Práctica 2. Exploración de datos II.

Universidad de Alcalá. Curso 2023-24.

Estadística (650008). Grado en Biología sanitaria.

Actualizado: 2023-10-05

Introducción

Trabajaremos con una tabla de datos procedente de una muestra de 1000 mujeres que participaron en un estudio sobre osteoporosis. El fichero contiene algunas variables auxiliares en las columnas iniciales, pero nosotros nos vamos a fijar en estas.

- edad (en años).
- peso (en kg).
- talla (altura en cm).
- imc (índice de masa corporal)
- bua (resultado de la exploración densitométrica)
- clasific (normal / osteopenia / osteoporosis)
- menarqui (edad primera menstruación, en años)
- edad_menop (edad inicio menopausia, en años)
- menopausia (sí, no)
- tipo de menopausia
- nivel educativo

Organiza tu entorno de trabajo

Empieza por descargar la tabla de datos desde aquí y guardarlos en la variable `osteo` (para que coincida con las soluciones). Si no recuerdas como hacerlo, se explica en los vídeos de trabajo previo a esta práctica.

Ejercicio 1

Calcula el peso medio de las 100 primeras mujeres de la tabla.

Ejercicio 2

Calcula el peso medio de las 100 primeras mujeres con entudios `SUPERIORES`

Ejercicio 3

Calcula la desviación típica (muestral) del `imc` de las mujeres con edad menor o igual que 50 años.

Ejercicio 4

Calcula la media de la variable `peso` para el 50% de las mujeres con menor `imc`

Ejercicio 5

Visualiza la `talla` y el `imc` de las mujeres que ocupan las posiciones 100 a la 105 y de la 120 a la 123.

Ejercicio 6

Visualiza la tabla con las columnas `imc`, `talla` y `nivel_ed` de las mujeres con menos de 48 años.

Ejercicio 7

Calcula la media y la cuasidesviación típica muestral de la variable `edad` para las mujeres con y sin menopausia. ¿A qué conclusión llegarías?

Ejercicio 8

Calcula la media de la variable `peso` en función de la variable `tipo_men`.

Ejercicio 9

Analiza los cuartiles de la variable `peso` en función de la variable `tipo_men`. Interpreta los resultados.

Ejercicio 10

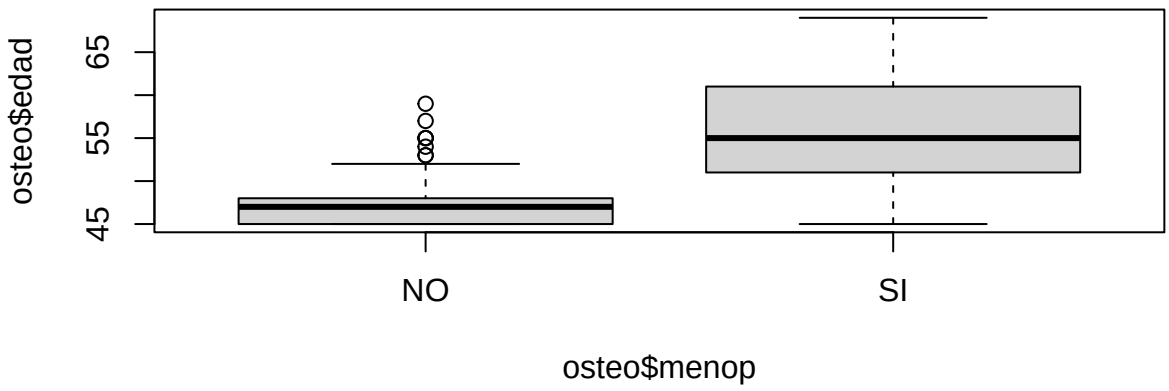
Utiliza boxplots para abordar gráficamente la cuestión anterior.

Ejercicio 11

Ejercicio guiado. Localiza los valores atípicos de la `edad` para cada nivel del factor `menop`, eliminalos y genera los boxplot correspondientes para concluir si las dispersones de la edad en cada nivel del factor son o no comparables.

Para localizar los valores atípicos tenemos que acceder a la información que genera `R` para construir el boxplot. Esto se hace guardando todo el boxplot en una variable, que llamaremos, por ejemplo, `bp`:

```
bp = boxplot(osteo$edad ~ osteo$menop)
```



Fíjate en que, en realidad, hemos generado 2 boxplots, uno para las mujeres que tienen la menopausia y otro para las que no. Puedes ver en el marco superior derecho *environment* que hay una nueva variable, `bp` que es una lista de 9 elementos. Uno de ellos contiene el valor que tienen los datos atípicos:

```
bp$out
```

```
## [1] 57 59 55 53 53 55 54 55
```

Lo siguiente es localizarlos en la tabla, para eso hay que usar la función `which`. Observa el siguiente código:

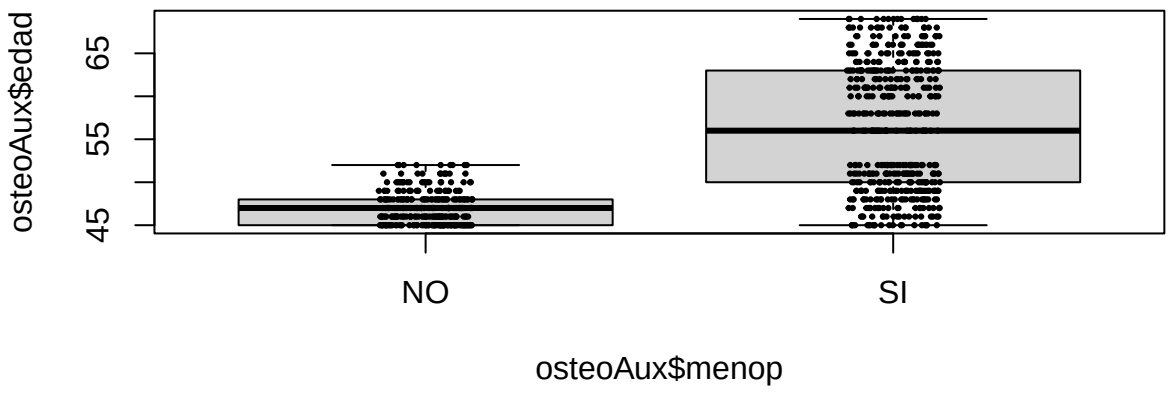
```
which(osteo$edad %in% bp$out)
```

```
##   [1]   1   4  10  12  20  22  30  36  38  39  40  41  43  47  48  50  52  58
##  [19]  63  80  84  89  92  94  95  96  97 102 105 106 112 120 122 124 126 128
##  [37] 131 133 135 137 138 139 147 154 160 162 169 185 192 194 195 199 202 204
##  [55] 205 206 208 210 212 213 215 219 221 222 223 227 232 238 239 240 246 256
##  [73] 257 258 261 262 263 281 283 285 290 292 293 296 316 319 326 327 328 331
##  [91] 338 341 343 344 345 349 351 352 353 360 362 364 368 370 372 373 378 380
## [109] 382 383 384 397 399 403 411 413 416 418 425 426 428 433 434 435 437 439
## [127] 440 441 450 456 461 474 475 478 487 501 506 515 522 523 526 530 533 541
## [145] 542 554 617 622 623 624 628 631 636 638 647 654 663 668 669 670 672 676
## [163] 678 680 687 694 698 705 709 711 722 723 726 728 733 735 760 761 767 768
## [181] 770 774 782 783 790 791 796 797 798 799 810 811 812 813 816 819 831 832
## [199] 833 834 835 864 890 898 981 998
```

Su traducción simultánea sería, literalmente: qué (`which`) mujeres tienen una edad (`osteo$edad`) que está entre (`%in%`) los valores atípicos de los boxplots (`bp$out`). Si te parece una sintaxis rara, te tenemos que dar la razón. Recuerda que no tienes que memorizarla; sólo que saber dónde encontrarla.

```
# crear nueva tabla "osteoAux" de la que se eliminan las filas de "osteo" que continen outliers
osteoAux = osteo[-which(osteo$edad %in% bp$out), ]

# reprentar los boxplots de la nueva tabla
boxplot(osteoAux$edad ~ osteoAux$menop)
stripchart(osteoAux$edad ~ osteoAux$menop,
           add = T, vertical = T, method = "jitter", pch = 19, cex = .3)
```



Aunque no es el caso, podrían aparecer nuevos outliers que, de ninguna manera, podrían ser ya considerados como tale. sería sólo un efecto secundario de haber eliminado la primera tanda de atípicos. Recuerda que no siempre es preciso eliminar los valores atípicos. Su presencia puede ser una pista inestimable en nuestro análisis.