Taller de Estadística

# Descripción de datos bivariantes

El objetivo de esta práctica es familiarizarse con las técnicas de descripción de datos bidimensionales y con algunas de las opciones del programa `SPSS/PC` relacionadas con estas técnicas. Para los ejemplos se utilizará el fichero `best-x3.2` correspondiente al estudio de la esperanza de vida en hombres y mujeres.

## 1. ¿Cómo introducir datos en SPSS/PC?

A continuación se ilustra como introducir los datos, para el ejemplo `best-x3.2`:

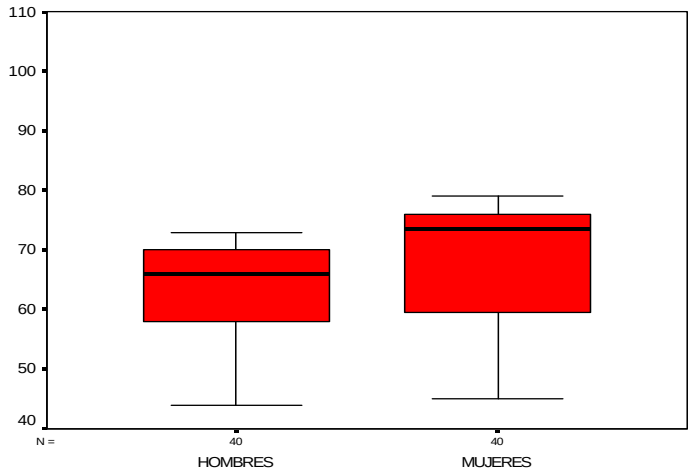
1. `Archivo` ↪ `Abrir` ↪ `Datos`.
2. En el cuadro de diálogo que aparece, en `Archivos de tipo`, elegir la extensión `Todos los archivos (*.*)`, seleccionar el archivo `best-x3.2` ↪ `Abrir`.
3. En el cuadro de `Asistente para importación de texto`, elegimos las opciones teniendo en cuenta que los datos están almacenado: 1 caso por línea, las variables están separadas por un espacio, y en el fichero no aparece el nombre de las variables.
4. Para cambiar el nombre a las variables, seleccionamos `Ver` ↪ `Variables`, y cambiamos `v1` por `pais`, `v2` por `hombres` y `v3` por `mujeres`.
5. Por último, `Archivo` ↪ `Guardar`.

## 2. Diagrama de cajas múltiple

Para estudiar la relación entre una variable cualitativa y una cuantitativa continua se utilizan los diagramas de cajas múltiples. En el ejemplo, vamos a estudiar la relación entre la esperanza de vida (variable cuantitativa continua) y el sexo (variable cualitativa). Los pasos a seguir son:

1. Se selecciona `Gráficos` ↪ `Diagramas de cajas`.
2. Se eligen las opciones `Simple` y luego `Resúmenes para distintas variables`, y finalmente `Definir`.
3. Para el cuadro `Las cajas representan:` se seleccionan las variables `hombres` y `mujeres`. Finalmente, se pulsa `Aceptar`.

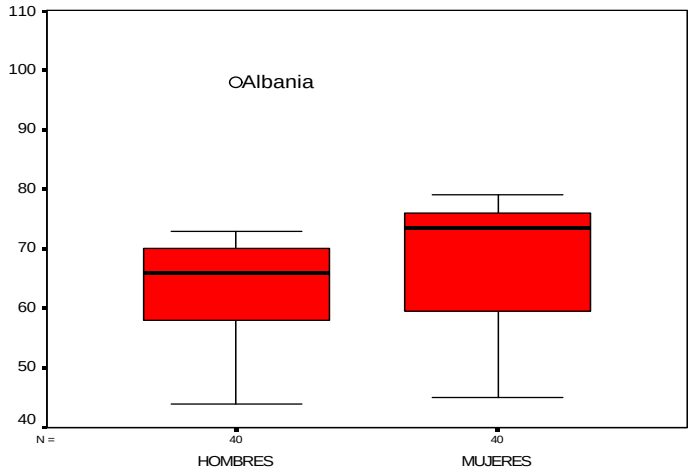
En el gráfico obtenido, podemos comparar la posición, la dispersión y la asimetría para la distribución de la esperanza de vida para los hombres y para las mujeres.




Si se desea identificar los casos en la muestra con alguna etiqueta, en el ejemplo podemos utilizar la variable `pais`, entonces modificamos el paso 3 anterior a:

3. Para el cuadro `Las cajas representan:` se seleccionan las variables `hombres` y `mujeres`, y para el cuadro `Etiquetar los casos mediante:` seleccionamos `pais`.

Para ilustrar la utilidad de este etiquetado, vamos a cambiar el dato correspondiente a la variable `hombres` en el primer país: $\boxed{68} \hookrightarrow \boxed{98}$, y obtenemos los "nuevos" diagramas de cajas:




# 3. Diagrama de dispersión y recta de regresión

Para estudiar gráficamente la relación existente entre dos variables los pasos a seguir son:

1. Se selecciona `Gráficos` ↪ `Dispersión`.

2. Como `Eje Y:` debemos elegir la variable que queremos explicar. En el ejemplo, tomaremos la esperanza de vida de los `hombres`. Como `Eje X:` elegimos las variables explicativas, `mujeres`.

3. Si queremos identificar cada punto en el gráfico debemos introducir la variable con las etiquetas que identifican a cada par de datos (en este caso `pais`), es decir `Etiquetar los casos mediante: pais`.

4. Si queremos dibujar la *recta de mínimos cuadrados* tenemos que editar el gráfico (hacer clic dos veces sobre él):

   - `Diseño` ↪ `Opciones` ↪ `Ajustar línea` ↪ `Total`.

5. Si queremos conocer la recta se edita el gráfico y se elige:

   - `Analizar` ↪ `Regresión` ↪ `Lineal`.

**Coeficientes[a]**

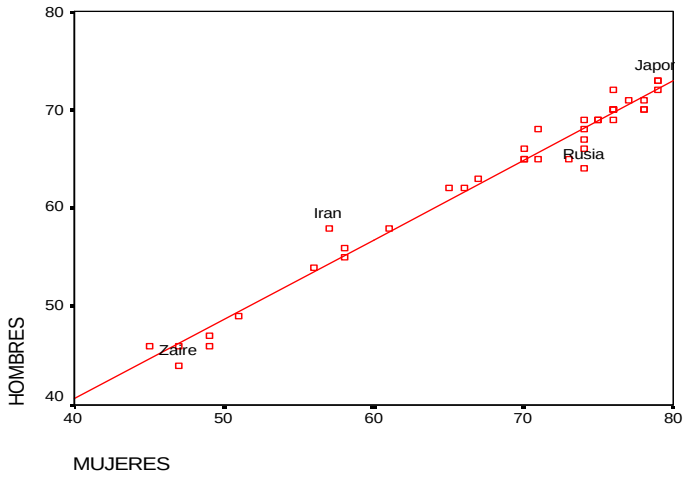
| Modelo | | Coeficientes no estandarizados | | Coeficientes estandarizados | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Error típ. | Beta | | |
| 1 | (Constante) | 8,314 | 1,458 | | 5,703 | ,000 |
| | MUJERES | ,807 | ,021 | ,987 | 38,180 | ,000 |

a. Variable dependiente: HOMBRES

Los valores del punto de corte con el eje de las ordenadas y la pendiente aparecen en la tabla `Coeficientes`, en la columna `B`. También aparece la pendiente que se obtendría si ajustáramos la recta tipificando previamente las variables.

6. Para identificar los puntos con su correspondiente etiqueta se edita el gráfico y se activa el botón de `Identificación de puntos:`

7. A continuación se sitúa el cursor sobre el punto que se quiere identificar y se hace clic. ¿Cuál es el país en el que la esperanza de vida es más baja, tanto para hombres como para mujeres? ¿Y el que tiene esperanza de vida más alta?

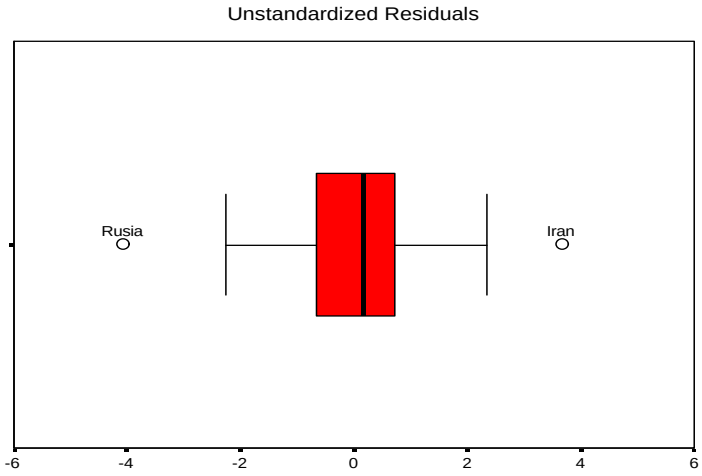



7. Si queremos obtener los residuos de la regresión, volvemos a:

   - `Analizar` ↪ `Regresión` ↪ `Lineal`.

   y en el cuadro de diálogo seleccionamos el botón `Guardar` y marcar en la columna de `Residuos` la opción `No tipificados`.

Cuando concluye el análisis debemos tener una nueva columna con los residuos en la ventana de datos. Representa los residuos mediante un diagrama de barras, ¿tienen distribución simétrica? ¿Cuánto vale la media de los residuos? ¿Cuánto vale la varianza residual? ¿Es mucho menor que la varianza de la variable respuesta `hombres`?



**Estadísticos descriptivos**

| | N | Media | Desv. típ. | Varianza |
|---|---|---|---|---|
| HOMBRES | 40 | 63,325 | 8,698 | 75,661 |
| Unstandardized Residual | 40 | -2,10E-15 | 1,3864450 | 1,922 |
| N válido (según lista) | 40 | | | |

**Ejercicio 1**. El fichero `best-x5.2` contiene tres variables de varios países: nombre del país, las emisiones de $CO_2$ per capita y el producto interior bruto (PIB) en dólares per capita, respectivamente.

**(a)** Construye el diagrama de dispersión para los datos de las emisiones y el PIB (elige como variable $y$ las emisiones de $CO_2$ per capita). Identifica los países atípicos en el gráfico.

**(b)** Dibuja la recta de mínimos cuadrados y calcula los valores de la pendiente y el término independiente.

**(c)** Calcula los residuos y construye un diagrama de dispersión de los residuos con la variable PIB. Representa un diagrama de barras de los residuos.

**(d)** Repite el ejercicio eliminando el dato correspondiente a EEUU.

En `http://halweb.uc3m.es/omar/taller/TALLER.html` dentro de Prácticas con SPSS se encuentran los ficheros de datos ASCII: `best-x3.2`, `best-x5.2` a utilizar en esta práctica.