Práctica 2

Estadística descriptiva (I)

2.1. Introducción

La estadística descriptiva comprende una serie de técnicas destinadas a organizar, resumir y presentar de manera adecuada la información contenida en grandes conjuntos de datos, obtenidos de una población, de manera que sea fácilmente comprensible.

El resumen de la información contenida en los datos se puede hacer de manera numérica, mediante estadísticos descriptivos y tablas, o mediante gráficas. Dedicaremos esta práctica a estudiar algunos de los estadísticos descriptivos y tablas más usuales, mientras que el resumen de la información mediante gráficas los estudiaremos en la práctica 3.

2.2. Estadísticos descriptivos: Medidas de centralización, localización, dispersión y forma

El menú `Estadísticos`→`Resúmenes`→`Conjunto de datos activos` proporciona un resumen básico del conjunto de datos activo en el que se muestra:

- Para las variables cuantitativas: mínimo (Min.), primer cuartil (1st Qu.), mediana ó segundo cuartil (Median), media (Mean), tercer cuartil (3rd Qu.) y máximo (Max.).
- Para las variables cualitativas (de tipo factor), se muestran los nombres de las modalidades de la variable así como su frecuencia absoluta.

En lo que sigue trabajaremos con el fichero *RCars.Rdata*, que puede encontrarse en Moodle. La descripción de las variables estudiadas en el fichero puede encontrarse en la práctica 1.

Ejemplo 1: Obtener e interpretar el resumen de los datos contenidos en el fichero *RCars*. Determinar la moda de las variables origen y cilindros.

```
    consumo          motor             cv             peso            acel
 Min.   : 5.00   Min.   :  66   Min.   : 46.00   Min.   : 244.0   Min.   : 8.00
 1st Qu.: 8.00   1st Qu.:1708   1st Qu.: 75.75   1st Qu.: 741.2   1st Qu.:13.62
 Median :10.00   Median :2434   Median : 95.00   Median : 936.5   Median :15.50
 Mean   :11.23   Mean   :3180   Mean   :104.83   Mean   : 989.5   Mean   :15.50
 3rd Qu.:13.00   3rd Qu.:4806   3rd Qu.:129.25   3rd Qu.:1203.8   3rd Qu.:17.07
 Max.   :26.00   Max.   :7456   Max.   :230.00   Max.   :1713.0   Max.   :24.80
 NA's   :8                      NA's   :6
      año           origen       cilindr
 Min.   :70.00   EE.UU.:253   Min.   :3.000
 1st Qu.:73.00   Europa: 73   1st Qu.:4.000
 Median :76.00   Japón : 79   Median :4.000
 Mean   :75.94   NA's  :  1   Mean   :5.469
 3rd Qu.:79.00                3rd Qu.:8.000
 Max.   :82.00                Max.   :8.000
 NA's   :1                    NA's   :1
```

Figura 2.1: Salida obtenida en el ejemplo 1.

Una vez cargado el archivo *RCars*, seleccionamos el menú `Estadísticos`→`Resúmenes`→`Conjunto de datos activos`, obteniendo la salida de la figura 2.1.

Consideremos una de las variables numéricas como es la variable *consumo*. Para esta variable obtenemos que el valor mínimo es 5 y el máximo 26 lo que significa que, en los coches de la muestra, el mínimo consumo observado es de 5 ℓ/100Km. y el mayor consumo es de 26 ℓ/100Km. Asimismo, observamos que el consumo medio es de 11.23 ℓ/100Km.

Por otra parte, el primer cuartil, la mediana (o segundo cuartil) y el tercer cuartil de los consumos son, respectivamente, 8, 10 y 13 ℓ/100Km. Esto significa que hay un 25 % de coches en la muestra cuyo consumo es igual o inferior a 8 ℓ/100Km., un 50 % cuyo consumo es igual o inferior a 10 ℓ/100Km., y un 75 % cuyo consumo es igual o inferior a 13 ℓ/100Km. En los datos de la salida correspondientes a la variable consumo también apreciamos que hay 8 valores de tipo `NA`, lo que significa que en 8 de los coches estudiados no se dispone de datos sobre el consumo. En cualquier caso, cuando se calculan las medidas descriptivas del conjunto de datos, los casos `NA` son desechados y los estadísticos descriptivos (media, varianza, cuartiles, etc.) se calculan considerando únicamente los restantes casos.

Para las variables de tipo factor el programa muestra los valores que toman (también llamados modalidades), así como las frecuencias de cada uno de estos valores. En el conjunto de datos *RCars* la única variable de tipo factor es la variable *origen*. Esta variable presenta tres modalidades: *EE.UU., Europa y Japón* y en uno de los casos del estudio se desconoce el origen del coche. Observando las frecuencias, podemos ver que en la muestra hay 253 coches de Estados Unidos, 73 de Europa y 79 procedentes de Japón, además de un caso del que se desconoce el origen.

Calcularemos ahora las *modas* solicitadas. En una variable estadística, de tipo factor, la moda es el valor que más veces se repite. En el caso de la variable *origen* podemos comprobar que la moda es el valor *EE.UU.*, que tiene frecuencia absoluta 253. Sin embargo, en el caso de la variable *cilindr*, podemos observar que el programa no muestra las frecuencias absolutas de cada valor. Esto es debido a que la variable *cilindr* es una variable de tipo numérico. Por tanto, si queremos que R nos muestre las frecuencias absolutas de los valores que toma esta variable, tenemos que convertirla en una variable de tipo factor. Para ello disponemos del menú `Datos`→`Modificar variables del conjunto de datos activo`→`Convertir variable numérica en factor`. En la ventana emergente, seleccionaremos la/s variable/s que queremos convertir en factor, en este caso, la variable *cilindr*. En las opciones `Niveles del factor` disponemos de dos opciones. La opción `Asignar nombres a los niveles` es útil en el caso en que queramos renombrar manualmente los niveles (modalidades) de la variable. Al elegir esta opción, si el número de modalidades no es muy elevado, se abrirá una ventana emergente en la que podremos hacer esto. Por el contrario la opción `Utilizar números` convertirá, de manera automática, cada valor numérico de la variable en una cadena de caracteres. De este modo, los valores que observaremos serán los mismos aunque el programa ya no los considerará números sino cadenas de caracteres.

Elegimos entonces la segunda opción, `Utilizar números`, y en la ventana `Nuevo nombre o prefijo para variables múltiples` le damos a la nueva variable el nombre *cilindr_fac*. Si solicitamos ahora el resumen del conjunto de datos podemos observar, en la salida correspondiente a la nueva variable, que el valor que más se repite (esto es, la moda de la variable) es el 4 y tiene frecuencia absoluta igual a 207.

El menú `Estadísticos`→`Resúmenes`→`Resúmenes numéricos` permite obtener estadísticos descriptivos de una o más variables, con la posibilidad de calcularlos sobre grupos de observaciones definidos a partir de los valores de otra variable de tipo factor. Los estadísticos que pueden obtenerse mediante esta opción son la media, desviación típica, coeficiente de variación, coeficiente de asimetría, coeficiente de apuntamiento o curtosis y cuantiles, entre otros. Nótese que los percentiles y los cuantiles proporcionan la misma información pero, mientras los percentiles se expresan en tantos por ciento, los cuantiles se expresan en tanto por uno. Así, el percentil 75 corresponde con el cuantil 0.75.

Ejemplo 2: Determinar la media y el percentil 30 de la potencia en c.v. y del peso de los coches, para cada uno de los orígenes definidos en el fichero. Repetir el estudio anterior para los grupos de coches definidos por el número de cilindros.

Calcularemos primero los estadísticos para cada uno de los orígenes. Para ello seleccionamos el menú `Estadísticos`→`Resúmenes`→`Resúmenes numéricos`. En la ventana emergente marcamos la/s variables/s sobre las que queremos calcular los estadísticos, en este caso, las variables *cv* y *peso*. Para seleccionarlas, dejamos pul-

```
Variable: cv
           mean 0%  25% 50%   75% 100% 30%   n NA
EE.UU. 119.60643 52 88.0 105 150.0  230  90 249  4
Europa  81.00000 46 69.5  77  90.5  133  71  71  2
Japón   79.83544 52 67.0  75  95.0  132  67  79  0

Variable: peso
           mean  0% 25%  50%    75% 100%   30%   n NA
EE.UU. 1122.1146 600 906 1126 1351.0 1713 953.4 253  0
Europa  810.1233 608 688  748  933.0 1273 705.0  73  0
Japón   740.0506 537 661  718  803.5  976 669.4  79  0
```

Figura 2.2: Salida obtenida en el ejemplo 2a.

sada la tecla `'Ctrl'` mientras pinchamos en las variables con el ratón. En el botón `Resumir por grupos` elegimos como variables de grupo *origen* y pulsamos en el botón `Aceptar`, volviendo a la ventana resúmenes numéricos. Finalmente, seleccionamos la pestaña `Estadísticos`, marcamos la casilla correspondiente a la media y, en la casilla `Cuantiles`, añadimos 0.3.

Obtenemos entonces la salida que aparece en la figura 2.2. En dicha salida vemos que la potencia media para los coches con origen *EE.UU.* es de 119.606 cv., para los coches con origen *Europa* es de 81 cv. y para los coches con origen *Japón* es de 79.835 cv.

En lo que respecta al peso, vemos que el peso medio de los coches con origen *EE.UU.* es de 1122.11 Kg., para los coches con origen *Europa* es de 810.12 Kg. y para los coches con origen *Japón* es de 740.05 Kg.

En consecuencia, la potencia media y el peso medio es mayor en los coches de EE.UU., mientras que los coches con origen japonés son los que presentan menor potencia y peso medio.

Finalmente, observando la columna del percentil 30 vemos que entre los coches con origen *EE.UU.* hay un 30 % que tienen potencia 90 cv. o menos, entre los coches con origen *Europa* hay un 30 % con potencia 71 cv. o menos y entre los coches con origen *Japón*, hay un 30 % cuya potencia es de 67 cv. o menos. Del mismo modo el percentil 30 de la variable *peso*, para los orígenes *EE.UU.*, *Europa* y *Japón*, es de 953.4 Kg., 705 Kg. y 669.4 Kg., respectivamente.

Repetimos ahora el procedimiento anterior para calcular la media y el percentil 30 para los grupos definidos por el número de cilindros. Nótese que en la ventana `Resúmenes numéricos`, cuando pulsamos el botón `Resumir según` para definir los grupos, sólo podemos elegir variables de tipo factor. En consecuencia, para hacer la agrupación, no es posible utilizar la variable *cilindr* original sino que es necesario

```
Variable: cv
       mean 0%    25%   50%    75% 100%   30%   n NA
3  99.25000 90  95.25  98.5 102.50  110  96.3   4  0
4  78.47030 46  68.00  78.0  88.75  115  70.0 202  5
5  82.33333 67  72.00  77.0  90.00  103  73.0   3  0
6 101.50602 72  92.50 100.0 110.00  165  95.0  83  1
8 158.13084 90 140.00 150.0 175.00  230 145.0 107  0

Variable: peso
       mean   0%     25%  50%     75% 100%    30%   n NA
3  799.0000  708  759.00  791  831.00  906  769.2   4  0
4  770.5507  537  681.50  744  857.50 1090  701.6 207  0
5 1034.0000  943  963.00  983 1079.50 1176  967.0   3  0
6 1065.7500  824  979.75 1067 1143.25 1302  993.4  84  0
8 1366.1028 1028 1266.00 1378 1460.00 1713 1288.6 107  0
```

Figura 2.3: Salida obtenida en el ejemplo 2b.

convertirla en factor, cosa que ya hicimos anteriormente. A la variable ya factorizada le habíamos dado el nombre de *cilindr_fac*, por lo que será esta la que utilicemos para definir los grupos. Obtenemos entonces la salida que podemos ver en la figura 2.3, en la que podemos encontrar los datos los datos necesarios para el estudio.

Ejemplo 3: Calcular la desviación típica y la varianza de los datos correspondientes a la aceleración de los vehículos.

La desviación típica y la varianza que calcula R se corresponden con la cuasidesviación típica y la cuasivarianza que se definen en los apuntes de Estadística Descriptiva. Sin embargo, cuando se solicite calcular la desviación típica o varianza de una variable, nos referiremos a las calculadas por R aunque tenemos que ser conscientes de que el valor que obtengamos no corresponderá a la desviación típica o la varianza reales. No obstante, cuando el conjunto de datos sea muy grande, que es lo habitual cuando se trabaja con un paquete de software estadístico, la diferencia no sera apreciable. Por el contrario, sí que tendremos que ser cuidadosos cuando estemos trabajando con un conjunto de datos pequeño.

La desviación típica se puede obtener marcando la correspondiente opción en la pestaña `Estadísticos` del menú `Estadísticos`→`Resúmenes`→`Resúmenes numéricos`. Obtenemos entonces que el valor de la desviación típica de la variable *acel* es igual a 2.820984 m/s^2.

Para el cálculo de la varianza podemos calcular el cuadrado de la desviación típica o bien podemos usar la función `var()`. En este caso, el argumento de la función será de la forma `nombre del conjunto de datos activo$nombre de la variable`. El nombre del conjunto de datos activo lo tenemos disponible en la pestaña `Conjunto de datos`. Además, tenemos que añadir a la función `var()` la opción `na.rm=TRUE` para que, antes de hacer el cálculo, elimine los casos no disponibles (los

de tipo NA). Así pues para calcular la varianza de la variable *acel* escribiremos, en la ventana de comandos de R-Commander, la instrucción `var(RCars$acel,na.rm=TRUE)` y pulsaremos el botón `Ejecutar`.

En la salida observamos que el valor de la varianza es 7.957951 m^2/s^4. Obsérvese que la unidad de la varianza es el cuadrado de la unidad de la magnitud bajo estudio.

La varianza y la desviación típica miden la dispersión de los datos en términos absolutos. Esto hace que no sean apropiadas para comparar la dispersión de dos conjuntos de datos cuando estos corresponden a magnitudes que se miden en unidades distintas. Cuando ocurre esto o incluso cuando las medias de los conjuntos de datos a comparar son muy diferentes, la dispersión de los conjuntos de datos puede compararse utilizando una medida de dispersión relativa como el coeficiente de variación de Pearson. Dicho coeficiente mide, en términos de porcentajes, la dispersion relativa del conjunto de datos. Así, cuanto menor es el valor del coeficiente, menor es la dispersión relativa de los datos y más representativa es la media. Una descripción más detallada de este coeficiente puede encontrarse en el tema de Estadística Descriptiva.

Ejemplo 4: Calcular la desviación típica y el coeficiente de variación de las variables *consumo* y *peso*. Comparar los resultados obtenidos.

Las desviaciones típicas de las variables *consumo* y *peso* son, respectivamente, 3.946 ℓ/100Km. y 283.277Kg. por lo que, en términos absolutos, la variable *peso* presenta una mayor dispersión de los datos observados, si bien el hecho de medirse en unidades distintas dificulta la comparación. Por otra parte, el coeficiente de variación de la variable *consumo* es del 35.14 %, mientras que el de la variable *peso* es del 28.63 %, por lo que la variable *peso* presenta una menor dispersión relativa.

Ejemplo 5: Determinar e interpretar los coeficientes de asimetría y de apuntamiento de la variable *peso*, para cada uno de los orígenes.

Aunque se pueden calcular con independencia del comportamiento de la variable bajo estudio, los coeficientes de asimetría y de apuntamiento (o curtosis) están diseñados para el estudio de distribuciones unimodales, campaniformes y moderadamente asimétricas. No obstante, dado que no veremos las representaciones gráficas hasta la práctica 3, nos limitaremos a calcular los coeficientes sin comprobar ninguna de esas condiciones.

Para calcular los coeficientes solicitados, volvemos a seleccionar el menú `Estadísticos`→`Resúmenes`→`Resúmenes numéricos`. Elegimos la variable *peso* y solicitamos resumir la información según la variable *origen*. En la pestaña `Estadísticos` seleccionamos `Asimetría` y `Apuntamiento` y, en `Tipo`, seleccionamos `Tipo 1`. R proporciona la posibilidad de calcular tres coeficientes de asimetría y apuntamiento (`Tipo 1, 2 y 3`), que son igualmente válidos, si bien por defecto R trabaja con el tipo 2. No obstante, mientras no se nos diga lo contrario, trabajaremos con el `Tipo 1` que es

el que corresponde a los coeficientes descritos en el tema de Estadística Descriptiva.

En la salida correspondiente podemos observar que, el valor del coeficiente de asimetría de la variable peso es de 0.055, 0.757 y 0.487, para los orígenes EE.UU., Europa y Japón, respectivamente. Por tanto, en todos los casos se trata de distribuciones sesgadas a la derecha o con asimetría positiva.

El valor del coeficiente de curtosis o apuntamiento, es de -0.958, -0.386 y -0.431, para cada uno de los tres orígenes. Por tanto, en los tres casos la distribución de los pesos es platicúrtica, esto es, menos apuntada que la distribución normal, que es la distribución que se toma como referencia para el estudio del apuntamiento y que se estudiará con detalle en el tema 3 de teoría.

Ejemplo 6: Determinar la media, desviación típica y mediana del consumo de los coches según su origen y número de cilindros.

Para calcular estadísticos descriptivos de una variable de tipo numérico, según las modalidades definidas por una o más variables de tipo factor, usaremos el menú `Estadísticos`→`Resúmenes`→`Tabla de estadísticas`. En la ventana emergente, el espacio `Factores` sirve para elegir las variables de tipo factor que se usarán para hacer la agrupación. En el espacio `Variables explicadas` elegiremos las variables para las que se desea calcular el `Estadístico`. Éste puede ser, la `Media`, la `Mediana`, `la Desviacion típica`, `el Rango intercuartílico` o cualquier otro que especifiquemos en la ventana `Otro`.

Así pues, seleccionamos el menú `Estadísticos`→`Resúmenes`→`Tabla de estadísticas`. En la ventana `Factores` elegimos las variables *cilindr_fac* y *origen* y, en la ventana `Variables explicadas`, seleccionamos la variable *consumo*. Como `Estadístico` a calcular elegiremos la `Media`. Nótese que sólo es posible calcular uno cada vez. Obtenemos entonces la salida, que refleja los consumos medios para los coches agrupados según el número de cilindros y su origen. Por ejemplo, vemos que el consumo medio de los coches con cuatro cilindros y origen EE.UU. es de 8.68 $\ell/100Km$.. Los valores NA aparecen en los grupos en los que no se dispone de ninguna observación. Por ejemplo, en la muestra no hay ningún coche con origen Europa que tenga 8 cilindros, de ahí que el valor obtenido sea NA.

Finalmente, repetimos los pasos anteriores para los estadísticos `Desviación típica` y `Mediana`.

Ejercicio 7: Determinar la media del consumo de los coches en función de su origen, del número de cilindros y del año de fabricación (nótese que para ello es necesario categorizar previamente la variable *Año*).

2.3. Distribuciones de frecuencias. Tablas de doble entrada

Veremos en este apartado cómo calcular *frecuencias absolutas* y *porcentajes* de las distintas modalidades de una variable de tipo factor. La frecuencia absoluta de una modalidad no es más que el número de veces que se observa dicha modalidad. La frecuencia relativa indica la proporción de veces que se observa dicha modalidad y se calcula como el cociente entre su frecuencia absoluta y el número de observaciones válidas de la variable. Hay que decir que R no calcula dicha proporción sino que nos devuelve el porcentaje de casos que corresponde a cada modalidad. No obstante, ambos valores son equivalentes y nos proporcionan la misma información ya que, para obtener los porcentajes basta con multiplicar las frecuencias relativas por 100. Nótese también que R calcula las frecuencias absolutas y los porcentajes, únicamente, para variables de tipo factor.

Ejemplo 8: Determinar cuantos coches de cada origen hay en la muestra. Determinar también qué porcentaje, del total de casos estudiados, representa cada origen.

Seleccionamos el menú `Estadísticos`→`Resúmenes`→`Distribución de frecuencias`. En la ventana emergente seleccionamos la variable *origen* y pulsamos en el botón `Aceptar`. En la salida se nos muestran las frecuencias absolutas (counts) para las modalidades de la variable *origen*. Vemos entonces que hay 253 coches con origen *EE.UU.*, 73 coches con origen *Europa* y 79 coches con origen *Japón*.

La salida también nos muestra los porcentajes (percentages). Encontramos que el 62.47 % de los casos corresponden a coches con origen Estados Unidos, el 18.02 % tienen origen Europa y el 19.51 % tienen por origen Japón.

En el ejemplo anterior hemos utilizado una única variable para realizar la agrupación de los datos y el posterior estudio de las frecuencias. También es posible calcular frecuencias absolutas y porcentajes de las modalidades definidas, de manera conjunta, por dos o más factores. Para ello disponemos del menú `Estadísticos`→`Tablas de contingencia`→`Tabla de doble entrada` y del menú `Estadísticos`→`Tablas de contingencia`→`Tabla de entradas múltiples`. El primero ofrece la posibilidad de calcular frecuencias mediante tablas en las que la agrupacion viene definida por dos factores, mientras que el segundo permite agrupar utilizando tres o más factores.

Ejemplo 9: Responder a las siguientes cuestiones:

a) ¿Cuántos coches europeos tienen 4 cilindros?

Seleccionamos entonces el menú `Estadísticos`→`Tablas de contingencia` →`Tabla de doble entrada`. Seleccionamos las variables *cilindr_fac*, como variable de fila, y *origen* como variable de columna. En la pestaña

`Estadísticos` marcamos la opción `Sin porcentajes`. En la salida observamos que hay 66 casos de coches que son europeos y tienen cuatro cilindros.

b) ¿Cuál es el número de cilindros más frecuente en los coches de EE.UU.?

Si en la salida obtenida en el apartado anterior nos restringimos a la columna de EE.UU., vemos que el valor de la variable *cilindr_fac* con mayor frecuencia es el valor 8.

c) ¿Qué porcentaje de los coches observados tienen 6 cilindros?

Repetimos los pasos del apartado a) pero, en este caso, en la pestaña `Estadísticos` marcamos la opción `Porcentajes totales`. Esta opción nos muestra: en la columna `Total` los porcentajes que representan los casos de cada fila, respecto al total de casos estudiados y, en la fila `Total`, los porcentajes que representan los casos de cada columna, respecto al total de casos estudiados. Finalmente, en cada intersección de fila y columna nos muestra qué porcentaje representan, con respecto al total de datos, los datos que cumplen la condición correspondiente a esa fila y esa columna.

La respuesta a la pregunta la encontramos en valor que se encuentra en la fila correspondiente a la modalidad 6 (cilindros) y la columna `Total`: los coches con 6 cilindros representan el 20.7 % del total de coches estudiados.

d) ¿Qué porcentaje de los coches observados son europeos?

Miramos ahora la columna `Europa` y la fila Total, que contiene los porcentajes que representan las frecuencias de cada modalidad de `origen` con respecto al total de casos estudiados: el 18 % de los coches de la muestra tienen como origen *Europa*.

e) ¿Qué porcentaje de coches son europeos y tienen 6 cilindros?

Nuevamente, usamos la última salida para responder a la pregunta. En la intersección de la fila correspondiente a la modalidad 6 y la columna de la modalidad *Europa*, encontramos que el 1 % de los coches son europeos y tienen 6 cilindros.

f) ¿En qué origen hay mayor porcentaje de coches con 4 cilindros?

Para responder a la pregunta usaremos la opción `Porcentajes por columnas` que podemos encontrar en la pestaña `Estadísticos` del menú `Estadísticos`→`Tablas de contingencia` →`Tabla de doble entrada`. Dicha opción nos muestra qué porcentaje representan las observaciones correspondientes a cada modalidad de la variable fila, cuando nos restringimos a cada una de las modalidades de la variable columna. Esto es, nos indica los porcentajes correspondientes a la modalidad de la variable fila, condicionada por la modalidad observada en la variable columna.

Al seleccionar la opción `Porcentajes por columnas` podemos ver que, si nos restringimos al origen *EE.UU.*, los coches con cuatro cilindros representan el 28.5 % de los coches con ese origen; dentro del origen *Europa*, los coches con cuatro cilindros representan el 90.4 % de los coches y, si consideramos el origen *Japón*, los coches con cuatro cilindros son el 87.3 %. Por tanto, el origen con mayor porcentaje de coches con cuatro cilindros es *Europa*, donde el 90.4 % de los coches tienen cuatro cilindros.

g) De los coches de 4 cilindros, ¿cuál es el porcentaje de coches japoneses?

En este caso, seleccionamos la opción `Porcentajes por filas`. Dicha opción nos muestra qué porcentaje representan las observaciones correspondientes a cada modalidad de la variable columna, cuando nos restringimos a cada modalidad de la variable fila.

Obtenemos en este caso que el 33.3 % de los coches con cuatro cilindros, tienen como origen *Japón*.