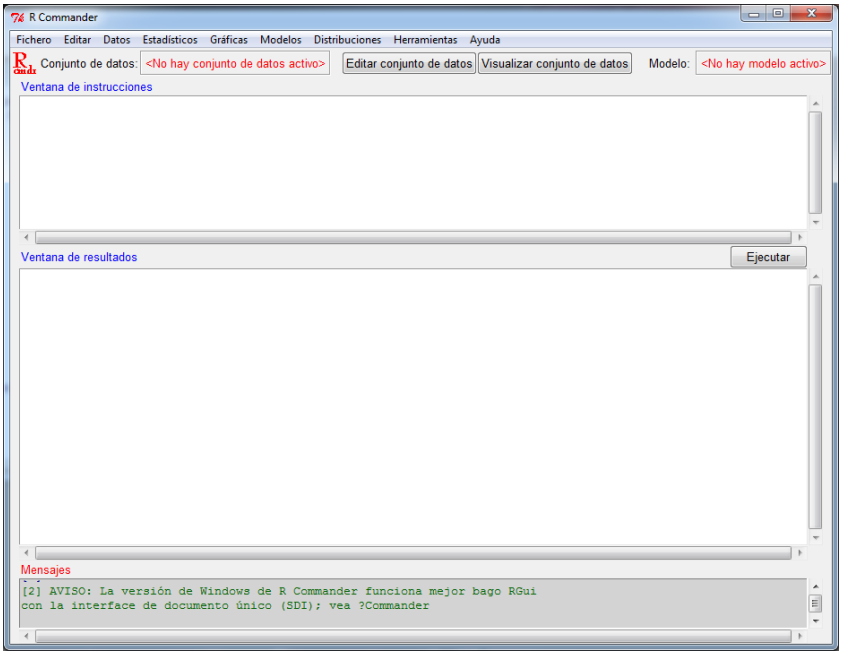
Otra alternativa es:

Seleccionar en el menú Paquetes → Cargar paquete... *y cargar el paquete 'Rcmdr'*

Nota

Si se ha cerrado R-Commander sin cerrar R, se puede reiniciar R-Commander introduciendo la instrucción 'Commander()' en la consola de R.

1.3.1 Partes de la Ventana de R-Commander

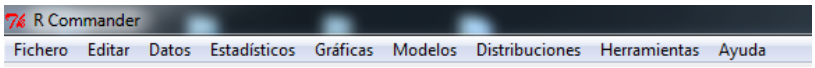


La ventana de R-Commander se divide de arriba a abajo en 5 partes:

1. Menú de R-Commander
2. Barra de Herramientas
3. Ventana de Instrucciones
4. Ventana de Resultados
5. Ventana de Mensajes

A continuación se hará una descripción de cada una de estas partes.

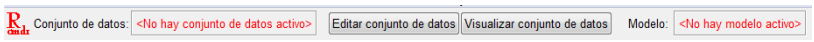
- La primera franja horizontal contiene el menú de la interfaz, pulsando sobre las diferentes opciones se despliegan los correspondientes menús.



Algunas de las opciones del menú son:

- Fichero: Permite guardar las instrucciones y los resultados de una sesión de trabajo. Además permite terminar la aplicación.
- Editar: Contiene las opciones habituales relacionadas con la edición: 'Cortar', 'Copiar', 'Pegar', 'Borrar', 'Buscar...', 'Seleccionar todo', 'Deshacer', 'Rehacer', 'Limpiar ventana'.
- Datos: Mediante las opciones de este menú se pueden cargar, editar y guardar datos. Además se puede acceder a los datos de ejemplo que vienen con R. Otras opciones de este mismo menú permiten operaciones con los datos, como por ejemplo, recodificación, tipificación, construcción de nuevas variables, ...
- Herramientas: Mediante este menú se pueden cargar paquetes, añadidos para R-Commander y se puede configurar la interfaz.
- Ayuda: Muestra ayuda sobre R-Commander.

- La segunda franja horizontal contiene la barra de herramientas. Dicha barra muestra información sobre los datos y el modelo en uso.

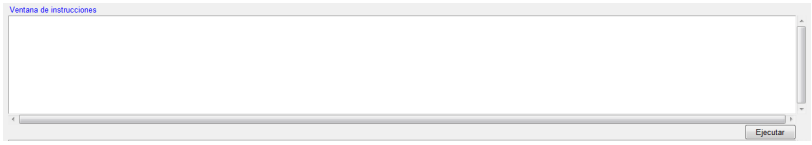


De izquierda a derecha se encuentran cuatro botones:

- Conjunto de datos: Este botón muestra el nombre del conjunto de datos activo o la leyenda 'No hay conjunto de datos activo' cuando todavía no se ha cargado o creado ningún conjunto de datos. Pulsando sobre este botón, se despliega un menú que permite activar otro conjunto de datos, entre los previamente cargados o creados.
- Editar conjunto de datos: Permite la edición del conjunto de datos activo en un entorno similar al de una hoja de cálculo. Durante la edición de los datos no es posible realizar ninguna otra operación con R. Por ello, es absolutamente imprescindible cerrar la ventana de edición de datos antes de intentar cualquier otra operación.

- Visualizar conjunto de datos: Muestra una ventana con los datos del conjunto de datos activo en formato similar al anterior. Esta ventana no permite la modificación de los datos, pero puede mantenerse abierta mientras se continúa haciendo operaciones.
- Modelo: La leyenda muestra el nombre del modelo activo o la leyenda 'No hay modelo activo' cuando no se ha construido ningún modelo previamente. La pulsación sobre dicho botón permite la selección del modelo en uso de entre los disponibles.

- En esta ventana se introducen las instrucciones de R para su evaluación. R-Commander funciona introduciendo en esta ventana las instrucciones necesarias para realizar los cálculos o gráficos correspondientes a las opciones seleccionadas en los menús. Esta ventana permite la modificación y la ejecución de código, tanto del introducido manualmente como del introducido por R-Commander.



- Es la ventana donde se copian las instrucciones ejecutadas seguidas de los resultados producidos. Esta ventana permite la modificación de su contenido pero no permite la ejecución de código. Las instrucciones se muestran en rojo y precedidas del símbolo '>'. Las salidas se muestran en azul.

- En esta ventana se muestran mensajes de información referidos a las instrucciones evaluadas.

El significado del mensaje se refuerza con un código de color

- Rojo (Error): Se ha producido un error en la evaluación de la expresión, por lo que no se obtiene ningún resultado. El mensaje informa del motivo del error.
- Verde (Aviso): La expresión ha sido evaluada, si bien el resultado podría no ser el esperado. El mensaje muestra información detallada del motivo del aviso.
- Azul (Información): Muestra información de carácter general.

1.3.2 Manipulación básica de datos

R viene acompañado de múltiples colecciones de datos preparados para su uso. Cada colección de datos en R se denomina 'conjunto de datos'. Antes de usar un conjunto de datos es necesario cargarlo. Después de iniciar R y R-Commander, observa la leyenda 'Conjunto de datos: <No hay conjunto de datos activo>' en la barra de herramientas. Esto indica que todavía no se ha cargado ningún conjunto de datos.

Rcmdr

Para cargar el conjunto de datos de ejemplo 'Chile', se procede de la siguiente forma:

1. *En el menú de R-Commander se seleccionan las opciones*
   Datos→ Conjuntos de datos en paquetes→ Leer conjunto de datos desde paquete adjunto
2. *En la parte izquierda del cuadro de diálogo selecciona 'car', haciendo una pulsación doble sobre 'car'.*
3. *En la parte derecha busca 'Chile' y haz una pulsación doble sobre él.*
4. *Pulsa el botón 'Aceptar'.*

Se puede observar que en la leyenda de la barra de herramientas aparece 'Chile' indicando que el conjunto de datos activos es 'Chile'.

Repite el proceso anterior y carga el fichero de datos de ejemplo "Baumann" del paquete "car". En este caso la leyenda ha cambiado a "Baumann" indicando que el nuevo conjunto de datos activos es él.

Activación de Conjunto de Datos

R-Commander puede cargar varios conjuntos de datos, pero opera únicamente sobre el conjunto de datos activo. Los conjuntos de datos cargados pueden activarse nuevamente.

Rcmdr

Para activar el conjunto de datos "Chile", se procede:

1. *En la barra de herramientas se pulsa sobre "Baumann"*
2. *En el cuadro se elige 'Chile' y se pulsa sobre "Aceptar"*

En la barra de herramientas aparece "Chile" porque es el que se ha activado nuevamente.

Visualización de Datos

Para visualizar el conjunto de datos activo, en la barra de herramientas se pulsa sobre el botón 'Visualizar conjunto de datos'. Los datos se muestran en formato tabla en una nueva ventana.

Los datos aparecen en una nueva ventana fuera de la ventana de R-Commander. Dicha ventana queda a veces oculta por debajo de otra ventana. En caso necesario, minimiza las otras ventanas para acceder a la ventana de datos.

La ventana de visualización de datos no permite la modificación de éstos, pero puede mantenerse abierta mientras se continúa haciendo operaciones.

Edición de Datos

Para editar el conjunto de datos activo, en la barra de herramientas se pulsa sobre el botón 'Editar conjunto de datos'. Los datos se muestran en formato tabla en una nueva ventana.

El editor de datos es muy simple y permite pocas operaciones, como contrapartida, R es capaz de importar datos desde un gran número de formatos, incluidas direcciones de internet.

Para cerrar la ventana de edición pulse sobre el botón 'Quit' o sobre el cuadro de cierre de dicha ventana. Más adelante, se volverá a tratar con más detalle la edición de datos.

Los datos aparecen en una nueva ventana fuera de la ventana de R-Commander. Dicha ventana queda a veces oculta por debajo de otras ventana. En caso necesario, minimiza las otras ventanas para acceder a la ventana de datos.

Nota muy importante

La ventana de edición de datos no puede mantenerse abierta mientras se continúa haciendo operaciones. R parece bloquearse, queda en espera, hasta que se cierre la ventana de edición de datos.

Guardado de Datos

Los cambios que se realizan afectan a los datos en memoria. Para que los cambios sean permanentes el conjunto de datos debe guardarse.

Para guardar el conjunto de datos activo en formato texto, en el menú de R-Commander, seleccionan las opciones Datos → Conjunto de datos activo → Exportar el conjunto de datos activo , y se marcan las casillas correspondientes al formato de salida deseado.

Si se desean guardar los datos en el formato nativo de R, se seleccionan las opciones Datos → Conjunto de datos activo → Guardar el conjunto de datos activo

La extensión .rda es la que usa R para los ficheros de datos y la extensión R se utiliza para los ficheros de instrucciones

1.4 Los datos: organización, carga y edición

1.4.1 Organización de los datos

La ingente cantidad de información que se genera, a veces de forma espontánea y otras veces intencionada, en nuestros distintos ámbitos de actuación, hace obligado el uso de distintos recursos y herramientas que nos faciliten su organización para posteriormente proceder al análisis.

Independientemente de como se obtenga la información, son dos los elementos centrales del proceso: los objetos estudiados y las características que se estudian de los mismos. Los objetos pueden ser personas, animales, plantas, cosas, etc., la naturaleza de las características tiene mucho que ver con su forma de obtención. De forma general nos referiremos a los objetos como individuos y a las características como variables.

Como acabamos de ver, la información tiene dos dimensiones, individuos y variables, una forma eficiente de organizar la información es colocar los individuos en filas y las variables en columnas. La estructura resultante es una tabla de doble entrada o matriz de datos. Las aplicaciones informáticas de bases de datos, hojas de cálculo y los editores de datos de los paquetes estadísticos reproducen este esquema.

1.4.2 Carga, importación y exportación

Rcmdr viene acompañado de múltiples colecciones de datos preparados para su uso, muchos de ellos históricos.

Cuando se hace una instalación básica, Rcmdr instala los paquetes 'car' y 'datasets', además algunos paquetes de R incorporan a su vez otros paquetes de datos. Cada colección de datos en R se denomina 'conjunto de datos'. Antes de usar un conjunto de datos es necesario cargarlo.

Estos ficheros empaquetados tienen la extensión de los ficheros de datos de R, que sabemos que es 'rda'.

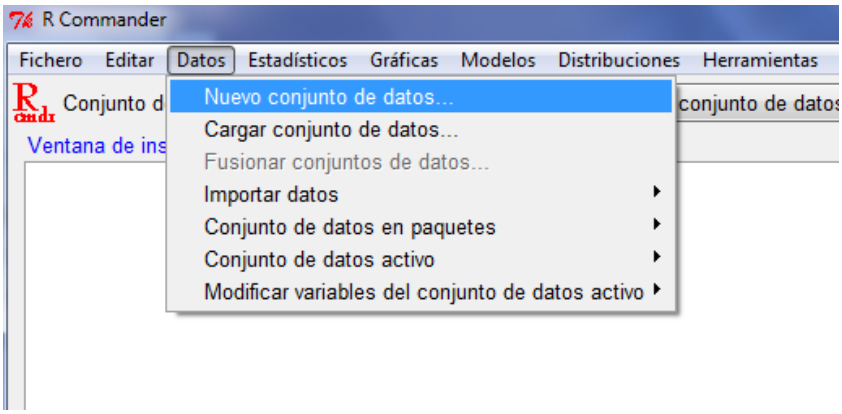
Rcmdr

Si quisiéramos cargar un fichero de datos 'rda' desde un directorio de nuestro ordenador bastaría seleccionar: Datos → Cargar conjunto de datos

y en el navegador de archivos que se nos abrirá solo aparecerán los ficheros con extensión 'rda'.

Cuando queramos cargar un fichero de datos que tenga una extensión distinta de 'rda', debemos hacer una importación. R permite importar ficheros de casi cualquier formato, por ejemplo, 'dat', 'txt', 'xls', 'sav', ...

Por último, un fichero con formato 'rda' es exportable mediante la secuencia: Datos → Conjunto de datos activo → Exportar el conjunto de datos activo



En realidad, cuando se cargan o importan datos, R crea un objeto de un tipo denominado 'data.frame' (marco de datos) con una estructura que contiene información sobre la matriz de datos que incluye el número de individuos (filas), el número de variables (columnas) y los nombres de las variables.

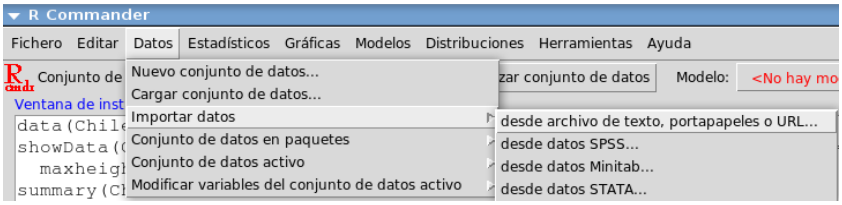
Nota

Un data.frame es un fichero de datos de R, tiene estructura de matriz y no puede contener celdas vacías. Si falta algún dato hay que poner en la celda correspondiente el código 'NA', de no disponible ('Not Available' en inglés).

Importar datos

R puede importar datos desde muy diversos formatos: bases de datos, hojas de cáculo, otros programas estadísticos, etc.,

Para lograrlo basta con usar la secuencia de opciones: Datos → Importar datos y pinchar el origen de los datos a importar.



Una de las operaciones más habituales es importar desde una hoja de cálculo o un archivo de texto.

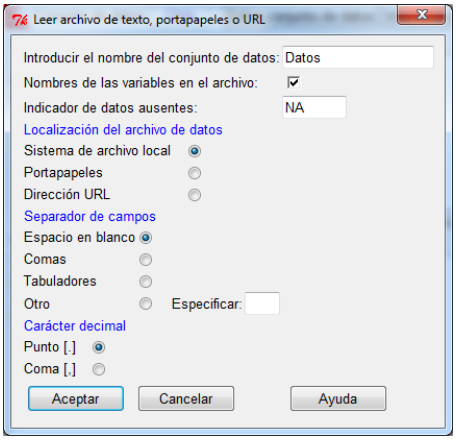
En estos casos, lo más fácil es:

Rcmdr

Marcar el conjunto de datos y copiarlo al portapapeles, después, marcamos la opción del menú:

... → desde archivo de texto → portapapeles o URL

y seleccionamos las opciones que interesen: el nombre del fichero, si los nombres de las variables están en el mismo, el orígen de los datos, el separador de campos y el decimal, como se puede ver en la captura de pantalla.

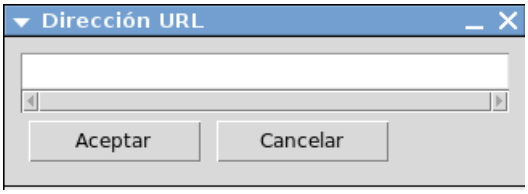


Si importa desde el portapapeles el fichero cebada que se encuentra en:
http://knuth.uca.es/repos/ebrcmdr/bases_datos/cebada.dat,
podrá observar que hay cuatro tipos de cebada.

Importar desde una URL

Otra de las posibilidades que ofrece Rcmdr es la de importar datos directamente desde una URL. Ya se ha visto en el ítem anterior como la ventana de importación nos daba la posibilidad de especificar, dentro de 'Localización del archivo de datos', una dirección URL.

Al hacerlo emerge la siguiente ventana en la que habría que escribir la dirección.



- Importa el conjunto de datos 'caracoles.dat' desde

  http://knuth.uca.es/repos/ebrcmdr/bases_datos/caracoles.dat

  utilizando para ello la opción de importar desde una URL.

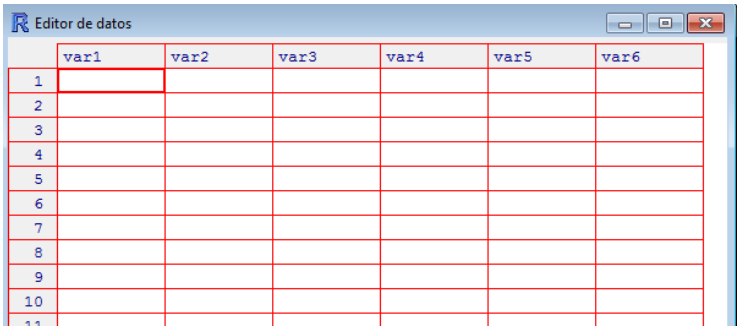
Creación de un fichero con el editor de datos de R

En ocasiones se necesita crear un pequeño fichero con pocas variables e individuos.

En este caso se podría usar el editor de Rcmdr mediante la secuencia de opciones:

Rcmdr

Datos → *Nuevo conjunto de datos* , *lo que nos dará acceso al editor:*



Como puede observarse, el editor no está muy depurado y solo debería usarse cuando sean pocos los datos del fichero.

1.5 Actividades propuestas

Ejercicios

1. Introducir los siguientes datos utilizando como editor una hoja de cálculo:

   Para realizar un estudio sobre la contaminación en la zona costera andaluza se ha dividido la costa andaluza en tres zonas que corresponden al Atlántico (A), Estrecho (E) y Mediterráneo (M) a dos profundidades A (10 metros) y B (20 metros)

Calcio	Cloro	Zona	Profundidad
0,245	5,75	A	A
0,174	7,55	A	B
0,275	10,25	A	A
0,164	6,3	A	B
0,21	7,5	A	A
0,143	4,05	A	B
0,203	5,3	E	A
0,104	5,15	E	B
0,0125	2,1	E	A
0,21	8,5	M	B
0,025	3,2	M	A
0,275	9,3	M	B
0,05	5,1	M	A
0,153	6,2	M	B

2. Utilizando el portapapeles del sistema operativo importa los datos en R-Commander, al nuevo fichero de datos lo llamaremos 'contaminacion'. Observa en la hoja de cálculo el separador decimal para completar correctamente la ventana de diálogo.
3. Se ha medido el peso de 40 personas que pertenecen a dos ciudades distintas A y B teniendo en cuenta su sexo (H-M). ¿Cómo se organiza esta información?
4. A 10 pacientes con altos niveles de colesterol se le suministra un fármaco con el objetivo de reducirlo. Para estudiar su efecto se realizan dos análisis de sangre, uno antes de tomar el fármaco y otro 8 días después. ¿Cómo se organizan los datos?

Parte II

Estadística Descriptiva

2

Análisis Unidimensional

Contenidos

En esta práctica se van a plantear y resolver todas las cuestiones referentes al análisis de un atributo y de una variable numérica. Finalmente se propondrán una serie de actividades similares que puedan facilitar al alumno la asimilación de los objetivos planteados.

2.1 Objetivos

- Diferenciar los distintos tipos de variables existentes, para poder analizar cada una de ellas de forma correcta.
- Aplicar adecuadamente las medidas de posición, dispersión y forma.
- Elaborar gráficos de representación adecuados a la distribución de datos.
- Ser capaz de interpretar y extraer conclusiones tanto del análisis numérico, como del análisis gráfico.

2.2 Descripción del conjunto de datos: La Cebada

La cebada es una planta monocotiledónea perteneciente a la familia de las gramíneas. Ésta está representada por dos importantes especies cultivadas: *Hordeum distinchon L.*, que es empleada en la elaboración de la cerveza, y *Hordeum hexastichon L.*, que se utiliza como forraje en la alimentación animal.

Porcentaje de humedad.

El conocimiento y control de la humedad contenida en el grano de cebada tiene importancia para el manejo y conservación de la misma, debido a su diferente comportamiento. Así, bajos niveles hídricos disminuyen los procesos bioquímicos en el grano y permiten su conservación por períodos largos sin pérdida apreciable de su poder biológico; por el contrario, niveles altos activan los procesos biológicos y facilitan la transformación de las reservas del grano. El agua también tiene gran importancia en las operaciones de compra-venta porque altera las características físicas consideradas en esta operación, como son volumen y peso específico del producto.

En la cebada, la humedad máxima permisible para garantizar una buena conservación del grano es de 13.5 por ciento, pero en ocasiones es necesario secar más, ya que en el almacén es posible que la humedad aumente en un 2 por ciento o más, razón por la cual aparecen focos de calentamiento e infestaciones por hongos y bacterias que se alimentan del grano.

En el laboratorio se ha procedido a evaluar, entre otros aspectos, la humedad de 24 muestras de grano de las dos variedades de cebada a través del **Método de la estufa**. En éste, para determinar la humedad de los granos, se somete una muestra de peso conocido al secado y se calcula el porcentaje de humedad a través del peso que se pierde durante el secado. Para obtener el porcentaje de humedad se divide la pérdida de peso de la muestra entre el peso original de ella y el resultado se multiplica por 100. Los resultados obtenidos han sido los siguientes:

Cebada

	Color_cascara	Tamaño	Peso_gr	Longitud_sin_borba	Variedad	Fibra	Almidón	Proteinas	Humedad	Longitud_factor
1	amarillo claro	Mediano	30	7	Hordeum distichon	3.1	55.4	10.1	6.1	7
2	amarillo claro	Mediano	32	8	Hordeum distichon	3.2	48.6	11.1	5.8	8
3	amarillo claro	Pequeño	43	6	Hordeum distichon	3.0	49.0	10.9	6.2	6
4	amarillo claro	Mediano	45	7	Hordeum distichon	4.0	55.3	10.2	5.9	7
5	amarillo claro	Grande	51	8	Hordeum distichon	3.6	56.9	9.8	4.9	8
6	amarillo pálido	Grande	55	8	Hordeum distichon	4.0	48.7	11.0	5.5	8
7	amarillo pálido	Grande	58	9	Hordeum distichon	4.3	48.9	11.2	6.0	9
8	amarillo pálido	Grande	60	9	Hordeum distichon	4.4	53.0	10.5	5.8	9
9	amarillo pálido	Pequeño	43	6	Hordeum distichon	4.1	53.3	10.4	4.0	6
10	amarillo pálido	Pequeño	41	6	Hordeum distichon	4.0	53.2	10.4	4.7	6
11	amarillo pálido	Pequeño	42	6	Hordeum distichon	3.9	53.1	10.6	5.3	6
12	crema claro	Mediano	52	7	Hordeum distichon	3.9	48.5	11.1	6.1	7
13	crema claro	Mediano	55	7	Hordeum distichon	3.8	59.9	9.5	5.5	7
14	crema claro	Mediano	41	8	Hordeum distichon	2.9	55.1	10.2	5.1	8
15	crema claro	Grande	45	9	Hordeum distichon	3.7	55.1	10.3	4.3	9
16	crema pálido	Pequeño	33	6	Hordeum hexastichon	5.1	46.0	12.9	5.7	6
17	crema pálido	Pequeño	31	7	Hordeum hexastichon	6.0	43.0	13.4	6.2	7
18	crema pálido	Mediano	48	7	Hordeum hexastichon	4.2	43.2	13.3	3.0	7
19	azul verdoso	Grande	60	8	Hordeum hexastichon	7.9	42.5	14.1	8.9	8
20	azul verdoso	Grande	61	8	Hordeum hexastichon	3.5	42.3	14.0	7.1	8
21	azul verdoso	Grande	58	9	Hordeum hexastichon	6.6	40.8	16.3	6.5	9
22	azul	Pequeño	29	7	Hordeum hexastichon	7.0	42.0	15.8	3.2	7
23	azul	Pequeño	41	6	Hordeum hexastichon	7.6	41.0	16.3	4.6	6
24	azul	Pequeño	35	6	Hordeum hexastichon	6.8	40.0	16.2	5.5	6

Para introducir el conjunto de datos *cebada.txt* en Rcmdr, se puede usar la opción del menú Datos → Importar datos → desde archivo de texto portapapeles o URL... , marcando la url: http://knuth.uca.es/repos/ebrcmdr/bases_datos/cebada.txt

2.3 Análisis descriptivo de un factor o variable cualitativa

Vamos a comenzar distinguiendo entre *factor no ordenado* y *factor ordenado*. Tanto el "Color_cascara" como el "Tamaño" son atributos, pero entre los valores de tamaño es posible establecer un orden, mientras que en el color de cáscara esto no es posible. Esta diferencia hace que la forma de analizar ambas sea también distinta.

2.3.1 Factor no ordenado

Comencemos analizando el factor "Color_cascara".

Para resumir la información proporcionada por los datos vamos a calcular la *moda*, a través de la tabla de frecuencias.

Rcmdr

Seleccionamos para ello Estadísticos → Resúmenes → Distribución de frecuencias...

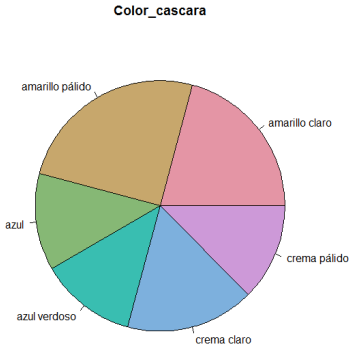
```
> .Table  # counts for Color_cascara

 amarillo claro amarillo pálido            azul    azul verdoso     crema claro     crema pálido
              5               6               3               3               4               3

> round(100*.Table/sum(.Table), 2)  # percentages for Color_cascara

 amarillo claro amarillo pálido            azul    azul verdoso     crema claro     crema pálido
          20.83           25.00           12.50           12.50           16.67           12.50
```

Vemos así que, el valor que más se repite es el color de cáscara “amarillo pálido”, que se repite en la muestra 6 veces, esto es, en el 25 % de los casos.

Para resumir gráficamente la información proporcionada por este factor, se realiza un diagrama de sectores, ya que en este gráfico no se establece ningún orden entre los distintos valores.

Rcmdr

Se accede a esta opción mediante la ruta *Gráficas* → *Gráfica de sectores...*



2.3.2 Factor ordenado

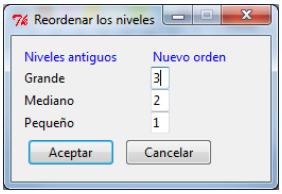
Al tratarse también en este caso de una variable cualitativa la única medida que vamos a calcular es la moda, a través de la distribución de frecuencias. En este caso

resulta ser el *Tamaño=Pequeño* el más frecuente.

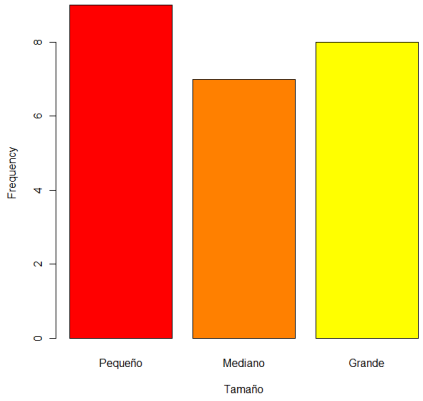
A la hora de analizar gráficamente el tamaño del grano se puede establecer un orden entre sus distintos valores, por lo que éstos podrían ser codificados como *Pequeño*$=1$, *Mediano*$=2$ y *Grande*$=3$. El gráfico recomendado en este caso sería el diagrama de barras, al existir un orden definido entre los elementos. Si realizamos este gráfico, sin codificar los datos previamente, RCommander ordena los distintos valores del atributo, en orden alfabético. Por ello, para que el gráfico aparezca en el orden correcto, debemos reordenar previamente los niveles del factor.

Rcmdr

Esta opción aparece en Datos→Modificar variables del conjunto de datos activos→Reordenar niveles de factor...



Ahora ya es posible realizar el gráfico de barras en el orden lógico. Seleccionamos para ello Gráficas→Gráfica de barras... *y, si además retocamos la instrucción añadiendo la opción col=heat.colors(3), se obtiene el siguiente gráfico:*

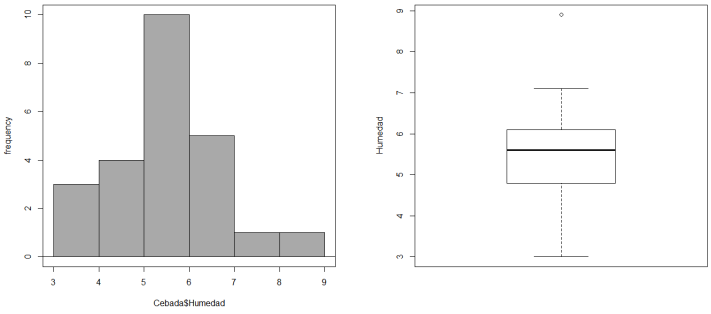


2.4 Análisis descriptivo de una variable

2.4.1 Variable continua

Se estudiará ahora el tratamiento de una variable continua. La variable continua que vamos a analizar es la *Humedad* ya que es esta la más importante en nuestro conjunto de datos. No podemos olvidar la importancia del grado de humedad en la conservación del grano de cebada en el supuesto planteado.

Para hacernos una idea de como se distribuyen los datos vamos a comenzar con el análisis gráfico de la variable. En este caso realizamos el histograma y el diagrama de caja y bigotes. Ambos están disponibles en el menú de *Gráficas*.



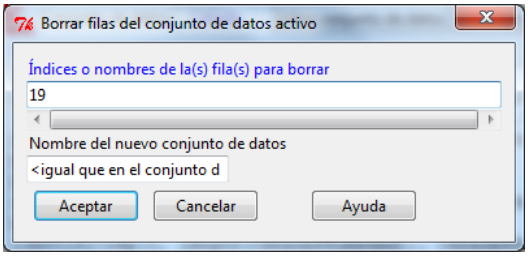
a) Histograma *b*) Diagrama de caja

Figura 2.1: Humedad

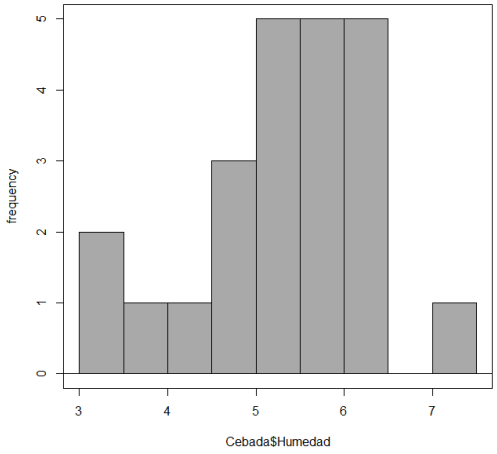
Vemos así que, en el gráfico de caja y bigotes, existe un valor atípico. Si volvemos a realizar el gráfico marcando la pestaña *Identificar atípicos con el ratón* y señalamos con el botón izquierdo del ratón el valor atípico en el gráfico, se puede ver que éste se encuentra en la fila 19 del conjunto de datos.

Vamos entonces a eliminar el valor atípico antes de comenzar el análisis.

Para ello en la opción del menú Datos → Conjunto de datos activos → Borrar fila(s) del conjunto de datos activo... *señalamos que es la fila 19 la que queremos eliminar.*



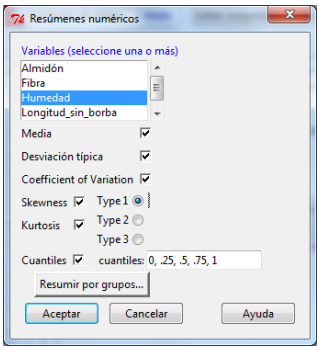
Ahora, repetimos el histograma, al haber modificado el conjunto de datos y se obtiene el siguiente grafico:



A la vista del gráfico la variable parece ser asimétrica negativa y algo leptocúrtica. Estos aspectos van a analizarse a continuación mediante el resumen numérico.

Rcmdr

Comenzamos con el análisis numérico en el que las medidas básicas se pueden calcular desde el menú de RCommander en Estadísticos → Resúmenes → Resúmenes numéricos *. Los parámetros de forma que se utilizan normalmente son están disponibles en la opción tipo 1.*




```
> numSummary(Cebada[,"Humedad"], statistics=c("mean", "sd", "quantiles", "cv",
+   "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="1")
     mean       sd        cv   skewness  kurtosis 0% 25% 50%  75% 100%  n
 5.347826 1.013486 0.1895137 -0.7840509 0.1511948  3 4.8 5.5 6.05  7.1 23
```

Con estos resultados podemos concluir que todas las muestras analizadas están por debajo de 13.5, que es el grado máximo de humedad permisible, y que la humedad media, que es 5.347826, representa más o menos bien al conjunto de datos, al ser $0{,}1 < CV < 0{,}5$. En lo referente a la forma de los datos se tiene que la distribución es asimetrica negativa y leptocúrtica al ser $skewness < 0$ y $kurtosis > 0$.

Por otro lado, sabemos que, en este conjunto de datos se encuentran mediciones de dos variedades de cebada que son utiizadas con dos finalidades muy diferenciadas, por lo que podría ser lógico pensar que entre ellas existirían diferencias significativas. Vamos por ello a analizar en nuestra muestra el grado de humedad, según la variedad de cebada.

Ésta opción se puede realizar a través de Estadísticos → Resúmenes → Resúmenes numéricos... marcando en la ventana emergente la opción *Resumir por grupos* y siendo la variable grupo la *Variedad*.

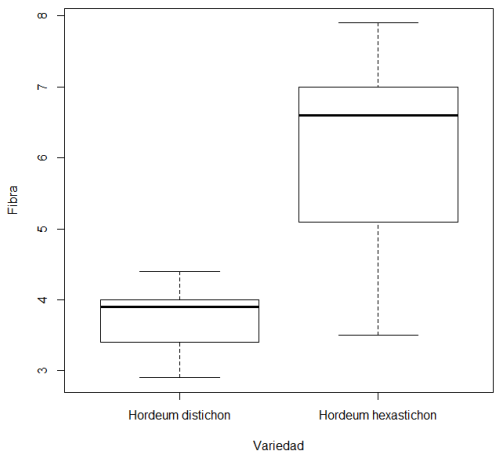
```
> numSummary(Cebada[,"Humedad"], groups=Cebada$Variedad, statistics=c("mean", "sd", "quantiles", "cv",
+   "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="1")
                     mean        sd        cv   skewness   kurtosis 0%  25% 50%   75% 100% data:n
Hordeum distichon   5.413333 0.6854265 0.1266182 -0.7267119 -0.6035435  4 5.00 5.5 5.950  6.2     15
Hordeum hexastichon 5.225000 1.5040422 0.2878550 -0.4418537 -1.1674297  3 4.25 5.6 6.275  7.1      8
```

A la vista de los resultados no parece que las diferencias entre los dos tipos sean demasiado grandes, en lo que al grado de humedad se refiere. Sin embargo, si realizamos un resumen por grupos de la cantidad de fibra por grano se obtienen los siguientes resultados:

```
> numSummary(Cebada[,"Fibra"], groups=Cebada$Variedad, statistics=c("mean", "sd",
+   "quantiles", "cv", "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1),
+   type="1")
                     mean        sd        cv   skewness   kurtosis  0% 25% 50% 75% 100% data:n
Hordeum distichon   3.726667 0.4712698 0.1264588 -0.5121549 -0.9382464 2.9 3.4 3.9   4  4.4     15
Hordeum hexastichon 6.077778 1.5172160 0.2496333 -0.5240288 -1.0105045 3.5 5.1 6.6   7  7.9      9
```

Así, el grano de la variedad *Hordeum hexastichon L.*, que es utilizado como forraje en la alimentación animal, tiene una cantidad media de fibra claramente superior a la contenida en la variedad usada para la elaboración de cerveza, siendo además esta última más homogenea.

Estas diferencias pueden observarse también a través de *diagrama de cajas*, usando la opción *Gráfica por grupos...*.

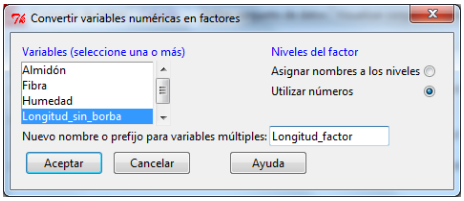


2.4.2 Variable discreta

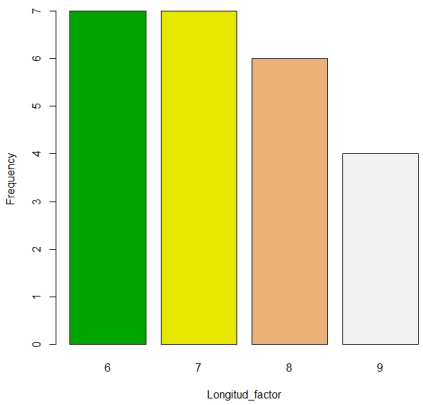
Aunque la variable *Longitud_sin_borba* es de tipo continua, ésta presenta muy pocos valores distintos y por ello vamos a realizar su análisis como si de una variable discreta se tratase. Así, al realizar el análisis gráfico de la variable es recomendable el uso del diagrama de barras. Sin embargo, en RCommander este gráfico sólo se puede realizar para variables de tipo factor, por lo que, un paso previo obligado es la conversión de la variable numérica en una de tipo factor.

Rcmdr

Esta opción está disponible en Datos → *Modificar variables del conjunto de datos activo* → *Convertir variable numérica en factor...*



Vamos a utilizar como valores del factor los mismos números y es importante, en este caso, almacenar el factor con un nombre distinto al de la variable original, ya que de lo contrario perderíamos la información proporcionada por la variable numérica. Ahora es ya posible realizar el gráfico de barras y si, como antes, modificamos la orden añadiendo por ejemplo col=terrain.colors(4), se obtiene el siguiente gráfico:



En lo referente al análisis numérico, además de los parámetros básicos que proporciona la opción *Resumen numérico* de la variable *Longitud_sin_borba*, es posible calcular también la moda a través de la *distribución de frecuencias*. En este caso, como ha ocurrido con el gráfico de barras, debemos hacer uso de la variable factor *Longitud_factor*.

```
> .Table <- table(Cebada$Longitud_factor)

> .Table  # counts for Longitud_factor

6 7 8 9
7 7 6 4

> round(100*.Table/sum(.Table), 2)  # percentages for Longitud_factor

    6     7     8     9
29.17 29.17 25.00 16.67
```

Se puede ver que la variable es bimodal y que estos valores de longitud, *6 y 7*, se repiten el *29.17* % de las veces.

Ejercicios

1. Analiza, en el conjunto de datos cebada.txt, las variables *peso* y *longitud_sinborba*.
2. ¿Existen diferencias entre las dos variedades de cebada en *nivel de almidón*?
3. El conjunto de datos *quakes* del paquete adjunto *datasets* contiene información sobre 1000 eventos sísmicos de $MB > 4{,}0$ ocurridos en los alrededores de las islas Fiyi desde 1964.

   Analiza en este conjunto las variables *mag*, que mide la magnitud del temblor en la escala de *Richter* y la variable *stations* que refleja el número de estaciones sísmicas en las que se sintió el temblor.

3

Ajuste y Regresión

Contenidos

En esta práctica se va a continuar el análisis del conjunto de datos *Cebada*. En concreto, se va a analizar la relación existente entre las variables *Almidón* y *Proteinas*.

3.1 Objetivos

- Evaluar la relación existente entre las variables del conjunto de datos.
- Encontrar el modelo que mejor explique la relación existente entre cada par de variables.
- Evaluar la bondad del modelo.
- Hacer uso del modelo establecido para realizar predicciones.
- Evaluar la bondad de las predicciones realizadas.

3.2 Descripción de las variables utilizadas

Como ya se dijo, la cebada es una planta monocotiledónea perteneciente a la familia de las gramíneas. Ésta está representada por dos importantes especies cultivadas: *Hordeum distinchon L.*, que es empleada en la elaboración de la cerveza, y *Hordeum hexastichon L.*, que se utiliza como forraje en la alimentación animal. En esta práctica se va a analizar la relación entre las variables *Almidón* y *Proteinas*, que vienen medidas en porcentajes.

Cebada

	Color_cascara	Tamaño	Peso_gr	Longitud_sin_borba	Variedad	Fibra	Almidón	Proteinas	Humedad	Longitud_factor
1	amarillo claro	Mediano	30	7	Hordeum distichon	3.1	55.4	10.1	6.1	7
2	amarillo claro	Mediano	32	8	Hordeum distichon	3.2	48.6	11.1	5.8	8
3	amarillo claro	Pequeño	43	6	Hordeum distichon	3.0	49.0	10.9	6.2	6
4	amarillo claro	Mediano	45	7	Hordeum distichon	4.0	55.3	10.2	5.9	7
5	amarillo claro	Grande	51	8	Hordeum distichon	3.6	56.9	9.8	4.9	8
6	amarillo pálido	Grande	55	8	Hordeum distichon	4.0	48.7	11.0	5.5	8
7	amarillo pálido	Grande	58	9	Hordeum distichon	4.3	48.9	11.2	6.0	9
8	amarillo pálido	Grande	60	9	Hordeum distichon	4.4	53.0	10.5	5.8	9
9	amarillo pálido	Pequeño	43	6	Hordeum distichon	4.1	53.3	10.4	4.0	6
10	amarillo pálido	Pequeño	41	6	Hordeum distichon	4.0	53.2	10.4	4.7	6
11	amarillo pálido	Pequeño	42	6	Hordeum distichon	3.9	53.1	10.6	5.3	6
12	crema claro	Mediano	52	7	Hordeum distichon	3.9	48.5	11.1	6.1	7
13	crema claro	Mediano	55	7	Hordeum distichon	3.8	59.9	9.5	5.5	7
14	crema claro	Mediano	41	8	Hordeum distichon	2.9	55.1	10.2	5.1	8
15	crema claro	Grande	45	9	Hordeum distichon	3.7	55.1	10.3	4.3	9
16	crema pálido	Pequeño	33	6	Hordeum hexastichon	5.1	46.0	12.9	5.7	6
17	crema pálido	Pequeño	31	7	Hordeum hexastichon	6.0	43.0	13.4	6.2	7
18	crema pálido	Mediano	48	7	Hordeum hexastichon	4.2	43.2	13.3	3.0	7
19	azul verdoso	Grande	60	8	Hordeum hexastichon	7.9	42.5	14.1	8.9	8
20	azul verdoso	Grande	61	8	Hordeum hexastichon	3.5	42.3	14.0	7.1	8
21	azul verdoso	Grande	58	9	Hordeum hexastichon	6.6	40.8	16.3	6.5	9
22	azul	Pequeño	29	7	Hordeum hexastichon	7.0	42.0	15.8	3.2	7
23	azul	Pequeño	41	6	Hordeum hexastichon	7.6	41.0	16.3	4.6	6
24	azul	Pequeño	35	6	Hordeum hexastichon	6.8	40.0	16.2	5.5	6

Porcentaje proteico.

La cebada es un grano con un contenido de energía similar y a veces hasta superior al del maíz y, aunque tiene algunas deficiencias en minerales, tiene niveles de proteína relativamente altos frente al resto de los granos de cereales. Estas condiciones nutritivas han determinado que se esté usando, en muchos países, en la suplementación de ganado lechero y en engorde de novillos.

Porcentaje de almidón.

La energía de la cebada procede principalmente de su riqueza en hidratos de carbono. Ésta es rica en azúcares o hidratos de carbono complejos, principalmente

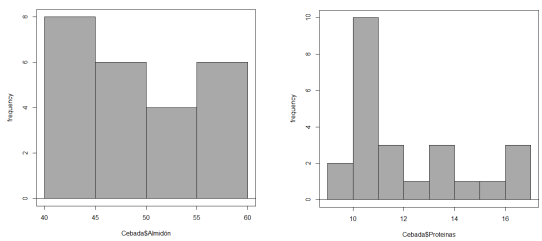
almidón y celulosa y cantidades menores de maltosa y rafinosa.

El almidón del grano de cebada es el principal responsable de que la cebada sea una alimento tan energético, útil para proveer de energía el organismo, tanto en las actividades diarias, como el gasto que el cuerpo experimenta en estado de reposo, lo que se conoce como metabolismo basal.

Además, se ha comprobado que, más interesante que su poder energético, es la forma en que este componente proporciona esta energía: La cebada contiene mucho almidón y el almidón es un hidrato de carbono complejo que precisa que el organismo realice una serie de transformaciones hasta convertirlo en glucosa para que puede ser absorbido y aprovechado.

3.3 Relación entre variables

En primer lugar, se realiza el histograma de las variables *Almidón* y *Proteínas*, con el objetivo de determinar posibles variables multimodales. Como paso previo a éste y, como se hizo en la práctica de *Análisis Unidimensional*, se analiza la existencia de valores atípicos. En este caso no existen en ninguna de las variables implicadas, por lo que se continúa con la realización de los histogramas:

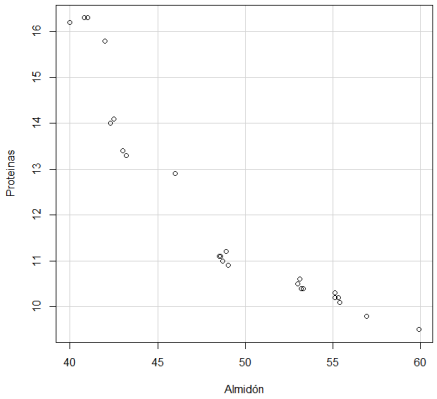


a) Almidón *b*) Proteinas

Figura 3.1: Histogramas

Rcmdr

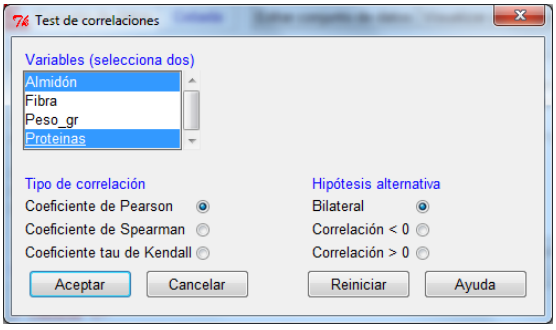
A continuación, al objeto de decidir el tipo de función de ajuste que convendrá utilizar, se representa el diagrama de dispersión, a través de la secuencia *Gráficas* → *Diagrama de dispersión...* *, resultando el gráfico:*



A la vista del gráfico parece existir una fuerte dependencia inversa. Para comprobarlo se va a comenzar calculando el coeficiente de correlación lineal de Pearson.

Rcmdr

Para ello, en el apartado Estadísticos → Resúmenes → Test de correlación , se seleccionan las variables Almidón y Proteinas, resultando ser $r = -0{,}9290365$.



3.4 Ajuste de modelos

Dado que, se conoce la existencia de relación entre las variables *Almidón* y *Proteinas*, se van a analizar a continuación los posibles modelos de ajuste.

- Modelo lineal

  Para realizar el modelo lineal se selecciona Estadísticos → Ajuste de modelos → modelo lineal... , tomando como fórmula del modelo la variable *Proteinas* en función del *Almidón*, esto es, `Proteinas~Almidón`, obteniéndose la siguiente salida:

```
Call:
lm(formula = Proteinas ~ Almidón, data = Cebada)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1547 -0.7698 -0.1373  0.4330  1.4331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.30821    1.47449   19.88 1.52e-15 ***
Almidón     -0.35223    0.02991  -11.78 5.69e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.863 on 22 degrees of freedom
Multiple R-squared: 0.8631, Adjusted R-squared: 0.8569
F-statistic: 138.7 on 1 and 22 DF,  p-value: 5.689e-11
```

  Se puede ver así que, la recta de mínimos cuadrados es:

  $Proteinas = -0{,}35223 * Almidón + 29{,}30821$, con un coeficiente de determinación $R^2 = 0{,}8631$, por lo que el $13{,}69\,\%$ de los datos no queda explicado mediante este modelo.

  A la vista de este resultado y del gráfico de dispersión se va a evaluar la conveniencia de usar un modelo parabólico.

- Modelo parabólico

  Para realizar el modelo parabólico se selecciona en *Rcmdr* Estadísticos → Ajuste de modelos → Modelo lineal... , tomando como fórmula del modelo:

  `Proteinas~Almidón+I(Almidón^2)`

```
Call:
lm(formula = Proteinas ~ Almidón + I(Almidón^2), data = Cebada)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78172 -0.33926  0.03696  0.29957  0.96382

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.571041   8.411121  10.055 1.76e-09 ***
Almidón      -2.642854   0.347248  -7.611 1.81e-07 ***
I(Almidón^2)  0.023393   0.003542   6.605 1.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5035 on 21 degrees of freedom
Multiple R-squared: 0.9555, Adjusted R-squared: 0.9513
F-statistic: 225.5 on 2 and 21 DF,  p-value: 6.399e-15
```

En este caso el coeficiente de determinación $R^2 = 0{,}9555$, por lo que el modelo ha mejorado considerablemente con respecto al modelo lineal, puesto que la variable *Proteínas* explica el $95{,}5\,\%$ de la variable *Almidón*, según el ajuste parabólico estimado.

En este caso la ecuación del modelo es:

$$\textit{Proteinas} = 0{,}023393 * \textit{Almidón}^2 - 2{,}642854 * \textit{Almidón} + 84{,}571041$$

- Modelo exponencial

Para realizar el modelo exponencial se selecciona Estadísticos → Ajuste de modelos → Modelo lineal... , tomando como fórmula del modelo:

`log(Proteinas)~Almidón`

```
Call:
lm(formula = log(Proteinas) ~ Almidón, data = Cebada)

Residuals:
      Min        1Q    Median        3Q       Max
-0.084336 -0.047704 -0.004681  0.031531  0.092201

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.856523   0.099167   38.89  < 2e-16 ***
Almidón     -0.028233   0.002011  -14.04 1.85e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05804 on 22 degrees of freedom
Multiple R-squared: 0.8996, Adjusted R-squared: 0.895
F-statistic:   197 on 1 and 22 DF,  p-value: 1.853e-12
```

Se descarta en este caso el modelo exponencial, ya que a penas mejora el modelo lineal, con un coeficiente de determinación $R^2 = 0{,}8996$. Se podrían probar otros modelos, como el hiperbólico, pero debido a la bondad del modelo parabólico no parece necesario.

3.5 Predicciones

Tras evaluar los distintos modelos y haber seleccionado el modelo parabólico, se va a proceder a la utilización de dicho modelo para realizar predicciones.

1. Se plantea en primer lugar, qué cantidad de proteínas contendría un grano de cebada que contiene un $54\,\%$ de almidón.

   Como en el apartado anterior se ha visto que el mejor modelo es el de ecuación: $\textit{Proteínas} = 0{,}023393 * \textit{Almidón}^2 - 2{,}642854 * \textit{Almidón} + 84{,}571041$, se tendría que el porcentaje de proteínas que contendría el grano de cebada sería de

   $\textit{Proteínas} = 0{,}023393 * \textit{54}^2 - 2{,}642854 * \textit{54} + 84{,}571041 = 10.07091$, esto es, aproximadamente el $10\,\%$ de proteínas.
2. Si ahora se plantea, qué cantidad de proteínas contendría un grano de cebada que contiene un $35\,\%$ de almidón. Se podría utilizar el modelo anterior, pero en este caso el valor obtenido para el porcentaje de proteínas no sería fiable ya que los datos utilizados para estimar el modelo contenían porcentajes de almidón

entre el 40 y el $59{,}9\,\%$, por lo que no se puede estimar la bondad de la predicción para un valor fuera de ese rango.

3. Si en algún momento se plantease qué porcentaje de almidón contendría un grano con un $15\,\%$ de proteínas, sería necesario estimar el modelo que mejor explica la variable *Almidón* en función de la variable *Proteínas*, y en este, sustituir el valor *Proteínas*$= 15\,\%$.

3.6 Actividades propuestas

Ejercicios

1. Estimar el modelo que mejor explica la variable *Almidón* en función de la variable *Proteínas*.
2. ¿Qué porcentaje de almidón contendría un grano con un porcentaje del $15\,\%$ de proteínas?
3. ¿Existe relación entre las variables *Fibra* y *Proteínas*? En caso afirmativo establece el mejor modelo que explique dicha relación.
4. A partir del ejercicio anterior, ¿qué porcentaje de fibra podría contener un grano de cebada con un $15\,\%$ de proteínas? Sería fiable esta estimación.

Parte III

Teoría de Probabilidad

4

Distribuciones

Contenidos

En esta práctica se van a plantear y resolver todas las cuestiones referentes al análisis de variables aleatorias discretas y continuas. Finalmente se propondrán una serie de actividades similares que puedan facilitar al alumno la asimilación de los objetivos planteados.

4.1 Objetivos

- Reconocer las distribuciones de probabilidades más conocidas.
- Determinar cuál es la probabilidad pedida en cada situación y ser capaz de calcularla haciendo uso del software *Rcmdr*.
- Reconocer y saber calcular cuantiles para las distribuciones conocidas a través del programa.

4.2 Distribuciones de probabilidad y descripción del entorno de trabajo

La existencia de fenómenos no determinísticos 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 fenómenos que poseen la característica anterior se denominan aleatorios y, la concreción numérica de estos fenómenos, mediante la asignación de valores con un cierto criterio, da origen a la *variable aleatoria*.

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 muchos modelos de la vida real. El estudio de estos modelos teóricos, incluyendo la caracterización de sus parámetros y el cálculo de probabilidades en sus distintos formatos van a ser objetos de estudio en este capítulo. En concreto, en esta práctica guiada se trabajará con datos y situaciones ambientales y, referentes a las ciencias del mar. Se tratarán modelos de distribuciones discretos y continuos.

4.3 Distribuciones discretas

4.3.1 Distribución binomial

La distribución binomial es la distribución de probabilidad de una variable aleatoria en la que *se mide el número de éxitos obtenidos al realizar un número fijo, n, de pruebas de Bernouilli independientes y con igual probabilidad de éxito.* Una prueba de Bernouilli es un experimento con sólo dos posibles resultados E y F, con probabilidades de ocurrencia p y q, respectivamente, que se mantienen invariantes a lo largo del proceso. Los dos resultados posibles suelen identificarse con el *éxito* y *fracaso* del experimento, verificándose que $p + q = 1$.

Se verá a continuación un ejemplo:

Los técnicos de cierta compañía que se dedican a asegurar superficies arboladas contra fenómenos meteorológicos extraordinarios han estimado que cada año mueren alrededor del $0{,}05\,\%$ de árboles a causa de este tipo de fenómenos. Si la

compañía tiene 10000 árboles asegurados este año, ¿Con qué probabilidad tendrán que pagar por más de 3 árboles?

Solución:

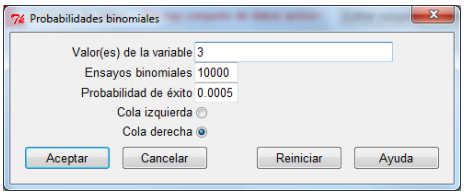
Cada árbol puede morir, o no morir, a causa de un fenómeno meteorológico extraordinario. Así se tienen los dos posibles resultados de una prueba de Bernouilli. Se llamará *E="Que un árbol muera por un fenómeno meteorológico extraordinario"* y *F="Que un árbol no muera por un fenómeno meteorológico extraordinario"*. Además se sabe que, según las estimaciones de los técnicos de dicha compañía la probabilidad de que un árbol muera es de $p(E) = \frac{0{,}05}{100} = 0{,}0005$.

Así la distribución de la variable X =*número de éxitos (árboles muertos por este tipo de fenómenos), entre los n=10000 asegurados por la compañía*, sigue una binomial, $B(n = 10000, p = 0{,}0005)$.

Ahora bien, en el problema se pide, ¿con qué probabilidad tendrán que pagar por más de 3 árboles?. Esto es lo mismo que decir cual es la probabilidad de que mueran más de 3 árboles. Así, la probabilidad pedida es: $P(X > 3)$. Ahora se va a resolver esta cuestión en *Rcmdr*:

Rcmdr

En primer lugar, como la probabilidad pedida es acumulada en el menú del programa seleccionamos Distribuciones→Distribuciones discretas→Distribución binomial→Probabilidades binomiales acumuladas... y, marcándose las opciones que se ven en la siguiente imagen, resulta que $P(X > 3) = 0{,}7350443$.



4.3.2 Distribución geométrica

Una variable aleatoria X definida como el número de fracasos antes de obtener el primer éxito en sucesivas repeticiones de experimentos de Bernouilli, independientes y con iguales probabilidad de éxito se dice que sigue una distribución de probabilidad geométrica de parámetro p, esto es $X \sim Ge(p)$.

Se verá a continuación un ejemplo:

Al inyectar a un delfín una dosis de cierto producto contra el virus "morbillivirus" que ha causado el deceso de los cientos de delfines en la costa norte del Perú, un biólogo marino estimó que la probabilidad de que un delfín se cure con una dosis es de 0,7. ¿Qué probabilidad hay de que un delfín infectado con este virus se cure con menos de 3 dosis del producto?

Solución:

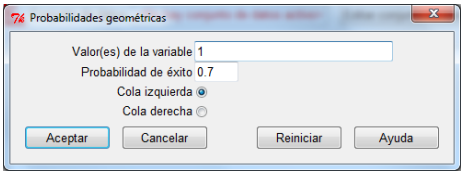
Cada dosis del producto puede curar, o no curar, a un delfín. Así se tienen los dos posibles resultados de una prueba de Bernouilli. Se llamará *E="Que el delfín se cure con la dosis"* y *F="Que no se cure con la dosis"*. Además, según los biólogos la probabilidad de que un delfín se cure con una dosis es de $p(E) = 0{,}7$. Si a cada animal se le aplican dosis hasta que resulta curado, la distribución de la variable X =*número dosis fallidas hasta conseguir la cura* es la de una geométrica, $Ge(p = 0{,}7)$.

Ahora bien, en el problema se pide, ¿qué probabilidad hay de que un delfín infectado con este virus se cure con menos de 3 dosis del producto? Esto es p(Intentos< 3)=p(Dosis fallidas+1 dosis acertada< 3)=p(Dosis fallidas< 2)=p($X < 2$).

Rcmdr

Para calcular en Rcmdr esta probabilidad se debe tener en cuenta que la opción cola izquierda calcula para un valor a, la probabilidad $p(X \leq a)$, así se debe considerar que, al tratarse de una variable discreta $p(X < 2) = p(X \leq 1)$ y tras seleccionar en el menú del programa Distribuciones→Distribuciones discretas→Distribución geométrica→Probabilidades geométricas acumuladas... *y marcar las opciones que se ven en la siguiente imagen, re-*

sulta que $p(X < 2) = 0{,}91$.



4.3.3 Distribución binomial negativa

Una variable aleatoria X definida como el número de fracasos antes de obtener el r−ésimo éxito en sucesivas repeticiones de experimentos de Bernouilli, independientes y con iguales probabilidad de éxito, se dice que sigue una distribución de probabilidad binomial negativa de parámetros r y p, esto es $X \sim BN(r, p)$.

Se verá a continuación un ejemplo:

Si de cierta población de animales se sabe que el $70\,\%$ son machos y se quieren extraer 3 hembras de esta especie, ¿qué probabilidad hay de tener que extraer al menos 7 animales de la población para conseguir las 3 hembras buscadas?

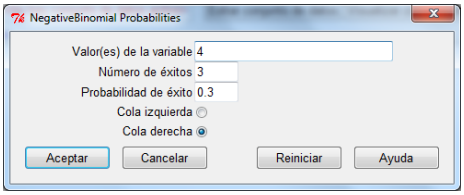
Solución:

En cada extracción de un animal de la especie puede ocurrir que sea hembra, o que no lo sea. Así se tienen los dos posibles resultados de una prueba de Bernouilli. Se llamará *E="Que el animal extraído sea hembra"* y *F="Que el animal extraído no sea hembra"*. Además, se conoce que la probabilidad de que un animal de dicha especie sea hembra es $p(E) = 1 - 0{,}7 = 0{,}3$. Si se extraen animales de la especie hasta conseguir las 3 hembras buscadas, la distribución de la variable X =*número machos extraídos hasta conseguir la tercera hembra (tercer éxito)* es la de una distribución binomial negativa, $BN(r = 3, p = 0{,}3)$.

Ahora bien, en el problema se pide, ¿qué probabilidad hay de tener que extraer al menos 8 animales de la población para conseguir las 3 hembras buscadas? Esto es p(Extracciones≥ 8)=p(Machos+3Hembras≥ 8)=p(Machos≥ 5)=p($X \geq 5$).

Rcmdr

Para calcular en Rcmdr esta probabilidad, se debe tener en cuenta que la opción cola derecha calcula para un valor a, la probabilidad $p(X > a)$*, así se debe considerar que, al tratarse de una variable discreta* $p(X \geq 5) = p(X > 4)$ *y tras seleccionar en el menú del programa* *Distribuciones* → *Distribuciones discretas* → *Distribución binomial negativa* → *Probabilidades binomiales negativas acumuladas...* *y marcar las opciones que se ven en la siguiente imagen, resulta que* $P(X \geq 5) = 0{,}6470695$.



4.3.4 Distribución de Poisson

La distribución de Poisson se obtiene como limite de la distribución binomial cuando el número de veces que se realiza el experimento, *n*, tiende a ∞, la probabilidad de éxito, *p*, tiende a cero y el número medio de éxitos se estabiliza alrededor de un número λ, que es el valor que caracteriza a la distribución. Cuando X$\sim Po(\lambda)$ X mide una eventualidad que se produce en un soporte continuo y que, de media en dicho soporte, se estabiliza en torno a el valor λ.

Se verá a continuación un ejemplo:

Se sabe que el número de bacterias por mm^3 de agua en un estanque es una variable aleatoria X con distribución de Poisson de parámetro $\lambda = 0{,}5$. ¿Cuál es la probabilidad de que en un mm^3 de agua no haya ninguna bacteria? ¿Y la de que en 2 mms^3 haya 2 a lo sumo?

Solución:

En este ejemplo el soporte continuo es el volumen, medido en mm^3, y el número de eventos que mide la variable Poisson X= *número de bacterias por* mm^3. Según el enunciado λ =*número medio de eventos (bacterias) por unidad* (mm^3)= $0{,}5$, luego $X \sim Po(\lambda = 0{,}5)$.

Ahora bien, en el problema se pide, en primer lugar, ¿cuál es la probabilidad de que en un mm^3 de agua no haya ninguna bacteria? Esto es p(X= 0).

Rcmdr

Para calcular en Rcmdr esta probabilidad puntual se selecciona Distribuciones→ Distribuciones discretas→ Distribución de Poisson→ Probabilidades de Poisson... y el programa devuelve una lista con todas las probabilidades puntuales, resultando que $p(X = 0) = 0{,}6065$.

```
> .Table
      Pr
0 0.6065
1 0.3033
2 0.0758
3 0.0126
4 0.0016
5 0.0002
```

Para responder a la pregunta "¿cuál es la probabilidad de que en 2 mms^3 haya 2 a lo sumo?" debemos tener en cuenta que si se sabe que,en 1 mms^3 de agua hay de media $\lambda = 0{,}5$ bacterias, al considerar 2 mms^3 de agua habrá de media $\lambda = 1$ bacteria.

Así pues, en este caso, $X \sim Po(1)$ y la probabilidad pedida es p(X$\leq$ 2) y, como la opción cola izquierda incluye el extremo, no se tiene mas que marcar la opción *cola izquierda* para el valor 2, dentro de la opción:

Distribuciones→Distribuciones discretas→Distribución de Poisson→Probabilidades de Poisson acumuladas... , resultando que $p(X \leq 2) = 0{,}9196986$.

4.4 Distribuciones continuas

4.4.1 Distribución exponencial

A partir de un proceso de Poisson, se define la distribución exponencial como la de aquella variable aleatoria X que mide el tiempo transcurrido entre la ocurrencia de dos éxitos consecutivos. Se denota $Exp(\lambda)$, siendo λ el número medio de éxitos por unidad de tiempo de la distribución de Poisson.

Se verá a continuación un ejemplo:

Si se sabe que la vida de cierto motor que funciona con biodiesel es de 15 años, ¿cuál es la probabilidad de que un motor de este tipo dure al menos 17 años? y, ¿cuántos años debería durar, al menos, un motor de este tipo para estar entre el 15 % de los más duraderos?

Solución:

Para la primera de las cuestiones se debe considerar en primer lugar que $\lambda = 1/15$ y que la probabilidad pedida es p$(X > 17)$.

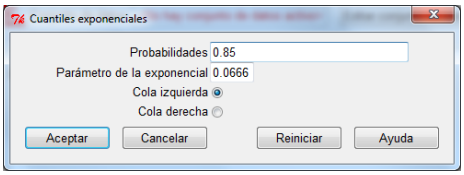
Rcmdr

Para calcular en Rcmdr esta probabilidad se selecciona Distribuciones→ Distribuciones continuas → Distribución exponencial → Probabilidades exponenciales... , resultando que $p(X > 17) = 0{,}3219583$.

En la segunda de las cuestiones planteadas, ¿cuántos años debería durar, al menos, un motor de este tipo para estar entre el 15 % de los más duraderos?, se pide el valor de la variable exponencial que deja por debajo el 85 % de los motores menos duraderos. Esto es, P_{85}.

Rcmdr

En Rcmdr se selecciona ahora Distribuciones→ Distribuciones continuas → Distribución exponencial → Cuantiles exponenciales... , resultando que el valor a tal que $p(X \leq a) = 0{,}85$ *es* $a = 28{,}4568$ *años.*



4.4.2 Distribución uniforme

Se define la distribución uniforme como la extensión natural de la distribución uniforme discreta, que es aquella que asigna igual probabilidad a todos los valores de un intervalo $[a, b]$. En el caso continuo, en lugar de considerar puntos en $[a, b]$, se consideran subintervalos de igual amplitud equiprobables.

Se verá a continuación un ejemplo:

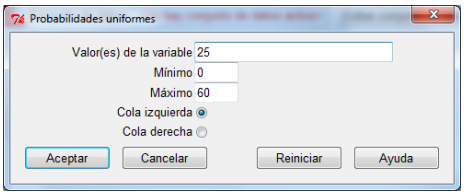
La alarma de incendios de cierta fábrica está averiada y salta todas las mañanas en cualquier momento entre las 6 y las 7. Si el vigilante necesita 20 minutos para desactivarla y, el horario laboral comienza a las 6:45, ¿qué probabilidad hay de que todo esté solucionado cuando comience el trabajo?

Solución:

Si se define X =Tiempo (en minutos) transcurrido desde las 6:00 hasta que suene la alarma, se tiene que $X \sim U(0, 60)$. Ahora bien, para que el vigilante tenga 20 minutos para desconectarla y esto ocurra como muy tarde a las 6:45, la alarma debe sonar, como muy tarde, a las 6:25. Así, la probabilidad pedida es $p(X < 25)$.

Rcmdr

En Rcmdr se selecciona Distribuciones → Distribuciones continuas → Distribución uniforme → Probabilidades uniforme... y, marcando las opciones siguientes:



Resulta que $p(X < 25) = 0{,}4166667$.

4.4.3 Distribución Normal

La distribución Normal, caracterizada por los parámetros μ y σ es la más importante de las distribuciones ya que, es la distribución límite de una amplia gama de sucesiones de variables aleatorias independientes y, además, la gran mayoría de las variables aleatorias que se estudian en experimentos físicos son aproximadas por una distribución Normal N(μ, σ).

Se verá a continuación un ejemplo:

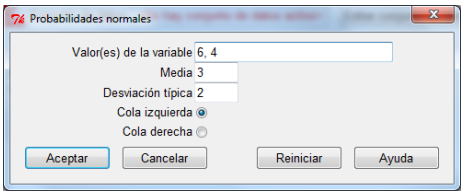
La concentración de cloro en un litro de agua viene dada por una variable aleatoria X, con distribución de media 3 y desviación típica 2. Se considera que un litro de agua es potable si su concentración de cloro está entre 4 y 6. ¿Qué probabilidad hay de que, seleccionado un litro de agua al azar, éste sea potable?

Solución:

En este caso $X \sim N(\mu = 3, \sigma = 2)$ y la probabilidad pedida es $p(4 < X < 6) = p(X < 6) - p(X < 4)$.

Rcmdr

En Rcmdr se selecciona Distribuciones → Distribuciones continuas → Distribución Normal → Probabilidades Normales... y, marcando las opciones siguientes:



Resulta que la probabilidad de que un litro de agua sea potable es:
$p(X < 6) - p(X < 4) = 0{,}9331928 - 0{,}6914625 = 0{,}2417303$.

4.4.4 Distribución χ^2

Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes idénticamente distribuidas según una $N(0,1)$, se tiene que $\sum_{i=1}^{n} {X_i}^2 \sim \chi^2_n$.

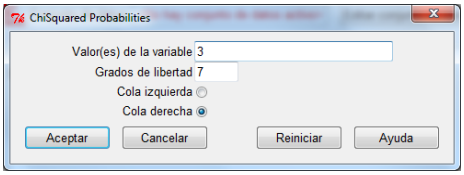
Se verá a continuación un ejemplo:

Si, el tiempo de funcionamiento en años de un generador eólico, se distribuye según una χ^2 con 7 grados de libertad. ¿Con qué probabilidad un generador cualquiera funcionará al menos 3 años?

Solución:
Se sabe que $X \sim \chi^2_7$ y la probabilidad pedida es $p(X \geq 3) = p(X > 3)$.

Rcmdr

En Rcmdr se selecciona Distribuciones→Distribuciones continuas→Distribución Chi−Cuadrado→Probabilidades Chi−Cuadrado... con las opciones siguientes:



Resulta así que $p(X \geq 3) = p(X > 3) = 0{,}8850022$

4.4.5 Distribución *t de Student*

Sean X e Y dos variables aleatorias independientes distribuidas según una $N(0,1)$ y χ^2_n, respectivamente, la variable aleatoria $T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$.

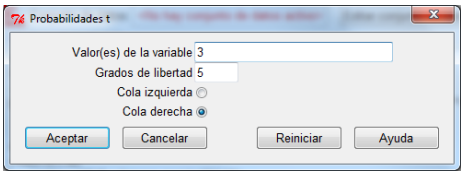
Se verá a continuación un ejemplo:

Se sabe que el error absoluto (en gramos) cometido por una balanza se distribuye según una t de Student con 5 grados de libertad. ¿Con qué probabilidad esta máquina cometerá, por exceso, un error de más de 3 gramos?

Solución: En este caso $X \sim t_5$ y la probabilidad pedida es $p(X > 3)$.

Rcmdr

En Rcmdr se selecciona Distribuciones→Distribuciones continuas→Distribución t→Probabilidades t... con las opciones siguientes:



Resulta así que $p(X > 3) = 0{,}01504962$.

4.4.6 Distribución *F de Snedecor*

Sean X e Y dos variables aleatorias independientes distribuidas según una χ^2, de n y m grados de libertad, respectivamente. Entonces la variable aleatoria $F = \frac{\frac{X}{n}}{\frac{Y}{m}} \sim F_{n,m}$.

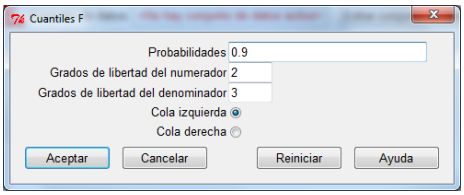
Se verá a continuación un ejemplo:

Cierto laboratorio farmacéutico está promocionando un nuevo analgésico. Si la variable X que mide el tiempo, en minutos, transcurridos desde la ingesta del analgésico hasta que se alivia el dolor sigue una distribución *F de Snedecor* de 2 y 3 grados de libertad. ¿Cuánto tiempo tarda a lo sumo el medicamento en hacer efecto en el $90\,\%$ de los casos?

Solución: Se sabe que $X \sim F_{2,3}$ y se pide "¿Cuánto tiempo tarda a lo sumo el medicamento en hacer efecto en el $90\,\%$ de los casos?", es decir P_{90}.

Rcmdr

En Rcmdr se selecciona Distribuciones → Distribuciones continuas → Distribución F → Cuantiles F... con las opciones siguientes:



Resulta así que el valor a tal que $p(X \leq a) = 0{,}9$ es $a = 5{,}462383$. Es decir, en el $90\,\%$ de los casos transcurren, como mucho, 5 minutos y medio, aproximadamente.

Ejercicios

1. Durante cierta epidemia de gripe, enferma el $30\,\%$ de la población. En un aula con 150 estudiantes, ¿cuál es la probabilidad de que haya exactamente 50 estudiantes con gripe?.
2. Cierta compañía se dedica a asegurar superficies arboladas contra fenómenos meteorológicos extraordinarios. Dicha compañía está vendiendo un producto que asegura los árboles que mueren antes de los 3 primeros años tras su plantación. Se conoce que la distribución de los años sobrevividos es normal con

media 8 años y desviación típica de 4 años. Si la compañía tiene 100000 árboles asegurados este año, ¿con qué probabilidad tendrán que pagar por más de 100 árboles?

3. Al suministrar una dosis de cierto medicamento a una especie marina contra una enfermedad que ha causado el deceso de la población de dicha especie marina, un biólogo marino estimó que la probabilidad de que un individuo se cure con una dosis es de $0,3$. ¿Qué probabilidad hay de que un individuo infectado se cure con menos de 8 dosis del medicamento?
4. Se sabe que el m^2 de coral por m^3 de agua en un arrecife sigue una variable aleatoria con distribución de Poisson $(\lambda = 0,2)$. ¿Cuál es la probabilidad de que en $5m^3$ de agua no haya coral? ¿Y la de que en $4m^3$ haya $1m^2$ de coral a lo sumo?
5. Un pájaro come mariposas de una población de gran tamaño. Dichas mariposas pueden comer de cierta planta venenosa, de forma que si en algún momento un pájaro come una mariposa envenenada, debido a la intoxicación, éste dejaría de comer durante ese día. Si se supone que el $40\,\%$ de la población de las mariposas come de la planta venenosa, calcular la probabilidad de que un pájaro coma en un día mas de 4 mariposas.
6. La concentración de cloro en un litro de agua viene dada por una variable aleatoria X, con distribución de media 3 y desviación típica 2. Se considera que un litro de agua es potable si su concentración de cloro está entre 4 y 6. ¿Si se han seleccionado 15 muestras de 1 litro de agua, ¿qué probabilidad hay de que al menos 6 sean potables?
7. Se sabe que el error absoluto (en gramos) cometido por una balanza se distribuye según una t de Student con 5 grados de libertad. ¿Cuál es el error máximo (por exceso) que comete la balanza en el $95\,\%$ de los casos?
8. Cierto laboratorio farmacéutico está promocionando un nuevo analgésico. Si la variable X que mide el tiempo, en minutos, transcurridos desde la ingesta del analgésico hasta que se alivia el dolor sigue una distribución *F de Snedecor* de 2 y 3 grados de libertad. ¿Con qué probabilidad tardará el analgésico menos de 5 minutos en hacer efecto?

Parte IV

Inferencia Estadística

5

Inferencia Paramétrica

Contenidos

En esta práctica se van a estudiar problemas que involucran a una o dos poblaciones. Se aceptará, a expensas de poder comprobarlo en la próxima práctica, 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.

5.1 Objetivos

- Construir intervalos de confianza para algún parámetro de la población.
- Plantear contrastes paramétricos.
- Ser capaz de interpretar y extraer conclusiones de los análisis realizados.

5.2 Descripción de los conjuntos de datos: parque_eolico, fenofibrato

Para una muestra y dos muestras independientes, el conjunto de datos con los que se va a trabajar en esta práctica queda recogido en un archivo de texto nombrado *parque_eolico.dat*, que resulta ser un conjunto que contiene datos de la velocidad del viento, registrados durante 730 horas de forma simultánea, en dos localizaciones alternativas, Parque1 y Parque2.

La estructura del conjunto de datos consta de dos columnas, correspondiendo la primera de ellas a las velocidades del viento recogidas en el Parque1 mientras que la segunda a las recogidas en el Parque2.

Para el caso de muestras pareadas se tomará el conjunto de datos *fenofibrato.dat* en el que se recoge el fibrinógeno de 32 individuos antes y después de un año tratados con fibrinógeno. La estructura del conjunto de datos consta de dos columnas, correspondiendo la primera de ellas al fibrinógeno de los individuos antes de recibir el tratamiento y la segunda después de recibirlo.

Los dos conjuntos se encuentran en:

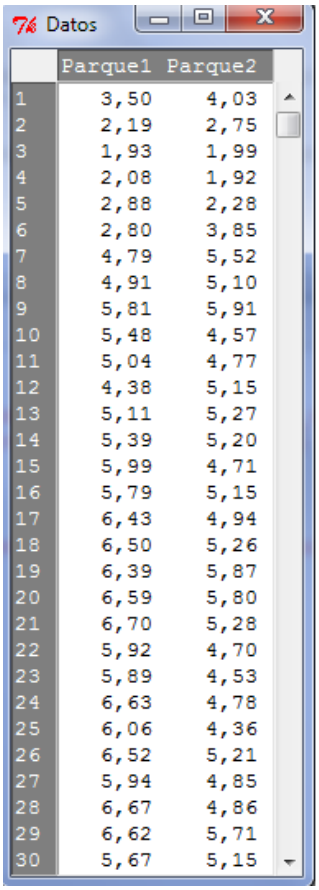
http://knuth.uca.es/repos/ebrcmdr/bases_datos/parque_eolico.dat

http://knuth.uca.es/repos/ebrcmdr/bases_datos/fenofibrato.dat

5.3 Inferencias sobre una población

Para abrir el conjunto de datos *parque_eolico.txt* en Rcmdr, se puede usar la opción del menú Datos → Importar datos → desde archivo de texto portapapeles o URL...

Al visualizar los datos se mostraría:



	Parque1	Parque2
1	3,50	4,03
2	2,19	2,75
3	1,93	1,99
4	2,08	1,92
5	2,88	2,28
6	2,80	3,85
7	4,79	5,52
8	4,91	5,10
9	5,81	5,91
10	5,48	4,57
11	5,04	4,77
12	4,38	5,15
13	5,11	5,27
14	5,39	5,20
15	5,99	4,71
16	5,79	5,15
17	6,43	4,94
18	6,50	5,26
19	6,39	5,87
20	6,59	5,80
21	6,70	5,28
22	5,92	4,70
23	5,89	4,53
24	6,63	4,78
25	6,06	4,36
26	6,52	5,21
27	5,94	4,85
28	6,67	4,86
29	6,62	5,71
30	5,67	5,15

En esta sección se abordará el estudio de la velocidad media de un parque, por ejemplo Parque1, de la que se dispone de una muestra aleatoria simple de tamaño 730. Dicha muestra se utilizará para estudiar los valores medios del Parque1.

Estimación puntual

Las características muestrales se obtienen como siempre en:

Estadísticos → Resúmenes → Resúmenes numéricos , seleccionando la correspon-

diente variable.

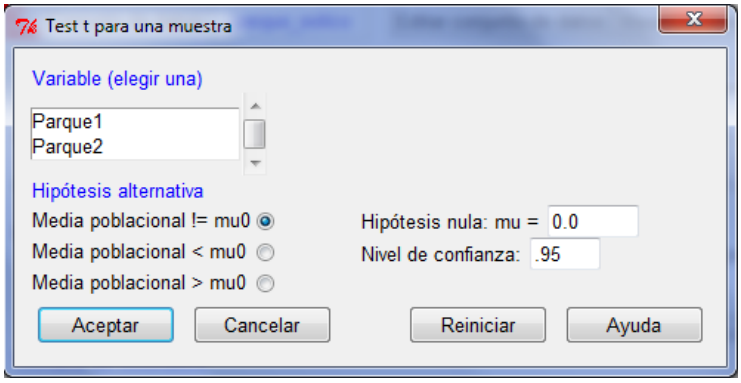
```
> numSummary(Parque_eolico[,"Parque1"], statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
     mean       sd   0%    25%  50%    75%  100%   n
 5.801795 3.241256 0.27 3.2725 5.31 7.9575 16.28 730
```

Intervalos de confianza.

A continuación se obtendrá el intervalo de confianza del 95 % para la velocidad producida en el Parque 1.

Rcmdr

Se marca Estadísticos → Medias → Test t para una muestra , seleccionando en la ventana de diálogo la variable que interesa, en este caso la Parque1, y comprobando que el nivel de confianza está fijado en el 0,95.



Los resultados que se generan son:

```
> t.test(Parque_eolico$Parque1, alternative='two.sided', mu=0.0, conf.level=.95)

        One Sample t-test

data:  Parque_eolico$Parque1
t = 48.3627, df = 729, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.566278 6.037311
sample estimates:
mean of x
 5.801795
```

De la salida interesa la parte que hace referencia al intervalo de confianza, la media de velocidad producida en el Parque 1 se encuentra dentro del intervalo $(5,57; 6,04)$ 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 el cuadro de diálogo.

En primer lugar se desea realizar el contraste:

$$\begin{cases} H_0 : \mu = 5{,}75 \\ H_1 : \mu \neq 5{,}75 \end{cases}$$

con un nivel de significación $\alpha = 0{,}01$.

Ejecutando el test con esos cambios se tiene:

```
> t.test(Parque_eolico$Parque1, alternative='two.sided', mu=5.75, conf.level=.99)

        One Sample t-test

data:  Parque_eolico$Parque1
t = 0.4317, df = 729, p-value = 0.6661
alternative hypothesis: true mean is not equal to 5.75
99 percent confidence interval:
 5.491976 6.111613
sample estimates:
mean of x
 5.801795
```

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*$= 0{,}667$, 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. Lo mismo ocurría en la salida anterior donde se había fijado el nivel de confianza en $0,95$, pues $5{,}75$ estaba dentro del intervalo.

Contraste unilateral.

Se plantea ahora la realización del contraste:

$$\begin{cases} H_0 : \mu \geq 6 \\ H_1 : \mu < 6 \end{cases}$$

con un nivel de significación $\alpha = 0{,}1$.

Rcmdr

Se marca Estadísticos → Medias → Test t para una muestra , seleccionando en la ventana de diálogo la variable que interesa, en este caso el Parque1, comprobando que el nivel de confianza está fijado en el $0{,}9$ *y cambiando la hipótesis alternativa a less*

Con estos parámetros, se ejecuta y se obtiene:

```
> t.test(Parque_eolico$Parque1, alternative='less', mu=6, conf.level=.9)

        One Sample t-test

data:  Parque_eolico$Parque1
t = -1.6522, df = 729, p-value = 0.04946
alternative hypothesis: true mean is less than 6
90 percent confidence interval:
     -Inf 5.955674
sample estimates:
mean of x
 5.801795
```

En este caso el `p-valor`=0,04 es menor que el nivel de significación y por tanto se rechaza la hipótesis nula. Igualmente se puede comprobar que 6 no pertenece al intervalo de confianza.

5.4 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 la próxima práctica, 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.

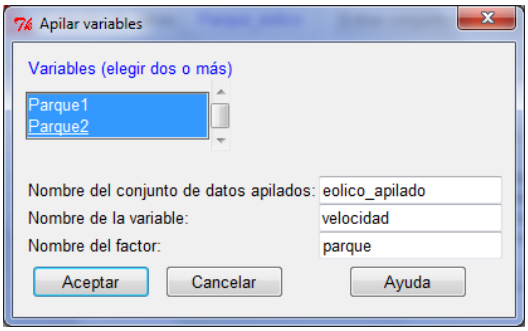
Si se miden el peso de 50 alevines de truchas antes y después de una cierta dieta alimenticia, ambas observaciones están relacionadas. La aplicación de dos pomadas en diferentes zonas de la piel de un individuo y la observación de ambas respuestas conduce a observaciones pareadas. A veces la dependencia 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.

5.4.1 Muestras independientes

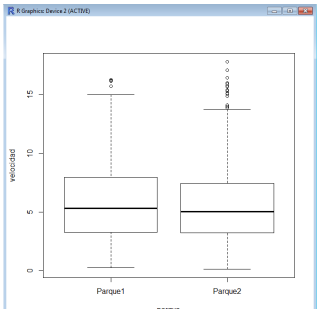
Para el caso de muestras independientes se usará el fichero *parque_eolico.dat*. Se tratará de establecer la localización más aconsejable para la instalación de un parque de producción de energía eólica.

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 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 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
```

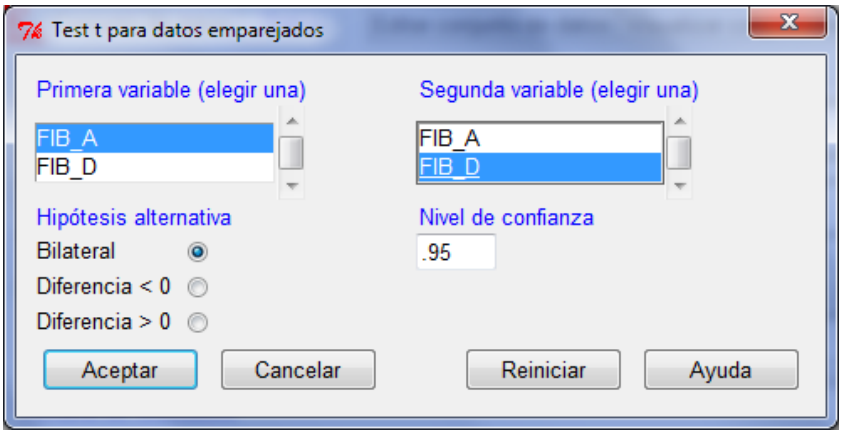
Como *p-valor*$= 0,9093 > 0,05$ no hay motivos para rechazar la igualdad de varianzas. Siendo así, como se supone que los datos están distribuidos normalmente y las varianzas son iguales, los dos parques eólicos serán igualmente productivos cuando la diferencia de sus medias no se separe significativamente de 0. Para realizar este contraste se selecciona Estadísticos$\rightarrow$Medias$\rightarrow$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.

```
> 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.

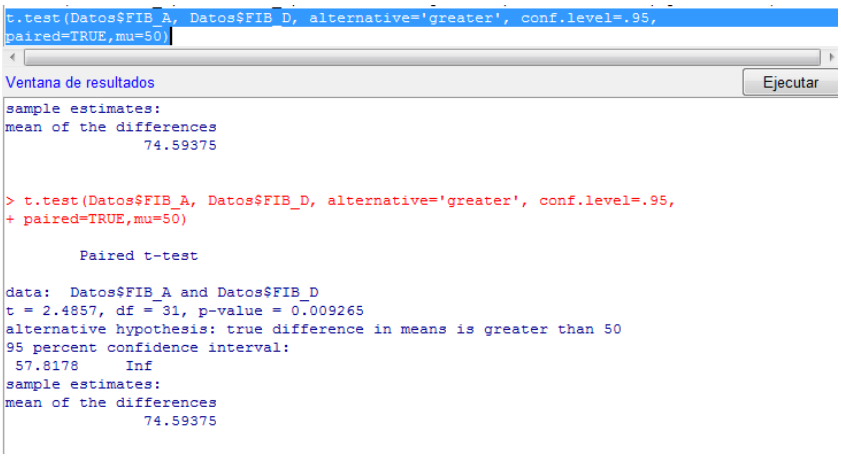
5.4.2 Muestras pareadas

Para el caso de muestras pareadas se tomará el conjunto de datos *fenofibrato.dat* . Se efectúa el *Test t* en Estadísticos $\rightarrow$ Medias $\rightarrow$ Test t para datos relacionados , realizando un contraste unilateral




```
> 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 $p - valor < 0,05$ 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:



De nuevo dado que $p < 0,05$ 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.

Ejercicios

1. Utilizando el fichero *cebada.txt*, mencionado en prácticas enteriores obtén un intervalo de confianza al 98 % para el porcentaje medio del almidón en grano

2. Siguiendo con el fichero del ejercicio anterior,¿ existen diferencias en el porcentaje protéico de las dos variedades de cebada?

3. Cargando el fichero 'datos de empleados.sav', que se encuentra en:

   http://knuth.uca.es/repos/ebrcmdr/bases_datos/, ¿hay diferencias significativas en el salario actual según el sexo?

6

Inferencia No Paramétrica.

Contenidos

En esta práctica se van a plantear y resolver algunas cuestiones referentes a los contrastes no paramétricos para una o dos muestras. En primer lugar se realizará un contraste sobre la calidad de la muestra, a continuación se realizará un test sobre la bondad de ajuste, y por último, se dará una alternativa no paramétrica para el caso en que las poblaciones no sean normales. Finalmente se propondrán una serie de actividades similares que puedan facilitar al alumno la asimilación de los objetivos planteados.

6.1 Objetivos

- Ser capaz de analizar la hipótesis de aleatoriedad de una muestra de datos.

- Plantear contrastes no paramétricos como alternativa para muestras no normales.
- Ser capaz de interpretar y extraer conclusiones de los análisis realizados.

6.2 Descripción del conjunto de datos: Contaminantes – NO2

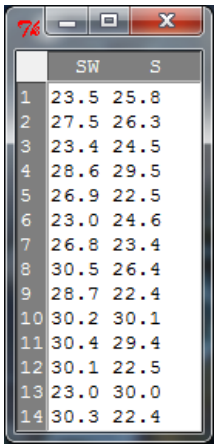
El conjunto de datos con los que se va a trabajar en esta práctica queda recogido en un archivo de texto nombrado *NO2.txt*. Dicho conjunto resulta ser una fracción concreta de una muestra realizada para un experimento que tiene como fin el estudio de la calidad del aire en un medio urbano.

El dióxido de nitrógeno u oxido nitrico (NO_2), es un compuesto químico formado por los elementos nitrógeno y oxígeno, y uno de los principales contaminantes entre los distintos óxidos del nitrógeno. Se forma como subproducto en los procesos de combustión a altas temperaturas, como en los vehículos motorizados y las plantas eléctricas. Por ello es un contaminante frecuente en zonas urbanas. Se presenta como un gas tóxico, irritante y precursor de la formación de partículas de nitrato. Afecta principalmente al sistema respiratorio. La exposición a corto plazo en altos niveles causa daños en las células pulmonares, mientras que la exposición a más largo plazo en niveles bajos de dióxido de nitrógeno puede causar cambios irreversibles en el tejido pulmonar similares a un enfisema.

En el conjunto de datos anterior los contrastes serán aplicados a un único contaminante: NO_2. La estructura del conjunto de datos consta de dos columnas, correspondiendo la primera de ellas a las cantidades de NO_2 para las muestras tomadas con el tipo de viento SW (Suroeste) y la segunda de ellas a las cantidades de NO_2 para las muestras tomadas con el tipo de viento S (Sur).

Para abrir el conjunto de datos *NO2.txt* en Rcmdr, se puede usar la opción del menú Datos → Importar datos → desde archivo de texto portapapeles o URL...

Al visualizar los datos se mostraría:



	SW	S
1	23.5	25.8
2	27.5	26.3
3	23.4	24.5
4	28.6	29.5
5	26.9	22.5
6	23.0	24.6
7	26.8	23.4
8	30.5	26.4
9	28.7	22.4
10	30.2	30.1
11	30.4	29.4
12	30.1	22.5
13	23.0	30.0
14	30.3	22.4

6.3 Prueba de aleatoriedad

Vamos a comenzar realizando un estudio sobre la calidad de la muestra extraída de la población, pues 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. Como en todo contraste, debe tenerse en cuenta que el test sólo desestimará la hipótesis en el caso de que la evidencia muestral en su contra sea de peso.

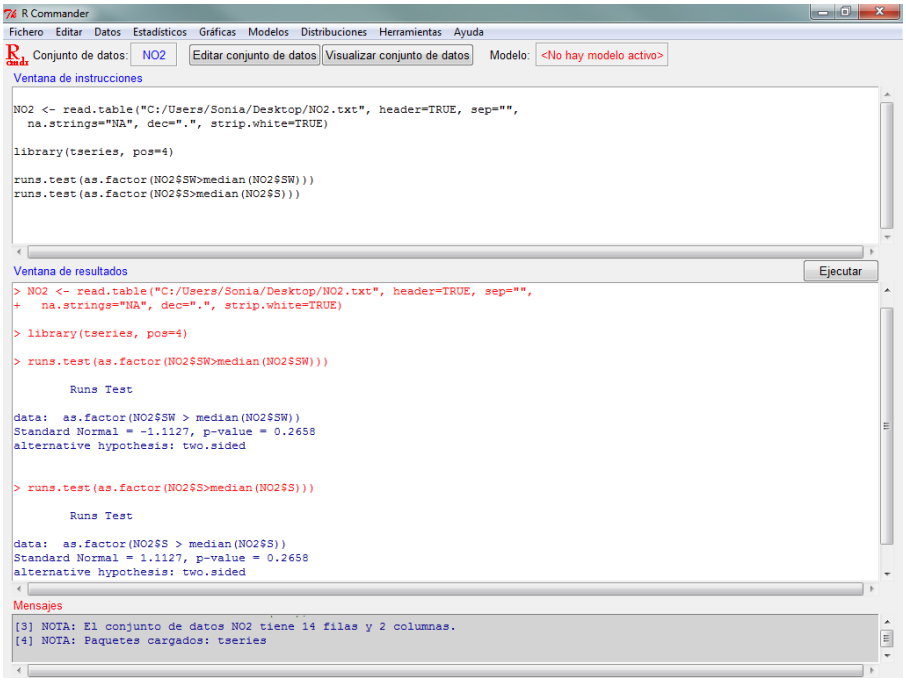
La perspectiva desde la que se analizará la aleatoriedad de la muestra es la de comprobar si existen rachas, entendiendo por racha al grupo de valores consecutivos iguales interrumpido por uno de signo distinto.

Rcmdr

Para aplicar el test de rachas, previamente se deberá cargar el paquete tseries, bien desde el menu de Rcmdr seleccionando para ello Herramientas → Cargar paquete(s)... → tseries , o ejecutando la instrucción "library(tseries, pos= 4)".

El test de rachas se aplicará tanto a la muestra del tipo de viento SW como a la muestra del tipo de viento S. Como ninguna de ambas variables son de tipo binario, se las transformarán primero para que lo sean asignando las clases de la dicotomía en función de que el elemento muestral quede por encima o por debajo de un determinado valor, típicamente la mediana. Para ello, en el argumento de la función relativa a cada contraste se podrían introducir las instrucciónes
"as.factor(NO2$SW>median(NO2$SW))", y
"as.factor(NO2$S>median(NO2$S))".

Las instrucciónes relativas a los tests de rachas serían entonces
"runs.test(as.factor(NO2$SW>median(NO2$SW)))" y
"runs.test(as.factor(NO2$S>median(NO2$S)))".



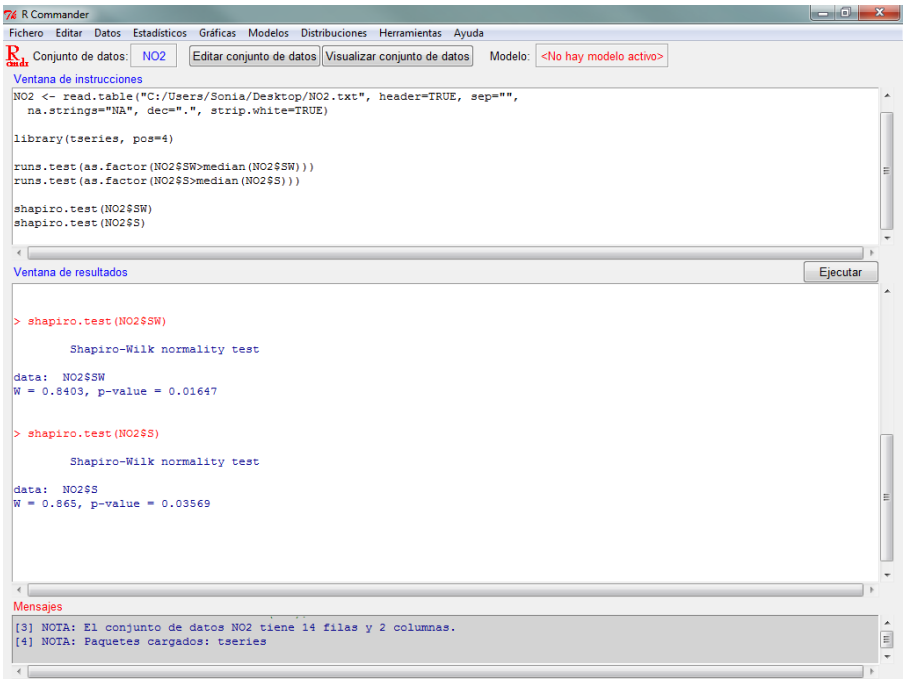
En ambos procedimientos, puesto que la salida indica que $p - value > 0{,}05$, no existirían evidencias para considerar los datos como no aleatorios.

6.3.1 Bondad de ajuste

En esta sección se contrastará si la estructura de la población analizada se ajusta a una determinada distribución, en particular a una distribución normal. Debido a que se está trabajando con muestras pequeñas, $n \leq 50$, se usará el contraste de Shapiro−Wilk.

Rcmdr

Para aplicar el test de Shapiro−Wilk, a ambas muestras, desde el menu de Rcmdr, se seleccionaría Estadísticos → Resúmenes → Test de normalidad de Shapiro−Wilk... *, y se eligiría una de las variables cada vez.*



En ambos procedimientos, puesto que la salida indica que $p - value < 0{,}05$, existirían evidencias para considerar que los datos no son normales.

6.4 Prueba de localización

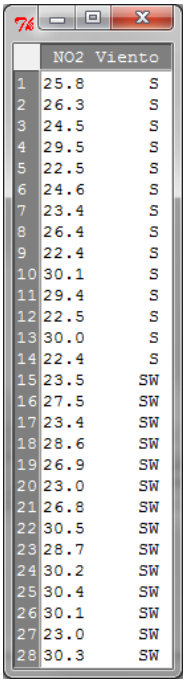
Si se desestima la hipótesis de normalidad de los datos, no son aplicables los test vistos en el capítulo de contrastes paramétricos, siendo necesario utilizar contrastes no paramétricos. Este tipo de test se basan en el análisis de la situación de los elementos de la muestra respecto a determinadas medidas de posición, muy en especial respecto a la mediana. Se estudia si los datos muestrales quedan por encima o por debajo de la mediana, o bien, se estudia la distancia ordenada a la que se encuentra de la mediana.

6.4.1 Wilcoxon para dos muestras.

En el caso tratado las dos muestras son independientes, puesto que provienen de medidas tomadas en distintos días, por lo que el test a aplicar será el de Wilcoxon para muestras independientes.

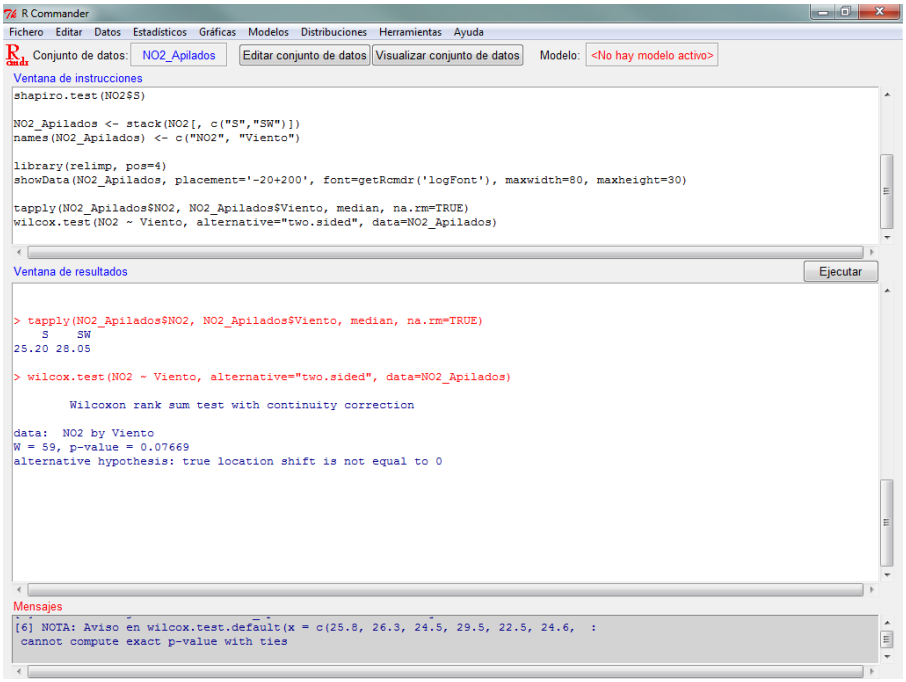
Rcmdr

Para aplicar el test de Wilcoxon para dos muestras independientes, previamente se deberán apilar los datos. Para ello, en el menu de Rcmdr se seleccionará Datos → Conjunto de datos activo → Apilar variables del conjunto de datos activo... *, eligiendo como variables para apilar las dos contenidas en los datos, SW y S. Se crearía así un nuevo conjunto de datos con la siguiente estructura:*



	NO2	Viento
1	25.8	S
2	26.3	S
3	24.5	S
4	29.5	S
5	22.5	S
6	24.6	S
7	23.4	S
8	26.4	S
9	22.4	S
10	30.1	S
11	29.4	S
12	22.5	S
13	30.0	S
14	22.4	S
15	23.5	SW
16	27.5	SW
17	23.4	SW
18	28.6	SW
19	26.9	SW
20	23.0	SW
21	26.8	SW
22	30.5	SW
23	28.7	SW
24	30.2	SW
25	30.4	SW
26	30.1	SW
27	23.0	SW
28	30.3	SW

Ahora si se puede aplicar el test de Wilcoxon para muestras independientes, y para ello, en el menu de Rcmdr se seleccionará Estadísticos → Test no paramétricos → Test de Wilcoxon para dos muestras... *, eligiendo la variable que contiene los tipos de vientos para los grupos y la variable numérica con las anotaciones de la cantidad de NO2 para la variable explicada.*



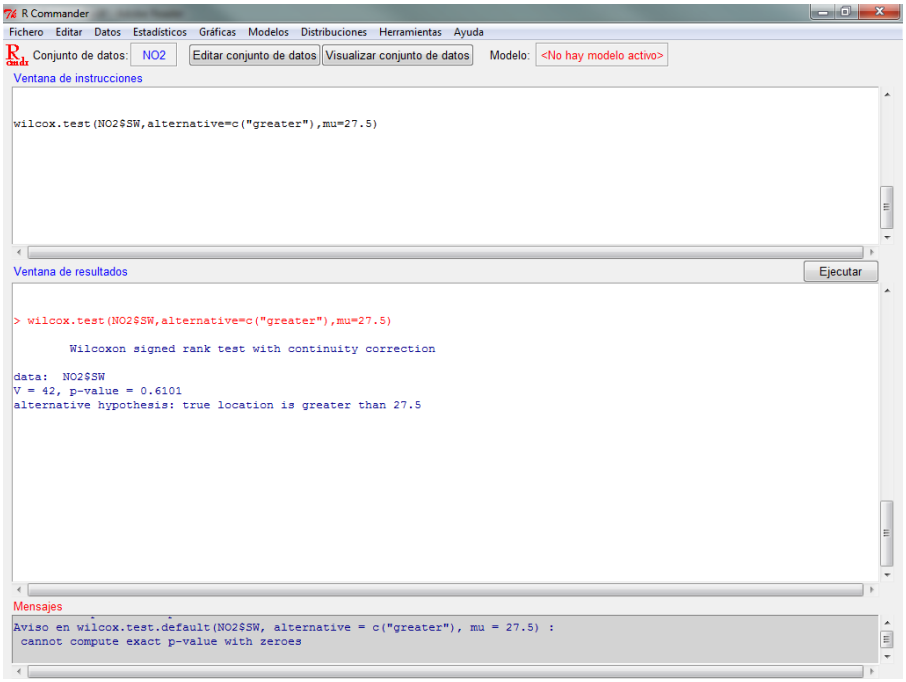
Puesto que la salida indica que $p - value > 0{,}05$, no existirían evidencias para rechazar la hipótesis nula de igualdad de medianas, siendo indiferente el tipo de viento para la distribución del contaminante.

6.4.2 Wilcoxon para una muestra.

Aplicaremos el test de Wilcoxon para una muestra, para estimar si la contaminación mediana de NO2 de la muestra para el viento suroeste es menor o igual que 27,5 mg/m^3.

Rcmdr

Para lo anterior, a partir del conjunto de datos original (sin apilar), se puede aplicar el test de Wilcoxon ejecutando directamente en la ventana de instrucciones la sentencia "wilcox.test(NO2$SW,alternative=c("greater"),mu= 27,5)".



Puesto que la salida indica que $p - value > 0{,}05$, no existirían evidencias para rechazar la hipótesis nula, por lo que no se rechazaría la hipótesis de que la contaminación mediana de NO2 de la muestra para el viento suroeste fuese menor o igual que $27{,}5\ mg/m^3$.

Ejercicios

1. Realizar un test de Wilcoxon en el conjunto de datos NO2.txt, suponiendo que ambas muestras son apareadas.
2. Comprobar la hipótesis de aleatoriedad y normalidad a la muestra completa en el caso de que no se distinguiesen las distintas medidas en función del tipo de viento.
3. Realizar un test de Wilcoxon para una muestra para estimar si la contaminación mediana de la muestra completa (como en el ejercicio anterior) es menor o igual que $27{,}5\ mg/m^3$.

Parte V

Análisis Multivariante

7

Análisis cluster

Contenidos

En la siguiente práctica se realiza una introducción al análisis cluster o análisis de conglomerados. El objetivo de dicho análisis es detectar las posibles agrupaciones de individuos en función de una serie de variables de interés y de una medida de similaridad que mida la semejanza entre individuos. En la presente práctica se estudiarán los aspectos más relevantes a la hora de realizar un análisis cluster a partir del conjunto de datos *mamiferos.Rdata* que representa las características de la leche de 22 mamiferos.

7.1 Objetivos

- Identificar y definir las disimilaridades adecuadas en función de las variables tratadas en el procedimiento de cluster.

- Aplicar adecuadamente el procedimiento de análisis jerárquico y analizar correctamente el dendograma.
- Aplicar adecuadamente el algoritmo de las k-medias para un análisis no jerárquico y analizar el diagrama de sedimentación.
- Ser capaz de caracterizar los grupos extraidos de ambos procedimientos.

7.2 Descripción del conjunto de datos: Leche Mamiferos

El conjunto de datos con el que se va a trabajar consta de un total de 22 individuos correspondiente al tipo de leche de 22 mamiferos y 5 variables cuantitativas continuas que miden la cantidad de *Agua*, *Grasa*, *Lactosa* y *Proteina* presentes en la leche. Se plantea establecer similitudes entre los distintos tipos de leche en base a las variables mencionadas con el fin de crear grupos de mamiferos con composición de leche materna parecida.

	Agua	Proteina	Grasa	Lactosa
Yegua	90.1	2.6	1.0	6.9
Burra	90.3	1.7	1.4	6.2
Ballena	64.8	11.1	21.2	1.6
Cebra	86.2	3.0	4.8	5.3
CerdaGuinea	81.9	7.4	7.2	2.7
Rata	72.5	9.2	12.6	3.3
Oveja	82.0	5.6	6.4	4.7
Rena	64.8	10.7	20.3	2.5
Mula	90.0	2.0	1.8	5.5
Cerda	82.8	7.1	5.1	3.7
Camella	87.7	3.5	3.4	4.8
Bufala	82.1	5.9	7.9	4.7
Zorra	81.6	6.6	5.9	4.9
Coneja	71.3	12.3	13.1	1.9
Llama	86.5	3.9	3.2	5.6
Cierva	65.9	10.4	19.7	2.6
Hipopotama	90.4	0.6	4.5	4.4
Bisona	86.9	4.8	1.7	5.7
Gata	81.6	10.1	6.3	4.4
Perra	76.3	9.3	9.5	3.0
Foca	46.4	9.7	42.0	0.0
Delfina	44.9	10.6	34.9	0.9

El conjunto de datos *Leche_mamiferos.RData* está disponible en la página del proyecto *http://knuth.uca.es/repos/p_innovacion/cuadernillo/guiones_practica/Datos*.

Una vez descargados los datos se pueden abrir en **Rcmdr** mediante la secuencia Datos → Cargar conjunto de datos...

7.3 Análisis descriptivo de las variables del conjunto de datos

En primer lugar, se realiza un análisis descriptivo de las variables, esto permitirá saber si existen distintas escalas entre las varibles. Es importante conocer si existe alguna variable cuya escala es mayor que las demás porque dicha variable tendrá una mayor influencia a la hora de calcular la disimilaridad entre dos elementos.

Para resumir la información proporcionada por los datos vamos a calcular una serie de medidas como: la *media*, *desciación típica*, *cuartiles*, *mínimo*, *máximo*, ... a través de la siguiente ruta en **Rcmdr**:

Rcmdr

Estadísticos → Resúmenes → Resúmenes numéricos...

```
> numSummary(mamiferos[,c("Agua", "Grasa", "Lactosa", "Proteina")],
+   statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
             mean        sd   0%    25%   50%    75% 100%  n
Agua     77.590909 13.217301 44.9 71.600 81.95 86.800 90.4 22
Grasa    10.631818 10.900781  1.0  3.675  6.35 12.975 42.0 22
Lactosa   3.877273  1.803682  0.0  2.625  4.40  5.200  6.9 22
Proteina  6.731818  3.573499  0.6  3.600  6.85 10.000 12.3 22
```

Se puede observar que las variables $Agua$ y $Grasa$ tienen una desviación típica mayor respecto de las variables $Lactosa$ y $Proteina$ cuyos datos están más concentrados alrededor de su media. Esto indica la presencia de una mayor escala en las dos primeras variables por lo que los grupos que se formen en el análisis cluster pueden responder a una influencia mayoritaria de estas dos variables.

Si no existe ningún motivo en especial para que una variable sea más impotante que otra y se desea que todas las variables tengan el mismo peso a la hora de realizar los grupos lo conveniente es tipificar las variables del conjunto de datos. La secuencia: Datos → Modificar variables del conjunto de datos activos → Tipificar variables... en **Rcmdr** añade al conjunto de datos cuatro nuevas variables: $Z.Agua$,

$Z.Grasa$, $Z.Lactosa$ y $Z.Proteina$, correspondientes a las variables originales tipificadas.

7.4 Elección de la disimilaridad apropiada

Dado un conjunto de individuos medidos a partir de una serie de n variables cuantitativas continuas, es posible hacer una identificación de cada individuo con un punto en el espacio vectorial real de dimensión n que generan dichas variables. En el contexto del problema, que dos mamiferos tengan una composición de leche similar equivale a decir que sus representaciones en dicho espacio vectorial están "cerca". Pero para decidir qué significa estar cerca se necesita de una medida que cuantifique esa cercanía o lejanía.

Es común usar para este fín, cuando las variables son cuantitativas continuas, distancias como: la distancia euclídea, de Mahalanobis, del supremo, Manhattan, etc. Aunque existen aplicaciones que no cumplen todos los requisitos de distancias que también son usadas para cuantificar lo diferente que son dos individuos, nos referimos a las *disimilaridades*.

En el caso de que las características de interés, en el conjunto de datos, sean de tipo cualitativo es posible usar disimilaridades como Cavalli-Sforza o Balakrishnan-Sanghvi.

En el caso de variables mixtas (cuantitativas y cualitativas) la medida de *disimilaridad* más conocida es la derivada a partir del coeficiente de Gower.

Volviendo al caso particular que nos ocupa y teniendo en cuenta la naturaleza de las variables, se propone la distancia euclídea como medida de similitud entre dos tipos de leche. Que dos mamiferos estén a una distancia euclidea pequeña significará que ambos poseen tipos de leche de características muy parecidas.

7.5 Análisis cluster jerárquico

Una vez elegida una medida de disimilaridad adecuada, se está en condiciones de poder realizar un análisis cluster. En primer lugar se estudiará un primer tipo de análisis cluster denominado jerárquico aglomerativo. La idea de este tipo de análisis es partir de todos los individuos, considerados como grupos de un único elemento, e ir uniendo los elementos hasta formar un único grupo con todos los

elementos. En una etapa inicial del algoritmo, únicamente es necesario tener en cuenta la matriz de distancias de los individuos para detectar qué dos elementos deben unirse por primera vez, pero en las etapas sucesivas es necesario medir de alguna forma la distancia entre un individuo y un grupo así como la distancia entre grupos.

En **Rcdmr** existen distintas opciones para medir la distancia entre grupos, obsérvese que la distancia de un individuo a un grupo se puede considerar una distancia entre dos grupos, uno de los cuales tienen un único elemento. Dichas medidas de enlace entre grupos se describen a continuación:

- **Método de Ward:** Este método selecciona los dos grupos que deben unirse en una etapa intermedia del clustering jerárquico a través de la mejora de una función objetivo. Se considera dicha función como la suma de los errores al cuadrado dentro de cada cluster, es decir, se calculan para cada grupo los errores al cuadrado como la suma de distancias al cuadrado de cada individuo al centroide de su grupo y posteriormente se suman todos esos errores. Partiendo de la etapa i con n_i grupos el método calcula todas las posibles clasificaciones uniendo pares de grupos y calcula el valor de la función objetivo sobre todas las clasificaciones eligiendo aquella clasificación que consigue un valor menor de la función.

- **Enlace simple:** En este caso la distancia entre dos grupos se toma como la mímima distancia entre elementos de ambos grupos. En cada etapa se unen los grupos más cercanos usando esta distancia.

- **Enlace completo:** Es el mismo método anterior usando como distancia entre dos grupos la máxima distancia entre los elementos de ambos grupos.

- **Enlace medio:** En este caso se mide la distancia entre grupos como la media de todas las distancias entre los elementos de ambos grupos.

- **Método de McQuitty:** Se define de forma recursiva, supóngase que el grupo C_M es la unión de dos grupos C_K y C_L. Se define la distancia de otro grupo C_J a C_M como:

$$D(C_J, C_M) = \frac{D(C_J, C_K) + D(C_J, C_L)}{2}$$

- **Enlace de medianas:** En el caso de considerar la distancia entre individuos como la distancia euclídea al cuadrado, el enlace de medianas se define de forma recursiva como:

$$D(C_J, C_M) = \frac{D(C_J, D_K) + D(C_J, C_L)}{2} - \frac{D(C_K, C_L)}{4}$$

- **Enlace de centroides:** La distancia entre dos grupos C_k y C_L se define como:

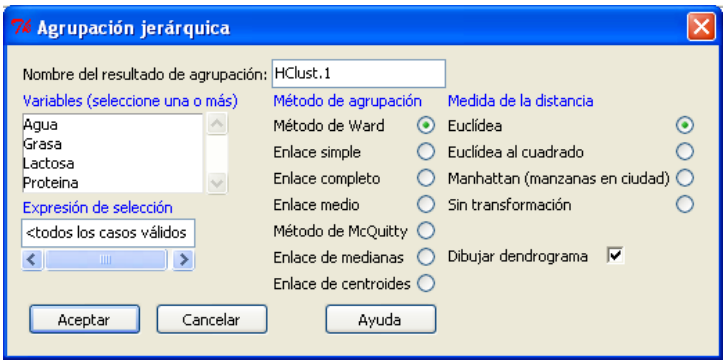
$$D(C_K, C_L) = \|\bar{x}_K - \bar{x}_L\|$$

donde $\bar{x}_K$ y $\bar{x}_L$ denotan al centroide del grupo C_K y C_L respectivamente.

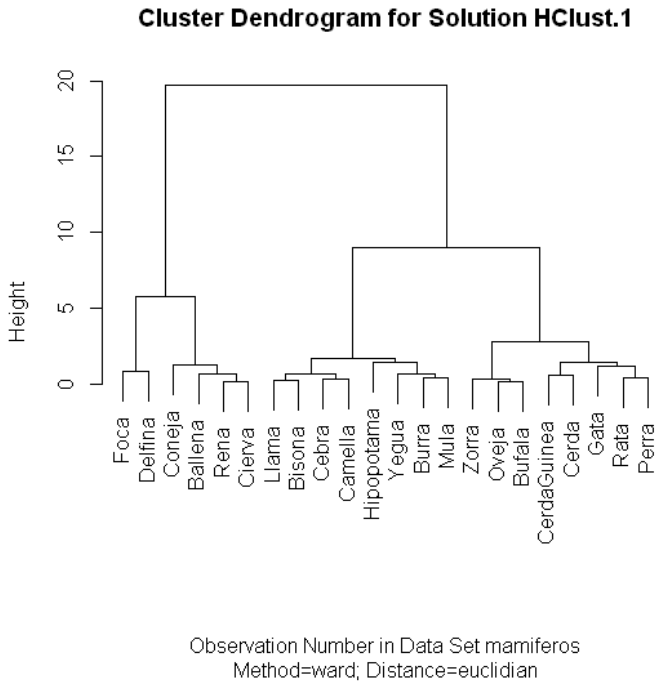
A continuación se detalla los pasos a seguir en **Rcmdr** para realizar un análisis cluster jerárquico aglomerativo:

Rcmdr

Estadísticos → Análisis dimensional → Análisis de agrupación → Agrupación jerárquica...



En la ventana emergente aparece en primer lugar una casilla para asignar el nombre que denotará el resultado del algoritmo jerárquico, las variables que se tendrán en cuenta, el método de enlace entre grupos, la distancia entre elementos, la posibilidad de tener en cuenta únicamente un subconjunto de la base de datos y la opción de dibujar el dendograma.



Analizando el dendograma se observa que se parte de todos los individuos cuando la altura en el eje OY vale cero y a medida que aumentamos dicha altura se van uniendo elementos en grupos. Para un valor y concreto del eje OY se unen todos aquellos elementos o grupos cuyas distancias entre sí sea inferior a y, de esa forma llegamos hasta un valor $y = 20$ a partir del cual todos los grupos se unen en uno solo.

La idea ahora es seleccionar el número adecuado de grupos y estudiar su posible caracterización en función de las variables originales. Una forma de estimar el número de grupos adecuado es teniendo en cuenta dos indicaciones:

1. Ver aquella configuración de grupos que permanece durante más tiempo constante si suponemos que la variable y se desplaza desde la posición $y = 0$ hasta

$y = 20$ a velocidad constante.

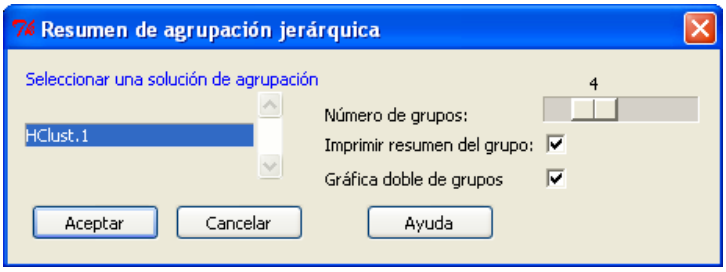
2. Aquella clasificación que explique lo mejor posible los grupos creados.

Se puede apreciar que al inicio del recorrido de la variable y se producen muchos cambios y a medida que y crece se estabilizan dichos cambios, si sólo tenemos encuenta la primera indicación sería aconsejable escojer una clasificación con dos grupos, pero es claro que 3 o incluso 4 dan una mejor idea de como se distribuyen los puntos, pues aportan una mayor información.

Supóngase que se opta por elegir 4 grupos, el siguiente paso es resumir la información de dicha clasificación mediante la siguiente opción en **Rcmdr**.

Rcmdr

Estadísticos→Análisis dimensional→Análisis de agrupación→Resumir la agrupación jerárquica...

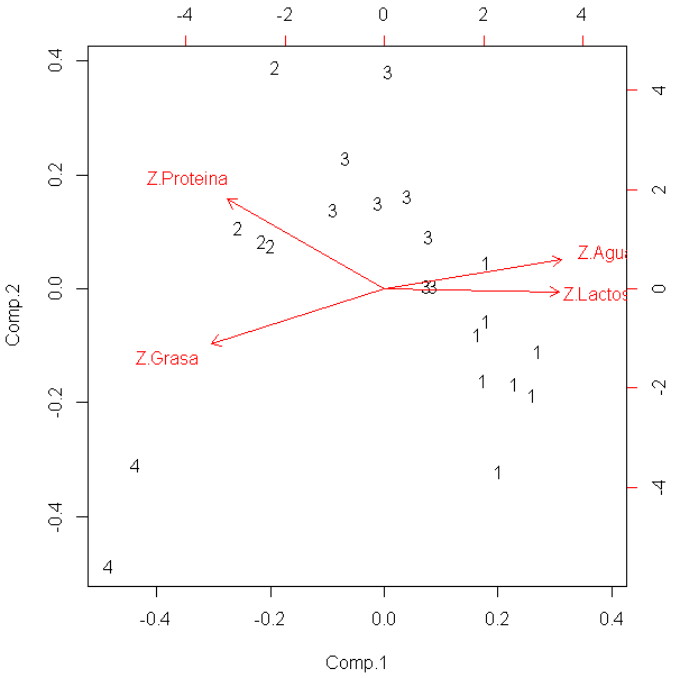


En la ventana de diálogo aparece en primer lugar el nombre del resultado del análisis cluster jerárquico, el número de grupos que se desea seleccionar, la opción de imprimir o no el resumen numérico y la posibilidad de realizar un gráfico biplot. En dicho gráfico se representa a la vez las proyecciones de los individuos y de las variables originales sobre el subespacio formado por las dos primeras componentes principales

```
> summary(as.factor(cutree(HClust.1, k = 4))) # Cluster Sizes
1 2 3 4
8 4 8 2
> by(model.matrix(~-1 + Z.Agua + Z.Grasa + Z.Lactosa + Z.Proteina, mamiferos),
+    as.factor(cutree(HClust.1, k = 4)), mean) # Cluster Centroids
INDICES: 1
    Z.Agua    Z.Grasa  Z.Lactosa Z.Proteina
 0.8263102 -0.7253442  0.9273959 -1.1107653
------------------------------------------------------------
INDICES: 2
    Z.Agua    Z.Grasa  Z.Lactosa Z.Proteina
-0.8239889  0.7286800 -0.9576371  1.2293784
------------------------------------------------------------
INDICES: 3
     Z.Agua     Z.Grasa   Z.Lactosa  Z.Proteina
 0.18983383 -0.27698181  0.02646103  0.25694199
------------------------------------------------------------
INDICES: 4
    Z.Agua    Z.Grasa  Z.Lactosa Z.Proteina
-2.4165984  2.5519438 -1.9001536  0.9565365
```

Si analizamos los resultados numéricos, aparece en primer lugar el tamaño de cada grupo. El grupo etiquetado con el 1 tiene 8 elementos, el grupo 2 esta formado por 4, el grupo 3 tiene 8 y el grupo 4 tiene únicamente 2 elementos. A continuación aparecen las coordenadas de los centroides de cada grupo. Si los individuos de cada grupo presentan poca dispersión respecto de su centroide entonces dichos centroides son representativos de cada grupo por lo que es posible analizar sus coordenandas para caracterizar a cada grupo.

Se observa que el grupo etiquetado como 1 lo forman individuos cuya leche tiene un alto contenido en agua y lactosa y poca grasa y proteinas. El grupo 2 lo forman individuos con caracteristicas opuestas, su leche tiene gran porcentaje en grasas y proteinas pero bajo nivel de agua y lactosa. El grupo 4 lo forman dos individuos con las mismas caracteristicas del grupo 2 pero mucho más acentuadas, sobre todo en el nivel de grasas, es un grupo con un número muy bajo de individuos pero merece la pena distinguirlo del grupo 2. El grupo 3 está formado por individuos con nivel equilibrado entre agua, lactosa y proteinas además de un nivel bajo en grasas.



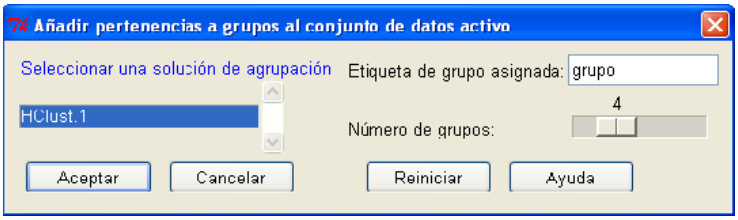
Este análisis se puede confirmar con el gráfico biplot, el cuál debe tomarse como una representación relativa de los individuos. Esto se debe a que el espacio original es de dimensión 4 mientras que el gráfico muestra el plano formado por las dos primeras componentes principales. La fiabilidad del gráfico se basa en el porcentaje de varianza explicada por ambas componentes. Además de los individuos, en el gráfico aparecen las proyecciones de las variables originales, con lo que podemos analizar que la primera componente está formada a partir de las variables *Agua* y *Lactosa* mientras que en la segunda componente tienen un mayor peso las variables *Grasa* y *Proteina*.

En este caso se puede realizar la misma lectura que se ha hecho con las coordenadas de los centroides y se aprecian las posiciones relativas de todos los individuos mediante la etiqueta de su grupo respectivo. El gráfico confirma la decisión de elegir 4 grupos pues el grupo 4 merece una especial mención como individuos cuya

leche contiene una mayor porcentaje de grasas.

Para finalizar, nos podemos plantear si todas las variables consideradas en el estudio tienen poder discriminatorio o si existen variables que no permiten distinguir entre los distintos grupos. Para ello realizamos un análisis ANOVA a partir de la variable que etiqueta a cada individuo.

En primer lugar, añadimos dicha variable al conjunto de datos mediante la siguiente secuencia: Estadísticos → Análisis dimensional → Análisis de agrupación → Agregar la agrupación jerárquica al conjunto de datos...



A continuación, realizamos un análisis de la varianza de un factor siguiendo las siguientes opciones: Estadístico → Medias → ANOVA de un factor... La idea es ver si existe alguna variable que no discrimine entre los 4 grupos que se han creado, para ello proponemos de manera secuencial para cada variable de estudio, un análisis ANOVA:

$$\begin{cases} H_0 \equiv & \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 \equiv & \text{existe al menos un grupo i con media distinta} \end{cases}$$

Empezamos dicho estudio con la variable *Agua*:

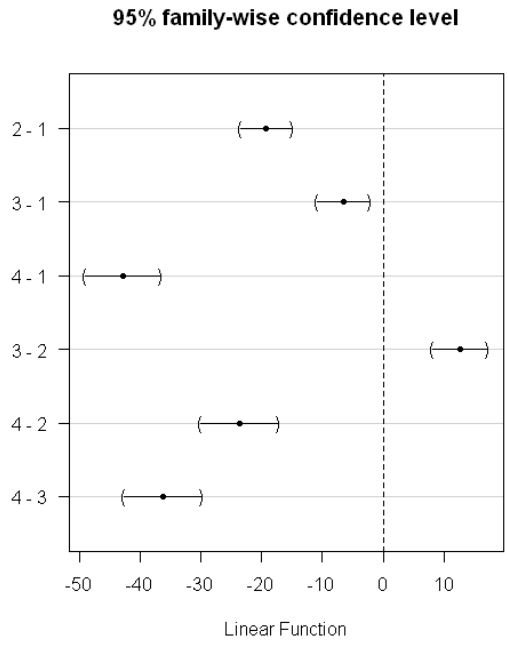
```
> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value    Pr(>F)
grupo        3   3527  1175.7   149.5 6.56e-13 ***
Residuals   18    142     7.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que el $p-valor = 6{,}56e^{-13}$ es menor que el nivel de significación, que por defecto es $0{,}05$. Esto significa que al menos uno de los grupos posee una media

distinta de los demás. Para ver esto con más precisión, estudiamos la comparación de medias por pares de grupos, bajo la variable *Agua*. Para obtener los resultados numéricos de los contrastes por pares, se debe indicar en el análisis ANOVA de un factor, marcando la pestaña correspondiente en la ventana de diálogo.

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Agua ~ grupo, data = mamiferos)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0  -19.246      1.514 -12.708  < 0.001 ***
3 - 1 == 0   -6.513      1.514  -4.300  0.00209 **
4 - 1 == 0  -42.863      2.217 -19.334  < 0.001 ***
3 - 2 == 0   12.733      1.619   7.865  < 0.001 ***
4 - 2 == 0  -23.617      2.290 -10.315  < 0.001 ***
4 - 3 == 0  -36.350      2.290 -15.876  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Según estos resultados, en todos los casos se rechaza la igualdad de medias (en la variable *Agua*) entre cada par de grupos. El caso más límite se presenta al comparar los grupos 1 y 3. Todo esto se puede ver gráficamente usando los intervalos de confianzas para la diferencias de medias.



Se observa que ningún intervalo contiene al origen y por lo tanto se pueden considerar medias distintas. De la misma forma, se repite el procedimiento para las tres variables restantes: *Proteinas*, *Grasas* y *Lactosa*. Se deja al lector la comprobación de que en todas las variables se obtienen los mismos resultados que en el caso del *Agua*.

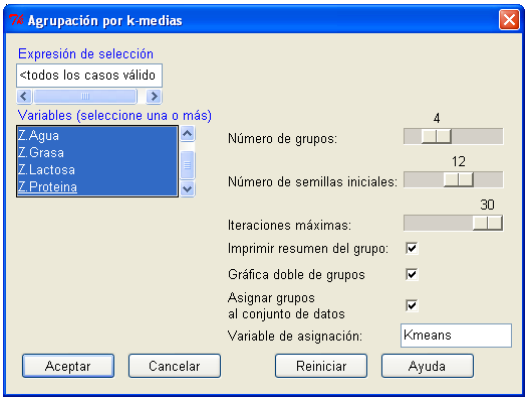
7.6 Análisis cluster no jerárquico: Algoritmo de las k-medias

El procedimiento de las *k-medias* es uno de los más conocidos entre los algoritmos no jerárquicos. La idea de este tipo de algoritmos es partir de un número k dado por el usuario y crear k grupos de manera que se optimice una función objetivo. En el caso de las *k-medias*, se pretende minimizar la suma de distancias al cuadrado entre cada representante de grupo y los individuos que componen dicho grupo. Existen distintas variantes del algoritmos de las *k-medias*, en *R* están implementadas las versiones de: *Lloyd*, *Forgy*, *McQueen* y *Hartigan-Wong*.

El algoritmo de las *k-medias* clásico consiste en seleccionar aleatoriamente k representantes del espacio de características y asignar cada individuo al representante más cercano. De esta forma, se tiene una primera partición del espacio y por lo tanto una clasificación de los individuos. El siguiente paso es calcular el valor de la función objetivo y calcular los nuevos centroides como aquellos puntos que se obtienen al hacer la media de cada grupo. Repetir los pasos anteriores hasta que se cumpla la condición de parada, que por regla general suele ser llegar a un valor de la función objetivo que sea inferior a cierta cota dada por el usuario o bien llegar al número máximo de iteraciones del algoritmo.

Como principal desventaja con respecto de los algoritmos jerárquicos se tiene que el número de grupos a realizar es dado por el usuario y por consiguiente necesitamos de algunas herramientas que nos indiquen el número adecuado a cada conjunto de individuos. Por otro lado, el comportamiento iterativo de los algoritmos hace que un individuo pueda pertenecer a distintos grupos en etapas intermedias del procedimiento, al contrario de lo que pasaba en el caso jerárquico.

Usando los datos trabajados en el caso jerárquico, repetiremos el análisis cluster, esta vez con el algoritmo de las *k-medias*. En primer lugar, seleccionamos en *R* la secuencia: Estadísticos → Análisis dimensional → Análisis de agrupación → Agrupación por k-medias



Se resumen a continuación las opciones del algoritmo de las *k-medias*:

- En primer lugar, seleccionamos las variables que van a ser usadas en el análisis.

- A continuación, seleccionamos el número de grupos. Basandonos en el dendograma del apartado anterior podemos seleccionar $k = 4$.

- La opción *número de semillas iniciales* tiene que ver con el primer paso del algoritmo de las *k-medias*, elegir de forma aleatoria k puntos del espacio, ya que el resultado final depende de la semillas seleccionadas. Por este motivo, se elige el número de veces que se repite el proceso de las *k-medias* obteniéndose distintas clasificaciones, *R* devuelve aquella clasificación que haya conseguido menor valor de la función objetivo. Elegir un número muy elevado puede hacer que el tiempo de ejecución se incremente considerablemente.

- La opción *iteraciones máximas* permite definir un criterio de parada del método.

- *Imprimir resumen del grupo* devuelve las coordenadas de los centroides de cada grupo así como el número de elementos que tiene cada grupo.

- *Gráfica doble de grupos* realiza un gráfico biplot, donde se representan a los individuos y las variables originales sobre las dos primeras componentes principales.

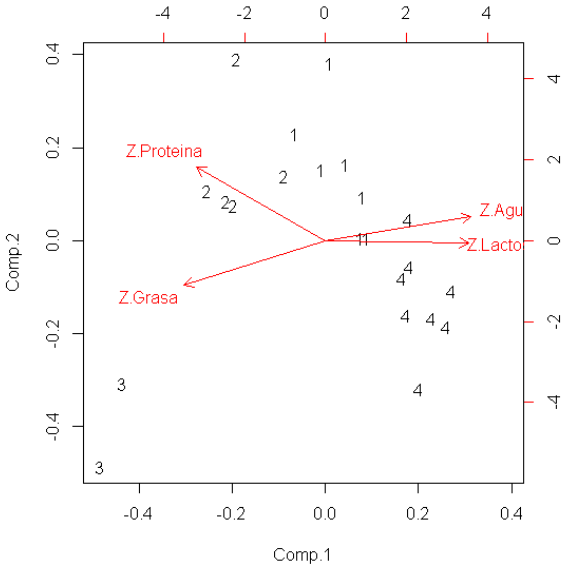
- *Asignar grupos al conjunto de datos* permite añadir una variable identificativa al conjunto de datos que identifica cada individuo con el grupo al que pertenece.

- *Variable de asignación* representa el nombre de la variable etiqueta del apartado anterior. Por defecto es *KMeans*.

Los resultados numericos son los siguientes:

```
> .cluster$size # Cluster Sizes
[1] 7 5 2 8

> .cluster$centers # Cluster Centroids
  new.x.Z.Agua new.x.Z.Grasa new.x.Z.Lactosa new.x.Z.Proteina
1    0.2719772    -0.3423441      0.07596294        0.1949779
2   -0.7362251     0.6190549     -0.83012017        1.1216408
3   -2.4165984     2.5519438     -1.90015364        0.9565365
4    0.8263102    -0.7253442      0.92739594       -1.1107653

> .cluster$withinss # Within Cluster Sum of Squares
[1] 3.0655878 1.8393871 0.3747602 2.5353643

> .cluster$tot.withinss # Total Within Sum of Squares
[1] 7.815099

> .cluster$betweenss # Between Cluster Sum of Squares
[1] 76.1849
```

En primer lugar, aparecen los tamaños de cada grupo, se observa que hay un grupo de dos elementos que aunque a priori parecería adecuado eliminarlo y volver a repetir el clustering con tres grupos, después se justificará su existencia. A continuación aparecen las coordenadas de cada representante de grupo o centroide, en el caso de que los grupos sean homogéneos se pueden usar las coordenadas de los centroides para ayudar a la caracterización de los grupos. Con el objetivo de medir lo homogeneo que son los grupos, se calculan la suma de cuadrado de distancias de cada individuo a su representante, esta medida no es absoluta y es sensible al número de elementos de cada grupo. Es preferible calcular la media de estos valores, dividiendolos por el tamaño de su grupo. El valor $7{,}815099$ representa el valor que ha alcanzado la función objetivo, que no es más que la suma de los valores de cada grupo que aparecen en la salida anterior. El último valor refleja lo distante que están los grupos, dicho valor es la suma de cuadrado entre los grupos.

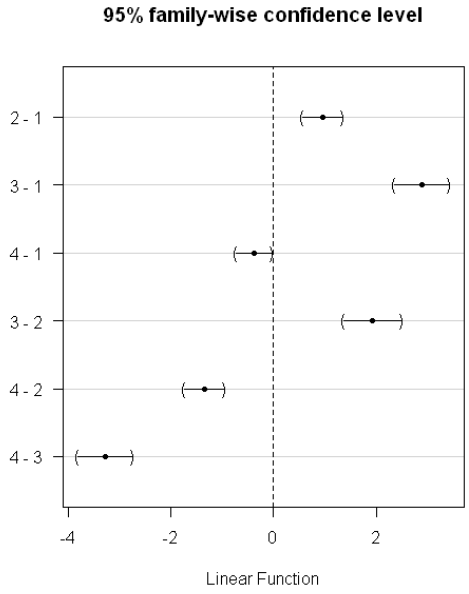


A continuación analizamos el gráfico biplot. Como es de esperar las variables quedan representadas de la misma forma que en el caso jerárquico. En cuanto a los individuos, vemos pocos cambios con respecto de la clasificación jerárquica aunque existen algunos individuos que han cambiado de grupo. Obsérvese que las etiquetas han sido modificadas con respecto a la clasificación jerárquica anterior pero esto no es más que un cambio en el nombre de los grupos. La interpretación y caracterización de los grupos es idéntica que en el caso jerárquico y por lo tanto no se va a reproducir.

De la misma forma que en el caso jerárquico, se estudia la capacidad discriminatoria de cada variable. Usando la variable etiqueta *KMeans* que se acaba de añadir al conjunto de datos. Se realizará un análisis de la varianza, eligiendo esta vez la variable *Grasa*.

```
> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value    Pr(>F)
KMeans       3  19.97   6.657   116.4 5.64e-12 ***
Residuals   18   1.03   0.057
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Según el $p-valor = 5{,}64e^{-12}$ podemos rechazar la hipótesis nula y por lo tanto sabemos que existe al menos dos grupos con medias distintas, en la variable *Grasa*. Podemos ver con más detalle el gráfico de intervalos de confianza que compara las medias de los grupos dos a dos.



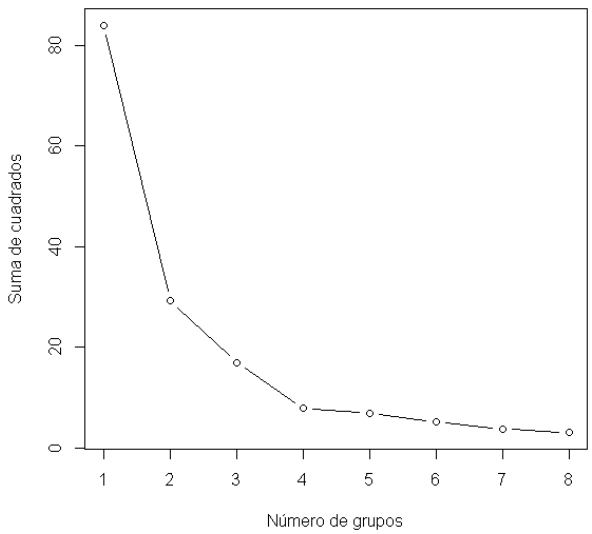
En este caso, se observa que los grupos 1 y 4 tienen medias distintas aunque están en el límite de tolerancia para tomar dicha decisión.

Para finalizar la práctica, mostraremos una herramienta muy usada a la hora de fijar el número de grupos adecuados antes de aplicar un análisis cluster no jerárquico. El diagrama de sedimentación se basa en el decrecimiento de la función

objetivo de las *k-medias* a medida que el número de grupos aumenta, la idea es ver cuando se estabiliza dicho decrecimiento y elegir como número adecuado aque a partir del cual la gráfica se estabiliza.

Las siguientes instrucciones permiten crear dicho diagrama de sedimentación. Obsérvese que en este caso el nombre del conjunto de datos es *mamiferos* y que las columnas que contienen las variables originales tipificadas son las quinta, sexta, séptima y octava.

```
wss<-numeric(8)
wss[1]<- (nrow(mamiferos)-1)*sum(apply(mamiferos[,5:8],2,var))
for(i in seq(2,8)){
wss[i]<- sum(kmeans(mamiferos[,5:8],centers=i)$withinss)
}
plot(1:8,wss,type="b",xlab="Número de grupos",
ylab="Suma de cuadrados")
```



Analizando el gráfico se confirma que el número $k = 4$ es una buena opción de partida.

Ejercicios

1. Considere el fichero *mtcars* del paquete *datasets*.
   - Realice un análisis cluster jerárquico eligiendo las variables apropiadas tipificadas. Use la distancia euclídea y el método de enlace de Ward. Interprete el dendograma.
   - A partir de la agrupación anterior construya cuatro grupos, identifique la variable grupo con el nombre *grupo_Hclust.1*. Compruebe la creación de dicha variable.
   - Utilice la opción del menú *Resumir la agrupación jerárquica* para obtener los centroides de los cuatros grupos y el gráfico *biplot*.
2. Aplique un cluster jerárquico sobre los datos del fichero *animals* del paquete *cluster*, partiendo de la matriz de distancias.
3. Realice una clasificación usando el algoritmo de las *k-medias* de los individuos del fichero *mtcars*. Decida el número de grupos óptimo y caracterice los mismos.

8 Componentes Principales

Contenidos

En la siguiente práctica se realiza una introducción al estudio de componentes principales. El objetivo es estudiar la reducción de la dimensión del conjunto de datos de forma óptima, en el sentido de mínima pérdida de información relevante para el estudio. Se estudiará un caso práctico usando el conjunto de datos *mamiferos.Rdata* que representa las características de la leche de 22 mamiferos.

8.1 Objetivos

- Identificar el grado de colinealidad de las variables de estudio.
- Aplicar la técnica de componentes principales usando el software libre *R*.
- Analizar el número adecuado de componentes a retener.

- Interpretar (si es posible) las componentes retenidas en función de las variables originales.

8.2 Descripción del conjunto de datos: Leche Mamiferos

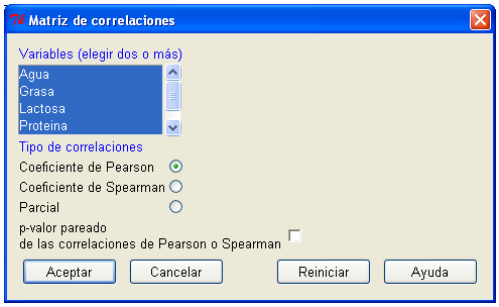
El conjunto de datos con el que se va a trabajar consta de un total de 22 individuos correspondiente al tipo de leche de 22 mamiferos y 5 variables cuantitativas continuas que miden la cantidad de $Agua$, $Grasa$, $Lactosa$ y $Proteina$ presentes en la leche. Se plantea establecer similitudes entre los distintos tipos de leche en base a las variables mencionadas con el fin de crear grupos de mamiferos con composición de leche materna parecida.

	Agua	Proteina	Grasa	Lactosa
Yegua	90.1	2.6	1.0	6.9
Burra	90.3	1.7	1.4	6.2
Ballena	64.8	11.1	21.2	1.6
Cebra	86.2	3.0	4.8	5.3
CerdaGuinea	81.9	7.4	7.2	2.7
Rata	72.5	9.2	12.6	3.3
Oveja	82.0	5.6	6.4	4.7
Rena	64.8	10.7	20.3	2.5
Mula	90.0	2.0	1.8	5.5
Cerda	82.8	7.1	5.1	3.7
Camella	87.7	3.5	3.4	4.8
Bufala	82.1	5.9	7.9	4.7
Zorra	81.6	6.6	5.9	4.9
Coneja	71.3	12.3	13.1	1.9
Llama	86.5	3.9	3.2	5.6
Cierva	65.9	10.4	19.7	2.6
Hipopotama	90.4	0.6	4.5	4.4
Bisona	86.9	4.8	1.7	5.7
Gata	81.6	10.1	6.3	4.4
Perra	76.3	9.3	9.5	3.0
Foca	46.4	9.7	42.0	0.0
Delfina	44.9	10.6	34.9	0.9

El conjunto de datos *Leche_mamiferos.RData* está disponible en la página del proyecto *http://knuth.uca.es/repos/p_innovacion/cuadernillo/guiones_practica/Datos*. Una vez descargados los datos se pueden abrir en **Rcmdr** mediante la secuencia Datos → Cargar conjunto de datos...

El siguiente paso es ver el grado de colinealidad que poseen las variables. Para ello podemos calcular la matriz de correlaciones y ver si hay variables con coeficientes cercanos a uno, eso significará que comparten un alto grado de información. Parece lógico en principio eliminar una de esas variables aunque la forma más inteligente de aprovechar la información de ambas es usar componentes principales.

Para obtener la matriz de correlaciones seguimos la siguiente secuencia Estadísticos → Resúmenes → Matriz de correlaciones...




```
> cor(mamiferos[,c("Agua","Grasa","Lactosa","Proteina")], use="complete.obs")
              Agua       Grasa    Lactosa   Proteina
Agua       1.0000000 -0.9816821  0.9029425 -0.7776543
Grasa     -0.9816821  1.0000000 -0.8908474  0.6825007
Lactosa    0.9029425 -0.8908474  1.0000000 -0.8156659
Proteina -0.7776543  0.6825007 -0.8156659  1.0000000
```

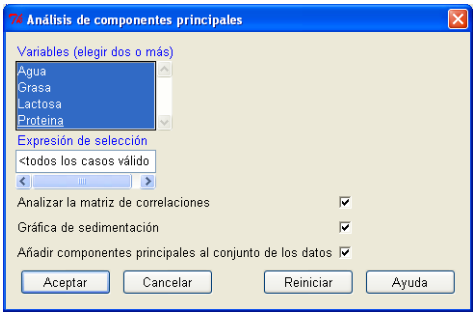
Los resultados muestran una gran correlación lineal entre la mayoria de las variables, el valor más pequeño lo alcanzan las variables *Grasa* y *Proteina* con un coefiente de Pearson de $0{,}6825$. Esto significa que muy posiblemente una única variable explique la mayoría de la información o variabilidad de los datos y que con dos de ellas se explique casi el $100\,\%$. Pero, ¿Qué variable elegir?, tal vez lo acertado sea elegir una combinación lineal de todas ellas de una forma inteligente.

8.3 Procedimiento para calcular las componentes principales usando *R*

El análisis de componentes principales *ACP* es una técnica estadística multivariante de simplificación o reducción de la dimensión, que permite transformar un conjunto de variables correlacionadas en otro conjunto de variables ortonormales

denominadas *componentes* o *ejes principales*. Para aplicar *ACP* se requiere que todas las variables de la matriz de datos sean cuantitativas o asimilables a éstas. El objetivo de *ACP* es encontrar un subespacio de menor dimensión que el original de manera que se retenga la mayor variabilidad posible de los datos.

En *Rcmdr* podemos acceder a dicho análisis mediante la ruta Estadísticos → Análisis dimensional → Análisis de componentes principales...



El primer resultado que se muestra es la matriz de correlaciones de las variables originales. Dicha matriz se ha analizado en el apartado anterior, por lo que pasaremos a comentar las siguientes salidas.

```
> unclass(loadings(.PC))  # component loadings
             Comp.1     Comp.2     Comp.3    Comp.4
Agua      0.5198985  0.2684160  0.3938995 0.7088735
Grasa    -0.5056630 -0.5002503 -0.2028317 0.6729887
Lactosa   0.5116260 -0.0262597 -0.8519732 0.1081247
Proteina -0.4607053  0.8228079 -0.2790057 0.1813658

> .PC$sd^2  # component variances
     Comp.1      Comp.2      Comp.3      Comp.4
3.532491829 0.357399346 0.103333242 0.006775583

> summary(.PC) # proportions of variance
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     1.879492 0.59782886 0.32145488 0.082313929
Proportion of Variance 0.883123 0.08934984 0.02583331 0.001693896
Cumulative Proportion  0.883123 0.97247279 0.99830610 1.000000000
```

La salida denotada por *component loadings* se refiere a las coordenadas de las componentes principales respecto de las variables originales. Cada columna es una

componente principal, se puede comprobar que estos cuatro vectores son ortonormales, es decir, son perpendiculares dos a dos y su norma vale uno. Posteriormente volveremos a referirnos a esta salida para intentar caracterizar dichas componentes en función de las variables originales.

La siguiente salida, se refiere a la varianza que captura cada componente, denotado generalmente por λ_i, son los autovalores extraidos de la matriz de correlación de las variables originales. Como el procedimiento en *R* ha utilizado la matriz de correlaciones para calcular los autovalores se tiene que la suma de la varianza de las variables originales (tipificadas) es cuatro que coincide con la suma de los cuatro autovalores. La idea es que la primera componente debe retener la mayor variabilidad posible y que lo que no explique la primera componente se explique por la siguiente, así sucesivamente hasta que las cuatro componentes expliquen toda la variabilidad de los datos.

La última salida, muestra la desviación típica, el porcentaje de varianza total y el porcentaje de varianza total acumulado de cada componente, respectivamente. Esta última parte es muy útil a la hora de decidir el número de componentes a retener.

8.4 Criterios para determinar el número de componentes adecuadas a retener

Existen diferentes criterios para decidir cuantas componentes son suficientes para mantener la mayor parte de la información de los datos sin distorsionar demasiado el estudio con la eliminación de dimensiones. A continuación resumimos los procedimientos más utilizados para este fin.

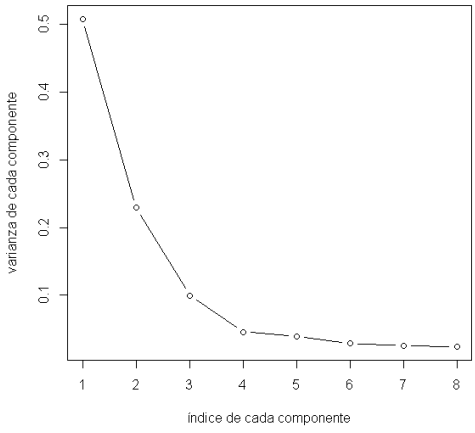
- *Varianza media* $\bar{\lambda}$. Supongamos que se puede repartir la varianza total entre todas las componentes de forma equitativa, el valor que le corresponde a cada componente será la varianza media. La idea es mantener aquellas componentes principales que superen a dicha varianza media, esto es, aquellas componentes principales tales que:

$$\lambda_i \geq \bar{\lambda} = \sum_{i=1}^{p} \frac{\lambda_i}{p}.$$

Donde p es el número de variables originales, que en nuestro caso es 4. Observemos que en el caso de extraer los autovalores de la matriz de correlaciones, se tiene que $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 4$, por lo que, $\bar{\lambda} = 1$ y el criterio se traduce en elegir aquellas componentes que cumplan $\lambda_i \geq 1$.

Volviendo al caso que nos ocupa, y observando la segunda salida, se tiene que según el criterio de la varianza media, lo correcto sería elegir la primera componente, pues la varianza del resto no superan la unidad.

- *Gráfico de sedimentación*. La idea de este procedimiento es representar de forma indexada los valores $\lambda_1 + \lambda_2 + \ldots + \lambda_p$ y unir los puntos por una poligonal. Típicamente, este gráfico muestra bruscos descensos en las primeras componentes que se van estabilizando a partir de un determinado índice. Si argumentamos que las componentes que se corresponden con la porción plana del gráfico representan componentes ruido no diferenciables del sistema, deberíamos elegir k como el primer índice donde la pendiente del gráfico se suaviza.

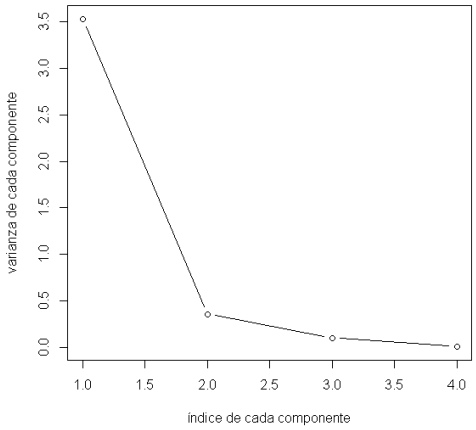


Obsérvese que en el caso de la figura anterior, se elegiría 4 componentes y dos de ellas tienen varianza menor que uno. Esto no significa que exista contradicción entre los criterios, simplemente se está usando otro prisma para analizar la

realidad y por lo tanto se debe tener en cuenta, a la hora de tomar la decisión acertada.

En el caso del conjunto de datos *mamiferos*, el gráfico de sedimentación es una salida más de la orden que calcula las componentes principales, pero viene expresado a partir de un diagrama de barras. Es posible la misma interpretación, pero si se desea representarlo mediante la poligonal que une los valores de cada autovalor, se pueden seguir las siguientes instrucciones:

```
indices  <- seq(1,4)
varianzas<- .PC^2
plot(indices,varianzas,type="b",xlab="índice de
cada componente",ylab="varianza de cada componente")
```



Como se puede observar en el gráfico, las varianzas empiezan a estabilizarse a partir de la segunda componente, por lo que según este criterio se debería retener las dos primeras componentes.

- *Porcentaje de varianza total* P_k. La proporción de varianza explicada por las primeras k componentes princiaples viene dada por:

$$P_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

En el caso de que los autovalores se extraigan de la matriz de correlaciones el porcentaje de la varianza total explicada por las primeras k componentes quedaría, $P_k = \frac{\sum_{i=1}^{k} \lambda_i}{p}$. Desde una óptica descriptiva, el *ACP* puede considerarse fiable cuando el conjunto de componentes retenidas explica al menos el $75\,\%$ de la varianza total. En el caso práctico que nos ocupa, la primera componente ya supera esa cantidad con una varianza acumulada del $0{,}883123\,\%$. Esto no significa que no podamos incluir una componente más, que no supone un aumento considerable de la dimensión, pues, por regla general, trabajar en el plano es algo común en cualquier estudio. Además se conseguirá explicar una mayor cantidad de variabilidad, en concreto para dos componentes principales se retiene el $0{,}97247279\,\%$.

8.5 Interpretación de las componentes principales

Una vez decididas las componentes sobre las que se va a trabajar, sería útil la interpretación de dichas componentes en función de las variables originales, pues el *ACP* no es por si mismo un fin, sino más bien una técnica que se utiliza previamente a cualquier otro análisis como la regresión multiple o el análisis cluster. Por otro lado, no siempre existe una interpretación clara de las componentes y puede dificultar en gran medida, por ejemplo, la caracterización de grupos en el análisis cluster. Si retornamos a la salida *component loadings*, las dos primeras componentes tienen unas coordenadas en función de las variables originales:

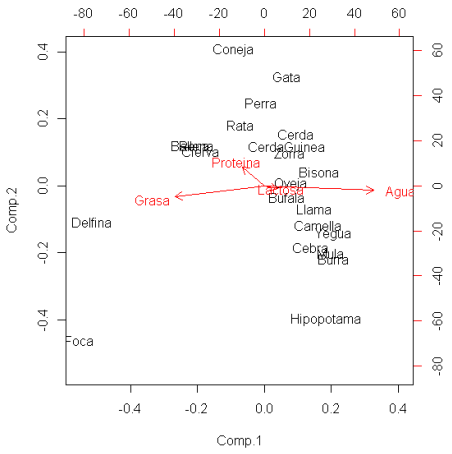
$$CP1 = 0{,}5199Agua - 0{,}5057Grasa + 0{,}5116Lactosa - 0{,}4607Proteina$$

$$CP2 = 0{,}2684Agua - 0{,}5002Grasa - 0{,}0262Lactosa + 0{,}8228Proteina$$

Estas coordenadas dan una idea del peso que tiene cada variable en la formación de la componente. Además las componentes tienen norma unidad por lo que coeficientes superiores a $0{,}5$ indican una presencia significativa de la variable original. En el caso de la primera componente, todas las variables son significativas en su formación, pero se puede interpretar que un aumento de la primera componente significa que existe un mayor aporte de agua y lactosa, mientras que individuos que tengan valores negativos en la primera componente significará que poseen una mayor cantidad de grasas y proteinas. Por otro lado, la segunda componente está influenciada por la variable proteina y en menor medida por las grasas. Individuos que consigan valores altos en la segunda componente significará que poseen un alto contenido en proteinas.

Para ver con más claridad estas interpretaciones sobre los individuos y las variables originales se realiza a continuación un diagrama biplot. A continuación se escribe la instrucción en *R*:

```
biplot(princomp(mamiferos[,1:4]))
```



Ejercicios

1. Utilizando el fichero de datos *UScereal*, que describe la composición de 65 tipos de cereales, incluido en el paquete *MASS*, responde a las siguientes cuestiones:
   - ¿Qué variables están más correlacionadas?.
   - ¿Tiene sentido aplicar la técnica de componentes principales?.
   - Utiliza los diferentes criterios para seleccionar el número adecuado de componentes a retener, justifica tu elección.
   - Interpreta (si es posible) las componentes seleccionadas para el estudio.
   - Representa un gráfico biplot de los individuos y las variables.
2. Responder a las mismas cuestiones del ejercicio anterior usando el fichero *Depredations* del paquete *car*. El conjunto de datos se corresponde con 434 granjas de *Minnesota* que sufrieron ataques de lobos durante el periodo $1976 - 1998$.

9

Series Temporales.

Contenidos

En esta práctica se van a plantear y resolver algunas cuestiones referentes a las series temporales. En primer lugar se realizará una visión genérica y gráfica de la serie de muestra, para a continuación hacer un análisis de la serie y estudiar

sus principales características (autocorrelación, tendencia, homocedasticidad y estacionalidad). Para terminar se transformará la serie en una serie estacionaria y se elegirá un modelo adecuado.

Finalmente se propondrán una serie de actividades similares que puedan facilitar al alumno la asimilación de los objetivos planteados.

9.1 Objetivos

- Conseguir visualizar conjuntamente los registros de una serie temporal.
- Manejar las funciones y los paquetes necesarios para el estudio de las series temporales.
- Analizar una serie temporal y sus principales características.
- Transformar una serie temporal en otra serie estacionaria.
- Interpretar y extraer conclusiones de los análisis realizados.
- Evaluar los distintos modelos y decidirse por un modelo apropiado a partir de los datos empíricos.

9.2 Descripción del conjunto de datos: Ipi inglés

El conjunto de datos con los que se va a trabajar en esta práctica queda recogido en un archivo de texto nombrado *Ipi_Ingles.txt*.

El IPI son las siglas del Índice de Producción Industrial, siendo este un indicador coyuntural que mide la evolución mensual de la actividad productiva de las ramas industriales. Mide, por tanto, la evolución conjunta de la cantidad y de la calidad, eliminando la influencia de los precios.

Habitualmente, el IPI de un país esta muy relacionado con su estado medioambiental. Lo ideal sería que un país con un elevado índice de producción mantenga

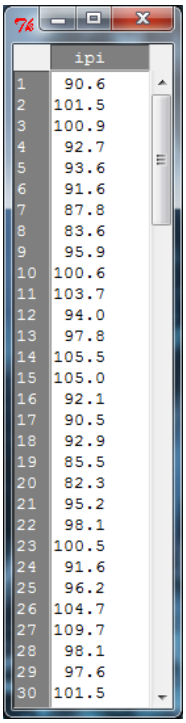
un notable estado medioambiental, que vendría a determinar que el desarrollo de dicho país resulta ser de naturaleza sostenible. La realidad presenta otro escenario, siendo los países con mayor índice de producción aquellos con mayor riesgo y deficiencias medioambientales.

Rcmdr

Para abrir el conjunto de datos Ipi_Ingles.txt en Rcmdr, se puede usar la opción del menú Datos → Importar datos → desde archivo de texto portapapeles o URL...

El histórico queda recogido en una única variable que contiene, con periodicidad mensual, cada uno de los valores del IPI inglés desde inicio de 1983.

Al visualizar los datos se mostraría:



	ipi
1	90.6
2	101.5
3	100.9
4	92.7
5	93.6
6	91.6
7	87.8
8	83.6
9	95.9
10	100.6
11	103.7
12	94.0
13	97.8
14	105.5
15	105.0
16	92.1
17	90.5
18	92.9
19	85.5
20	82.3
21	95.2
22	98.1
23	100.5
24	91.6
25	96.2
26	104.7
27	109.7
28	98.1
29	97.6
30	101.5

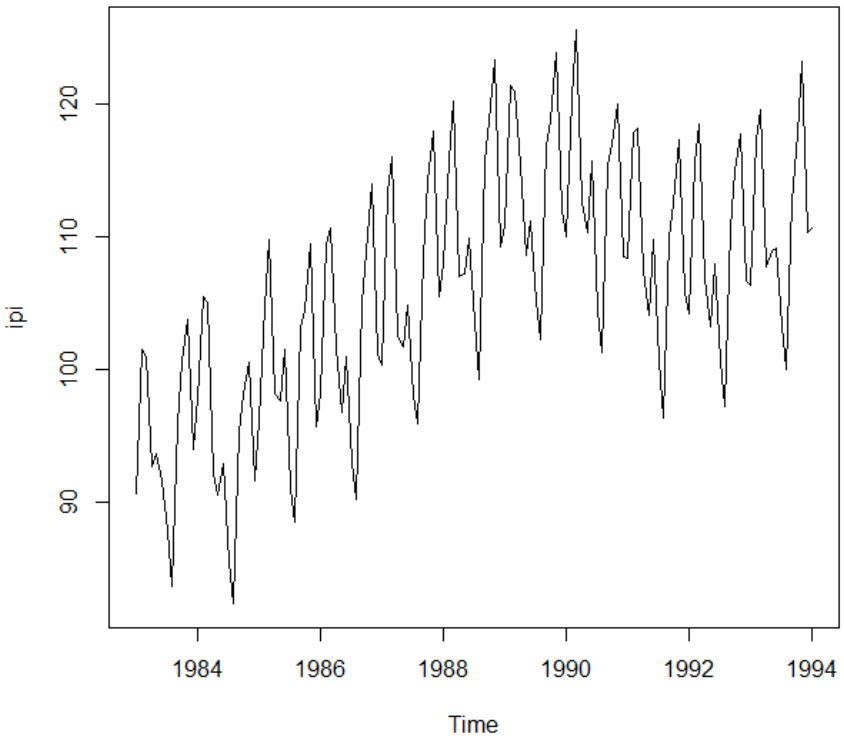
9.3 Presentación de los datos y representación gráfica

Se comenzará creando una variable, que se llamará *ipi*, con estructura de serie temporal de frecuencia mensual. A continuación se representará gráficamente la serie en un gráfico temporal y en un gráfico por periodos. A partir del gráfico temporal, se puede observar la existencia de una tendencia alcista en el largo plazo, y una tendencia alcista que pasa a ser bajista y al final vuelve a ser alcista en el medio plazo. En ambas gráficas se puede detectar la existencia de estacionalidad en la serie.

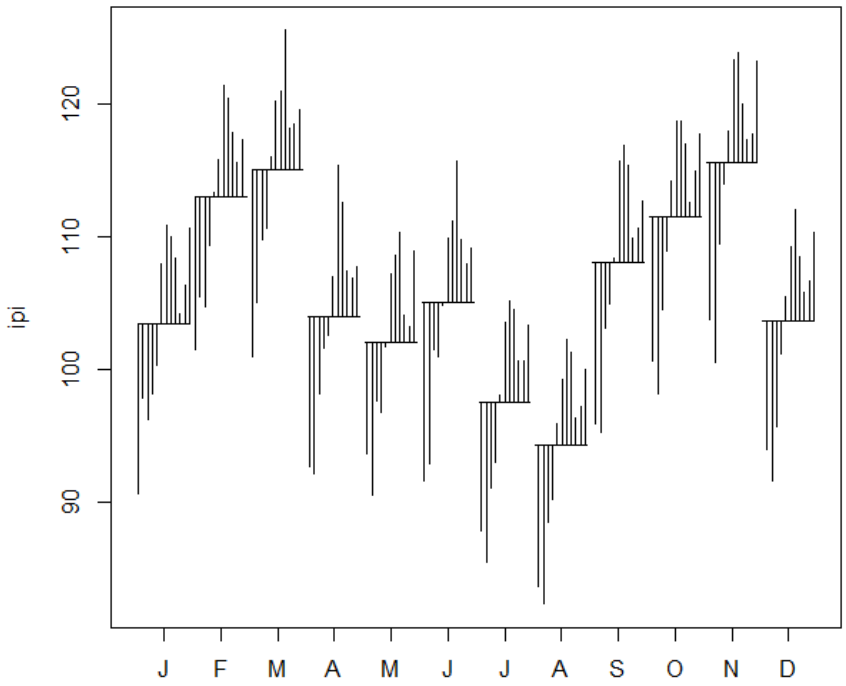
Rcmdr

Para crear en R una variable con estructura de serie temporal de frecuencia mensual, se podría ejecutar la instrucción
"ipi< −ts(Ipi_ingles$ipi,start= 1983,freq= 12)".

Para representar la serie con un gráfico temporal, se puede ejecutar la instrucción
"plot(ipi)".



Para representar la serie con un gráfico por periodos, se puede ejecutar la instrucción
"monthplot(ipi, type="h")".



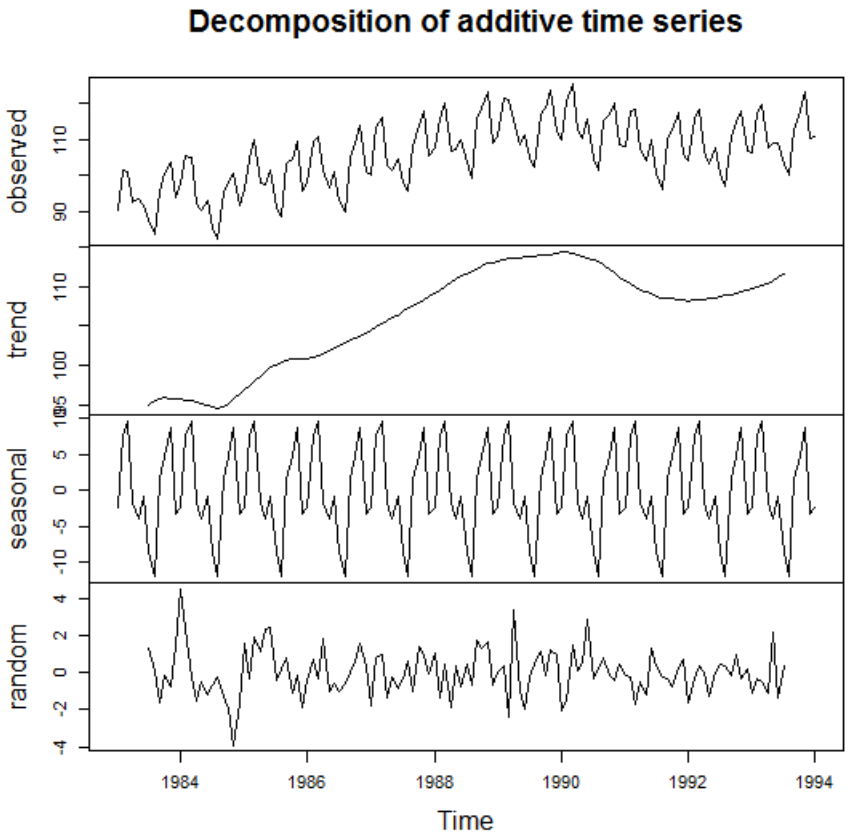
9.4 Descomposición de la serie

Existe un análisis implementado en R que automáticamente analiza y descompone la serie observada en sus componentes de tendencia, estacionalidad y ruido.

Rcmdr

Para descomponer la serie en R y guardar todas las componentes en una variable, se puede ejecutar la instrucción
"ipidesc< −decompose(ipi)".

La instrucción anterior permite extraer cada una de las componentes por separado. También permite visualizar las componentes gráficamente ejecutando la instrucción
"plot(ipidesc)".


Decomposition of additive time series
observed
trend
seasonal
random
1984
1986
1988
1990
1992
1994
Time

9.5 Análisis de la autocorrelación

En el análisis de la autocorrelación se pasará a calcular los coeficientes de autocorrelación simple y los coeficientes de autocorrelación parcial. Para mayor comodidad, se presentarán gráficamente. También se aplicará un test para los coeficientes de autocorrelación.

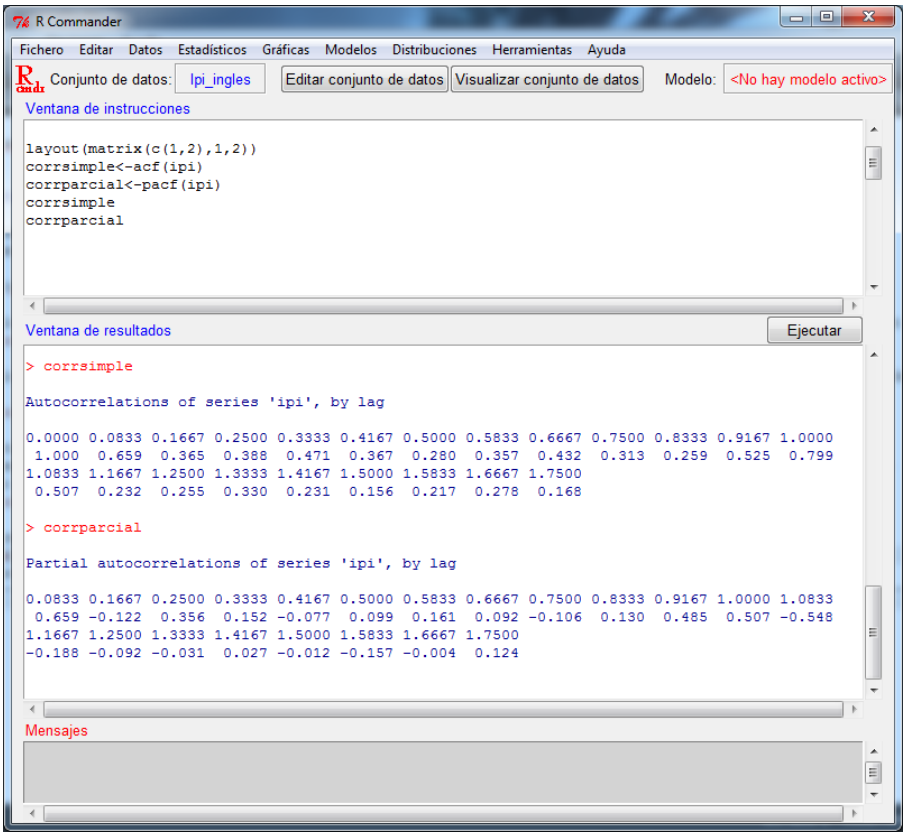
Rcmdr

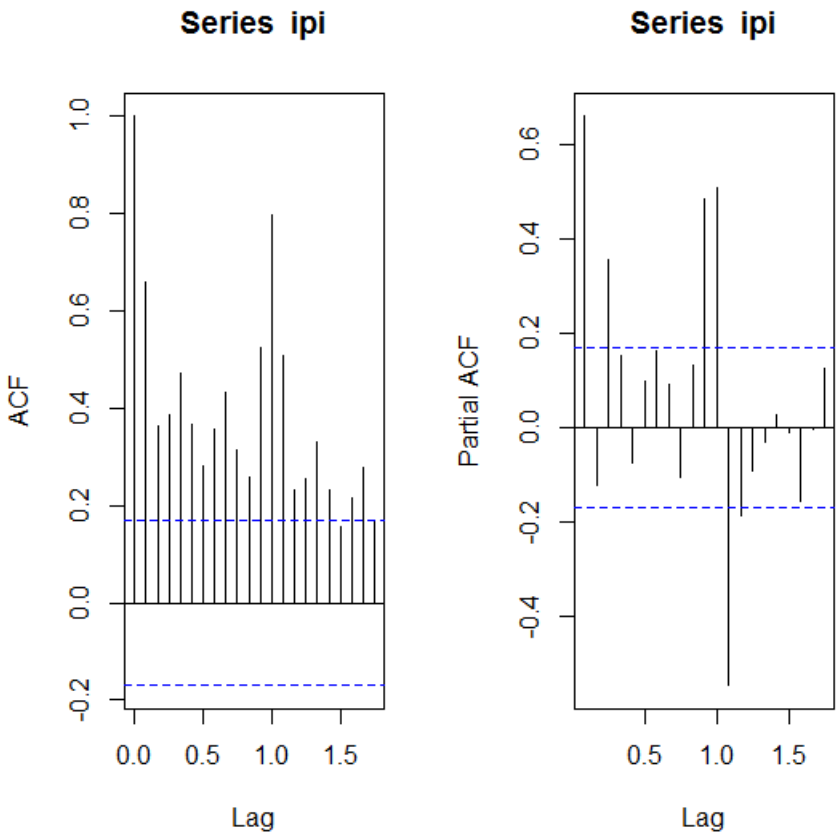
Para calcular los coeficientes de autocorrelación simple de la serie con R, se puede ejecutar la instrucción
"acf(ipi)".

Para calcular los coeficientes de autocorrelación parcial de la serie con R, se puede ejecutar la instrucción
"pacf(ipi)".

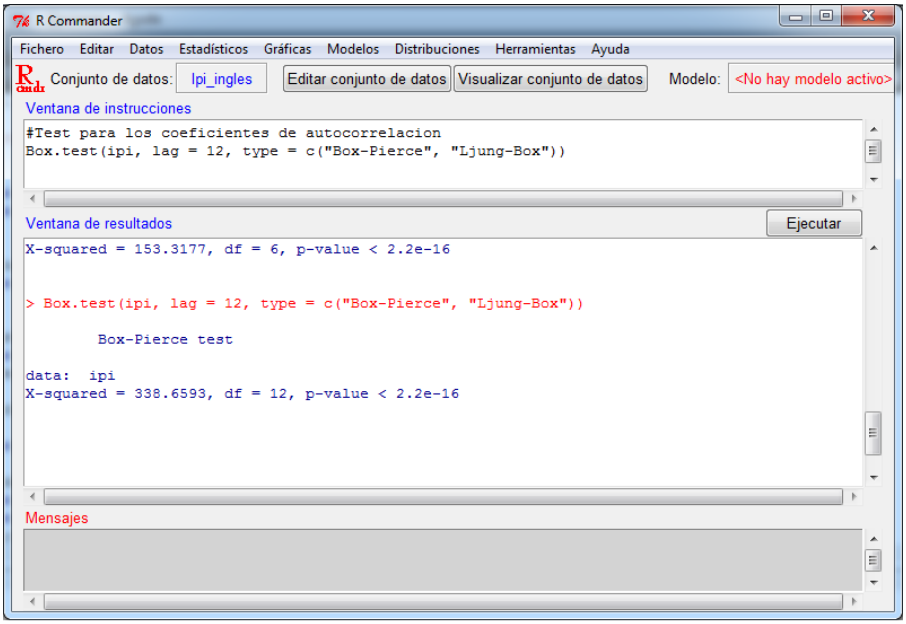
Las instrucciones anteriores permiten visualizar las gráficas de los coeficientes, pero si se introducen en variables, los valores serán almacenados(y se podrá acceder a ellos posteriormente). Una serie de instrucciones para el proceso completo podría ser la que sigue
"layout(matrix(c(1,2),1,2))
corrsimple<-acf(ipi)
corrparcial<-pacf(ipi)".

La función "layout" sirve para dividir el dispositivo gráfico en tantas filas y columnas como se indican en la dimensión de la matriz, con las anchuras de columna y las alturas de fila especificadas en los argumentos respectivos.


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Conjunto de datos: Ipi_ingles
Editar conjunto de datos
Visualizar conjunto de datos
Modelo: <No hay modelo activo>
Ventana de instrucciones
layout(matrix(c(1,2),1,2))
corrsimple<-acf(ipi)
corrparcial<-pacf(ipi)
corrsimple
corrparcial
Ventana de resultados
Ejecutar
> corrsimple
Autocorrelations of series 'ipi', by lag
0.0000 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000
1.000 0.659 0.365 0.388 0.471 0.367 0.280 0.357 0.432 0.313 0.259 0.525 0.799
1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500
0.507 0.232 0.255 0.330 0.231 0.156 0.217 0.278 0.168
> corrparcial
Partial autocorrelations of series 'ipi', by lag
0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167 1.0000 1.0833
0.659 -0.122 0.356 0.152 -0.077 0.099 0.161 0.092 -0.106 0.130 0.485 0.507 -0.548
1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500
-0.188 -0.092 -0.031 0.027 -0.012 -0.157 -0.004 0.124
Mensajes



Para aplicar el test para los coeficientes de autoorrelación, se puede ejecutar la instrucción
"Box.test(ipi, lag = 6, type = c("Box-Pierce", "Ljung-Box"))".



En los gráficos de ACF y PACF se pueden observar evidencias de una posible existencia de autocorrelación en las observaciones con separación de 12 retardos. Al aplicar el test para la autocorrelación de orden 12, se obtiene en la salida de R que $p - value < 0{,}05$, por lo que se rechaza la hipótesis nula, luego existe autocorrelación.

9.6 Tendencia

Como ya se comentó, con una simple vista al gráfico temporal se detectan periodos claros de tendencia. Dichos periodos pueden ser visualizados con mayor

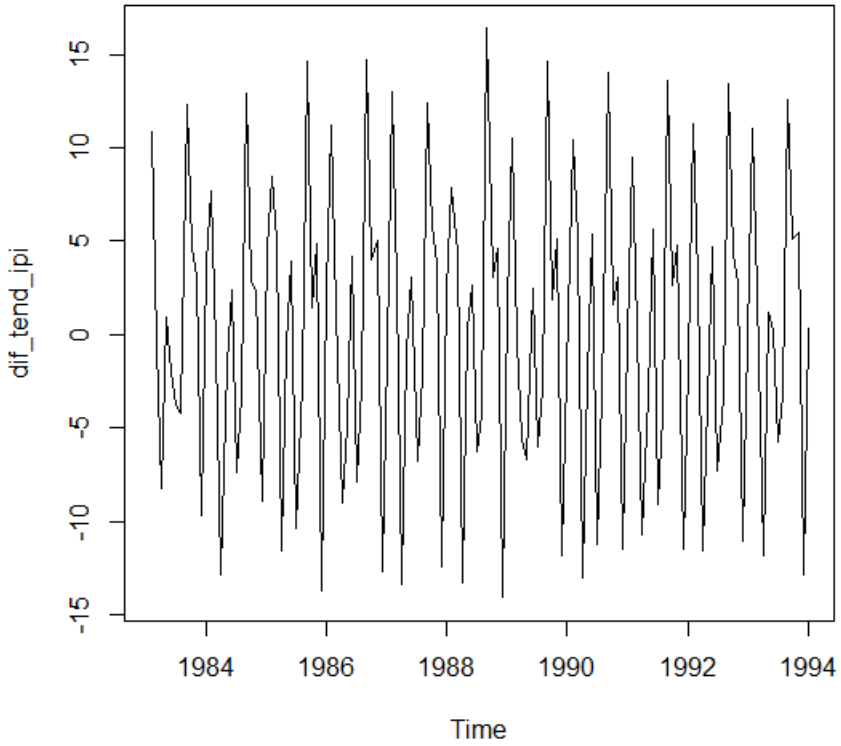
claridad en la componente de tendencia (*trend*) extraida por la función "*decompose*".

Para eliminar la tendencia de la serie, se pueden aplicar diferencias en la serie original. En este caso basta con una diferencia de orden 1.

Rcmdr

Para aplicar diferencias de orden 1 *a la serie con R, se podría ejecutar la instrucción*
"dif_tend_ipi< −diff(ipi, lag= 1)".

Para extraer el gráfico de la nueva serie, se puede ejecutar la instrucción
"plot(dif_tend_ipi)".



Se puede observar en la última gráfica que, aplicar diferencias de orden 1, ha sido suficiente para eliminar la tendencia de la serie.

9.7 Estacionalidad

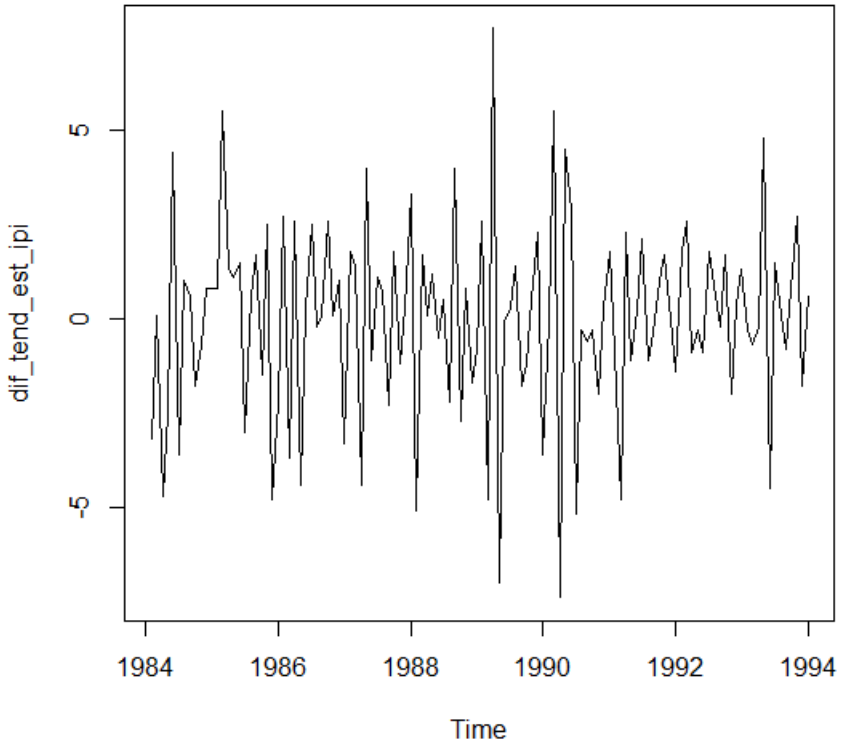
Se puede observar que, aunque haya desaparecido la tendencia, se conserva el ciclo. El ciclo se puede apreciar en el gráfico temporal de la serie, y en el ACF, ya que los palos separados por 12 retardos no terminan de bajar.

Para intentar eliminar la estacionalidad, se podría intentar aplicar diferencias de orden 12 en la serie.

Rcmdr

Para aplicar diferencias de orden 12 a la serie con R, se podría ejecutar la instrucción
"dif_tend_est_ipi< −diff(dif_tend_ipi, lag= 12)".

Para extraer el gráfico de la nueva serie, se puede ejecutar la instrucción
"plot(dif_tend_est_ipi)".



Se puede observar en la última gráfica que aplicar diferencias de orden 12 ha sido suficiente para eliminar la estacionalidad de la serie.

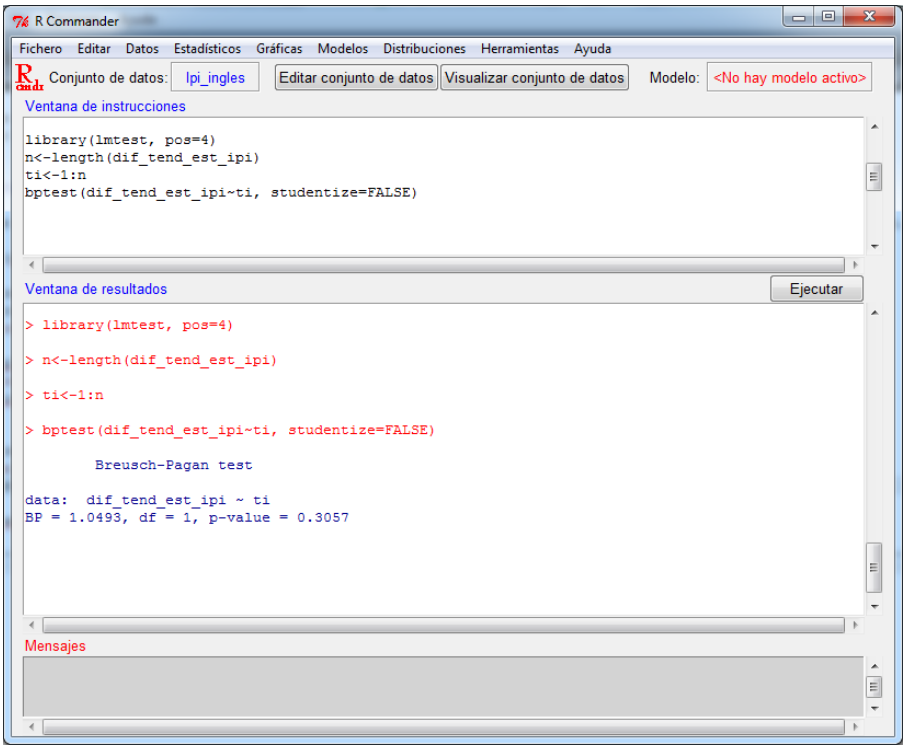
9.8 Homocedasticidad

Ahora se pasará a revisar la homocedasticidad de la serie. Para ello se usará el contraste de Breusch−Pagan.

Rcmdr

El contraste de Breusch−Pagan se encuentra en el paquete de R lmtest, que se podría cargar usando la instrucción
"library(lmtest, pos= 4)".

Para aplicar el contraste anterior en R, se puede ejecutar la siguiente serie de instrucciones
"n< −length(dif_tend_est_ipi)
ti< −1:n
bptest(dif_tend_est_ipi∼ti, studentize=FALSE)".



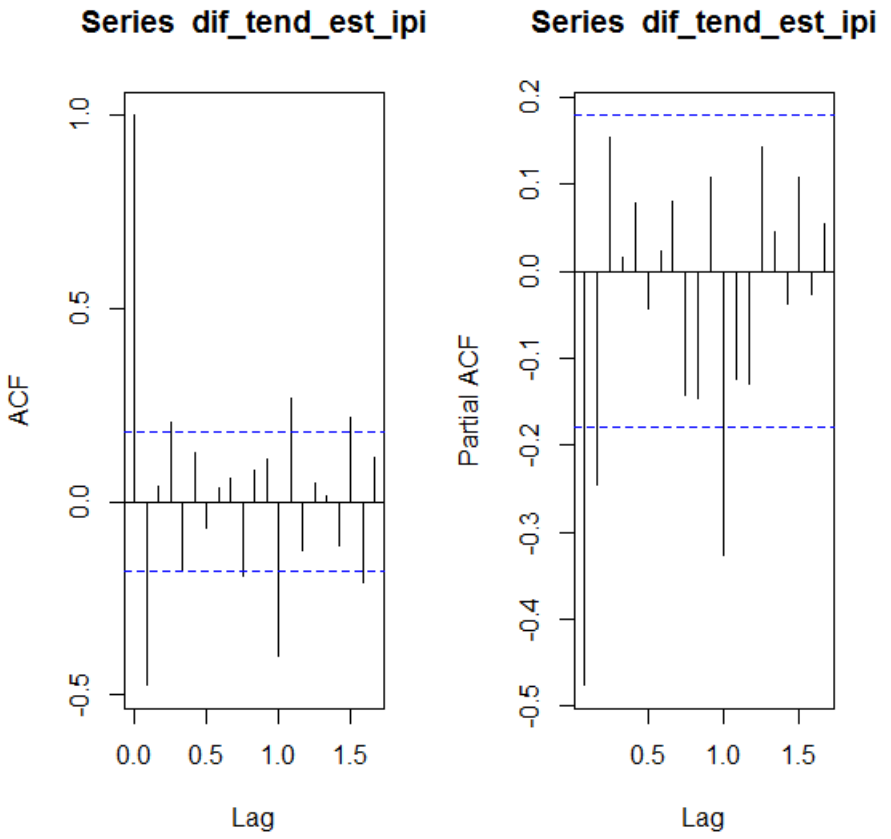
Por lo anterior, se rechaza la hipótesis de heterocedasticidad.

9.9 Elección del modelo

Para comenzar, se volverán a revisar los gráficos de autocorrelación simple y autocorrelación parcial.

Rcmdr

Para calcular los coeficientes de autocorrelación simple y autocorrelación parcial de la nueva serie con R, se pueden ejecutar las instrucciones "layout(matrix(c(1, 2), 1, 2))
acf(dif_tend_est_ipi)
pacf(dif_tend_est_ipi)".

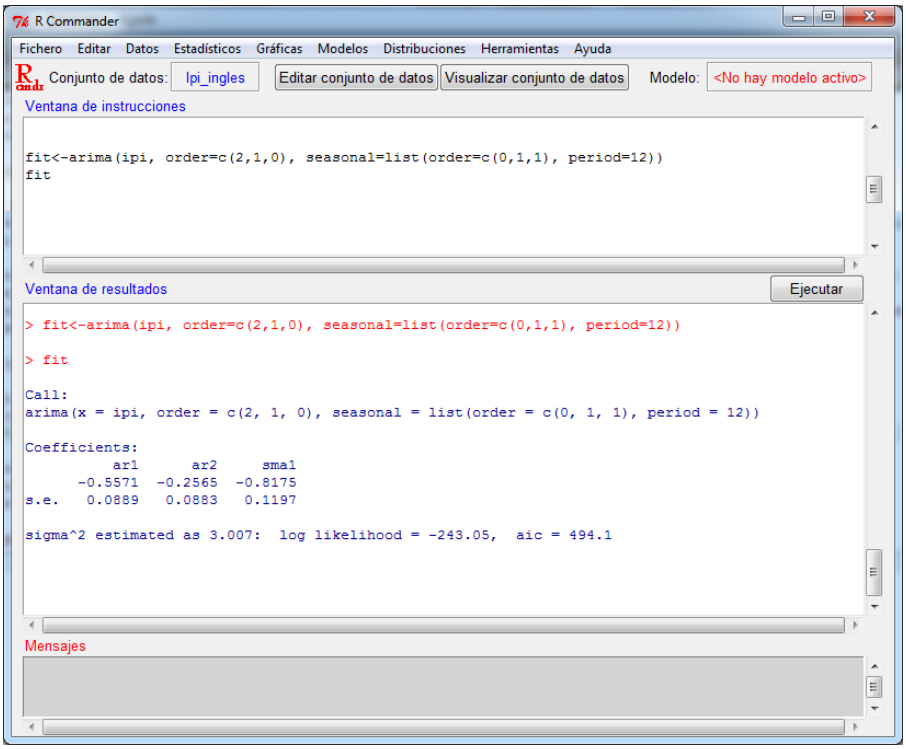


En la gráfica de ACF se pueden observar grandes palos hasta el quinto retardo, y en la gráfica de la PACF se pueden ver sólo dos palos significativos. Por lo tanto, la situación podría corresponderse con la de un modelo AR(2) en la parte regular.

A continuación, observando los palos estacionales, en la gráfica de ACF se puede observar que destaca el palo del nivel 12, pero no el palo que corresponde al nivel 24, ni el que corresponde al nivel 36. Por otro lado, en la gráfica de la PACF se aprecian el palo 12 y el palo 24. Por lo tanto, la situación podría corresponderse con la de un modelo MA(1)12.

Rcmdr

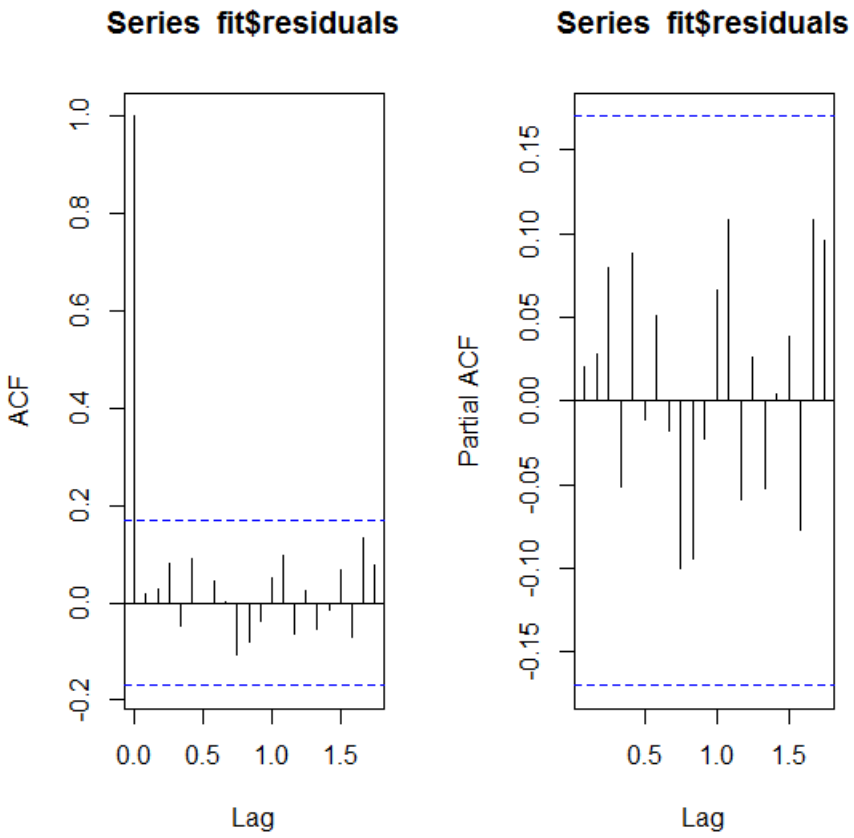
Para calcular el modelo ARIMA comentado con R, se puede ejecutar la instrucción
"fit<−arima(ipi, order= $c(2,1,0)$*, seasonal=list(order=* $c(0,1,1)$*, period=* 12*))"*.



Se puede ver que los residuos forman un auténtico ruido blanco.

Rcmdr

Para analizar los residuos con R, se pueden ejecutar las instrucciones "acf(fit$residuals) pacf(fit$residuals)".



Parece que las gráficas de ACF y PACF no muestran ningún patrón significativo.

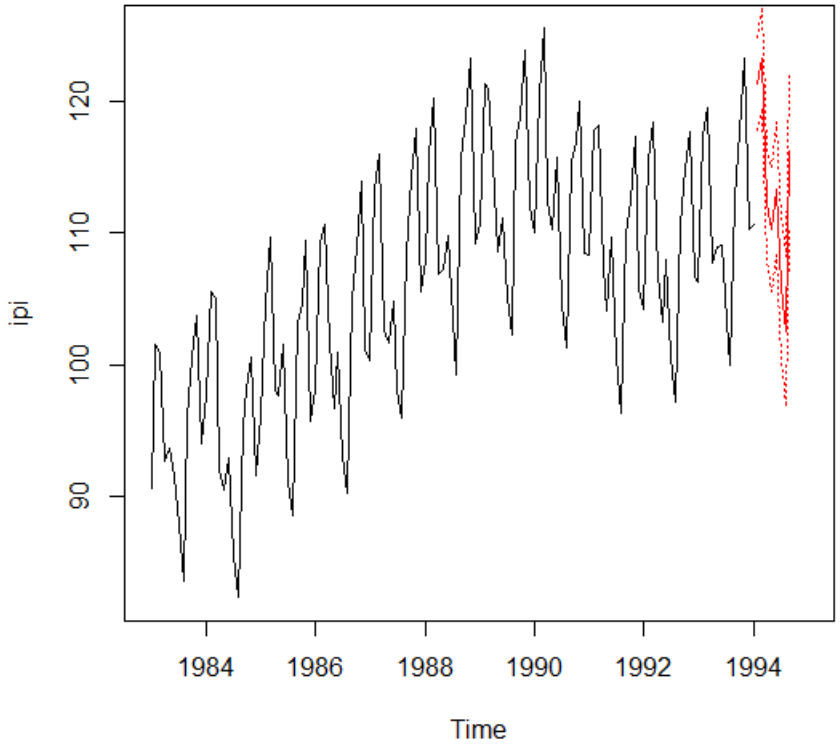
9.10 Predicciones

Tras la elección del anterior modelo de ajuste, se pueden realizar predicciones para valores futuros de la serie.

Rcmdr

Para realizar las predicciones y guardarlas en una variable con R, se puede ejecutar la instrucción
"ipi.pred< −predict(fit,n.ahead= 8)".

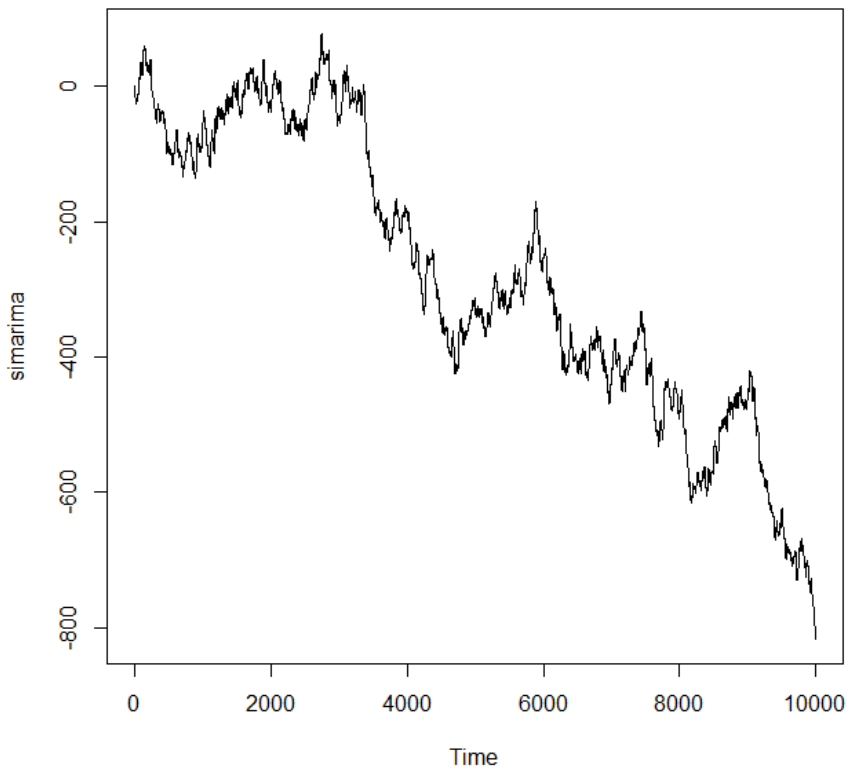
Para incorporar las predicciones realizadas, incluyendo bandas de confianza, en el gráfico temporal de la serie, se podrían ejecutar las siguientes instrucciones con R "plot(ipi,xlim= c(1983, 1995))
lines(ipi.pred$pred,col="red")
lines(ipi.pred$pred+2∗ipi.pred$se,col="red",lty= 3)
lines(ipi.pred$pred−2∗ipi.pred$se,col="red",lty= 3)".



9.11 **Simulación**

Rcmdr

Para simular con R una serie que siga un modelo ARIMA prefijado, se puede ejecutar la instrucción
"simarima< −arima.sim(10000,*model=list(order=* $c(2,1,0)$, *ar=* $c(0{,}7,0{,}1)$, *sd=* $1)$)*"*.



Ejercicios

1. Realizar un estudio completo de los datos de la serie temporal AirPassengers.txt.
2. Realizar un estudio completo de los datos de la serie temporal Ipi_Aleman.txt.
3. Realizar un estudio completo de los datos de la serie temporal Paro_españa.txt.