Práctica 3

Estadística descriptiva (II)

3.1. Introducción

Como ya se explicó en la práctica anterior, la estadística descriptiva tiene como objetivo el análisis de conjuntos de datos, con el objetivo de describir las características y el comportamiento de los mismos, mediante estadísticos descriptivos, tablas y gráficos.

En esta práctica veremos algunas de las representaciones gráficas más habituales a la hora de resumir la información contenida en los datos. R-Commander nos permite realizar fácilmente varios tipos de gráficos tanto para variables de tipo factor como para variables de tipo numérico. Seguiremos trabajando con el archivo `Rcars.RData`.

3.1.1. Gráfico de sectores y de barras

Ejemplo 1: Vamos a construir un gráfico de sectores para resumir la información contenida en la variable *peso*. Para ello utilizamos el menú `Gráficas`→`Gráfica de sectores`. En la ventana variable seleccionamos la variable que contiene la información a resumir. Nótese que, inicialmente, sólo aparece la variable *origen*. Esto es debido a que la gráfica de sectores sólo puede dibujarse para variables de tipo factor. Cerramos entonces la ventana `Gráfica de sectores` y procedemos a *factorizar* la variable *peso*. Para ello seleccionamos el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Convertir variable numérica en factor`. En la ventana `Variables` elegimos la variable *peso* y en `Niveles del factor` seleccionamos la opción `Utilizar números`. Como nombre de la nueva variable indicamos *peso_fact*, pulsamos en el botón `Aceptar` y, a continuación, construimos el gráfico de sectores para la nueva variable *peso_fac* como se indicó anteriormente. Observamos entonces que el gráfico resultante es ininteligible debido a que la variable seleccionada presenta una gran cantidad de valores distintos.

Si queremos resumir los datos de la variable *peso*, mediante un gráfico de sectores, necesitaremos agrupar los datos de la variable en intervalos. Si solicitamos un

resumen del conjunto de datos activo, vemos que los valores mínimo y máximo de la variable *peso* son 244 y 1713. Vamos a usar 5 intervalos de igual amplitud para agrupar los datos. Si tomamos como extremo izquierdo del primer intervalo el valor 244 y como extremo derecho del quinto intervalo el valor 1713 obtendremos intervalos de amplitud (1713-244)/5=293.8, que tendrían una expresión poco "amigable".

Siempre que tengamos que hacer una agrupación en intervalos podemos comenzar en un valor ligeramente inferior al menor valor de la variable y terminar en un valor ligeramente superior al mayor valor de la variable. Además los intervalos deben ser exahustivos, esto es, tienen que cubrir todos los valores entre el mínimo y el máximo. Teniendo en cuenta esto, para que queden números mas "redondos", comenzaremos el primero de los cinco intervalos en 225 y terminaremos el último en 1725. De este modo, se obtienen los intervalos, [225,525], (525,825], (825,1125], (1125,1425], (1425,1725].

Procedemos entonces a recodificar la variable *peso*. Para ello seleccionamos el menú `Datos`→`Modificar variables del conjunto de datos activo`→`Recodificar variables`. En la ventana emergente, seleccionamos la variable a recodificar, en este caso *peso* y, como nombre de la variable recodificada, indicaremos *peso_rec*.

La información relativa a la codificación de los valores se debe escribir en la ventana `Introducir directrices de recodificación`, siguiendo el formato *valor_antiguo=valor_nuevo*. Si los valores antiguos o nuevos son cadenas de caracteres se pondrán entre comillas. Los rangos de valores que se utilizarán 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*. Teniendo en cuenta esto, en dicha ventana introducimos la siguente información:

225:525="[225,525]"
525:825="(525,825]"
825:1125="(825,1125]"
1125:1425="(1125,1425]"
1425:1725="(1425,1725]"

Nótese que cuando un valor forma parte de dos directrices de recodificación, se clasificará en la primera que hayamos introducido en la ventana. De este modo, el valor 525, se clasifica en el rango 225:525, y no en 525:825. Por este motivo los intervalos son de la forma $[\cdot,\cdot]$, el primero, y $(\cdot,\cdot]$, el resto.

Finalmente, pulsamos el botón `Aceptar` y, a continuación, construimos el gráfico de sectores de la variable *peso_rec*. Observamos en el gráfico que los sectores con mayor frecuencia son los que corresponden a los intervalos (525,825] y (825,1125]. El intervalo con menor frecuencia es el intervalo [225,525]. El gráfico no muestra información sobre el porcentaje o el número de casos que corresponden a cada sector. Para obtener esta información vamos a recurrir al menú `Estadísticos`→`Resúmenes`→`Dis-`

```
counts:
peso_rec
(1125:1425] (1425:1725]   (525:825]  (825:1125]   [225:525]
          91          39         145         130           1

percentages:
peso_rec
(1125:1425] (1425:1725]   (525:825]  (825:1125]   [225:525]
      22.41        9.61       35.71       32.02        0.25
```

Figura 3.1: Frecuencias de la variable *peso_rec*

`tribución de frecuencias`. En nuestro caso, obtenemos la salida de la figura 3.1, en la que podemos ver que los intervalos (525,825] y (825,1125] agrupan, respectivamente, al 35.71 % y al 32.02 % de los casos y tienen frecuencia absoluta 145 y 130, respectivamente. También vimos en el gráfico de sectores que el intervalo con menor frecuencia es el intervalo [225,525]. Nuevamente, en la figura 3.1, vemos que este intervalo corresponde al 0.25 % de los casos y tiene frecuencia absoluta igual a 1.

Los gráficos de barras permiten resumir la información de manera similar a los gráficos de sectores pero, mientras en éstos las frecuencias absolutas se representan mediante las áreas de los sectores circulares, en los gráficos de barras la información sobre el número de veces que se observa cada valor en los datos (su frecuencia absoluta) viene representada por la altura de las barras. De este modo, la modalidad que más veces se observa (con mayor frecuencia absoluta) será aquella que esté asociada a una barra más alta.

Ejercicio 2: Construir e interpretar un gráfico de barras para la variable *origen*. Indicar las frecuencias absolutas que corresponden a cada barra.

3.1.2. Histograma

Un histograma es una representación gráfica que se utiliza para obtener una representación visual de distribuciones de datos de variables estadísticas cuantitativas y continuas (como la longitud o el peso). Nos permite observar la posible tendencia, por parte de los datos, de ubicarse hacia una determinada región dentro del conjunto de posibles valores que puede tomar la variable. De este modo, el histograma nos permite definir el comportamiento de los datos y observar su grado de dispersión así como estudiar la forma de la distribución de los datos.

Construiremos el histograma correspondiente a la variable *peso*. Para ello, utilizamos el menú `Gráficas`→`Histograma`. Seleccionamos la variable peso y pulsamos en el botón `Aceptar`. En el histograma resultante podemos observar que la distribución es unimodal y sesgada a la derecha. Si calculamos el coeficiente de asimetría

de tipo 1, para esta variable, obtenemos como el valor 0.4662921 que, efectivamente, nos indica que la distribución es sesgada a la derecha. Rescuérdese también que en este tipo de distribuciones hay mayoría de casos por debajo de la media lo que se traduce en nuestro caso que hay mayoría de coches que tienen peso inferior al peso medio.

También es posible construir los histogramas de una variable agrupando los datos de acuerdo a los valores que toma una variable de tipo factor. Por ejemplo, podemos construir el histograma de la variable peso para cada uno de los orígenes. Para ello, repetimos los pasos anteriores pero, tras pulsar el botón `Gráfica por grupos`, seleccionamos la variable *origen*. Podemos ver entonces que el comportamiento de la distribución del peso no es el mismo para los tres orígenes: mientras en el caso de los coches de Estados Unidos la distribución es simétrica, los datos con origen *Japón* y *Europa* siguen distribuciones con asimetría positiva.

Finalmente, indicaremos que es posible forzar la elección de unos intervalos concretos para la construcción del histograma. Si observamos la ventana de comandos, la instrucción con la que se ha generado el histograma de la variable *peso* es:

```
with(RCars,Hist(peso,scale="frequency",breaks="Sturges",col="darkgray"))
```

El modo en que se calculan los puntos que delimitan los intervalos se indican en la opción *breaks* de la instrucción `Hist`. Si deseamos definir unos intervalos concretos tenemos que sustituir dicha instrucción por un vector que contenga los extremos de los intervalos. Los vectores en R se definen mediante la función `c(·)` teniendo como argumento las componentes del vector a definir.

Ejemplo 3: Construir el histograma de la variable *motor* utilizando cuatro intervalos, de la misma amlitud, que comiencen en 0 y terminen en 8000.

En la pestaña de instrucciones, de la ventana de R-Commander, escribimos la siguiente instrucción (o bien modificamos una de las generadas anteriormente):

```
with(RCars, Hist(motor, scale="frequency", breaks=c(0,2000,4000,6000,
8000), col="darkgray"))
```

Dejamos el cursor en la misma línea de la instrucción y pulsamos el botón `Ejecutar`, obteniendo de esta forma el histograma deseado.

3.1.3. Gráfica de las medias

La gráfica de las medias permite comparar el efecto que producen las modalidades de una o dos variables de tipo factor, en el comportamiento de una variable cuantitativa. Disponemos de esta herramienta en el menú `Gráficas`→`Gráfica de las medias`.

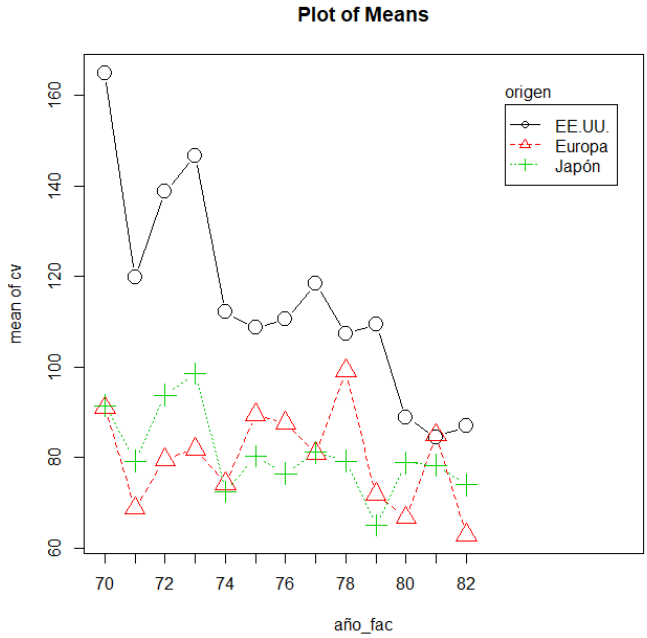

Figura 3.2: Gráfica de las medias

Ejemplo 4: Estudiar, mediante una gráfica de medias, el valor medio de la variable *cv* según el *origen* y el *año* de fabricación del modelo.

La variable numérica a estudiar es la variable *cv*. Los factores que determinan el comportamiento de la variable bajo estudio son *año* y *origen*. Sin embargo, la variable *año* es una variable de tipo numérico que, previamente, debemos convertir en factor, en una variable a la que daremos como nombre *año_fac*. Una vez hecho esto seleccionamos el menú `Gráficas→Gráfica de las medias` y, en la ventana emergente, elegimos los factores *año_fac* y *origen*. En la ventana `Variable explicada`, seleccionamos la variable *cv*. Para una mayor claridad de gráfico que obtendremos, en la pestaña `Opciones`, seleccionamos la opción `Sin barras de errores` y pulsamos el botón `Aceptar`. Obtendremos entonces una gráfica similar a la que se muestra en la figura 3.2.

En el gráfico resultante observamos tres líneas, una por cada modalidad de la variable origen. Consideremos, a modo de ejemplo, la línea que corresponde al origen *EE.UU.*. La forma en la que se construye dicha línea es la siguiente: para cada valor del año, se consideran los casos que corresponden únicamente a ese año y al origen *EE.UU.*, se calcula el valor medio de la variable *cv* para esos casos y sobre la etiqueta correspondiente al año se dibuja una marca a una altura igual al valor medio obtenido. Así, la primera marca nos dice que la potencia media en caballos, de los coches de la muestra con origen *EE.UU.* y que se construyeron en 1970, es de algo más de 160 cv. La segunda marca, nos indica que la potencia media en caballos,

de los coches de la muestra con origen *EE.UU.* y que se construyeron en 1971 es de, aproximadamente, 120 cv.

El procedimiento se repite para los otros dos orígenes: *Europa* y *Japón*. De este modo, las líneas representan la evolución en el tiempo de la potencia media de los vehículos procedentes de cada origen. Observando el gráfico vemos que los coches fabricados en EE.UU. tienen una mayor potencia media, si bien esta se ha ido reduciendo con el paso de los años hasta casi llegar al nivel de los cohes europeos y japoneses. Por otra parte, los datos disponibles sobre coches europeos y japoneses muestran un potencia media en cv, similar en ambos casos, que se ha mantenido más o menos estable, en el entorno de los 80 cv., desde 1970 hasta 1982.

3.1.4. Gráfica XY

La gráfica XY permite representar la relación entre variables cuantitativas con la posibilidad de realizar el estudio según los niveles establecidos por una o más variables de tipo factor.

Ejemplo 5: Construir una gráfica XY para representar la relación entre el consumo y el peso de los vehículos del archivo RCars.

Seleccionamos el menú `Gráficas`→`Gráfica XY`. Dado que el consumo se explica, al menos parcialmente, por el peso del vehículo, para representar la relación entre las variables indicadas seleccionamos la variable *peso* en la ventana `Variables explicativas` y la variable *consumo*, en la ventana `Variables explicadas`. Finalmente, pulsamos en el botón `Aceptar`.

La gráfica que obtenemos es una nube de puntos en la que cada punto representa el peso y el consumo del correspondiente caso. Podemos observar que existe una dependencia directa entre las variables de manera que, cuanto mayor es el peso del vehículo, mayor es también el consumo de combustible.

La ventana `Grupos` del menú `Gráfica XY`, permite representar, en una misma gráfica, la relación entre dos variables cuantitativas pero agrupando los datos según los valores que toma una variable de tipo factor. Los datos correspondientes a los distintos grupos se representan mediante diferentes colores.

Ejemplo 6: Construir una gráfica XY para representar la relación entre el consumo y el peso de los vehículos del archivo RCars, para los grupos definidos por el número de cilindros.

En primer lugar, creamos una variable de tipo factor que contenga la información relativa al número de cilindros. A dicha variable le daremos como nombre *cilindr_fac*. Repetimos ahora los pasos realizados en el ejemplo anterior pero ahora selecciona-

mos, además, la variable *cilindr_fac*, en la ventana `Grupos`. Finalmente, pulsamos el botón `Aceptar`.

Obtenemos una nube de puntos similar a la del ejemplo anterior pero ahora vemos que los puntos presentan distintos colores en función del número de cilindros de los casos a los que representan. Vemos así que predominan los coches con 4, 6 y 8 cilindros y que los coches con pesos y potencias más bajas son, mayoritariamente, coches con 4 cilindros, que los coches con pesos y potencias medias son, principalmente, coches con 6 cilindros, y que los coches con mayores potencias y pesos son los coches con 8 cilindros.

Por último, es posible realizar estos gráficos para grupos de casos definidos por los valores que toma una variable de tipo factor. Para ello basta elegir dicha variable en la ventana `Condiciones`. Al hacer esto, obtendremos una representación para cada uno de los valores de la variable.

Ejemplo 7: Construir gráficas XY que representen, para cada valor de *origen*, la relación entre el consumo y el peso de los vehículos del archivo RCars, para los grupos definidos por el número de cilindros.

Repetimos los pasos de los ejemplos anteriores seleccionando la variable *peso* en la ventana `Variables Explicativas`, la variable *consumo* en la ventana `Variables Explicadas`, la variable *cilindr_fac* en la ventana `Grupos` y la variable *origen* en la ventana `Condiciones`.

3.1.5. Gráfica de cajas

Los *diagramas de cajas y bigotes* se utilizan para representar de manera simultánea la dispersión y la simetría de los datos, permitiendo comparar dos o más conjuntos de datos. Por defecto, R-Commander dibuja el diagrama con orientación vertical. De este modo, el valor de las observaciones corresponden con la escala marcada en el eje vertical.

El gráfico contiene un rectángulo en el que el lado inferior y el superior corresponden con el primer y tercer cuartil, respectivamente. El rectángulo se divide con una línea horizontal que corresponde con la mediana. Finalmente se traza una línea vertical, desde la mitad del lado superior del rectángulo, hasta el tercer cuartil más 1.5 veces el recorrido intercuartílico (la diferencia entre el tercer cuartil y el primero). Si este valor fuera mayor que el valor máximo de los datos entonces el "bigote"sólo se extiende hasta el valor máximo. Para construir el bigote inferior se traza una línea vertical, desde la mitad del lado inferior del rectángulo, hasta el primer cuartil menos 1.5 veces el recorrido intercuartílico (la diferencia entre el tercer cuartil y el primero). Si este valor fuera menor que el valor mínimo de los datos entonces el "bigote"sólo se extiende hasta el valor mínimo.

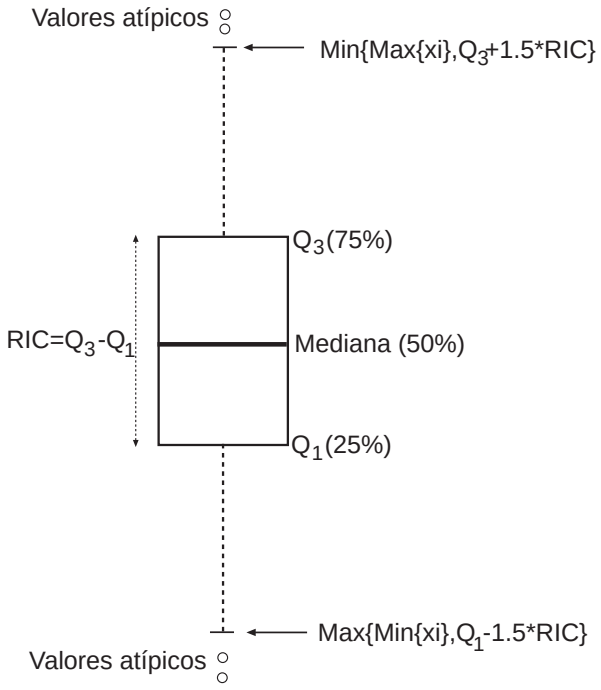

Figura 3.3: Gráfico de cajas y bigotes

Si hay datos por encima del bigote superior o por debajo del inferior se consideran valores atípicos que deben ser estudiados con más detenimiento, pues son datos que corresponden con casos que podrían presentar un comportamiento distinto al resto, o bien tratarse de valores que se han medido de manera errónea. La figura 3.3 muestra la estructura de un gráfico de cajas y bigotes.

Ejemplo 8: Construir e interpretar el gráfico de cajas y bigotes de la variable consumo, para cada uno de los orígenes.

Para construir el gráfico solicitado, seleccionamos el menú `Gráficas`→`Diagrama de caja`. En la pestaña `Datos` de la ventana emergente seleccionamos *consumo* como `Variable`. Pulsamos el botón `Gráfica según:` y elegimos la variable *origen* como variable de agrupación. Finalmente, cerramos la ventana `Grupos` y la ventana `Diagrama de caja`, pulsando en los respectivos botones `Aceptar`.

La gráfica de cajas y bigotes nos proporciona información sobre la simetría del conjunto de datos. Si la línea correspondiente a la mediana no se encuentra centrada en la caja o si las longitudes de los bigotes, junto con los correspondientes valores atípicos, son distintas el gráfico indica que la distribución de datos no es simétrica.

De este modo, en el gráfico podemos ver que en el caso del origen *EE.UU.* la línea de la mediana se encuentra centrada en la caja pero el bigote superior es ligeramente más largo que el inferior, lo que indica una distribución casi simétrica, aunque con una ligera asimetría positiva. Para el origen *Europa*, el comportamiento es similar, dado que la longitud del bigote superior junto con los posibles valores atípicos es mayor que la longitud del bogite inferior.

En el caso del origen *Japón* vemos que el bigote superior, junto con el valor atípico, tiene mayor longitud que el inferior y que la línea de la mediana no sólo no se encuentra centrada en la caja, sino que coincide con su lado inferior. Esto nos indica que los datos presentan una clara asimetría positiva (distribución sesgada a la derecha) dado que la mediana se encuentra desplazada hacia los valores más pequeños de la distribución.

Se aprecia también en el gráfico que la mediana de los consumos es superior en el origen *EE.UU.*, seguida de la mediana del origen *Europa* y que los coches japoneses son los que tienen una mediana del consumo inferior.

La dispersión de los datos, medida a través del rango de los mismos (diferencia entre el mayor y el menor valor), es mucho mayor para el origen *EE.UU.* y similar para *Europa* y *Japón*. También el rango intercuartílico ($Q_3 - Q_1$) es mayor en el caso de *EE.UU.* y más parecido en los orígenes *Europa* y *Japón*. En conclusión, obtenemos que los valores del consumos presentan mayor dispersión para el origen *EE.UU.* que para *Europa* y *Japón* y que la dispersión para estos dos orígenes es muy parecida.

Por último, vemos que al considerar los orígenes *Europa* y *Jápón* encontramos algunos valores que son considerados *atípicos*. Corresponden a los casos 283, 285 y 219, para *Europa* y al caso 119, para *Japón*. En el gráfico no se aprecia bien cuáles son los casos atípicos, por estar algunos de ellos solapados, pero también podemos encontrarlos en la ventana `Salida` de R-Commander.

Que hayamos encontrados valores atípicos no significa que haya que eliminarlos del conjunto de datos, sino que esos valores proceden de casos que quizás no se ajustan al patrón de comportamiento del resto de casos, o bien son valores procedentes de medidas erróneas. En cualquier caso, lo que habría que hacer es examinar con detenimiento las medidas que se han recogido en esos casos y estudiar cuál puede se el posible problema, si es que hubiera alguno.