Estadística. Práctica 1

1.1 Iniciando R

R es un paquete estadístico al tiempo que un lenguaje de programación. Estas características dotan a R de una gran versatilidad.

R es software libre con licencia GNU. Existen distintas interfaces gráficas (GUI) que simplifican su manejo. Una de estas interfaces gráficas es R-Commander y será la interface que usaremos en las prácticas del curso.

R-UCA es una instalación de R y R-Commander conjuntamente que ha sido desarrollado por un equipo de la Universidad de Cádiz. La versión con la que trabajaremos este curso puede descargarse de

http://knuth.uca.es/R/R-UCA-3.6.2.exe

1.1.1 Entorno básico de R

Cuando ejecutamos R aparecen la "R Console", que nos permite trabajar en modo consola y una ventana correspondiente a "R Commander", que es la que usaremos principalmente.

En R Commander podemos encontrar, además de la barra de menús, tres ventanas:

- Instrucciones (R-script).
- Resultados (salida).
- Mensajes.

1.1.2 Operadores de R

En ocasiones, como veremos más adelante, hay que utilizar algún operador al trabajar con R. Los más usuales son los siguientes:

Aritméticos	**Comparativos**	**Lógicos**
Suma: +	Igualdad: ==	Y lógico: &
Diferencia: -	Distinto: !=	No lógico: !
Producto: *	Menor que: <	O lógico: |
División: /	Mayor que: >	
Potencia: ^	Menor o igual: <=	
	Mayor o igual: >=	

1.2 Introducción directa de datos

Se realiza a través de `Datos` → `Nuevo conjunto de datos`. A continuación, tras darle nombre al conjunto de datos que se va crear, aparece una especie de hoja de cálculo donde se nombran las variables, se elige el tipo (numérico o carácter) y se introducen los datos.

Ejercicio: Introducir la siguiente información relativa a un grupo de 8 personas:

Estado civil	Edad
soltero/a	32
soltero/a	28
casado/a	41
separado/a	45
viudo/a	52
casado/a	62
casado/a	35
separado/a	40

Para guardar los datos introducidos seleccionar `Datos` → `Conjunto de datos activo` → `Guardar el conjunto de datos activos`. A continuación le damos un nombre (con extensión .RData) y un directorio donde se almacenará.

1.3 Trabajar con un archivo ya existente

Podemos acceder a ficheros bastante más amplios e interesantes creados previamente a través de `Datos` → `Cargar conjunto de datos`, a través de `Datos` → `Importar datos` o a través de `Datos` → `Conjunto de datos en paquetes` (son conjunto de datos que vienen con la instalación de R y que pueden ser utilizados. Por ejemplo, abrir el archivo **PlantGrowth** del paquete **datasets**.).
En moodle están colgados dos archivos con los mismos datos pero con distinto formato. En primer lugar los descargamos de moodle y los guardamos en el escritorio, por ejemplo. El archivo **world95** es un archivo de SPSS por lo que para abrirlo desde R habría que importarlo. El archivo **Rworld95** es un archivo de R por lo que se puede cargar directamente.

Observación: Los datos denotados por NA son *datos no disponibles o perdidos* (non available).

1.4 Filtrar el conjunto de datos activo

A través de `Datos` → `Conjunto de datos activo` se puede gestionar el conjunto de datos activo. Una de las opciones es `Filtrar el conjunto de datos activo`. Con esta opción podemos extraer del archivo los datos que nos interesen para trabajar, guardndose en un nuevo conjunto de datos.

Ejemplo: Queremos extraer los datos de la variable *densidad* (n° de habitantes por km^2) de los países de Europa Oriental. Para ello seleccionamos `Datos` → `Conjunto de datos activo` → `Filtrar el conjunto de datos activo`. Desmarcamos `Incluir todas las variables`, seleccionamos la variable *densidad* y en `Expresión de selección` escribimos `región=="Europa Oriental"` (entre comillas al no ser un número)y le damos nombre al nuevo conjunto de datos, por ejemplo, **filtrado1**.

Ejercicio: Extraer los datos de los países con tasa de mortalidad (variable *tasa_mor*) inferior al 10%.

Ejercicio: Extraer los datos de las variables *alfabfem* y *alfabmas* (tasa de alfabetización femenina y masculina, respectivamente) de los países que no son de la región OCDE.

1.5 Calcular una nueva variable

A través de `Datos` → `Modificar variables del conjunto de datos activo` se puede realizar modificaciones en los datos. Nos centraremos ahora una de las opciones disponibles: `Calcular una nueva variable`, que permite obtener nuevas variables a partir de las existentes. Las otras opciones de este menú las iremos conociendo y utilizando en otras prácticas.

Ejemplo: Queremos calcular la diferencia entre la esperanza de vida femenina (*espvidaf*) y masculina (*espvidam*). Para ello seleccionamos `Datos` → `Modificar variables del conjunto de datos activo` → `Calcular una nueva variable`. Tras darle nombre a la nueva variable, por ejemplo *dif_ev*, introducimos en `Expresión a calcular` la definición de tal variable, en este caso será *espvidaf-espvidam*. Aceptando se obtiene la nueva variable en la última columna de la tabla de datos.

1.6 Estadística descriptiva de una variable: tablas de frecuencias y gráficos para datos no agrupados

A través de `Estadísticos` → `Resúmenes` → `Distribución de frecuencias` podremos construir una tabla de frecuencias de una o varias variables. Por ejemplo, construir tablas de frecuencias para las variables *región* y *clima*. Observamos que se construye también una tabla de porcentajes para cada variable.

Mediante el menú `Gráficas` → `Gráfica de barras` y el menú `Gráficas` → `Gráfica de sectores` podemos obtener, respectivamente, gráficas de barras y de sectores para las variables anteriores. Los gráficos aparecen en la ventana "R-Graphics".

Si queremos una tabla de frecuencias o un gráfico de barras o de sectores para una variable numérica, ésta hay que convertirla previamente en una variable cualitativa a través de `Datos` → `Modificar variables del conjunto de datos activo` → `Convertir variable numérica en factor`. Al realizar este procedimiento conviene guardar la nueva variable con otro nombre para conservar la variable numérica original.

Por ejemplo, construir una tabla de frecuencias y gráficos de barras y sectores para la variable *calorías* (ingesta diaria de calorías) tras convertirla en una variable factor de nombre *factorcalorías*. Observamos que ni la tabla ni los gráficos obtenidos son adecuados. Ello es debido a que la variable *calorías* presenta una gran cantidad de valores distintos por lo que para estudiarla adecuadameante habría que agruparla en intervalos de valores. En la próxima práctica veremos cómo hacerlo.