INTRODUCCIÓN A R-Commander y a la ESTADISTICA DESCRIPTIVA

La estadística es la ciencia que se encarga de extraer información de los datos. A tal efecto se encarga de los métodos para recoger esos datos (muestreo), de analizar y describir los datos (estadística descriptiva) y de extraer conclusiones de los datos (inferencia estadística).

La población es el conjunto de individuos (persona, objeto o cosa) donde se van a medir las variables bajo estudio. Una variable estadística es cualquier medición en los individuos de la población. Por ejemplo la población son vehículos de gama alta y en ellos podemos medir variables como la cilindrada (medida en cm^3), potencia (medida en CV), la tracción (4x4 o trasera), el combustible (gasolina, diésel, eléctrico), etc

Una base de datos estadística es un hoja de trabajo donde las filas se corresponden con diferentes individuos y las columnas con la variables que se van a medir en los individuos. En la celda (i,j) se tiene la medición de la variable j en el individuo i. Por ejemplo la hoja de trabajo Gama alta son los datos de 37 vehículos de gama alta (individuos) a los que se han medido las variables (marca, tipo, potencia, cilindrada, precio, tracción y combustible).

Las variables en el análisis estadístico se pueden dividir en cualitativas y cuantitativas. Las variables cualitativas son las que miden el grado en que un individuo posee una cualidad (tipo, marca, tracción y cilindrada). Las variables cualitativas a su vez se dividen en nominales y ordinales. Se dicen nominales cuando no existe un orden natural entre los valores de la variable cualitativa y ordinales cuando si existe. Ejemplo de variables nominales son el sexo (Hombre y Mujer), la tracción (4X4, trasera), tipo (turismo, monovolumen, todoterreno, deportivo), etc y ejemplos de variables ordinales son mes (Enero, Febrero, Marzo, Abril, Mayo, Junio, Julio, Agosto, Septiembre, Octubre, Noviembre, Diciembre), nota cualitativa en un examen (suspenso, aprobado, notable, sobresaliente, Matrícula de honor). Las variables cuantitativas son las que toman valores numéricos. Las variables cuantitativas se clasifican en discretas y continuas. Las variables discretas son las que toman un número finito o infinito numerable de valores por ejemplo la edad y el número de goles recibidos en una temporada por un equipo. Las variables continuas son las que toman valores en intervalos de los números reales o conjuntos que se expresan en términos de ellos por ejemplo la Altura de un conjunto de individuos o su Peso. Las variables cuantitativas suelen ser mediciones de longitud, peso, tiempo, etc. Las variables cualitativas y cuantitativas continuas requieren de diferentes tipos de análisis estadístico.

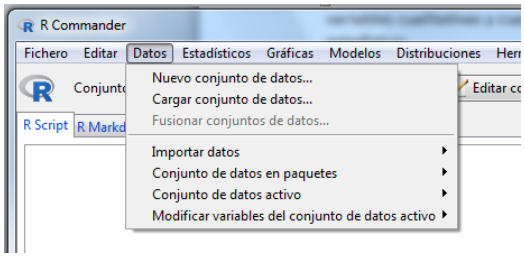
Las variables cuantitativas discretas que toman pocos valores distintos se pueden trabajar tanto con las técnicas de las variables cualitativas como con las de las

cuantitativas continuas. Cuando las variables cuantitativas discretas toman bastantes valores distintos sólo se pueden trabajar con las herramientas de las variables cuantitativas continuas.

Una variable que toma dos valores se denomina dicotómica, por ejemplo Sexo (Hombre, Mujer) y Fumador (Sí, No).

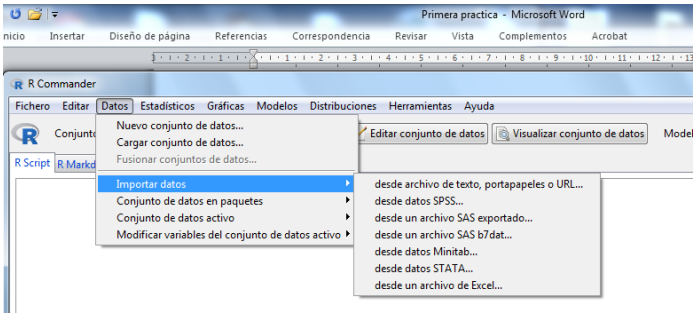
Primeros pasos en R-commander

En R-Commander se pueden trabajar tanto con bases de datos con formato de R que son aquellos que tienen la extensión .Rdata como con bases de datos en una hoja de cálculo tipo Excel. También es posible crear la propia base de datos en R-commander. Las opciones para las anteriores posibilidades están en la siguiente pestaña de R-commander

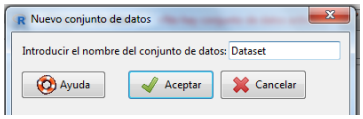


Siendo **Nuevo conjunto** de datos para crear una base de datos en R-commander, **Cargar conjunto de datos** para cargar un conjunto de datos en R-commander e **Importar datos** para importar una base de datos con formato de hoja de Excel, texto o de conocidos paquetes estadísticos.

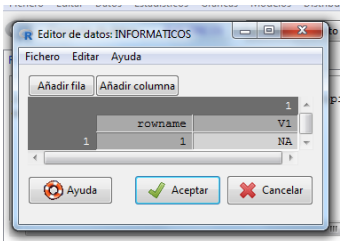
La opción **Importar datos** incluye las siguientes opciones



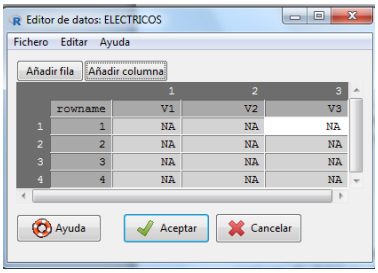
Si creamos una nueva base de datos tenemos que seleccionar la opción N**uevo conjunto de datos** de la pestaña de **Datos** y nos aparece el siguiente cuadro de dialogo



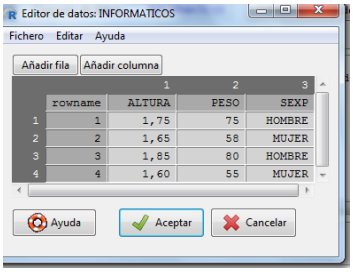
En el espacio donde está escrita la palabra Dataset pondremos el nombre del nuevo conjunto de datos sin dejar espacios ni utilizar caracteres especiales. Si por ejemplo ponemos INFORMATICOS y le damos a Aceptar obtenemos el siguiente cuadro de dialogo



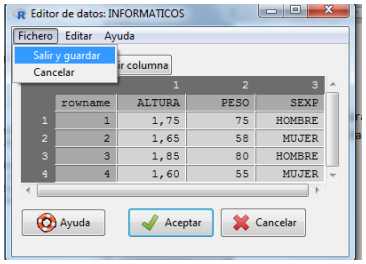
En esta ventana si damos a la opción **Añadir fila** creamos una nueva fila y a la de **Añadir columna** creamos una nueva columna. Así por ejemplo si deseamos crear una base de datos con cuatro individuos y 3 variables clicando en **Añadir fila** 4 veces y en **Añadir columna** 3 veces obtenemos:



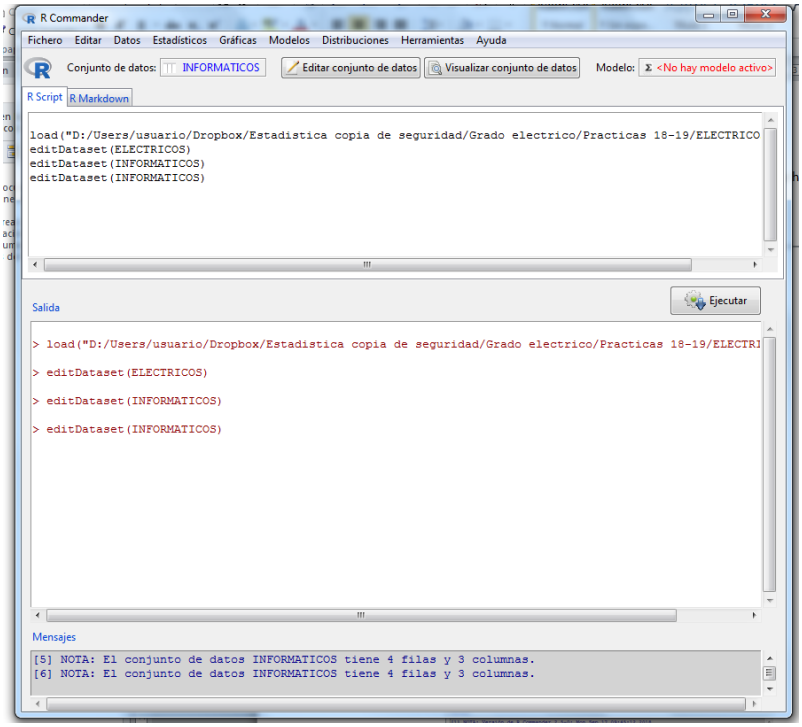
Ahora si las variables son ALTURA, PESO y SEXO si se ponen los nombres de las variables en V1, V2 y V3 y los datos en las celdas donde pone NA se obtiene:



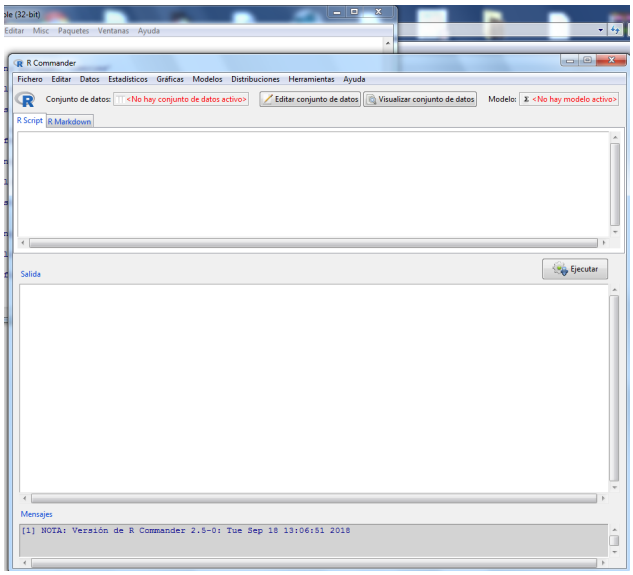
NA es el acrónimo en inglés de no disponible. Para salir de esta ventana hay que ir a **Fichero** y clicar en **Salir y Guardar**



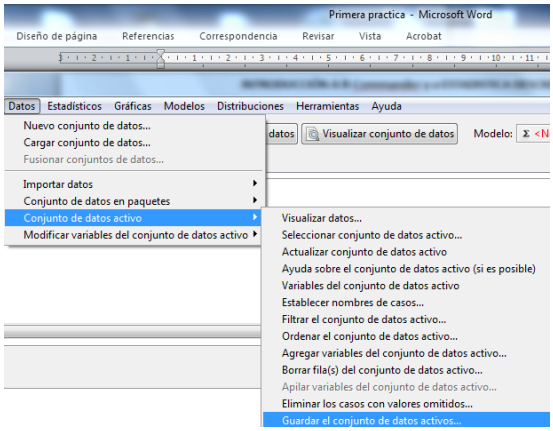
Obsérvese que el nombre del fichero está en azul en la parte de arriba



Si no hay fichero la misma zona está en color rojo No hay conjunto de de datos activo como aquí

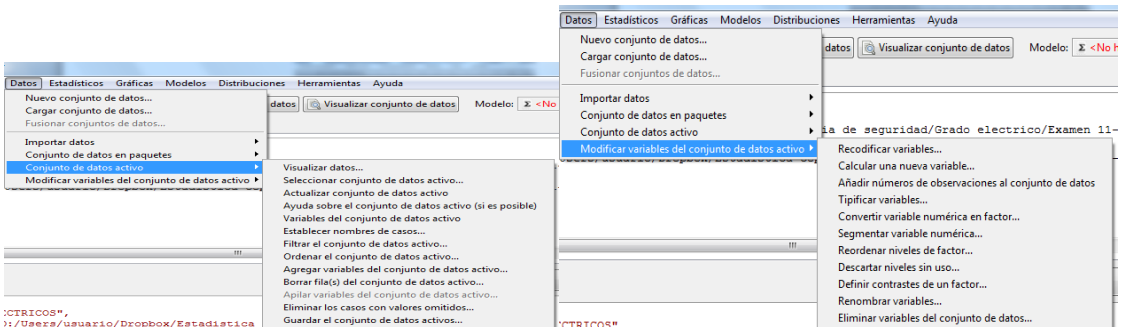


Si se quiere guardas el fichero INFORMATICOS como un fichero *.Rdata hay que ejecutar las siguientes opciones

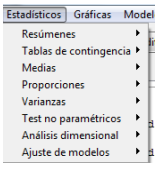


Si en el interface de R-commander aparecen mensajes en azul en la parte de abajo donde pone mensajes eso quiere decir que las instrucciones que estamos dando son correctas. Si los mensajes son en rojo o en verde lo que estamos haciendo es incorrecto.

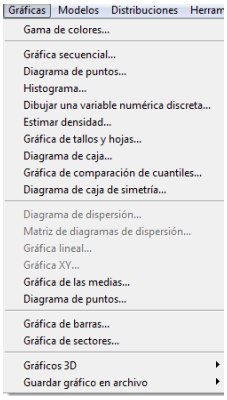
Para manipular los ficheros y sus variables se utilizan las opciones de la pestaña de **Datos Conjunto de datos activo** y **Modificar variables del conjunto de datos activo l**as cuales presenta diferentes opciones



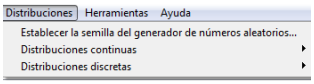
Ls pestaña **Estadísticos** con su despegable se emplea para realizar los análisis estadísticos de la variables del fichero



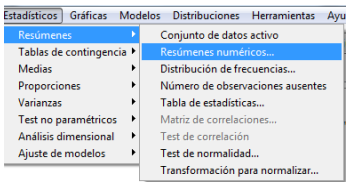
La pestaña **Gráficas** con su despegable se emplea para obtener los gráficos estadísticos de interés

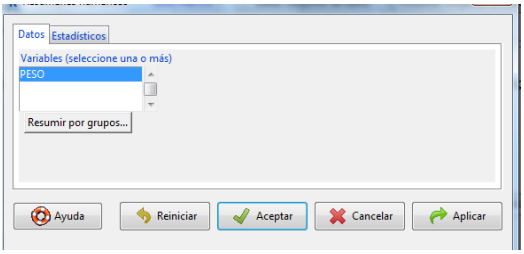


La pestaña **Distribuciones** permite hacer cálculos, gráficos y simulaciones de las distribuciones de probabilidad más habituales.

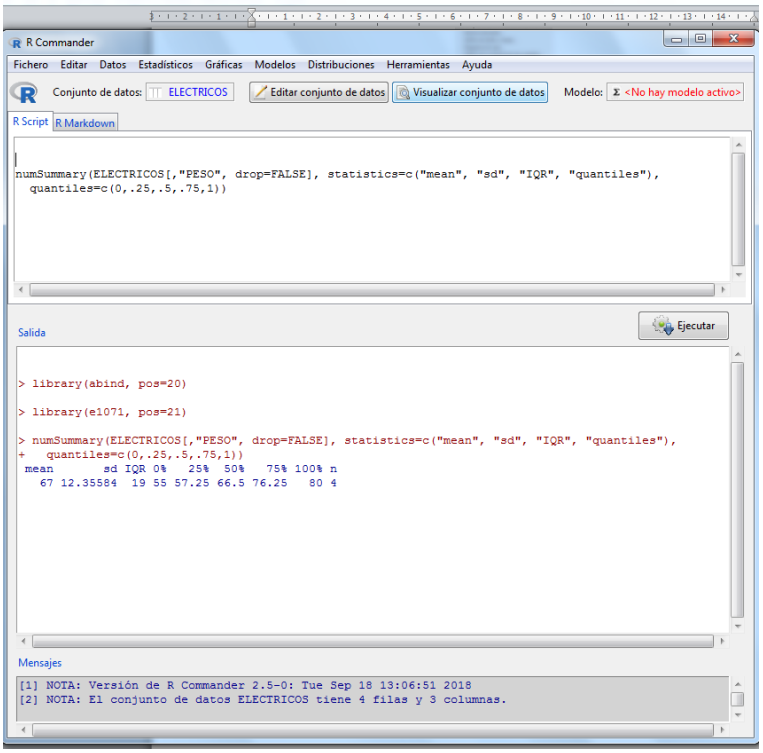


Cuando se realiza un cálculo aparecen en la parte de arriba de la pantalla las órdenes del lenguaje R que lo ejecutan y en la parte de abajo los cálculos o en una ventana separada si se trata de gráfico por ejemplo si se ejecuta para el fichero ELECTRICOS las siguientes opciones

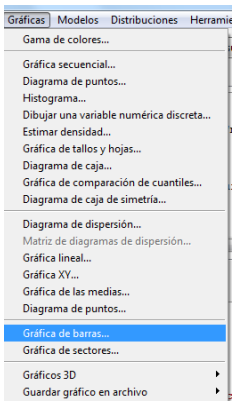


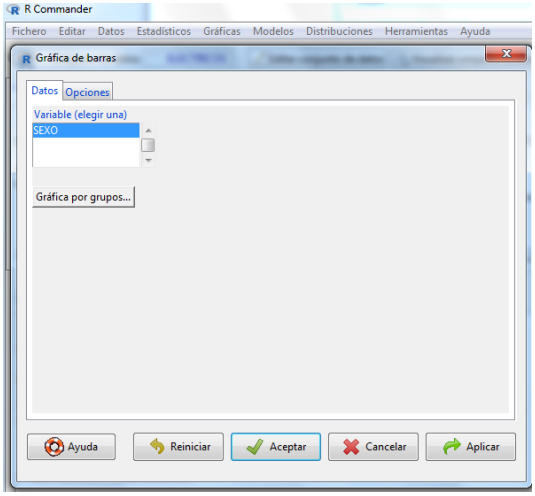


nos aparece en R-commander la siguiente salida de órdenes y análisis descriptivos

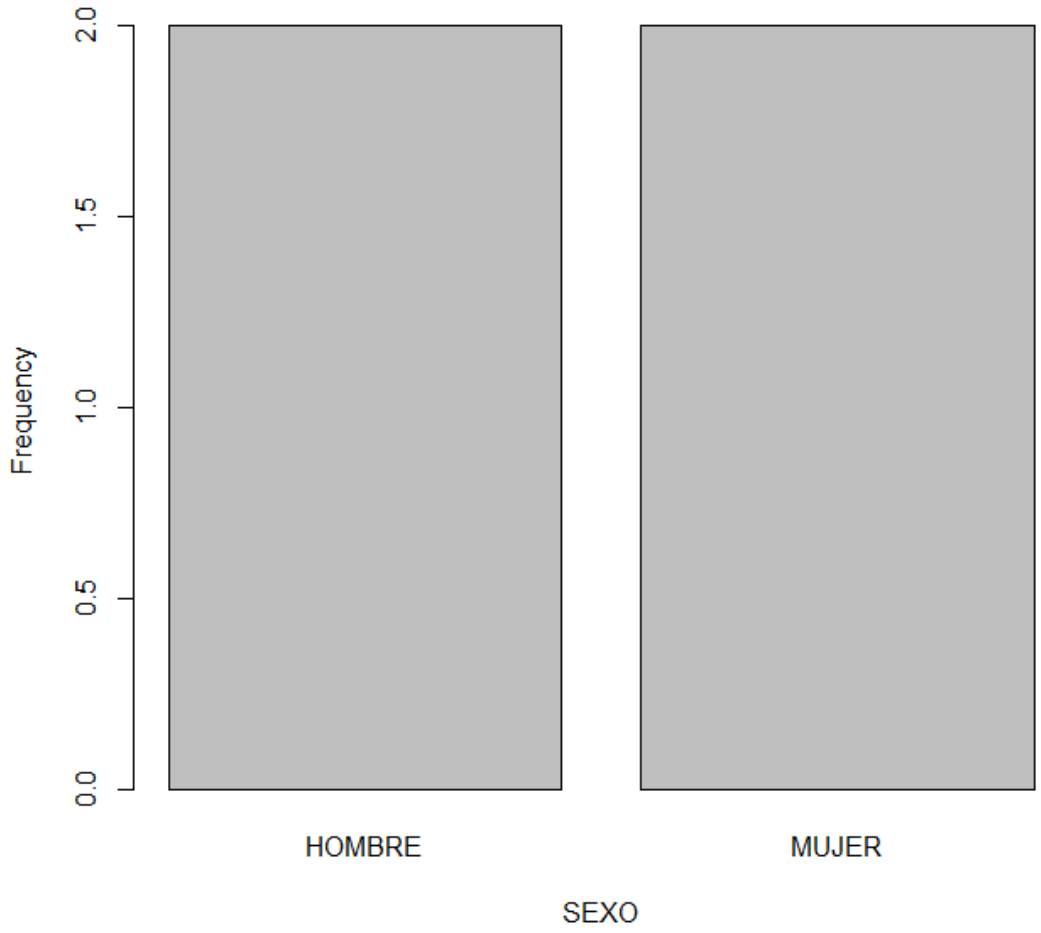


Análogamente para obtener el diagrama de barras de la variable SEXO del fichero eléctricos tenemos que ejecutar





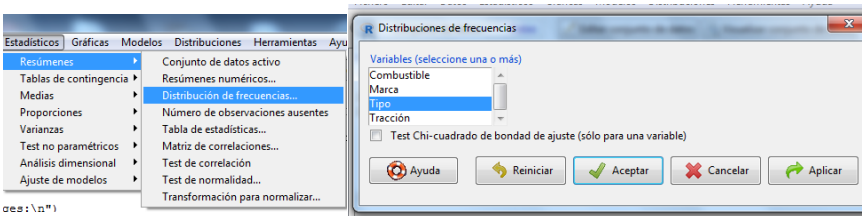
y nos aparece



Hay que tener cuidado en R-commander con el hecho de que la variable sea cualitativa o cuantitativa pues algunas opciones son solo para variables cualitativas (**Estadísticos Resúmenes Distribución de frecuencias** o **Gráficas Diagrama de barras)** o solo para variables cuantitativas (**Estadísticos Resúmenes Resúmenes numéricos**).

Análisis estadísticos de las variables cualitativas

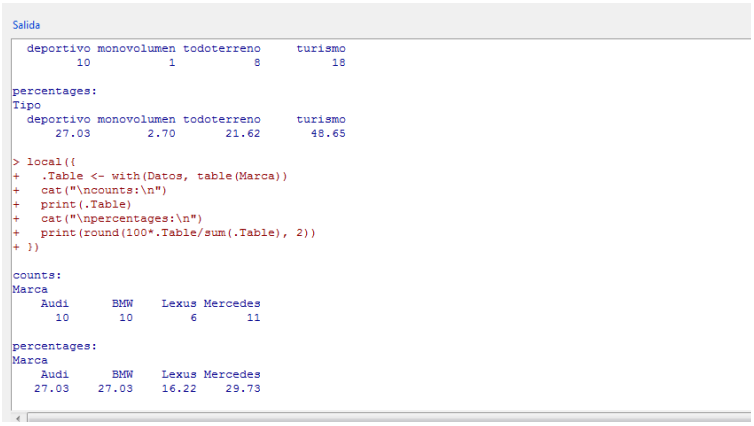
Una variable cualitativa se describe mediante la denominada distribución de frecuencias. La distribución de frecuencias consiste en contar cuantas veces aparece cada valor de la variable cualitativa y el porcentaje sobre el total de valores que representa el valor de la variable. En R-commander está opción está en la pestaña **Estadísticos Resúmenes Distribución de frecuencias**



La salida de R-commander de la distribución de frecuencias de la variable Tipo (deportivo, monovolumen, todoterreno, turismo) del fichero gamaalta.Rdata es

```
Salida

percentages:
    Audi      BMW    Lexus Mercedes
   27.03    54.05    70.27   100.00

> with(Datos, Hist(Precio, scale="frequency", breaks="Sturges", col="darkgray"))

> with(Datos, Hist(Precio, groups=Combustible, scale="frequency", breaks="Sturges", col="darkgray"))

> local({
+   .Table <- with(Datos, table(Tipo))
+   cat("\ncounts:\n")
+   print(.Table)
+   cat("\npercentages:\n")
+   print(round(100*.Table/sum(.Table), 2))
+ })

counts:
Tipo
  deportivo monovolumen todoterreno     turismo
         10           1           8          18

percentages:
Tipo
  deportivo monovolumen todoterreno     turismo
      27.03        2.70       21.62       48.65
```

Vemos que de los 37 coches 18 son turismos y representan el 48.65% de los vehículos, mientras que los vehículos de tipo deportivo y todoterreno representan el 27.03% y el 21.62%, respectivamente. Vehículos de tipo monovolumen sólo hay uno y representa el 2.70% del total. La salida de la distribución de frecuencias de la variables Marca (Audi, BMW, Lexus, Mercedes) del mismo fichero es



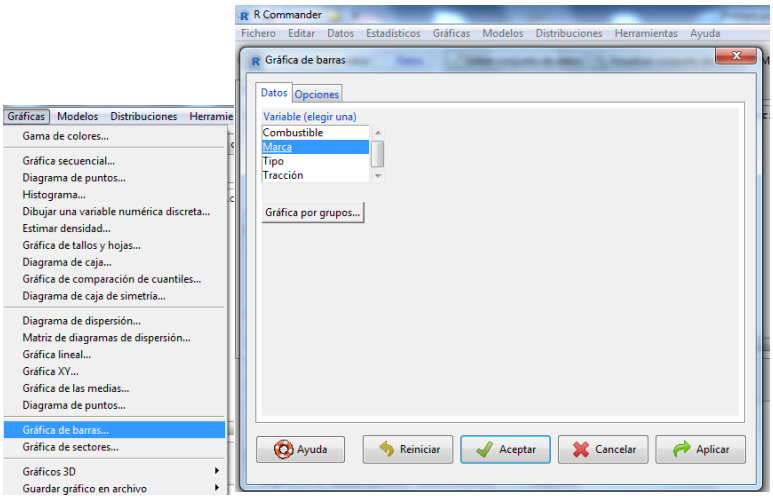
```
Salida

  deportivo monovolumen todoterreno     turismo
         10           1           8          18

percentages:
Tipo
  deportivo monovolumen todoterreno     turismo
      27.03        2.70       21.62       48.65

> local({
+   .Table <- with(Datos, table(Marca))
+   cat("\ncounts:\n")
+   print(.Table)
+   cat("\npercentages:\n")
+   print(round(100*.Table/sum(.Table), 2))
+ })

counts:
Marca
    Audi      BMW    Lexus Mercedes
      10       10        6       11

percentages:
Marca
    Audi      BMW    Lexus Mercedes
   27.03    27.03    16.22    29.73
```

Se observa que la frecuencia más elevada de Marca es la de Mercedes que representa casi el 30% de vehículos (29.73%) seguida muy de cerca por la de Audi y BMW que tienen ambos la misma frecuencia (27.03%). La frecuencia más baja es la de vehículos Lexus que es un poco más que la mitad de la de Mercedes (16.22%).

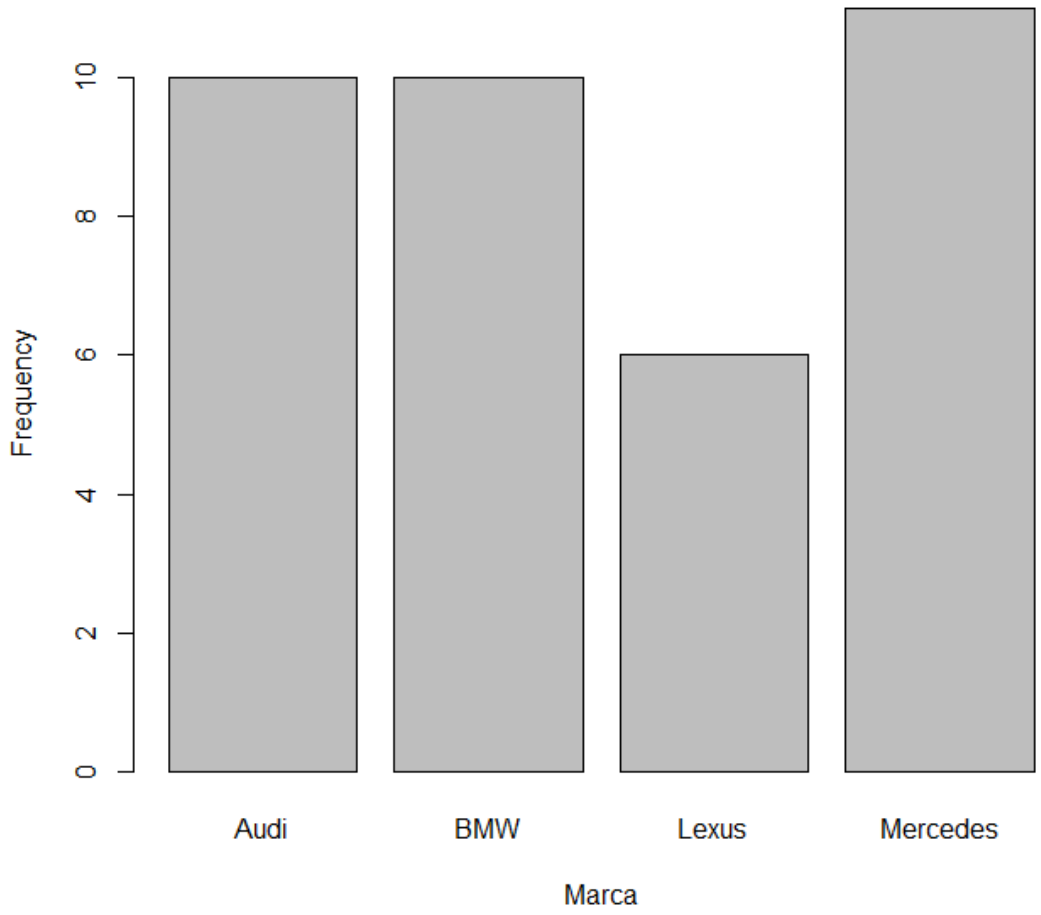
La moda de una variable cualitativa es el valor con mayor frecuencia. En los anteriores ejemplos la moda de Tipo es turismo y la de Marca Mercedes.

Si queremos representar gráficamente una variable cualitativa debemos utilizar el diagrama de barras o el diagrama de sectores. En el diagrama de barras se representan en el eje horizontal los valores de la variable cualitativa y a cada valor de la variable cualitativa se le asocia una barra vertical de altura la frecuencia absoluta (número de

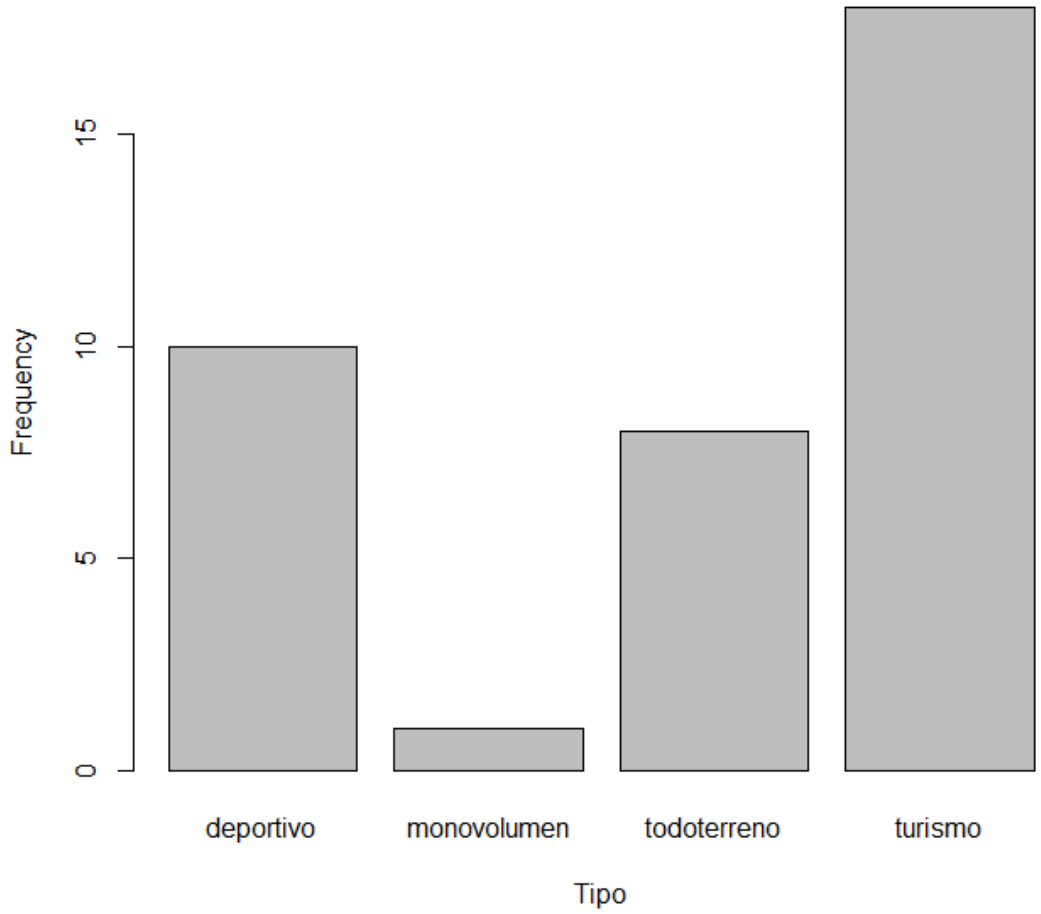
veces que aparece el valor) de la misma. En R-commander para realizar esto tenemos que utilizar las siguientes opciones



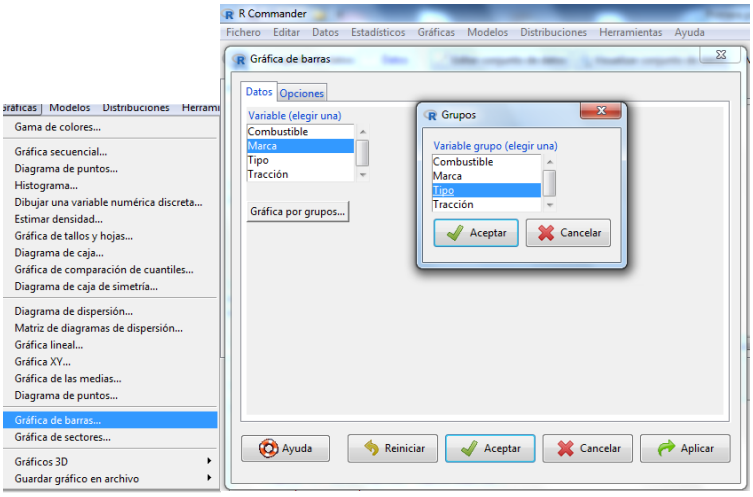
El diagrama de barras de la variable Marca del fichero gammaalta.Rdata es



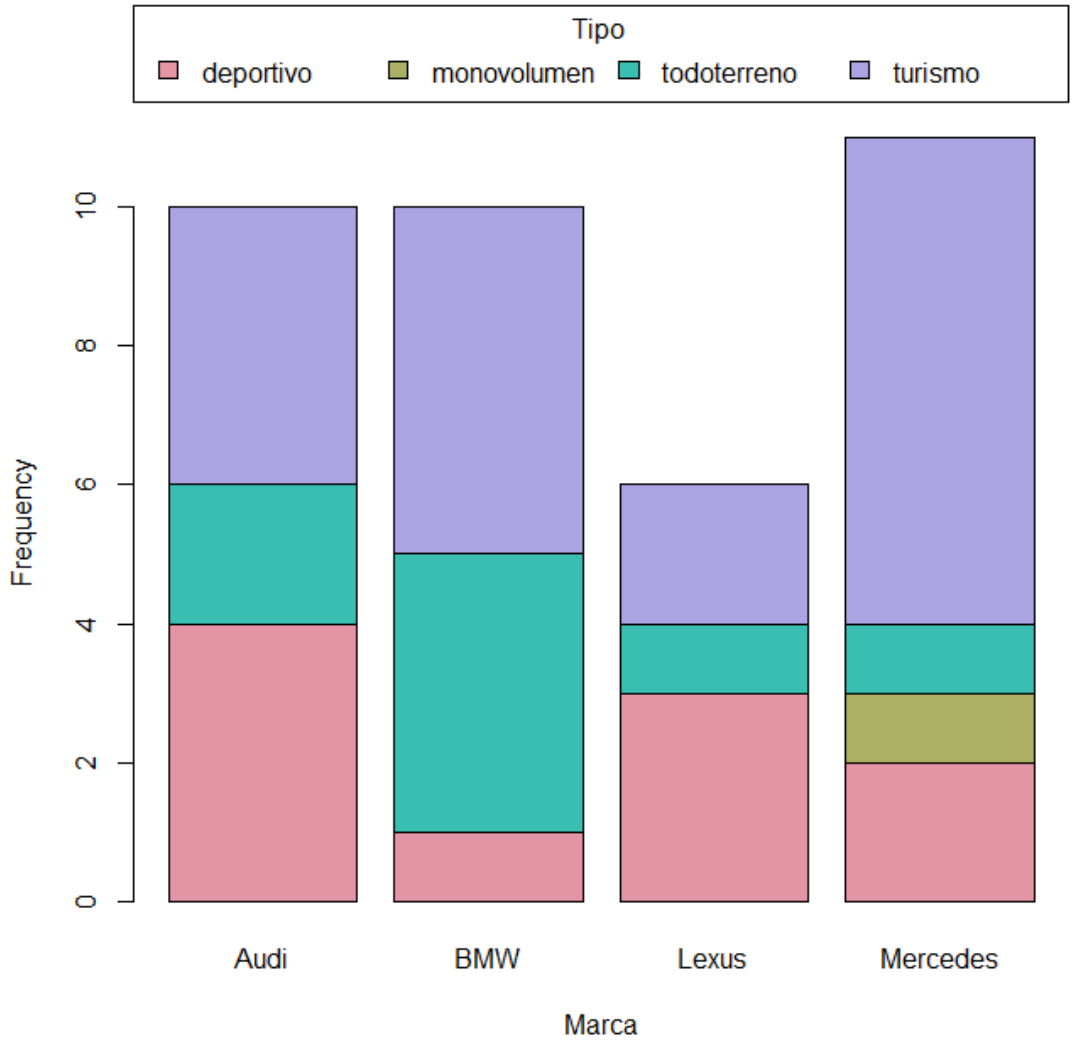
y el de la variable Tipo del mismo fichero es



En un mismo gráfico de barras podemos representar dos variables cualitativas. Esto se hace con las opciones

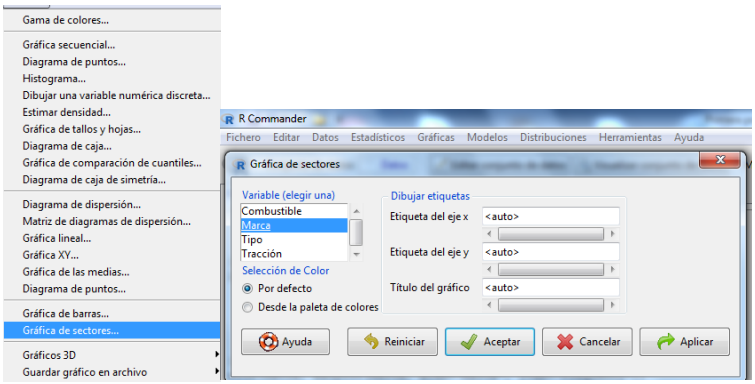


En el caso de las variables Marca y Tipo el gráfico de R-commander es

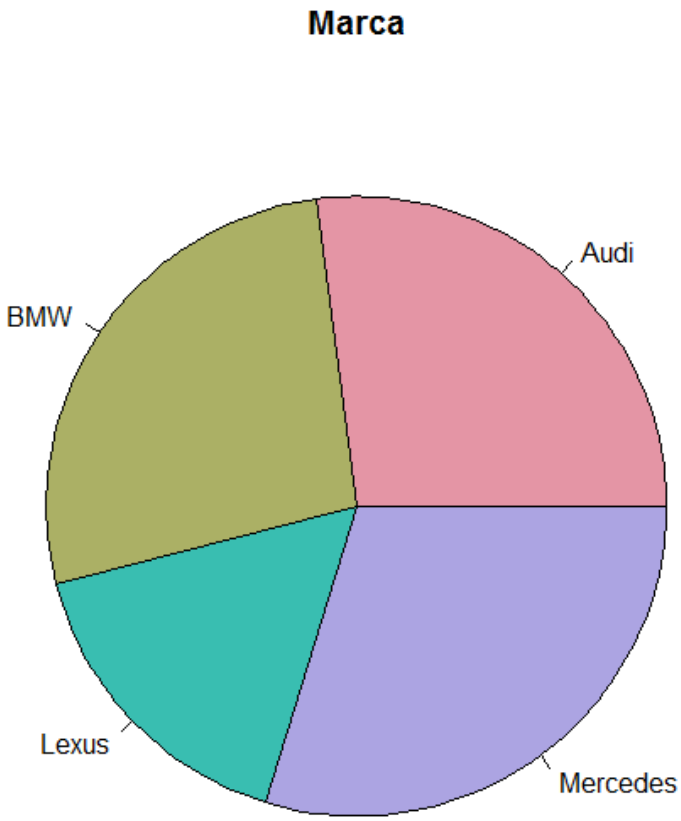


El anterior gráfico se interpreta como que de los 10 Audi 4 eran deportivo, 2 todoterreno y 4 turismo y de los 11 Mercedes 2 eran deportivo, 1 monovolumen, 1 todoterreno y 5 turismo. El anterior gráfico se denomina gráfico de barras apilado.

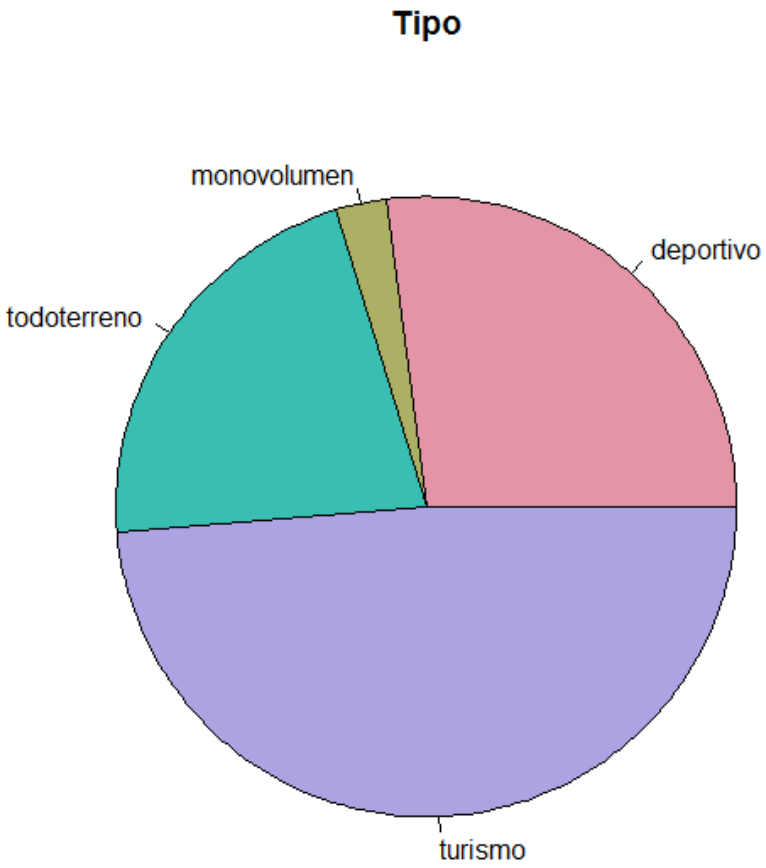
El diagrama de sectores es una representación gráfica de una variable cualitativa donde cada valor se representa en un sector circular de área proporcional a su frecuencia. En R-commander se hace con las opciones



En el caso de la variable Marca el gráfico de sectores de R-commander es

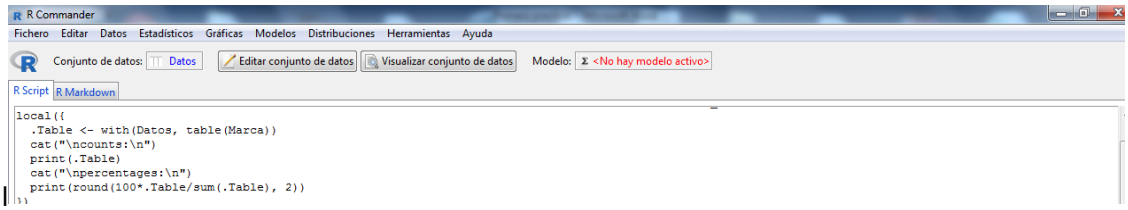


Y para la variable Tipo es

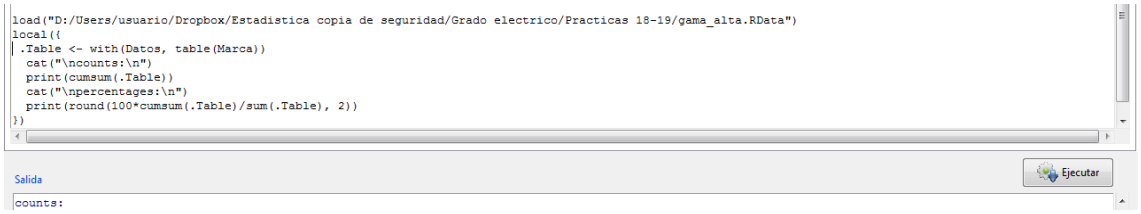


Con una programación básica en R se pueden disfrutar de más utilidades estadísticas de R-commander. Un de ellas de la tabla de frecuencias acumuladas con la orden **cumsum**. Al objeto de obtener la tabla de frecuencias acumuladas de la variable Marca del fichero gamaalta.Rdata. Se hace la distribución de frecuencias normal mediante el procedimiento habitual para Marca y las ordenes de R para hacer eso en R son las siguientes;

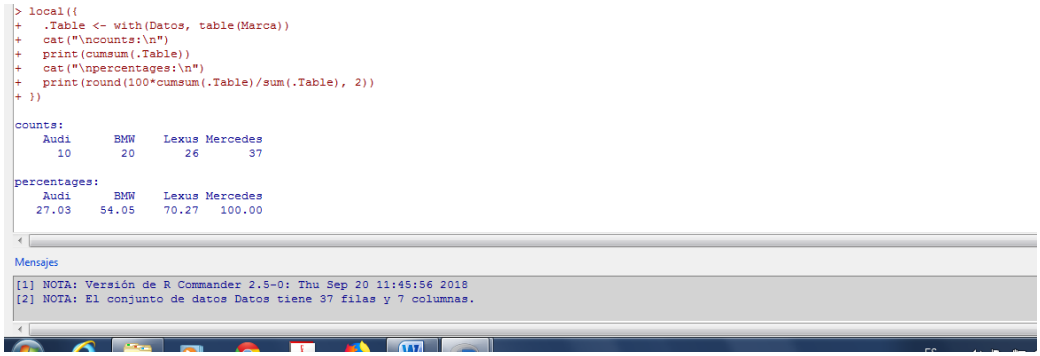
load("D:/Users/usuario/Dropbox/Estadistica copia de seguridad/Grado electrico/Practicas 18-19/gama_alta.RData")



A continuación se copian esas mismas órdenes a continuación y donde esta escrito .Table se sustituye por cumsum(.Table) resultando



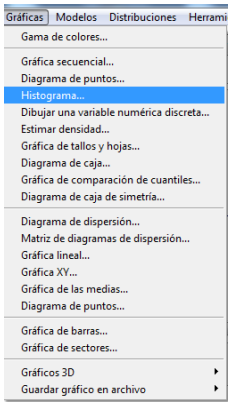
Se seleccionan esas órdenes y a continuación se clica en Ejecutar resultando:

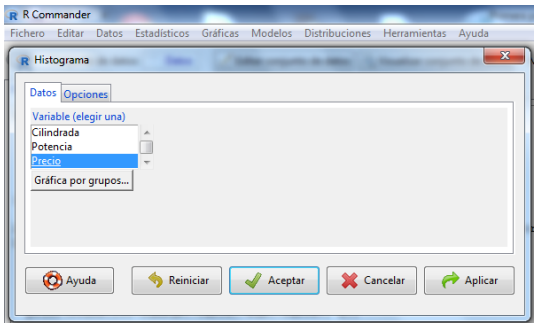


La cúal es la tabla de frecuencias acumuladas de la variable Marca.

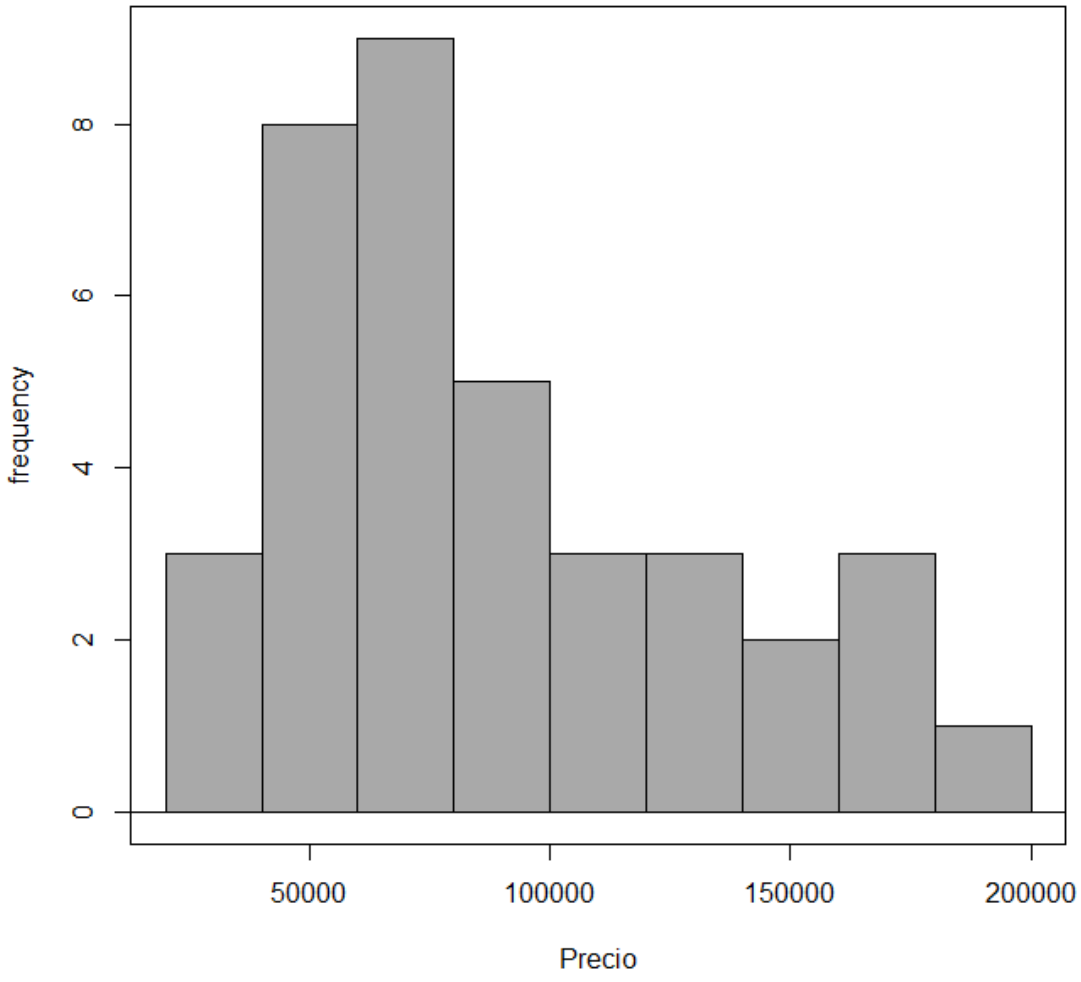
Análisis de las variables cuantitativas

Las variables cuantitativas continuas no se pueden representar gráficamente mediante un diagrama de barras o un diagrama por sectores pues habría muchos valores que representarían un porcentaje pequeño del total de valores. Las variables cuantitativas continuas se representan mediante el histograma o el diagrama de caja. El histograma es una representación gráfica para lo cual hay que dividir el rango de los datos en intervalos. Estos intervalos suelen ser de igual longitud, aunque en algunos casos se pueden considerar intervalos de diferente amplitud. Un ejemplo de variable para la que convendría dividir el rango en intervalos de diferente amplitud es la renta de las familias españolas en el año 2017, pues para las rentas altas se deberían hacer intervalos de mayor amplitud. Al objeto de dividir un conjunto de n datos en r intervalos de igual amplitud, la longitud de cada intervalo es $h = \frac{M-m}{r}$ siendo M el máximo valor de los n datos y m el mínimo y los intervalos serían $I_{j=[m+(j-1)h,m+jh]}$, j=1,2, … ,r. Los programas informáticos profesionales estadísticos emplean fórmulas complejas para determinar el número r de intervalos para un conjunto de n datos. Obviamente r debe ser una función creciente de n. Una regla empírica razonable es considerar que el número de intervalos r debe ser del orden de $\sqrt{n}$. Para construir el histograma se cuentan el número de datos n_j en el intervalo I_j y se representan los intervalos en el eje horizontal y a cada intervalo se le asocia un rectángulo de base el intervalo I_j y de altura I_j de modo que el área es proporcional a la frecuencia del intervalo. Si los intervalos son de diferente longitud hay que refinar esta metodología pues la longitud de los intervalos al ser diferente produce que si asociamos un rectángulo de altura n_j no sean proporcionales a la frecuencia del intervalo, para solventar esto bastaría con levantar una altura $n_j / long I_j$ donde el término en el denominador es la longitud del intervalo I_j. El histograma de la variable Precio del fichero gamaalta.Rdata se obtiene con



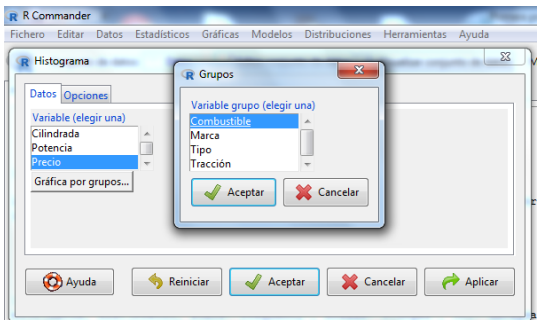


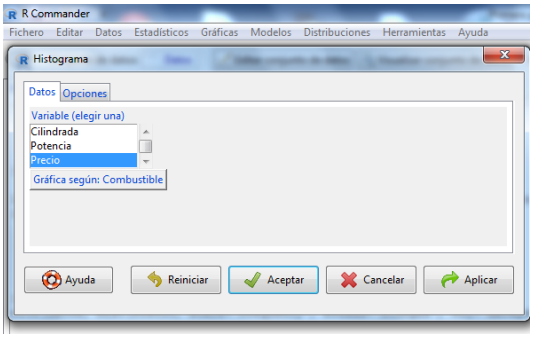
y produce el siguiente gráfico



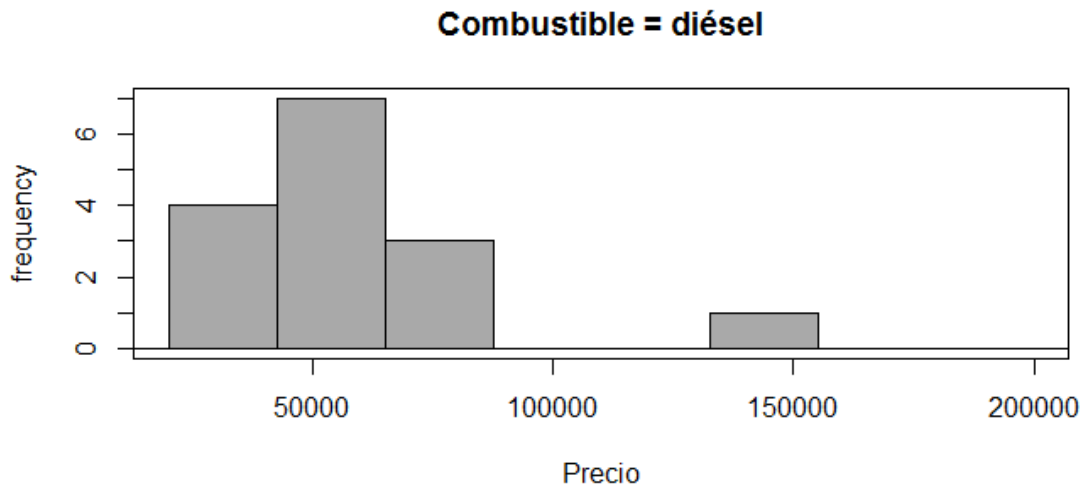
que es un histograma con 9 intervalos de igual amplitud para 37 datos. El histograma está sesgado para la derecha (más adelante veremos que significa cuando hablemos de la simetría de los datos).

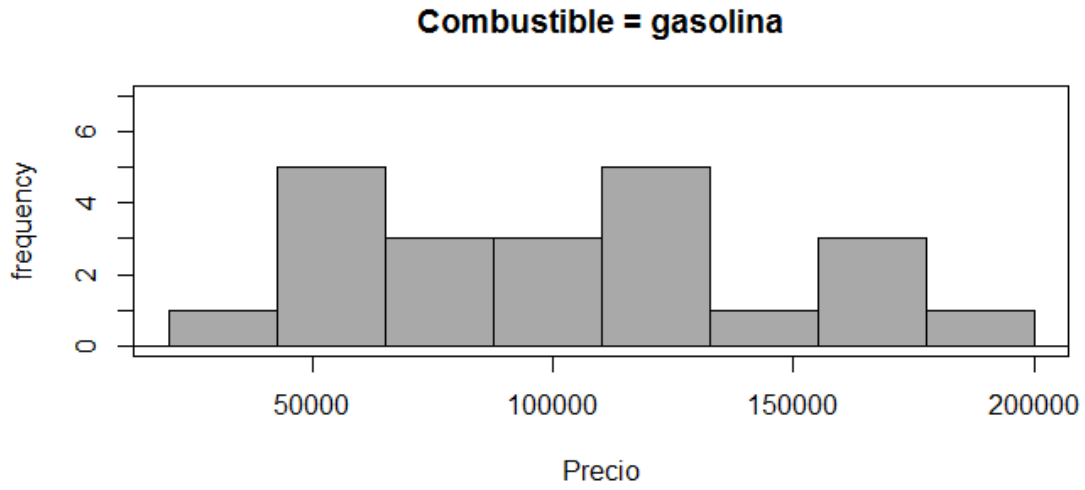
Se puede representar el histograma de Precio para cada valor de una variable cualitativa. Por ejemplo el combustible (gasolina, diésel) con la siguiente opción (donde se ha seleccionado **Gráfica por grupos**)





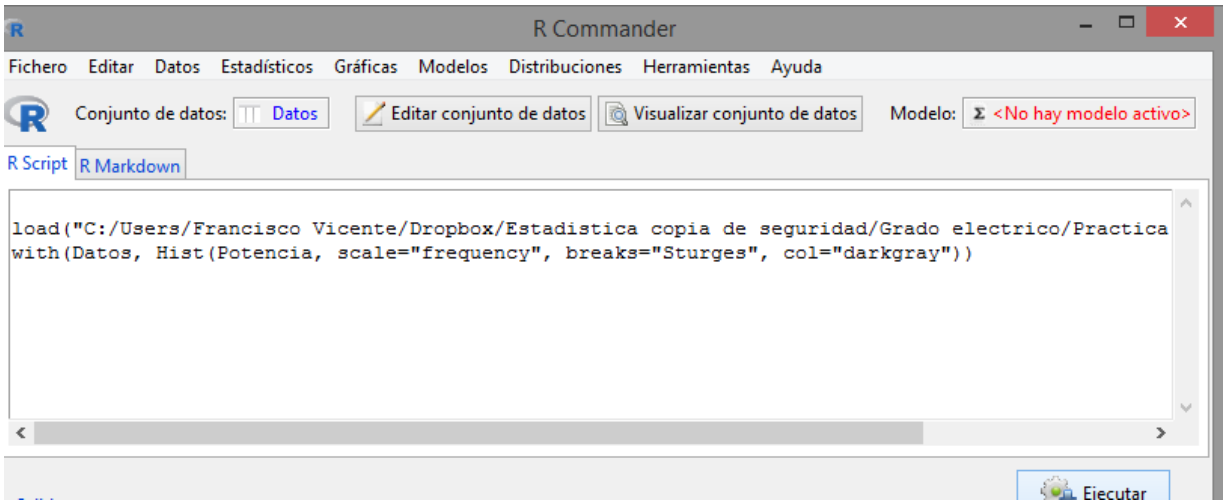
Y se obtiene el siguiente gráfico



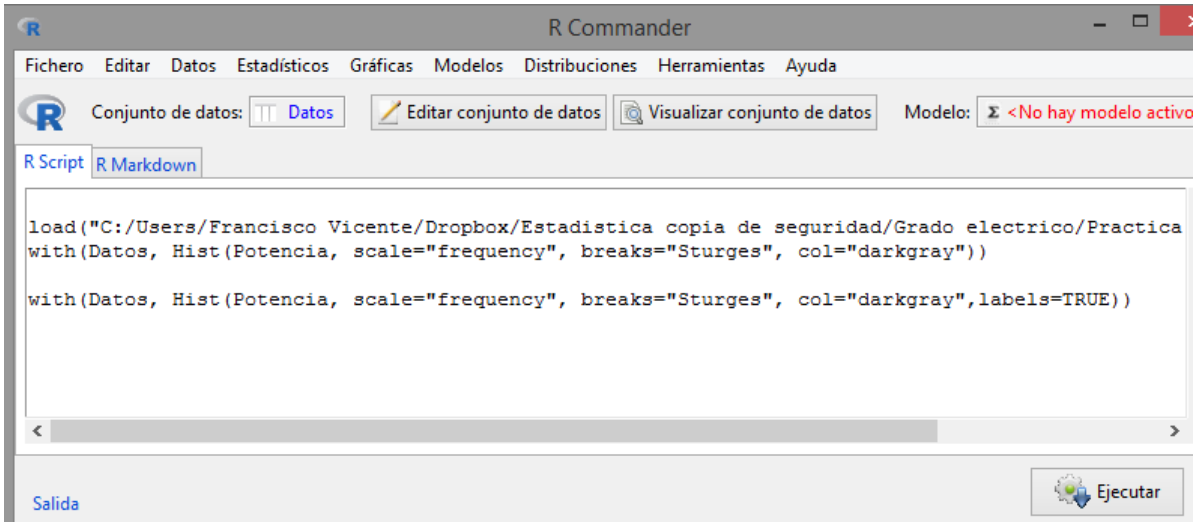


Se observa que el gráfico del Precio para los vehículos de gasolina es más simétrico que el gráfico del Precio para los vehículos diesel que está algo sesgado a la derecha por la presencia de un vehículo con precio muy grande (tal dato es un dato atipico del cual hablaremos más adelante).

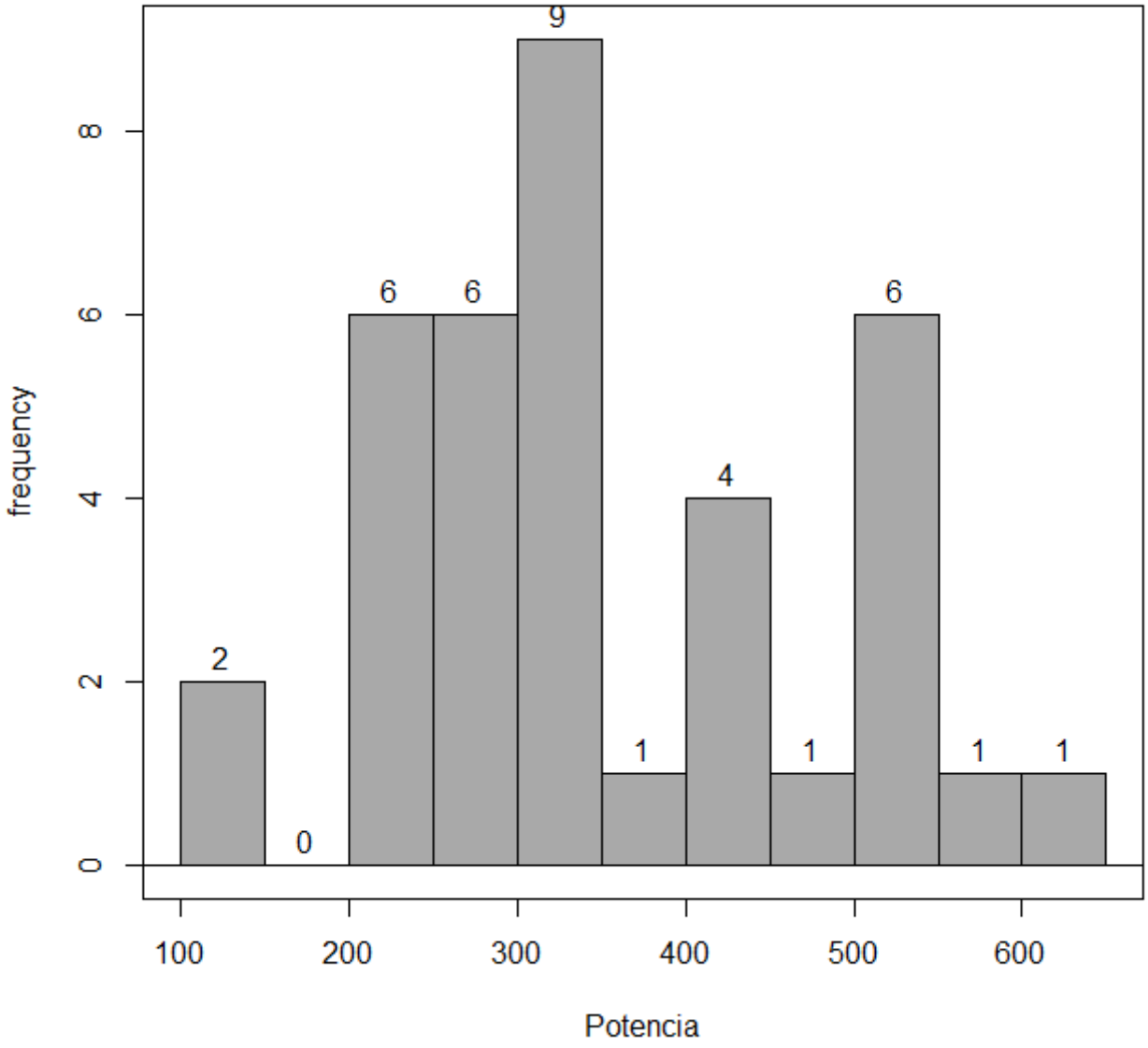
Programando un poco en R podemos conseguir que en el histograma aparezcan el número de datos del histograma. Se consigue añadiendo a las órdenes para obtener un histograma labels=TRUE. Las órdenes de un histograma de la variable Potencia son



La orden se cambia añadiendo labels=TRUE



Se selecciona y se le da a ejecutar y obtenemos el histograma de potencia con el número de datos en cada intervalo



Medidas númericas descriptivas

Al objeto de describir mejor las variables cuantitativas asociaremos números a sus valores que nos proporcionarán información relevante sobre la variable. Estos números nos informarán acerca de la tendencia central de la variable, la posición, la variabilidad o dispersión y la forma (simetría y apuntamiento)

Las medidas de tendencia central son la media, la mediana, la moda y las medias recortadas.

La media denotada $\bar{X}$ se define como la media aritmética de los n valores de la variable, es decir,

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

La media es la mejor medida de tendencia central y es un valor entorno al cúal tienden a agruparse los valores de la variable. El único incoveniente que tiene la media es que cuando el número de datos n de la variable es pequeño un dato que en estadística llamaremos atípico (muy grande o muy pequeño con respecto a un criterio que definiremos más adelante) puede distorsionar el valor de la media aumentándolo o disminuyéndola mucho. Si un conjunto de datos es simétrico respecto a un valor (al representar los datos los datos a ambos lados del punto de simetría coinciden) la media es el punto de simetría.

Un alternativa a la media que se ve afectada por la presencia de datos atípicos en menor medida que la media es la mediana. La mediana es el valor central de los datos ordenados. Es el valor que deja el 50% de los datos de la variable por encima y el 50% por debajo. El cálculo de la mediana se realiza ordenando los datos de la variable de menor a mayor. En el caso de un número de datos n sea impar con n=2m+1 el dato que ocupa la posición m+1 en la ordenación es la mediana. Si el número de datos n es par con n=2m la media aritmética de los datos que ocupan la posición m y m+1 en los datos ordenados es la mediana. Si un conjunto de datos es simétrico alrededor de un valor la mediana es el punto de simetría.

Una medida muy básica de tendencia central que realmente es sólo interesante para variables que toman pocos valores es la moda. La moda es el valor que más se repite de una variable estadística.

Otra medida muy interesante y robusta frente a los datos atípicos son las medias recortadas al 5%. La media recortada al 5% es la media aritmética del 90% de los datos centrales de la variable, es decir, en el cálculo de la media se eliminan el 5% de los datos extremos por la izquierda y por la derecha.

Las medidas de posición nos indican la posición que ocupan los datos. El percentil de pX100% con $0 \leq p \leq 1$ es el valor de la variable que deja el pX100% de los datos por debajo y el (1-p)X100%. Si el percentil del 90% de la altura de los niños de 5 años es 1,20 metros esto quiere decir que el 90% de los niños de 5 años miden menos de 1,20 metros y el 10% miden más. Unos percentiles de interés son los cuartiles. Los cuartiles son 3 puntos que dividen los datos en 4 zonas con el mismo porcentaje de datos, es decir, el 25%. Los 3 cuartiles se denotan Q1, Q2 y Q3. El Q1 se corresponde con el percentil del 25%, el Q2 con el percentil del 50% o mediana y el Q3 con el percentil del 75%. El percentil del 0% se corresponde con el mínimo de los datos y el del 100% con el máximo de los datos.

La media, mediana, moda, medias recortadas y percentiles se miden en la mismas unidades que la variable.

La variabilidad es un concepto básico en estadística que mide como varian los valores de la variable alrededor de los valores centrales. En el caso de que todos los valores sean iguales no hay variabilidad y cuanto más varian los valores del valor central mayor variabilidad o dispersión. La medida de dispersión más básica de dispersión es el rango que se define como la diferencia entre el valor máximo y mínimo de la variable. Si el rango vale 0 todos los valores de

la variable son iguales y en otro caso significa que la variable tiene valores distintas. El rango no es una medida muy informativa de la dispersión. La medida por autonomasia de dispersión es la varianza muestral denotada s^2 y que se define como

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Obsérvese que la anterior cantidad es mayor cuanto más difieran los valores de la media y que la medida es no negativa siendo cero cuando todos los valores coinciden con la media. La ponderación de las distancias a la media al cuadrado es n-1 en vez de n por una razón técnica que se justicará en el apartado de inferencia estadística del programa de la asignatura. Las unidades de medición de s^2 son las de la variable al cuadrado por ello para medir la dispersión o variabilidad se utiliza más su raíz cuadrada denotada por s y que se denomina desviación típica. Su fórmula es

$$s = \sqrt{s^2}$$

y se mide en las mismas unidades que la variable.

Al objeto de comparar la dispersión de dos variables en diferentes unidades se emplea el coeficiente de variación denotado como CV y que se define como

$$CV = \frac{s}{|\bar{x}|}$$

cuando $\bar{x} \neq 0$. El coeficiente de variación de una variable no posee unidades de medición (adimensional). Además de comparar variables medidas en diferentes unidades sirve para comparar la dispersión de variables medidas en las mismas unidades con un gran valor de diferencia en la media.

Otra medida de dispersión es el error típico de la media que mide la dispersión de la media de n valores de una variable se de denota SE$\bar{x}$ y se define como

$$\text{SE}\bar{x} = \frac{s}{\sqrt{n}}$$

Nótese que la variabilidad de la media de los n elementos dada por SE$\bar{x}$ es menor que la variabilidad de la variable dada por s.

Otra medida de dispersión es el rango intercuartílico denotado IQR por su acronimo en inglés y que se define como la diferencia entre el tercer cuartil y el primer cuartil, es decir,

$$IQR = Q3 - Q1$$

La justificación de esta medida es la siguiente. Entre el tercer y primer cuartil se encuentran el 50% de los valores centrales de la variable pues la mediana o Q2 se encuentra allí y, por tanto, cuanto mayor sea el IQR mayor es la dispersión.

La forma de los datos hace referencia a su simetría y al apuntamiento. La simetría mide lo simétricos que son los datos de una variable. El ejemplo más claro de simetría de los datos es cuando lo simétrico que son los datos alrededor de un valor. La simetría de los datos se mide con el coeficiente de asimetría denotado como CA y que se define como

$$CA = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}$$

El coeficiente de asímetría es adimensional. Obsérvese que el signo del CA depende del signo de cada uno de sus términos, verificándose que un término contribuye positivamente si es mayor que la media el valor y negativamente si es menor. Si cada término positivo tiene asociado un término negativo a una distancia parecida de la media el CA será cercano y los datos se dicen simétricos. Si los términos positivos están mas alejados de la media que los términos negativos el CA es positivo y los datos se dicen asimétricos positivo o sesgados a la derecha. Si los términos negativos están mas alejados de la media que los términos positivos el CA es negativo y los datos se dicen asimétricos negativo o sesgados a la izquierda. R-commander se refiere al coeficiente de asimetría con su palabra en inglés que es skewness.

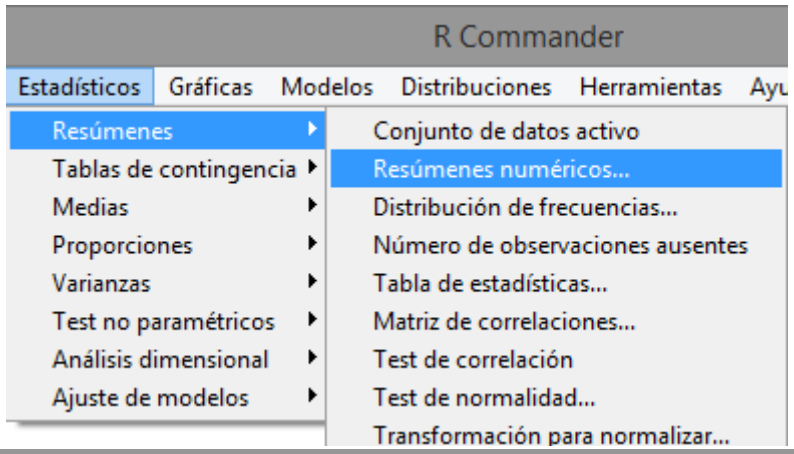
En los datos con coeficiente de asimetría cercano a 0 (simétricas) la media es muy parecida a la mediana. En los datos asimétricos positivos o sesgados a la derecha (coeficiente de asimetría claramente mayor que 0) la media es más grande que la mediana. En los datos asimétricos negativos o sesgados a la izquierda (coeficiente de asimetría claramente menyor que 0) la media es más pequeña que la mediana.

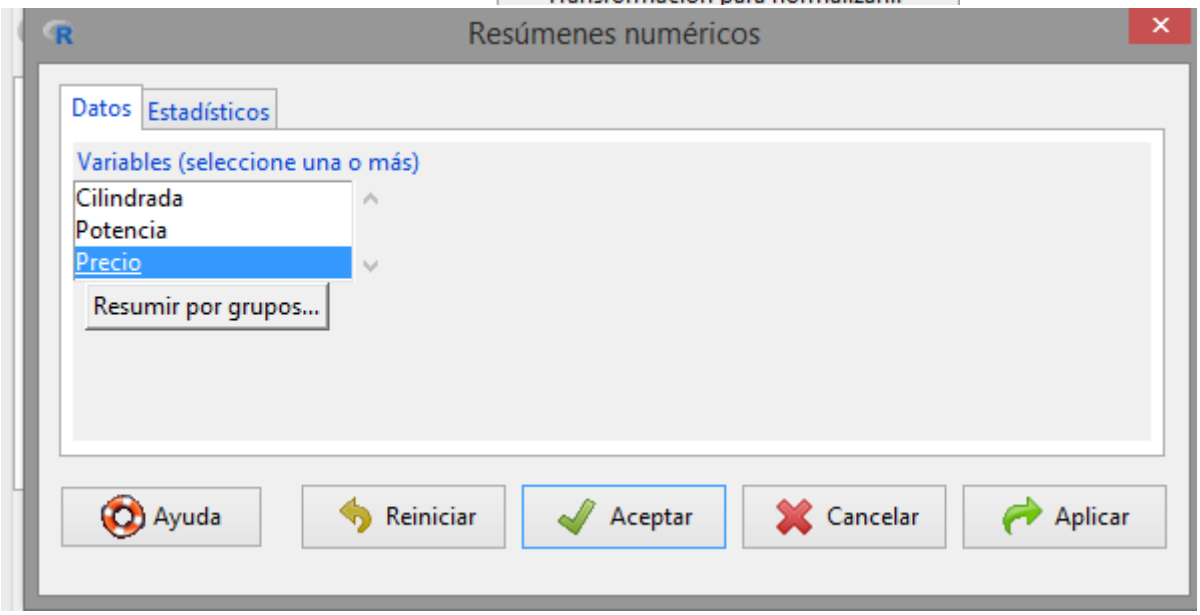
Para distinguir datos con la misma media, desviación típica que son simétricos se emplea el coeficiente de apuntamiento o curtosis (en inglés kurtosis) que se denota CAp y que se define como

$$CAp = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

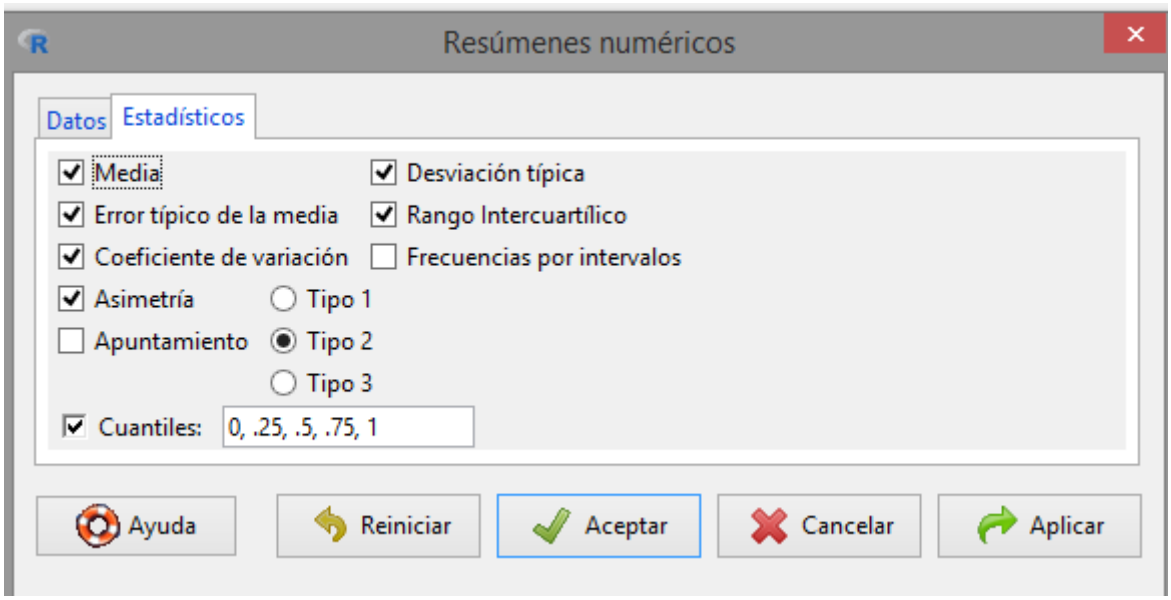
El CAp es como el CA adimensional. El valor 3 es debido a que los datos normales que son simétricos con es media y desviación típica tienen CAp=0. El equilibrio para datos simétricos con la misma media y desviación típica para repartir los datos entre el centro y los extremos lo proporcionan los datos normales con CAp=0. Unos datos con CAp cercano a 0 se dicen mesocúrticos El hecho de que el CAp es mucho mayor que 0 significa que esos datos concentran más en el centro que en los extremos que unos datos normales de esa media y desviación típica y se llaman leptocurticos. Si el CAp es muy negativo esos datos dan más peso a los extremos que al centro que unos datos normales de esa media y desviación tipica y se denominan platicúrticos. R-commander se refiere al coeficiente de apuntamiento con su palabra en inglés que es kurtosis.

Las medidas numéricas descriptivas de las variables cuantitativas se obtienen en R-commander como





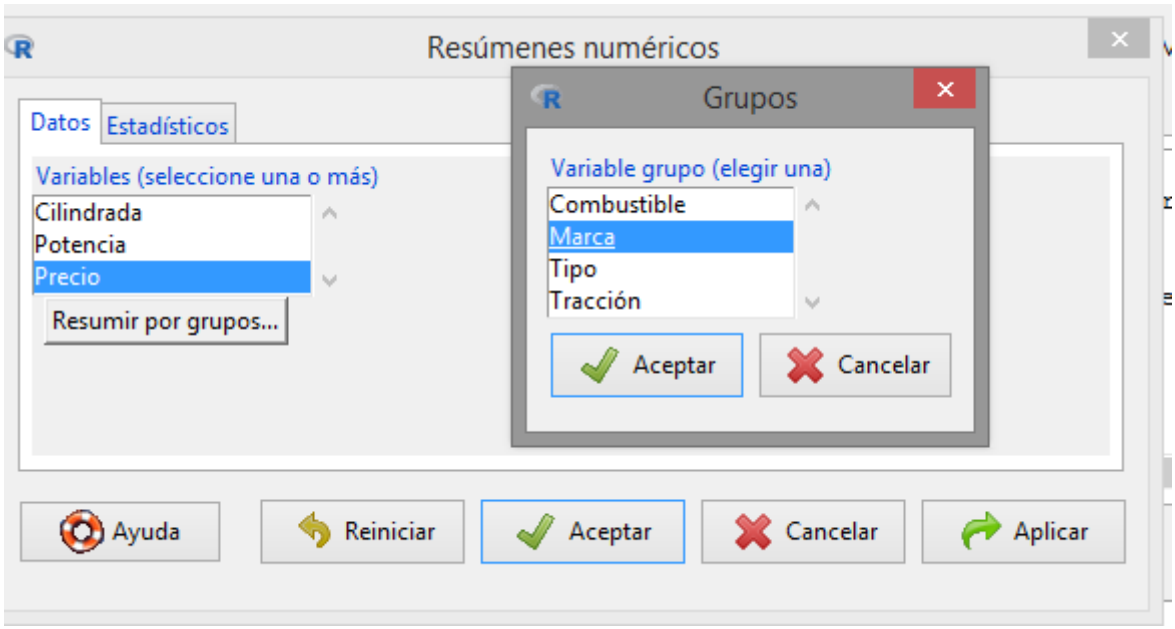
Si pulsamos en la pestaña Estadísticos podemos seleccionar las medida numéricas descriptivas que calcula R-commander

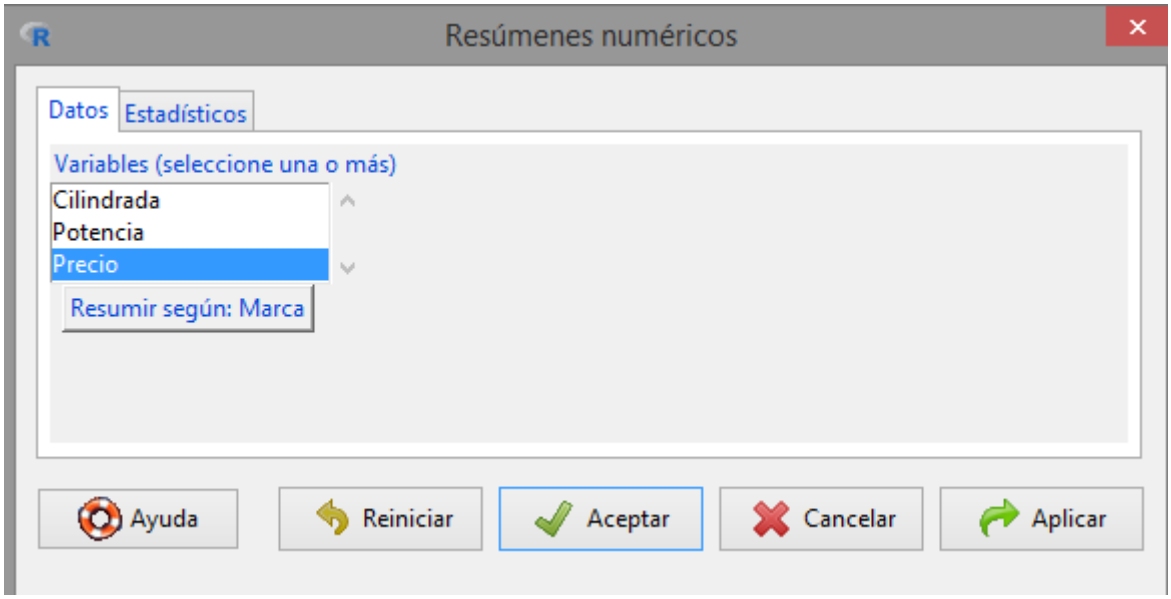


La salida de R-commander una vez seleccionadas las opciones anteriores es

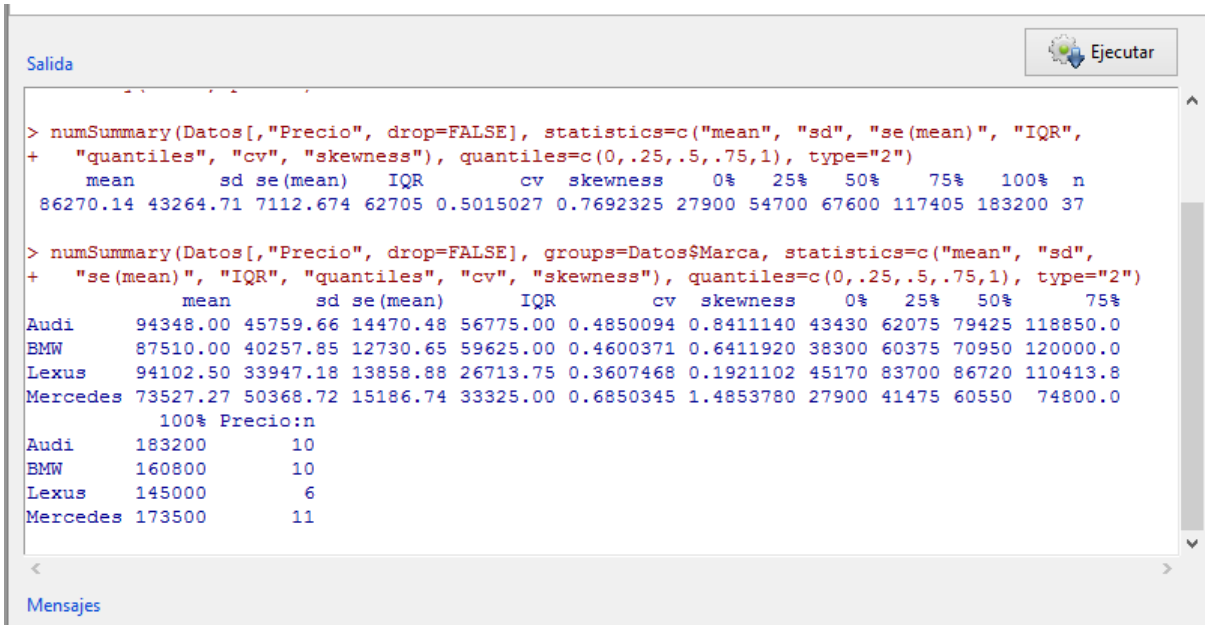
Se tiene que la media del precio de los vehículos es 86270.14 con una desviación típica de 43264.71. La media es más grande que la mediana (67600) lo que es síntoma de asimetría positiva o que los datos están sesgados a la derecha. El rango intercuartílico del precio es 62705. El mínimo precio es 27900 y el máximo 183200. El coeficiente de asimetría vale 0.7692325, luego ligera asimetría positiva. Los 3 cuartiles del precio son Q1=54700, Q2=Mediana=67600 y Q3=11740Si se quiere obtener el percentil 90 del Precio (en el cuadro de diálogo) de la pestaña Estadístios en cuantiles se selecciona (cuantiles 0,.25,.5,.75,1,.9) y nos aparecería en las salida el percentil 90 del Precio. Otras medidas que aparecen en la salida son el error típico de la media (se (mean)) y el coeficiente de variación (CV).

Si queremos ver las medidas numéricas descriptivas del Precio según una variable cualitativa como la Marca al objeto de estudiar la relación entre la variable cuantitativa (Precio) y la variable cualitativa (Marca) tenemos que seleccionar las órdenes siguientes:





La correspondiente salida de R-commander es

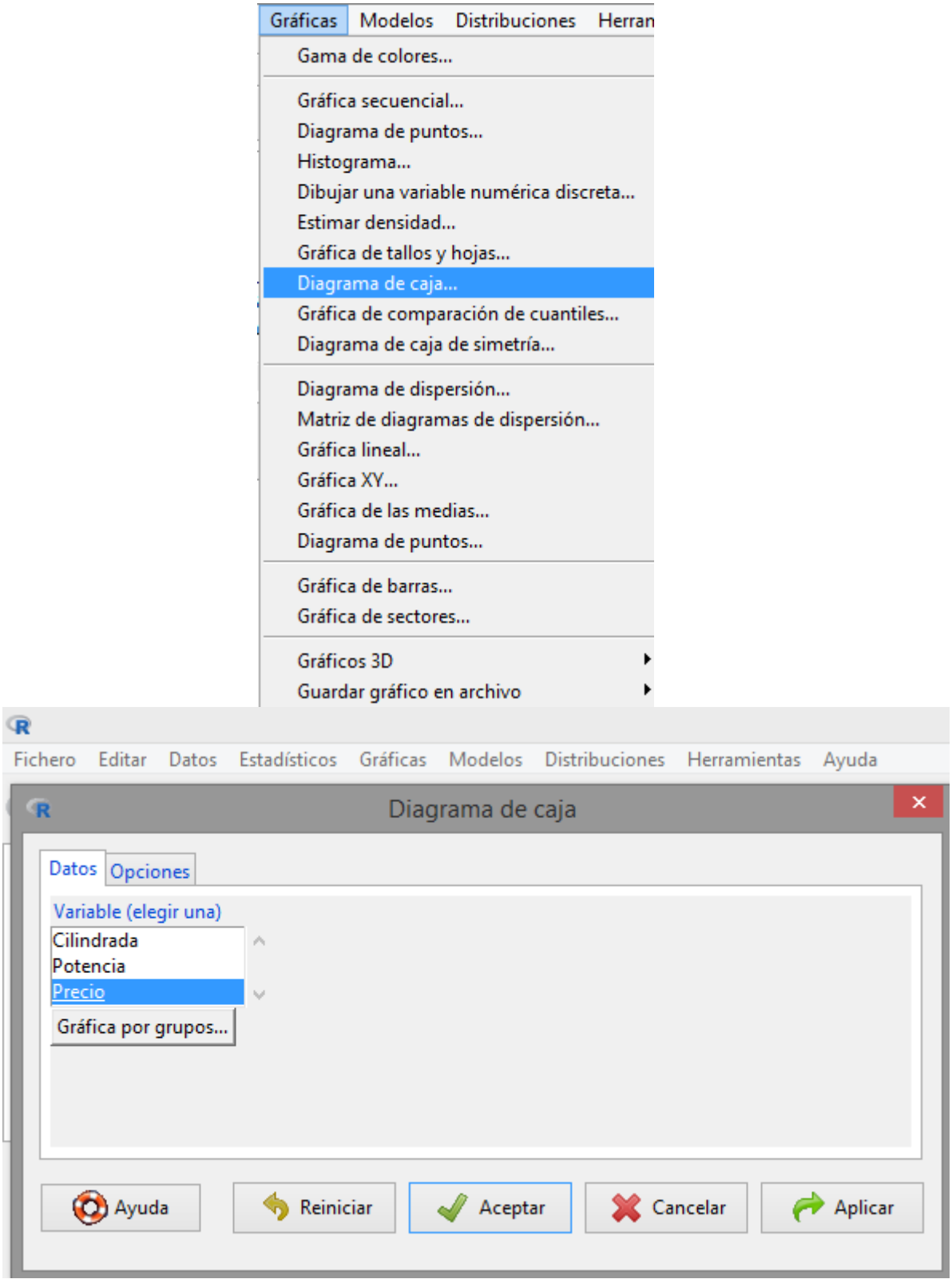


Se observa que la media del Precio según la Marca más alta es la de los Audi quel 94348 siendo este precio medio muy parecido al de los Lexus (94102.50). El precio medio de marca más bajo es el de los mercedes (73527.27). Las desviaciones típicas del precio según la marca también son diferentes siendo la más grande la de los mercedes (50368.72) por el contrario respecto al rango intercuartílico el varlor más alto es el de los BMW (56775.00) y el más bajo el de los Mercedes (33325.00). El precio de todas las marcas de coche es asimétrico positivo siendo la más sesgada a la derecha la marca mercedes con un coeficiente de asimetría (skewness) de 1.4853780. Los cuantiles (Mínimo, 3 cuartiles y máximo) también tienen diferente patrón según la marca de coche. El anális revela que el Precio depende de la Marcha siendo en promedio lo coches más caros los Audi y los Lexus y los más baratos los mercedes.

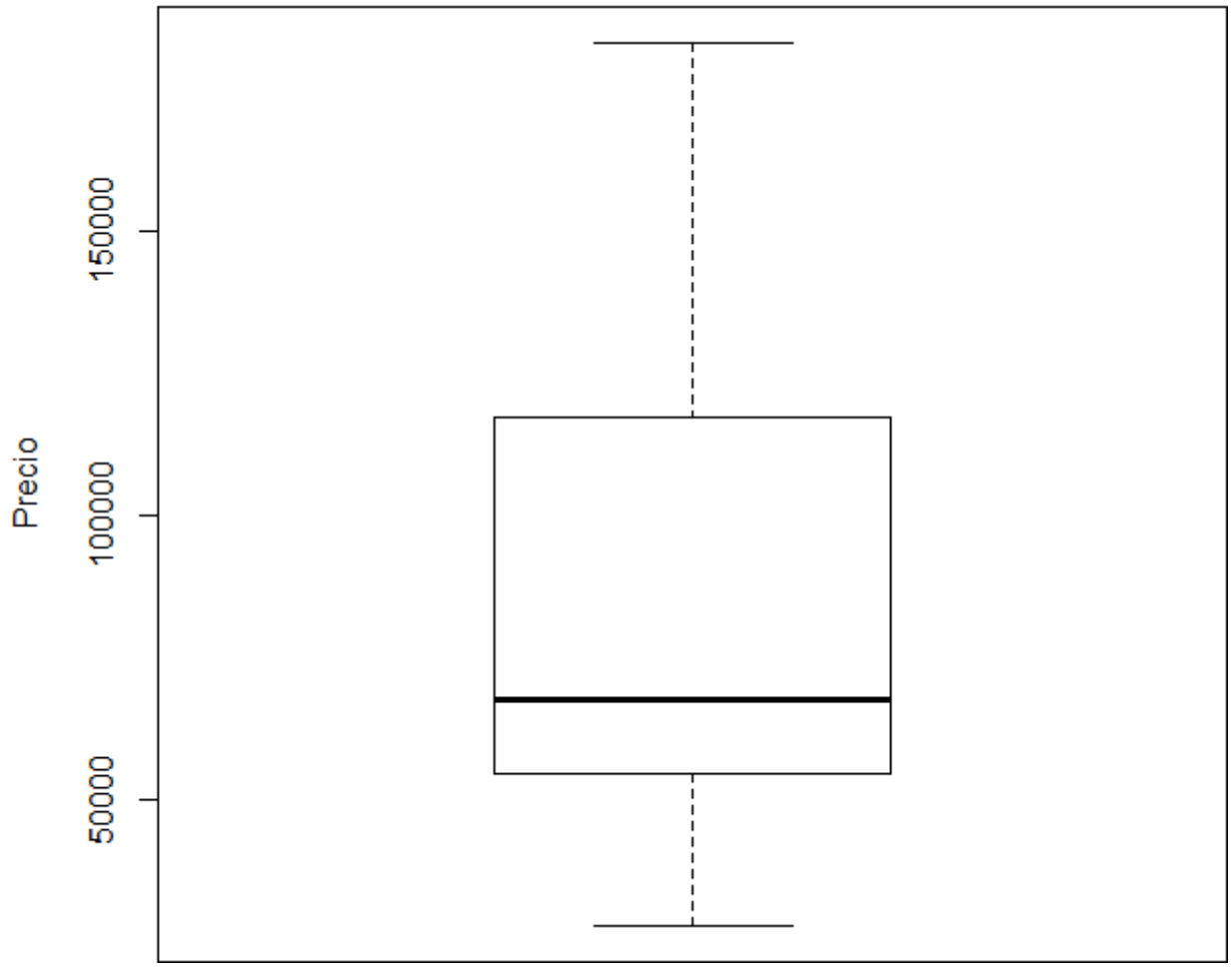
Los datos atípicos en una variable es un término estadístico para referirse a datos muy alejados de los valores centrales. Un criterio de uso común es decir que un dato es atípico cuando está a más de 3 desviaciones típicas de la media o sea x_i es atípico si $|x_i - \bar{x}| > 3s$. El problema del anterior criterio es que tanto la media como la desviación típica se pueden ver muy afectados por los datos atípicos que van a calificar. El criterio más empleado para calificar un dato como atípico es el basado en los cuartiles y el IQR. De acuerdo con tal criterio un dato x_i es atípico si $x_i > Q3 + 1{,}5IQR$ ó $x_i < Q1 - 1{,}5IQR$. Está calificación es la que emplea R-commander y de acuerdo con ella considera que un dato es atípico. Los datos atípicos de acuerdo con este criterio se marcan con un círculo en el diagrama de caja de R-commander que veremos a continuación.

El diagrama de caja es una representación gráfica de los cuartiles y la mediana en una caja de donde procede su nombre. La longitud de la caja es el rango intercuartílico. De los extremos de la caja salen unas líneas que se denominan bigotes cuya longitud va hasta el máximo por arriba o al mínimo por abajo si no hay datos atípico en le caso de haber datos atípicos estos se denotan con un círculo. Cuando hay datos atípicos por encima de la mediana el bigote va hasta Q·+1.5IQR y si hay datos atípico por debajo de la mediana el bigote va hasta Q1- 1.5IQR.

Las opciones de R-commander para obtener el diagama de caja son

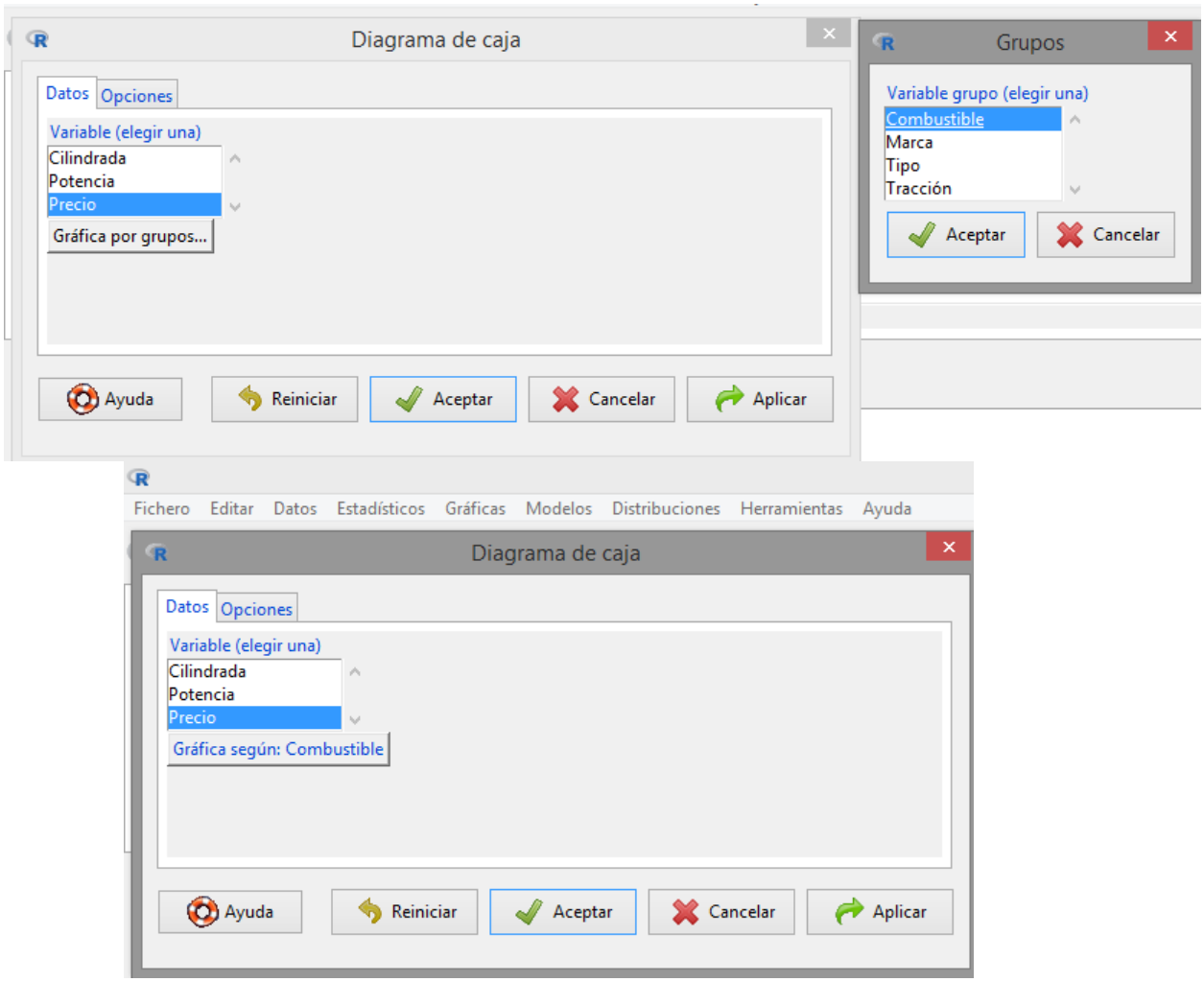


Obteniéndose el siguiente diagrama de caja de la variable Precio del fichero gamaalta.Rdata

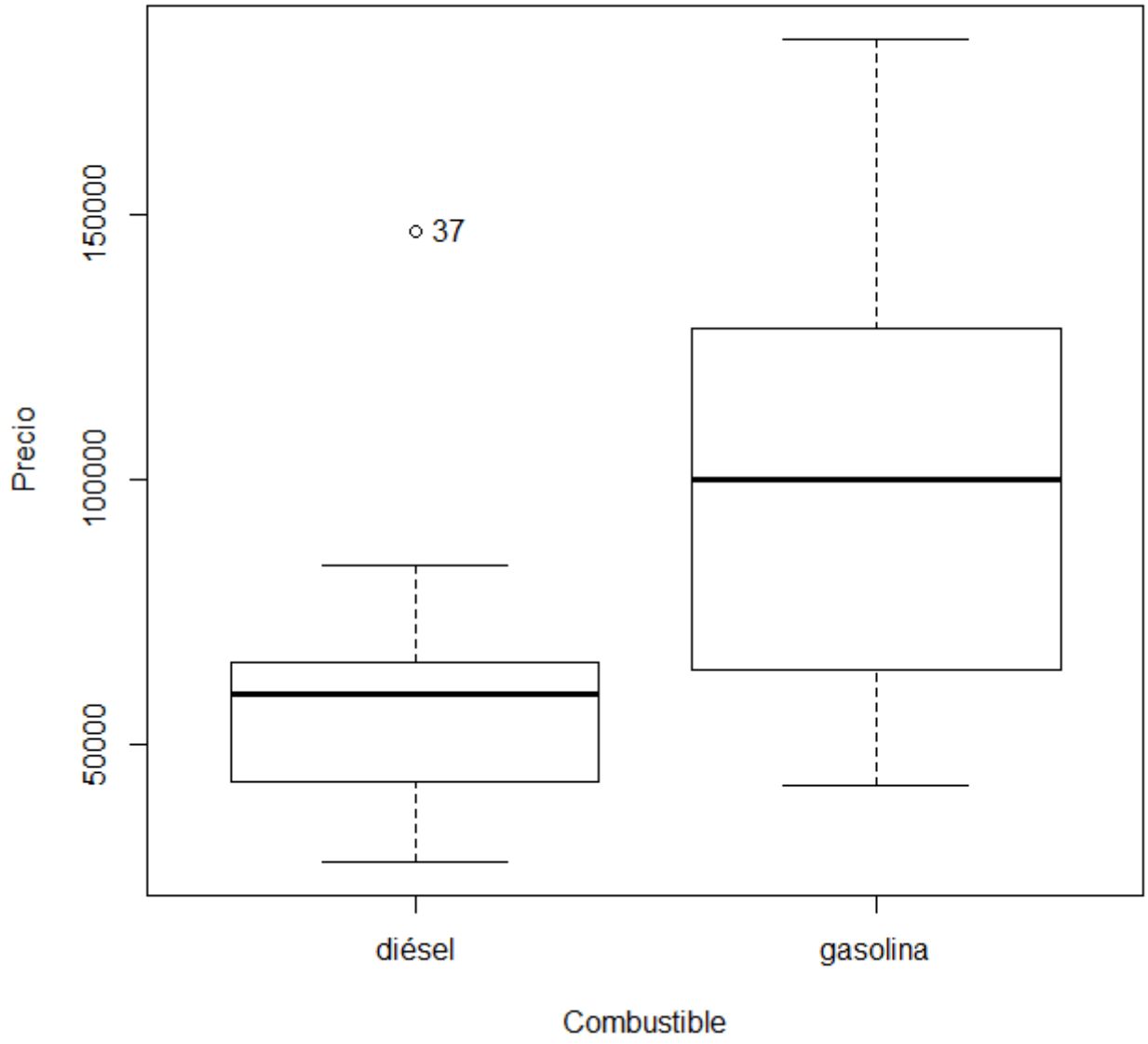


No se observan datos atípico pues no hay circulos. En breve cuando veamos la simetría de una variable diremos que de acuerdo con la simetría la variable Precio es asimétrica positiva o sesgada a la derecha pues el gráfico de caja es mucho más grande de la mediana para arriba que para abajo.

Se pueden obtener el diagrama de caja del Precio para cada valor de una variable cualtitativo como por ejemplo el Combustible. Esto se hace en R-commander con las opciones



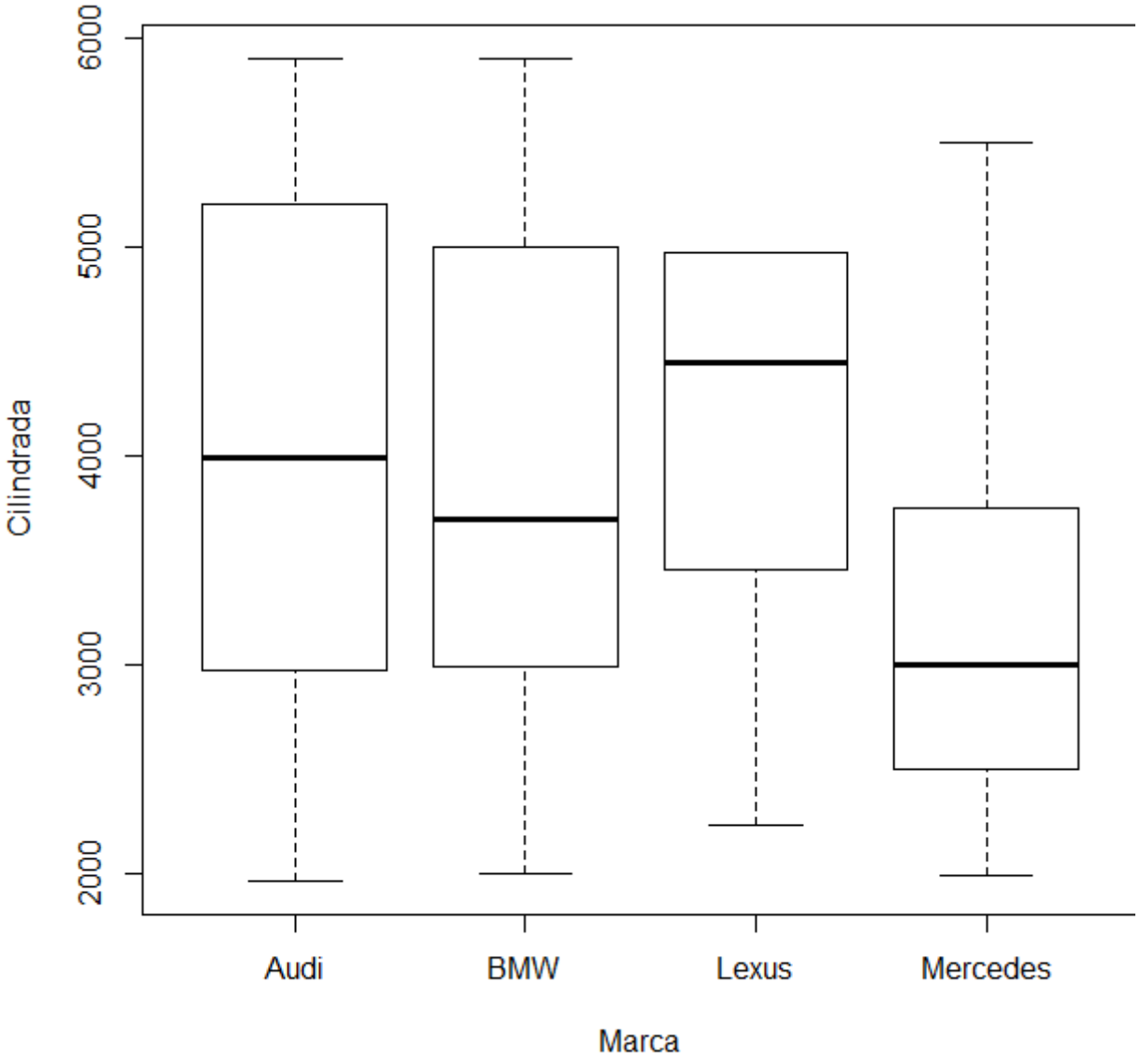
Obteniéndose



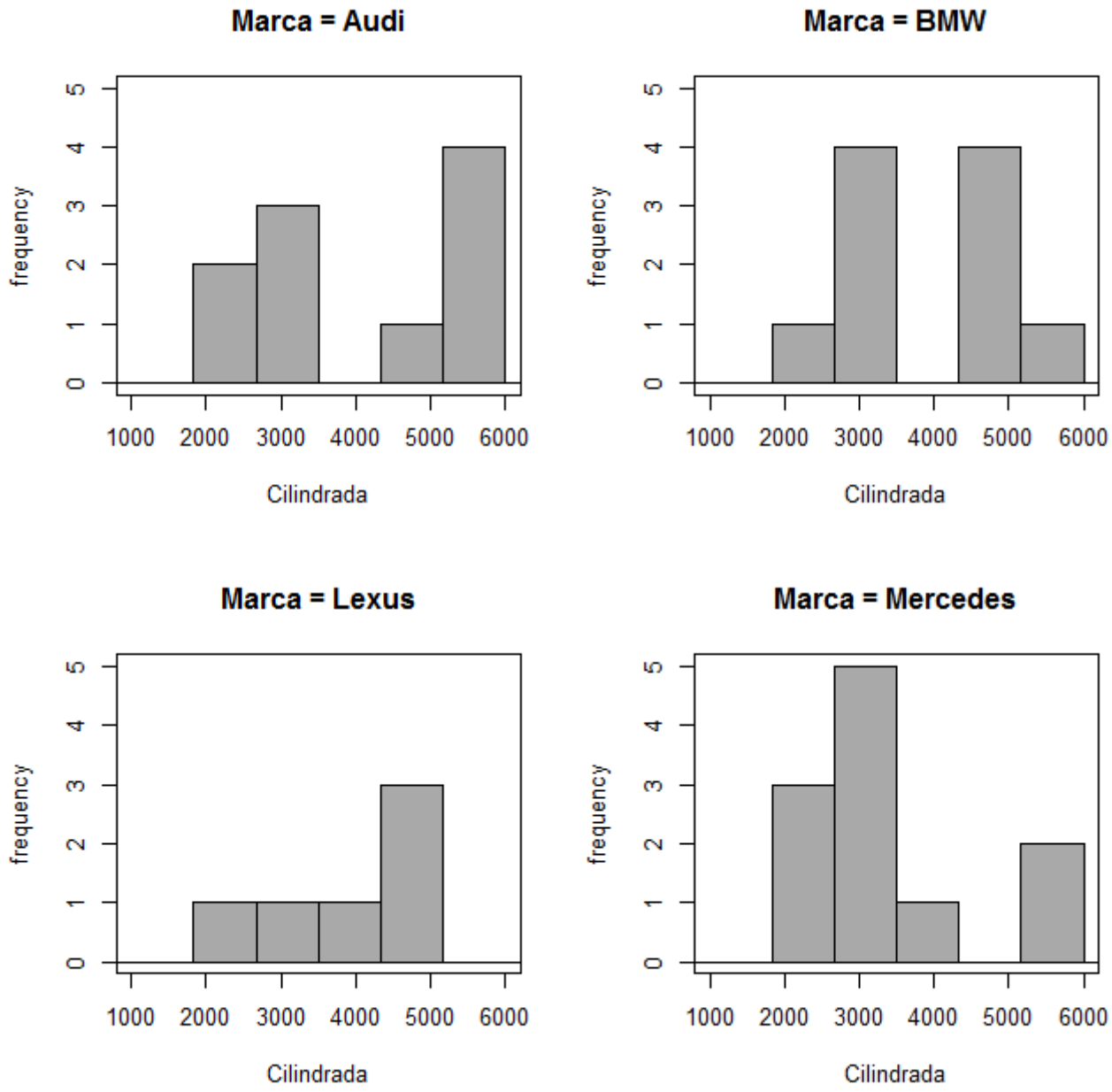
Se observa que la mediana del Precio de los vehículos diesel es menor que la de los de gasolina. Asimismo hay un dato atípico en el Precio de combustible diesel que es la observación 37. La variables Precio Diesel es ligeramente asimetrica negativa y la variable Precio gasolina es algo más asimétrico positiva (lo veremos a continuacion).

La simetría de los datos se puede analizar el diagrama de caja. La manera de hacerlo es doblar el gráfico del diagrama de caja por la línea de la mediana si la parte de arriba (encima de la mediana) es parecida (más grande, más pequeña) se tiene que los datos son simetricos (asimétricos positivos o sesgados a la derecha, asimétricos negativo o sesgados a la izquierda). En el diagrama de caja siguiente de la cilindrada según el precio se tien que los datos de cilindrada para los Audi son bastante simétricos, en el caso de los BMW y mercedes son

ligeramente sesgados a la derecha y en el caso de los Lexus son ligeramente asimétricos negativos.

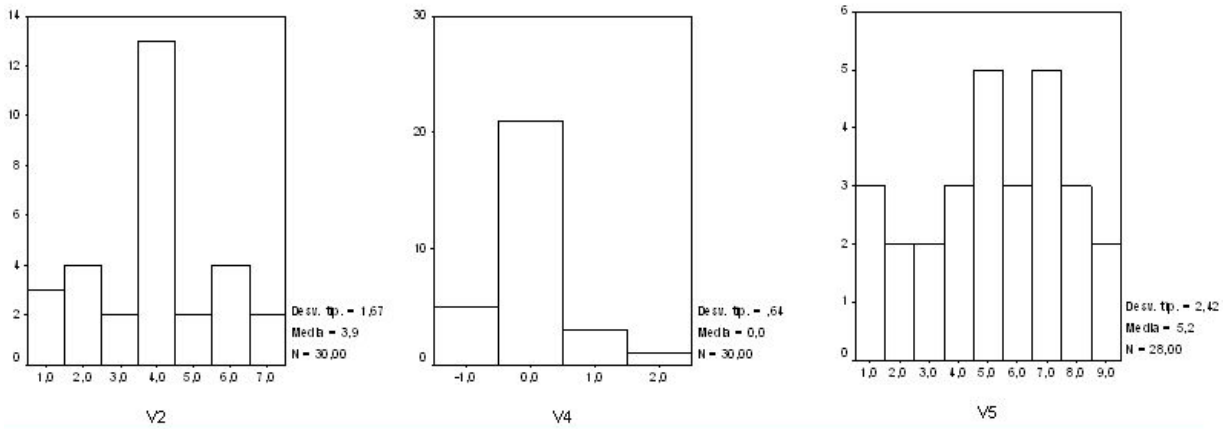


En relación con los histogramas la simetría ocurre el histograma es a groso modo simétrico, en el caso de asimetría positiva el histograma está sesgado a la derecha y en el caso de asimetría negativa sesgado a la izquierda. Los histogramas del precio según la marca en R-commander son



Los histogramas de la cilindrada para las Audi y BMW son bastante simétrico. El histograma de la cilindrada de los Lexus está sesgado a la izquierda y el de los mercedes a la derecha.

En el gráfico de debajo se ven tres histogramas con diferente grado de apuntamiento o curtosis. El grafico de la derecha es platicúrtico (curtosis <0 menos apuntados que unos datos normales de esa media y desviació típica), el de la iquierda es leptocúrtico (curtosis >0, más apuntado que unos datos normales de esa media y desviación tipica) y el del centro es mesocúrtico (curtosis cercana a cero, aproximadamente igual de apuntado que una normal de esa media y desviación típica).



Descipción de los ficheros de datos:

gamaalta.Rdata datos de caracteristicas de 37 coches de gama alta. Las carácterísticas o variables que se estudia con la marca, el tipo, el precio, la cilindrada, la potencia, el combustible, etc

ALTURA_PESO_TELECOS_90_14.xl datos de alturas y pesos de estudiantes de ingeniería de telecomunicaciones

Gallego.Rdata mediciones de caracteristica del agua del río Gallego en dos estaciones: Ardisa y Anzanigo.

SeriesTV.Rdata diferentes variables medidas a 46 series de televisión.

Baloncesto.Rdata datos de 105 jugadores de la liga ACB la temporada 2009

Pisosventa.Rdata datos de 78 pisos en venta en Zaragoza.

Supermercado.Rdata datos de 402 clientes de un supermercado