Práctica 2: Análisis descriptivo de variables. Parte I.

1. Análisis de variables cualitativas

Para empezar recuerda que tienes que crear una carpeta de trabajo con nombre, por ejemplo, practica2. Coloca en ella el archivo galton.rda que puedes encontrar en MiAulario.

1.1. Tablas de frecuencias

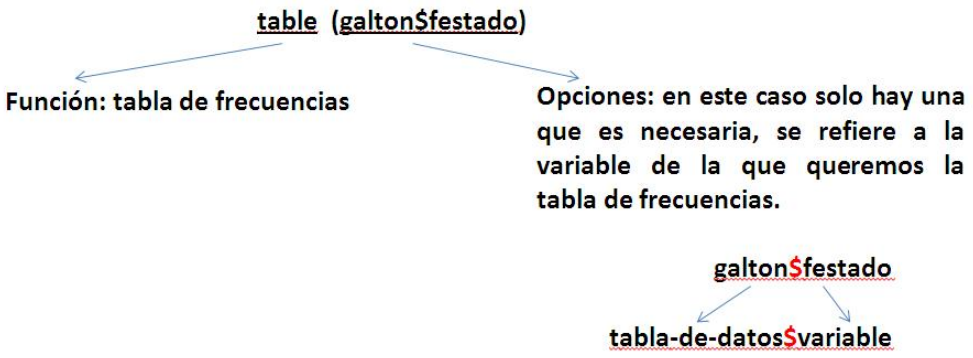
Ruta de menús: *Estadísticos–Resúmenes-Distribución de frecuencias*

1. Crea la tabla de frecuencias del estado civil.

   Observa que en la parte de arriba de la ventana (R Script), reservada para los comandos de R, se han pegado 4 comandos.

   La función ***table(galton$festado)*** es la que realiza la tabla de frecuencias.

   En general el lenguaje de R se escribe mediante funciones y entre parentesis las diferentes opciones separadas por comas (algunas necesarias y otras opcionales) para esa función.



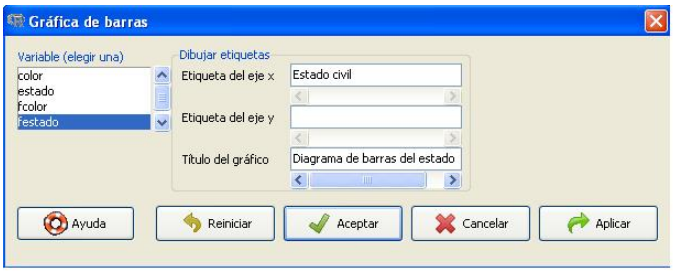
   Con los menús de R Commander nos evitamos escribir las funciones pero viene bien saber algunos elementos básicos.

2. Crea, de una en una, las tablas de frecuencias de todas las variables cualitativas.

1.2. Diagramas de barras y Sectogramas: manipulación de comandos

Ruta de menús: *Gráficas–Gráfica de barras (o Gráfica de sectores)*

1. Crea un diagrama de barras de la variable estado civil.



Observa el comando en la ventana R Script,

barplot(table(galton$festado), xlab="Estado civil", ylab="", main="Diagrama de barras del estado civil")

• Función: barplot.

• Opciones necesarias: table(galton$festado), la función barplot actua sobre una tabla de frecuencias (no lo hace directamente sobre la variable).

• Opciones no necesarias: xlab="título eje horizontal", ylab="título eje vertical", main="título principal del gráfico".

Se puede añadir esta opción para el color: col=c(color1,color2). Elige un color verde para casados y otro azul para solteros.

Nota: para elegir colores en R tienes en MiAulario la tabla *Colores en R* y el documento *su uso*.

2. Crea el sectograma de la variable estado civil escribiendo solo un título principal.

3. Identifica el comando en la ventana R Script que corresponde a esta orden y cambia los colores del mismo modo que en el diagrana de barras.

4. Crea un sectograma para otra variable cualitativa.

1.3. Ventana de gráficos

Como se ha podido comprobar al pedir un gráfico se abre una ventana específica de gráficos que se coloca sobre la ventana RGui. Es habitual al pedir un gráfico que

éste se quede en un segundo plano detrás de la ventana R Commander, para verlo hay que pasar a primer plano la ventana RGui.

El gráfico se puede copiar y pegar en un editor de texto (como el Word) pinchando sobre él con el botón derecho y eligiendo "*copy as metafile*".

También se puede guardar en un fichero con el botón derecho que ofrece dos opciones de formato para guardarlo. Hay más opciones de formato (como jpg) en el menú de la ventana de gráficos. Este menú está disponible en los menus de RGui cuando la ventana de gráficos está seleccionada.

Si la ventana de gráficos permanece abierta cada vez que se pide un gráfico se borra el anterior.

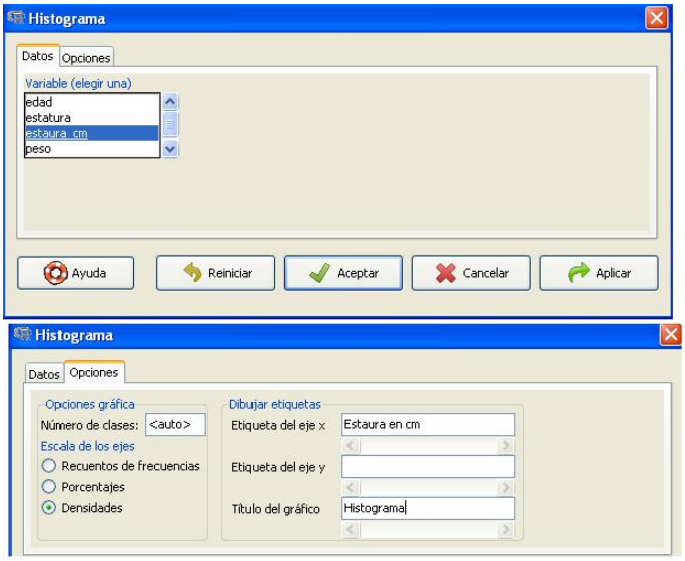
Se puede dividir la ventana de gráficos en partes para recibir más de un gráfico. La orden ***par(mfrow=c(n,m))*** prepara una ventana de gráficos para recibir una matriz de gráficos de n filas y m columnas. Por ejemplo, ***par(mfrow=c(1,2))***, prepara dos espacios para dos gráficos, en un fila y dos columnas. Los gráficos que se pidan a continuación ocuparán secuencialmente los espacios.

2. Análisis de variables cuantitativas (Primera parte: distribución de frecuencias)

2.1. Histogramas

Ruta de menús: *Gráficas–histograma*

1. Crea un histograma de la variable estatura en cm.



El número de clases (o intervalos) se puede elegir o dejar por defecto el criterio elegido por R (este criterio responde a una fórmula denominada de Sturges). La amplitud de los intervalos será la misma salvo que se determine otra cosa modificando el comando.

Cuando la amplitud de los intervalos es la misma se puede elegir cualquiera de las opciones en escala de los ejes: recuentos de frecuencias para cada clase, porcentajes para la frecuencia relativa en tanto por cien de cada clase y densidad para la frecuencia relativa de cada clase. Nosotros siempre utilizaremos densidades.

Si se toman intervalos de distinta amplitud solo se puede utilizar densidades.

2. La opción **breaks** del comando define los puntos de corte del intervalo. Observa en la linea de comandos que se ha escrito la opción *breaks="Sturges"*.

   Repite el histograma eligiendo 5 intervalos de igual amplitud. Observa que R finalmente se decide por dibujar 6 intervalos ya que entiende la petición como una sugerencia.

   Para definir con exactitud los puntos de corte de los intervalos se modifica la opción *breaks* desde la linea de comandos escribiendo una concatenación de los puntos de corte. Por ejemplo modifica la opción así **breaks=c(140,150,160,170,180,190,220)** y observa el cambio en el gráfico. Ten en cuenta que los intervalos no tienen la misma amplitud así que se necesita la opción densidades.

3. Crea un histograma de la variable peso, con un título principal "Histograma del

Peso", título en el eje horizontal "Peso en kg", sin título en el eje vertical, de color verde musgo, con puntos de corte de 5 en 5 kg desde 45kg hasta 90kg y de 10 en 10 kg de 90 a 110kg.

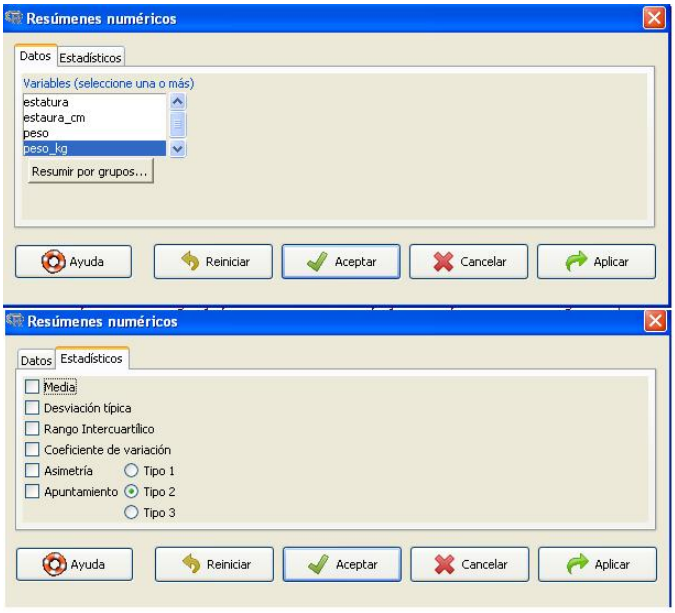
4. Se pueden modificar los límites de representación de los ejes. Mediante la opción $\boldsymbol{xlim = c(a, b)}$ el gráfico mostrará en el eje-x un valor mínimo en a y su máximo en b. Lo mismo para el eje-y con la opción $\boldsymbol{ylim = c(a, b)}$.

   Añade en la linea de comandos las opciones necesarias para que en el histograma anterior represente el eje-x desde 30 hasta 130kg.

2.2. Percentiles y diagramas de cajas

Ruta de menús, **percentiles**, *Estadísticos–Resúmenes-resúmenes numéricos.*

1. Calcula e interpreta los percentiles 10, 25, 50, 75 y 90 de la variable peso-kg.



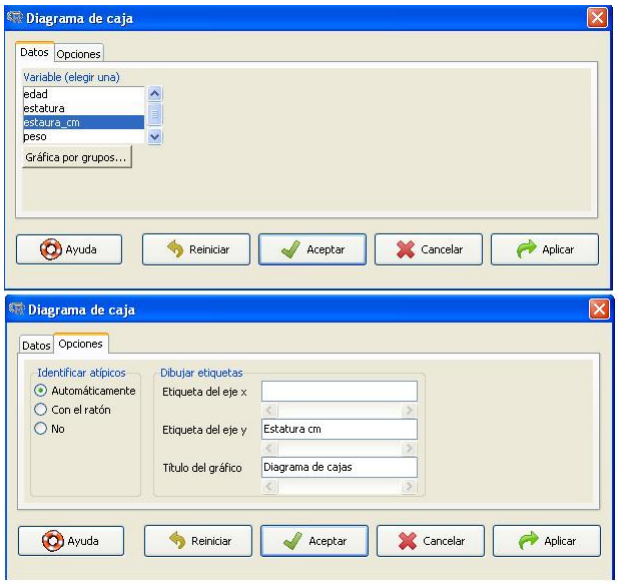
   Observa que en la ventana no se pueden elegir los percentiles esto es porque salen por defecto.

   En la linea de comandos se observa la opción *quantiles=c(0,.25,.5,.75,1)*, que significa que se han solicitado los percentiles: $P_0, P_{25}, P_{50}, P_{75}$ y P_{100}

   Modifica la linea de comandos para obtener los percentiles pedidos.

   Ruta de menús, **diagrama de cajas** *Gráficas–diagrama de caja.*

2. Crear un diagrama de cajas de la variable estatura y marca la opción de identificación de valores atípicos de modo automático. Si se hace con el ratón entonces sobre el gráfico pulsando el botón izquierdo aparece la identificación (con el derecho puedes paralizar la opción).



3. Si el valor extremo queda muy cerca del encuadre del gráfico, posiblemente no aparezca el valor. En tal caso, modifica el comando mediante la opción $ylim = c(a, b)$ para ampliar los límites de representación del eje de ordenadas.

2.3. Recodificar una variable cuantitativa en tramos de igual amplitud o de igual frecuencia

Recordemos de la práctica 1, que en ocasiones puede ser de interés recodificar una variable cuantitativa en tramos obteniendo una variable cualitativa o factor. Por ejemplo, la variable estatura se recodificó en tramos de modo que se obtuvo un factor ordinal de tres niveles de estatura: baja, normal y alta.

Si en la recodificación se pretende considerar los puntos de corte de modo que los distintos tramos contengan igual número de elementos, es decir utilizando percentiles, es más rápido utilizar:

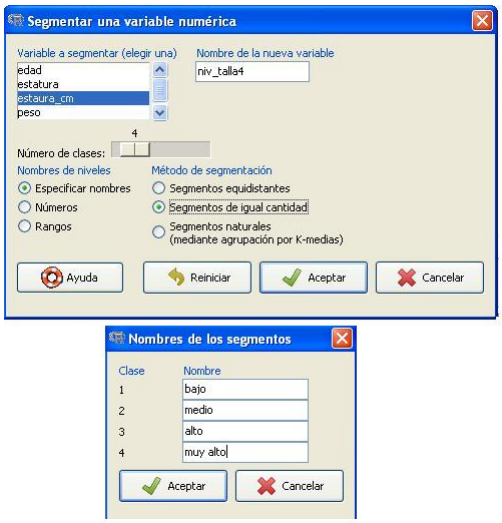
Ruta de menús: *Datos–Modificar variables...-segmentar variable numérica.*

Tiene la ventaja de que los niveles del nuevo factor salen en el orden adecuado y que

en la ventana de diálogo se puede elegir entre escribir los nombres de cada nivel o asignar como nombre el rango del cada tramo.

Este menú también permite dividir en tramos equidistantes.

1. Recodifica la variable estatura en una nueva variable "niv_talla4"con cuatro tramos de modo que cada tramo tenga el mismo número de elementos, etiqueta cada tramo (nivel) con: bajo, medio, alto, muy alto.



2. Observa con una tabla de frecuencias que los niveles del nuevo factor están el orden adecuado.

3. Recodifica la variable peso en un factor ordinal "niv_peso1"de 4 niveles con el mismo número de individuos, asignando como nombre de cada modalidad el rango del correspondiente tramo.

4. Recodifica la variable peso en un factor ordinal "niv_peso2"de 4 niveles equidistantes, asignando como nombre de cada modalidad el rango del correspondiente tramo.

5. Haz una tabla de frecuencias y un diagrama de barras de las variables "y niv_peso1"y "niv_peso2".

3. Análisis comparativo de una variable cuaNtitativa en distintos subgrupos (niveles del factor).

A menudo resulta de interés estudiar el comportamiento de **UNA VARIABLE para TODOS los NIVELES de un factor**. El estudio permite establecer comparaciones y detectar diferencias en su comportamiento según los niveles del factor.

1. Se pretende estudiar el peso en kg según el tipo de residencia: por un lado para los que viven en ciudad, por otro lado para los que viven en el campo y por otro lado para los que viven en suburbios. La variable que clasifica los subgrupos debe ser un factor.

   En R-commander este tipo de estudios es inmediato haciendo un **resumen numérico**, como ya hemos visto, cuya ventana ofrece el botón ***Resumir por grupos***

   Haz la comparación del peso en kg según la zona de residencia.

   Nota: para volver a hacer un resumen numérico de todos los datos, es decir sin distinción de grupos, ***Reiniciar***

2. En estas comparaciones, es muy apropiado comparar el gráfico de diagrama de cajas y bigotes para los distintos subgrupos. En R Commandet también es inmediato, el menú para el diagrama de cajas ofrece el botón ***Grafica por grupos***.

   Haz un gráfico de diagrama de cajas para el peso en kg según la zona de residencia.

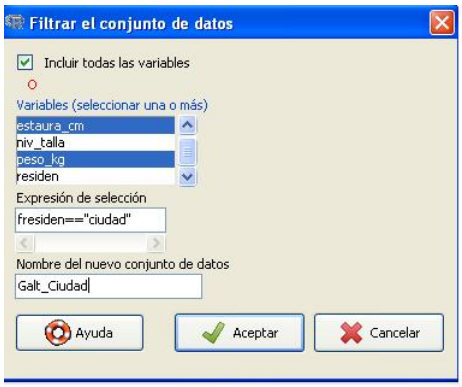
4. Filtrado de datos

En ocasiones nos interesan los resúmenes estadísticos y gráficos para **TODAS las VARIABLES de SÓLO un SUBCONJUNTO** de los datos y que queda definido por un valor que toma una de las variables categóricas (o por alguna condición para los valores de una variable numerica). Por ejemplo, interesa hacer una estadística de la variables estatura en cm y peso en kg sólo para los que viven en una ciudad. Esto puede hacerse creando un nueva conjunto de datos eligiendo algunos casos y algunas variables (que se pueden ser todas).

Ruta de menús: datos–Conjunto de datos activo–Filtrar el conjunto de datos activo.

• filtro por categoría

Crea un subgrupo formado sólo para los casos que viven en ciudad eligiendo las variables *peso_kg* y *altura_cm*. Al nuevo conjunto de datos se le denomina *Galt_Ciudad*.



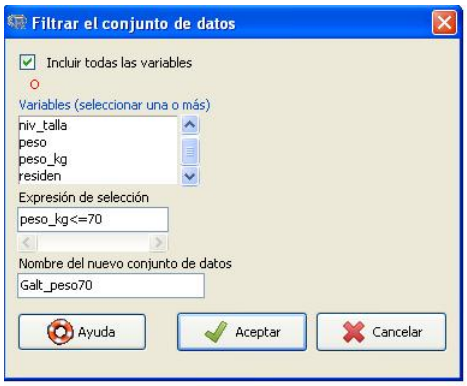
Observa que como ciudad es un valor no numérico se escribe entre comillas.

Guarda el nuevo conjunto de datos en la carpeta de trabajo .

• filtro por condición en variable numérica

El subconjunto también puede quedar definido por el valor que toma una de las variables numéricas. Por ejemplo, queremos estadísticas de todas las variables pero sólo para los casos en que el peso no supere a 70kg.

Crea el subgrupo formado por todas las variables pero sólo para los casos que peso_kg$\leq$ 70. Al nuevo conjunto de datos se le denomina *Galt_peso70*.



Observa que como 70 es un valor numérico se escribe sin comillas.

Guarda el nuevo conjunto de datos en la carpeta de trabajo.

5. Ejercicio

Abre el fichero "coches.rda", que se elaboró en el último ejercicio de la práctica 1. A partir de estos datos se pide:

1. Convertir la variable peso a toneladas con un decimal.

   Nota: la fórmula round(expresión, n) crea el valor de la expresión con n decimales. Por ejemplo con la fórmula round(consumo/100, 1) se obtendría el valor del consumo por kilómetro con un decimal.

2. Crear una variable de ratio de consumo por peso: consumo (l/100km) por tonelada.

3. Crear una nueva variable factor que recoja la información de la variable consumo en tres tramos, donde cada tramo tenga el mismo número de automóviles.

4. Elegir dos variables cualitativas y analizarlas. Introduce algún cambio en los gráficos.

5. Construye el histograma de la variable peso en kg en color rojo, con 5 intervalos, un título principal (Distribución del peso), otro en el eje x (Peso en kg) y sin título en el eje y.

6. Calcula los cuartiles de la variable consumo e interprétalos. Construye el diagrama de cajas de esta variable.

7. Analiza las variables peso total y acelaeración según la variable factor creada en el apartado 3. Obtén las medidas más relevantes y los diagramas de cajas asociados a cada variable según el grupo.

8. Recuerda antes de salir, que debes guardar los cambios del conjunto de datos con extensión .rda para poder acceder a los datos que has preparado en futuras sesiones de R.