Estadística. Práctica 2

Trabajaremos también en esta práctica con el archivo **Rworld95**.

2.1 Estadística descriptiva de una variable: tablas de frecuencias y gráficos para datos agrupados

Cuando una variable, aún siendo discreta, presenta una gran cantidad de valores distintos es necesario agruparla en intervalos tanto para construir una tabla de frecuencias como para obtener un gráfico adecuado. Por ejemplo, si, tras convertirla en factor, construimos la tabla de frecuencias de la variable *calorías* (ingesta diaria de calorías) observamos que presenta demasiados niveles y si solicitamos un diagrama de barras o de sectores se obtiene un gráfico ininteligible. El gráfico adecuado en este caso es el histograma, obtenido al agrupar en intervalos los valores de la variable.

Tras un análisis previo, decidimos agrupar en 5 intervalos de igual amplitud comenzando en 1500 y terminando en 4000. Tendremos pues intervalos de amplitud 500.

Para construir el histograma, primero dibujamos el que se obtiene por defecto seleccionando `Gráficas` → `Histograma` y elegiendo la variable para el gráfico. Ahora si queremos reflejar la agrupación anterior debemos modificar, en la ventana instrucciones (R-Script), la instrucción correspondiente cambiando `breaks="Sturges"` por `breaks=c(1500,2000,2500,3000,3500,4000)`. A continuación seleccionamos toda la instrucción modificada (toda la línea que ocupa) y pulsamos el botón `Ejecutar`.

Para obtener una tabla de frecuencias que refleje la agrupación que hemos elegido, debemos recodificar la variable a través de `Datos` → `Modificar variables del conjunto de datos activos` → `Recodificar variables`. Elegimos la variable *calorías* para recodificarla y asignamos nombre a la nueva variable, por ejemplo *ingcalr*.

A la hora de recodificar variables, la información relativa a la codificación de los valores se debe indicar en el apartado `Introducir directrices de recodificación` en la forma **valor antiguo = valor nuevo**. Si los valores antiguos o nuevos son cadenas de caracteres se pondrán entre comillas. En este caso, los rangos de valores para la construcción de los intervalos se consideran como valores antiguos y se deben escribir en la forma: **extremo izquierdo del intervalo:extremo derecho del intervalo**. Como valores nuevos asignaremos, por ejemplo, los puntos medios de los intervalos. Así pues, nuestras directrices de recodificación son:

1500:2000=1750
2000:2500=2250
2500:3000=2750
3000:3500=3250
3500:4000=3750

Observación: R considera el primer intervalo cerrado y el resto los considera abiertos por la izquierda y cerrados por la derecha.

Para la nueva variable, construir la tabla de frecuencias, así como el diagrama de barras y de sectores.

Observación: A través de `Datos` → `Modificar variables del conjunto de datos activo` → `Segmentar variable numérica` se pueden agrupar los datos de una variable numérica en intervalos de valores, pero este procedimiento no sustituye al que hemos realizado antes, ya que aunque se pueda elegir, como en el histograma, que se haga mediante intervalos iguales (equidistantes), no se puede indicar en qué valor ha de comenzar la segmentación. Además, se puede observar que la segmentación en intervalos equidistantes en realidad no es exactamente así. Para comprobarlo hacer la segmentación en 5 intervalos equidistantes de la variable *calorías* seleccionando `rangos` en el apartado `nombres de niveles` y construir a continuación una tabla de frecuencias de esa nueva variable para observar la segmentación realizada.

2.2 Estadística descriptiva de una variable: medidas descriptivas

A través de `Estadísticos` → `Resúmenes` → `Conjunto de datos activo` obtenemos un resumen del conjunto de datos activo en el que se muestran:

- Para las variables cuantitativas: mínimo, primer cuartil, mediana, media, tercer cuartil, máximo y número de datos perdidos.
- Para las variables cualitativas: nombres de las modalidades de la variable con su respectiva frecuencia absoluta y también el número de datos perdidos.

El cálculo de las principales medidas descriptivas de una variable numérica lo realizaremos a través de `Estadísticos` → `Resúmenes` → `Resúmenes numéricos`. A continuación, se elige(n) la(s) variable(s) bajo estudio y se seleccionan las medidas (estadísticos) que interesen. Se tiene además la posibilidad de obtenerlas según las modalidades definidas por otra variable.

Observación: La desviación típica y la varianza que calcula R se corresponden con la cuasidesviación típica y la cuasivarianza definidas en los apuntes de Estadística Descriptiva. En lo que sigue, cuando se solicite calcular desviación típica o varianza de una variable, nos referimos a las calculadas por R. Respecto a los coeficientes de asimetría y apuntamiento, los que se corresponden con los de los apuntes son los de `tipo 1`.

Ejercicio: *Calcular e interpretar las siguientes medidas descriptivas de las variable calorías: media, mediana, moda*[1]*, cuartiles, percentil 30(equivalentemente, cuantil 0.3), varianza*[2]*, desviación típica, coeficiente de variación, mínimo, máximo, asimetría y apuntamiento.*

Los resultados que se obtienen son:

- Media=2753.827.
- Desviación típica=567.8277.
- Varianza=567.8277^2=322428.3.
- Coeficiente de variación=0.2061959. Como está entre 0.1 y 0.3, la media de la variable *calorías* es moderadamente representativa.
- Mediana=Percentil 50=2653. Por tanto, el 50% de los datos de la variable *calorías* son menores o iguales que 2653, es decir, el 50% de los países tiene una ingesta diaria de calorías menor o igual que 2653.
- Percentil 30=2316.2. Por tanto, el 30% de los datos de la variable *calorías* son menores o iguales que 2316.2.
- Cuartiles. El segundo cuartil es la mediana. El primer cuartil es el percentil 25 cuyo valor es 2256 y el tercero es el percentil 75, que es 3226. Así pues, el 25% de los datos de la variable *calorías* son menores o iguales que 2256 y el 75% de los datos de la variable *calorías* son menores o iguales que 3226.
- Moda. Según la tabla de frecuencias de la variable calorías obtenida anteriormente, la moda es 2375 porque es el valor que se repite más veces aunque es más adecuado utilizar la agrupación de la variable en intervalos y decir que la moda está en el intervalo (2000,2500] ya que ese es el intervalo de mayor altura en el histograma.
- Coeficiente de asimetría=0.1670104. Al ser mayor que 0, se tiene que la gráfica de la variable *calorías* es sesgada a la derecha.
- Coeficiente de apuntamiento=-1.206553. Al ser menor que 0, se tiene que la gráfica de la variable *calorías* es menos apuntada que la gráfica de la Normal con esa media y esa varianza.
- Mínimo=1667.
- Máximo=3825.

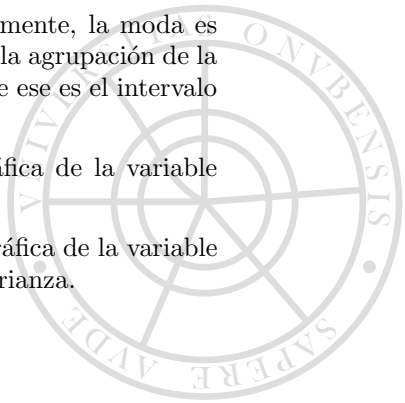

Diagrama de caja (y bigotes)

El diagrama de caja (y bigotes) es un gráfico que se construye a partir de los cuartiles y en el que se puede observar cómo se distribuyen los datos así como la existencia o no de valores atípicos. Por ejemplo, construir el diagrama de caja de las variables *calorías* y *mortinf* (mortalidad infantil).

[1]Para obtener la moda de una variable cuantitativa hay que convertirla en factor y obtener su distribución de frecuencias.

[2]La varianza se obtiene elevando al cuadrado el valor de la desviación típica.