Práctica 2. Estadística Descriptiva

El principal objetivo de la práctica es realizar con R Commander análisis descriptivos.

Sin embargo, antes de realizar un análisis de datos con técnicas estadísticas es fundamental determinar si los datos se refieren a toda la población o solo a una muestra, reconocer cuáles son los individuos del estudio y clasificar las variables. Recuerda que en *R* las variables se clasifican es *numéricas* y *factores*. Son *numéricas* aquellas cuyos valores figuran en el archivo de datos con números y son *factores* aquellas que figuran con caracteres alfanuméricos. No hay que olvidar, que en Estadística clasificamos las variables según su naturaleza en *cualitativas nominales u ordinales* y *cuantitativas discretas o continuas*, independientemente de cómo aparezcan representados sus valores en el archivo de datos.

Ejercicio 1. *El archivo ZonasProtegidas contiene de cada una de las comarcas aragonesas: la provincia a la que pertenece la mayor parte de su territorio, su superficie en* km^2*, su población en* 2011 *y el número de espacios naturales protegidos, lugares de importancia comunitaria (LIC) y zonas de especial protección para las aves (ZEPA) registrados en* 2014*, así como la superficie que ocupan (en hectáreas).*

1. *En la práctica anterior se definió la población objeto de estudios, los individuos de la población, si se trataba de un censo o una muestra y se clasificaron las diferentes variables. Recuerda estos conceptos.*

2. *Calcula el porcentaje de superficie de cada comarca que está ocupado por los distintos tipos de zonas protegidas en cada una de las comarcas.*

   AYUDA: Observa que la superficie que ocupan los espacios naturales protegidos, LIC y ZEPA está expresada en ha, mientras que la superficie de la comarca está expresada en km^2*. Cambia las unidades de medida de esta última a ha antes de calcular el porcentaje de superficie.*

3. *Calcula la densidad de población de cada comarca (medida en* hab/km^2*).*

4. *Clasifica las nuevas variables que has calculado.*

5. *Guarda el archivo con las variables calculadas para no perder los cambios. Modifica el nombre, para mantener también el archivo original sin las nuevas variables calculadas.*

Datos → Conjunto de datos activos → Guardar el conjunto de datos activo

Las nuevas columnas deberían tener el siguiente aspecto:

ZonasProtegidas

	Superficie_ha	SuperficieTotalZonasProtegidas	Porcentaje_ZonasProtegidas	DensidadPoblacion
1	187770	187538.906	99.876927	9.680993
2	135982	56074.396	41.236631	10.705829
3	226220	225348.569	99.614786	3.387410
4	246033	149808.326	60.889525	5.298883
5	306244	43695.009	14.268038	10.613106
6	252558	121918.472	48.273455	26.893228
7	117032	65686.961	56.127351	20.771242
8	57305	5156.486	8.998318	41.935259
9	73386	7655.480	10.431799	25.688823
10	276445	111915.395	40.483783	7.444519

1. Análisis de una variable cualitativa

En esta sección, vamos a ver cómo describir con R Commander cualquier variable de tipo cualitativo (variables de tipo factor para R Commander). Las tablas de frecuencias, los diagramas de sectores y los diagramas de barras son las herramientas básicas para describirlas.

1.1. Resumen tabular o tabla de frecuencias

Para obtener la tabla de frecuencias se debe proceder de la siguiente manera:

Estadísticos → Resúmenes → Distribución de frecuencias

En el cuadro de diálogo que aparece se elige la variable o variables para las que quieres obtener la tabla de frecuencias.

En la ventana de resultados aparecen en primer lugar los valores de la variable con sus frecuencias absolutas (*counts*) y en segundo lugar los valores de la variable con sus frecuencias relativas (*percentajes*).

Ejercicio 2. *Con los datos del archivo ZonasProtegidas, obtén la tabla de frecuencias de la variable que indica la provincia a la que pertenece la comarca. ¿Es equitativa la división de provincias en comarcas?*

```
counts:
Provincia
  Huesca   Teruel Zaragoza
      10       10       13

percentages:
Provincia
  Huesca   Teruel Zaragoza
   30.30    30.30    39.39
```

A la vista de los resultados obtenidos, la división de provincias en comarcas es casi un 10 % superior en Zaragoza en comparación con Huesca y Teruel. Además Huesca y Teruel tienen el mismo número de comarcas.

1.2. Resumen gráfico

Un resumen tabular es altamente recomendable que vaya acompañado de un resumen gráfico. Ya hemos comentado que los gráficos más adecuados para las variables cualitativas son el diagrama de sectores, y el diagrama de barras.

1.2.1. Diagrama de Sectores

El diagrama de sectores es uno de los gráficos más utilizados cuando se estudia una variable cualitativa nominal. Para obtener este gráfico debemos proceder de la siguiente manera:

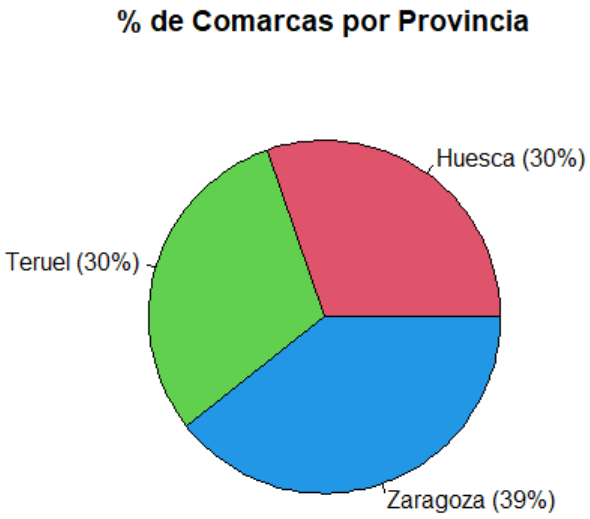
Gráficas → Gráfica de Sectores

En el cuadro de diálogo que aparece se elige la variable que se quiere representar.

En la opción de *Incluir en etiquetas de segmento*, se selecciona si se quiere visualizar las frecuencias absolutas, los porcentajes correspondientes o nada.

Además, se debe escribir el *Título del gráfico* y opcionalmente cambiar el nombre que queremos que aparezca en el eje x y eje y del diagrama.

Ejercicio 3. *Con los datos del archivo ZonasProtegidas, obtén el diagrama de sectores de la variable que indica la provincia a la que pertenece la comarca. ¿Es equitativa la división de provincias en comarcas?*



NOTA: La ventana de gráficos de R Commander sólo admite un gráfico activo. Por tanto, una vez realizado un gráfico conviene copiarlo o guardarlo, pues al realizar el siguiente gráfico el anterior se pierde. Además de copiar un gráfico a *Word* con la opción ***Archivo*** → ***Copiar para el área de transferencia*** → ***Como un Metafile***, R Commander permite guardar un gráfico en diferentes formatos (pdf, metafichero, png, jpg, etc.) para poder utilizarlo posteriormente con otras aplicaciones. Para guardar un gráfico, hay que utilizar la opción ***Archivo*** → ***Guardar como***, disponible en la ventana de gráficos.

1.2.2. Diagrama de Barras

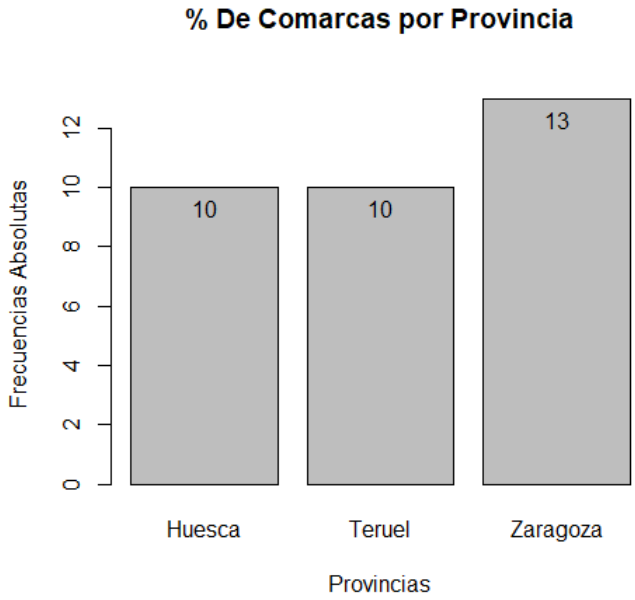
Si tenemos una variable cualitativa ordinal, el gráfico más utilizado es el diagrama de barras. Sin embargo, este gráfico también puede ser utilizado para variables cualitativas nominales con pocos valores distintos. Para obtener este gráfico debemos hacer:

Gráficas → Gráfica de barras

En el cuadro de diálogo que aparece en la pestaña *Datos*, se debe seleccionar la variable a representar. En la pestaña *Opciones* se debe elegir:

- La *Escala de los ejes*, si queremos que en el eje vertical se representen las frecuencias absolutas (*recuentos de frecuencias*) o las frecuencias porcentuales (*Porcentajes*)
- Si queremos que se *muestre el conteo o los porcentajes en barras*
- Los colores para las distintas categorías de la variable
- Indicar un título al gráfico, y las correspondientes etiquetas a los ejes x e y.

Ejercicio 4. *Con los datos del archivo ZonasProtegidas, obtén el diagrama de barras de la variable que indica la provincia a la que pertenece la comarca. ¿Es equitativa la división de provincias en comarcas?*



NOTA: Observa que R muestra las categorías de la variable en orden alfabético, tanto en la tabla de frecuencias como en el diagrama de barras. Si se desea cambiar el orden, se deben reordenar los niveles del factor, con el comando ya visto anteriormente:

Datos → Modificar variables del conjunto de datos activo → Reordenar niveles de factor

Observación: Las opciones *distribución de frecuencias*, *gráfico de sectores* y *gráfico de barras* solo están disponibles para variables almacenadas en el archivo de datos como factores. Por esta razón en el cuadro de diálogo solo aparecen las variables del archivo que cumplen esta condición.

2. Análisis de una variable cuantitativa

2.1. Variables cuantitativas discretas

Cuando una variable cuantitativa discreta toma pocos valores distintos, la tabla de frecuencias y el gráfico de barras nos pueden ser útiles para describir la variable.

Notar que para realizar la tabla de frecuencias y el gráfico de barras será necesario que nuestra variable sea de tipo *factor*, ya que sino, no será posible obtener dicha información.

Para convertir una variable numérica a tipo factor, ya hemos visto que se realiza de la siguiente manera:

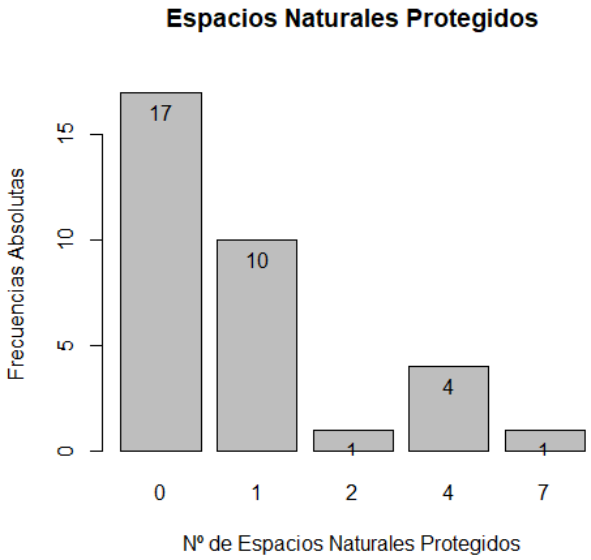
Datos → Modificar variables del conjunto de datos activo → Convertir variable numérica en factor ...

En el caso de que la variable sea cuantitativa discreta, es conveniente mantener la variable numérica y la variable factor, por ello al convertir la variable numérica en factor, le asignaremos un nombre distinto.

Ejercicio 5. *Las variable del archivo ZonasProtegidas que indican el número de espacios naturales protegidos, lugares de importancia comunitaria (LIC) y zonas de especial protección para las aves (ZEPA) registrados en* 2014 *en cada comarca son variables cuantitativas discretas.*

1. *Intenta obtener la tabla de frecuencias o el diagrama de barras del número de espacios naturales protegidos y observa que en la lista de variables solo aparecen aquellas cuyos valores contiene caracteres alfanuméricos.*
2. *Convierte la variable Espacios.Naturales.Protegidos en factor creando una nueva variable para así mantener la información numérica de la variable original. Esta nueva variable debe tener el nombre: ENPFactor.*
3. *Representa la nueva variable mediante una tabla de frecuencias y su representación gráfica más adecuada.*

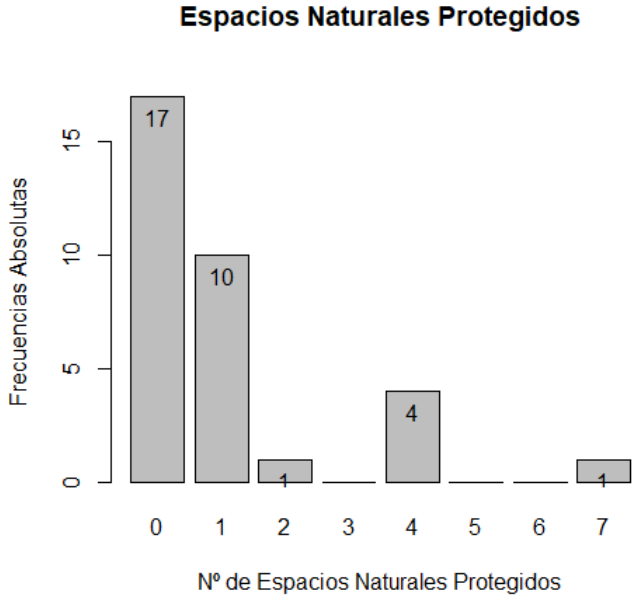
4. *¿Qué porcentaje de comarcas no tienen ningún espacio natural protegido? Entre las comarcas que sí tienen espacios naturales protegidos, ¿qué cantidad es la más frecuente? ¿Cuál es el número máximo de espacios protegidos que tienen las comarcas de Aragón?*



Observad que en el diagrama de barras obtenido se han representado los valores 0,1,2, 4 y 7. Una representación más adecuada debería incluir también los valores 3, 5 y 6 con frecuencia igual a 0. Para ello escribe en la venta *R Script* las órdenes que se muestran a continuación:

```
NuevaLista.ENP <- as.factor(seq(0,7,1))
ZonasProtegidas$ENPFactorBueno <- factor(ZonasProtegidas$ENPFactor, levels(NuevaLista.ENP))
```

Si dibujamos ahora el diagrama de barras de la nueva variable, obtenemos el siguiente gráfico:



Ejercicio 6. *Con los datos del archivo ZonasProtegidas, responde a las siguientes preguntas:*

1. *¿Qué porcentaje de comarcas no tienen ninguna zona de especial protección para las aves (ZEPA)? Entre las comarcas que sí tienen ZEPAs, ¿qué cantidad es la más frecuente? ¿Cuál es el numero máximo de ZEPAs que tiene una comarca en Aragón?*
2. *¿Qué se puede decir del número de lugares de importancia comunitaria (LIC)?*
3. *Calcula una nueva variable que indique el número total de zonas protegidas en cada una de las comarcas de Aragón. ¿Existe alguna comarca que no tenga ninguna zona protegida? ¿Es el diagrama de barras la representación gráfica más adecuada para esta nueva variable?*

2.2. Variables cuantitativas continuas

Si se desea construir la *tabla de frecuencias* de una variable cuantitativa con muchos valores distintos, el primer paso consiste en recodificar dicha variable (como se vio en la Práctica 1) agrupando sus valores en intervalos. Es conveniente almacenar la recodificación en una nueva variable para mantener la información numérica de la original, así como marcar la opción de convertir la nueva variable en factor.

Otra opción es seguir los siguientes pasos:

Estadísticos → Resúmenes → Resúmenes numéricos

y en estadísticos marcar únicamente la opción de *recuentos de frecuencias*.

Ejercicio 7. *Recodifica la variable número total de zonas protegidas calculada en el ejercicio anterior, en 8 intervalos de idéntica amplitud empezando por 0 y acabando en 40. Obtén la tabla de frecuencias de la variable recodificada. Obtén también la tabla de frecuencias a través de los resúmenes numéricos.*

2.2.1. Calculo de las frecuencias acumuladas

En las tablas de frecuencias de variables cualitativas ordinales o de variables cuantitativas suelen aparecer las frecuencias acumuladas. Para obtener las frecuencias acumuladas, debemos proceder con una variable tipo factor (es decir, con la primera opción explicada para construir la tabla de frecuencias para una variable cuantitativa continua).

La tabla de frecuencias de la variable que indica el número de espacios protegidos (convertida en factor) ha sido calculada con la opción:

Estadísticos → Resúmenes → Distribución de frecuencias

En la ventana *R Script*, el código que se ha generado ha sido el siguiente:

```
local({
  .Table <- with(ZonasProtegidas, table(TotalZonasProtegidasAgrupacion))
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
})
```

Para obtener las frecuencias absolutas y porcentuales acumuladas basta añadir las siguientes líneas al código anterior:

```
local({
  .Table <- with(ZonasProtegidas, table(TotalZonasProtegidasAgrupacion))
  cat("\ncounts:\n")
  print(.Table)
  cat("\npercentages:\n")
  print(round(100*.Table/sum(.Table), 2))
  cat("\nFrecuencias Absolutas Acumuladas:\n")
  print(cumsum(.Table))
  cat("\nPorcentajes Acumulados:\n")
  print(round(cumsum(100*.Table/sum(.Table)),2))
})
```

2.3. Histograma

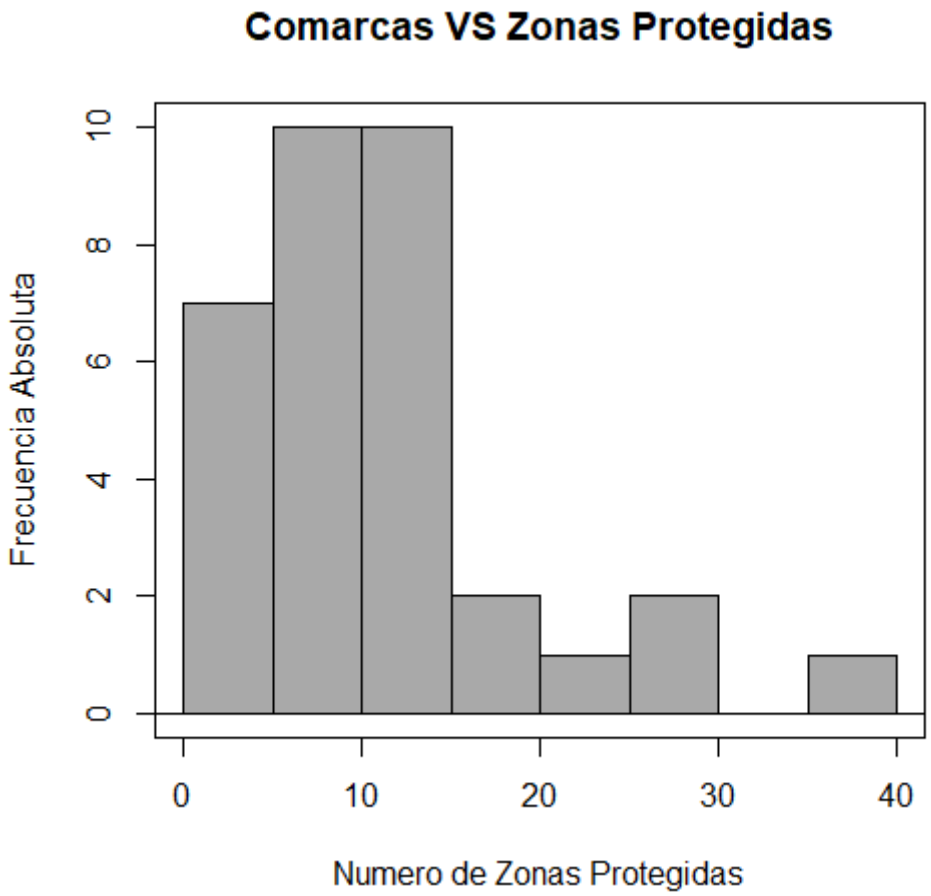
Para representar variables cuantitativas, ya sean discretas con muchos valores distintos o continuas, el gráfico más adecuado es el *histograma*.

Gráficas → Histograma

En la pestaña *Datos* se debe seleccionar la variable a representar. En el listado están disponibles las variables de tipo numérico, tanto si su naturaleza es cualitativa como cuantitativa, y en este segundo caso, tanto si presentan pocos como muchos valores distintos.

En la pestaña *Opciones*, se puede elegir el número de clases en las que quieres que se agrupe la variable (aunque se recomienda no escribir nada y dejar que *R Commander* elija el número que considere más adecuado). También se le debe indicar las frecuencias que se desean representar, así como dar título a los ejes y al gráfico.

Ejercicio 8. *Dibuja el histograma de la variable número total de zonas protegidas que se ha calculado en el ejercicio 6.*



NOTAS: Como hemos observado, *R Commander* dibuja los intervalos por defecto de la forma $(a, b]$, si lo deseamos, estos se pueden cambiar a la forma $[a, b)$ que es la forma más habitual de trabajar. Para ello en la ventana de *RScript* de R Commander habrá aparecido la siguiente instrucción cuando hemos creado el histograma:

```
with(ZonasProtegidas, Hist(TotalZonasProtegidas, scale="frequency", breaks="Sturges",
  col="darkgray", xlab="Numero de Zonas Protegidas", ylab="Frecuencia Absoluta",
  main="Comarcas VS Zonas Protegidas"))
```

Para modificar los intervalos debemos añadir: ***right = FALSE***

```
with(ZonasProtegidas, Hist(TotalZonasProtegidas, scale="frequency", breaks="Sturges",
  col="darkgray", right = FALSE, xlab="Numero de Zonas Protegidas", ylab="Frecuencia Absoluta",
  main="Comarcas VS Zonas Protegidas"))
```

Observando ambos histogramas, podemos apreciar las diferencias existentes.

Por otro lado, si se quiere modificar el **número de clases** que tiene R Commander por defecto hay que ser cauto, puesto que una excesiva partición (o un número excesivamente pequeño de clases) da lugar a gráficos poco informativos. Si le indicamos un número de clases a R Commander, por ejemplo breaks = 8, hay que tener en cuenta que R hace un redondeo, y no suele dibujar exactamente el número de clases que pedimos, ya que prefiere que los extremos de los intervalos sean número enteros. Otra posibilidad es indicarle

explícitamente los extremos de los intervalos, por ejemplo, en el ejercicio anterior se podría escribir: breaks = c(0,5,10,15,20,25,30,35,40) que es equivalente a breaks = c(seq(0,40,5)) (de 0 a 40 de 5 en 5).

Hemos visto que en la distribución de frecuencias no podemos establecer el número de clases deseadas, sin embargo en el histograma sí. Existe una manera de que la información obtenida con el histograma la podamos almacenar y consultar posteriormente. Para ello, en la ventana de RScript debemos introducir el siguiente comando:

```
HistTotalZonasProtegidas <- with(ZonasProtegidas, Hist(TotalZonasProtegidas, scale="frequency", breaks="Sturges",
  col="darkgray", right = FALSE, xlab="Numero de Zonas Protegidas", ylab="Frecuencia Absoluta",
  main="Comarcas VS Zonas Protegidas"))
HistTotalZonasProtegidas
```

Si ejecutamos la instrucción *HistTotalZonasProtegidas*, en la ventana de resultados aparece la información con la que se ha construido el histograma: Particiones (breaks), frecuencias absolutas (counts), densidades (frecuencia relativa) y marca de clase (mids).

```
> HistTotalZonasProtegidas
$breaks
[1]  0  5 10 15 20 25 30 35 40

$counts
[1]  4 13 10  2  1  2  0  1

$density
[1] 0.024242424 0.078787879 0.060606061 0.012121212 0.006060606 0.012121212 0.000000000 0.006060606

$mids
[1]  2.5  7.5 12.5 17.5 22.5 27.5 32.5 37.5

$xname
[1] "x"
```

3. Medidas de Síntesis o Resúmenes Numéricos

Las *medidas de posición, dispersión y forma* se obtienen con la opción:

Estadísticos → Resúmenes → Resúmenes numéricos

En la pestaña *Datos* se eligen la variable o variables para las que se desea obtener las medidas. Observa que en el listado solo aparecen variables cuantitativas. En la pestaña *Estadísticos* aparecen las medidas: *media, desviación típica, rango intercuartílico, coeficiente de variación, asimetría y cuantiles*, entre otras.

El *coeficiente de asimetría de Fisher* se calcula seleccionando la opción *Tipo 1*. En el cuadro junto a *Cuantiles* aparecen los valores $0, .25, .5, .75$ y 1 que corresponden al valor *mínimo, primer cuartil Q1, segundo cuartil Q2 (mediana), tercer cuartil Q3* y *valor máximo*, respectivamente. Si se desea calcular otros cuantiles, basta escribir su orden k en dicho cuadro.

Ejercicio 9. *Con los datos del archivo ZonasProtegidas:*

1. *Calcula todas las medidas de posición, dispersión y forma que conozcas de la densidad de población de las comarcas. Interpreta los resultados ayudándote de la representación gráfica más adecuada para esta variable.*

2. *¿Qué densidad de población tienen como máximo el* 10 % *de las comarcas con menor densidad de población? ¿Qué densidad de población tienen como mínimo el* 10 % *de las comarcas con más densidad?*

3.1. Diagrama de Caja

Como ya se vio en teoría, el gráfico más adecuado para representar los cuartiles de una variable continua, es el diagrama de caja. Además a través de este gráfico también se pueden sacar conclusiones acerca de la asimetría de la variable. Para realizar este gráfico debemos proceder de la siguiente manera:

Gráficas → Diagrama de caja

En la pestaña *Opciones*, se puede elegir la identificación automática de los datos atípicos o manual. Se recomienda la opción *Automáticamente*. Al igual que en los gráficos vistos anteriormente, se pueden elegir las etiquetas de los ejes y un título para el gráfico.

Ejercicio 10. *Con los datos del archivo ZonasProtegidas, dibuja el diagrama de caja de la densidad de población. Interpreta el gráfico. ¿Qué comarcas son las que mayor densidad de población tienen? ¿Pueden considerarse datos atípicos?*

3.2. Comparación de grupos definidos por una variable de tipo factor

Cuando se elige la opción:

Estadísticos → Resúmenes → Resúmenes numéricos

en la pestaña *Datos* aparece el botón *Resumir por grupos...* que permite realizar el estudio según los grupos definidos por los valores de una variable de tipo factor.

Ocurre lo mismo para el caso del histograma y el diagrama de caja, aparece la opción de *Gráfica por grupos...* que tiene la misma función.

Ejercicio 11. *Con los datos del archivo ZonasProtegidas,*

1. *Compara las medidas de posición, dispersión y forma de la densidad de población según la provincia a la que pertenece la comarca. Interpreta los resultados.*

2. *¿Qué provincia tiene la densidad de población más homogénea? ¿Se puede decir que hay mas variabilidad en Huesca que en Teruel?*

3. *Representa gráficamente la variable densidad de población según la provincia a la que pertenece la comarca mediante un histograma.*

   ***AYUDA**: Ten en cuenta que para comparar las gráficas debes representar en el eje Y las frecuencias y tanto en el eje Y como en el eje X el mismo rango de valores. Observa que R hace esto último por defecto.*

4. *Dibuja el diagrama de caja de la densidad de población según la provincia a la que pertenece la comarca.*

5. *Utilizando las medidas y los gráficos interpreta los resultados.*

6. *¿Qué provincia tiene más comarcas con densidad de población baja? ¿A qué provincia pertenecen las comarcas con mayor densidad de población?*

4. Tablas de doble entrada o tablas de contingencia

Como se ha comentado en las sesiones de teoría, las tablas de doble entrada se utilizan para estudiar conjuntamente el comportamiento de dos variables. *R Commander* solo permite su construcción para variables de tipo factor, por tanto, si alguna de las variables es cuantitativa tomando pocos valores distintos basta convertirla en factor y, si toma muchos valores distintos, es necesario agruparla previamente en intervalos.

La orden para generar una tabla de doble entrada es:

Estadísticos → Tablas de contingencia → Tabla de doble entrada

En la pestaña *Datos* se elige la variable que se desea representar en las filas de la tabla y la que se desea representar en las columnas. En la pestaña *Estadísticos* se elige el tipo de porcentajes que se quiere incluir:

- **Porcentaje por filas** para obtener las distribuciones de frecuencias de la variable que está representada en las columnas según los valores de la variable que está representada en las filas.
- **Porcentajes por columnas** para obtener las distribuciones condicionadas de la variable que está representada en las filas según los valores de la variable que está representada en las columnas
- **Porcentajes totales** para obtener las distribuciones de frecuencias porcentuales de las dos variables conjuntamente.
- **Sin porcentajes** para obtener únicamente la tabla de doble entrada de frecuencias absolutas.

Por el momento se deben dejar sin marcar las opciones correspondiente a Test de hipótesis que se estudiarán en la parte de Inferencia Estadística.

Ejercicio 12. *Con los datos del archivo submuestra_pisa.RData:*

1. *Calcula la distribución conjunta de frecuencias absolutas y de frecuencias porcentuales de las variables Ordenador y Género, y responde a las siguientes preguntas:*

   a) *¿Cuántos estudiantes disponen de ordenador en la vivienda? ¿Y cuántos no disponen de ordenador?*

   b) *¿Cuántos estudiantes son mujeres? ¿Y hombres?*

   c) *¿Cuántos estudiantes disponen de ordenador y son mujeres?*

   d) *¿Qué porcentajes de estudiantes no disponen de ordenador y son hombres?*

2. *Clasifica las variables Matemáticas y Ciencias en intervalos, de tal manera que se pueda calcular la tabla de doble entrada.*

   a) *De los estudiantes que tienen una calificación superior en matemáticas a* 500 *puntos, cuántos tienen una calificación superior en ciencias a* 600 *puntos?*

   b) *¿Cuantos estudiantes tienen una calificación inferior a* 400 *puntos en ciencias si en matemáticas tienen una calificación entre* 400 *y* 600 *puntos?*

3. *¿Cuantos hombres tienen padres que hayan estudiado a niveles Universitarios? ¿Y mujeres cuyos ambos padres no tengan niveles de estudios superiores a Bac/FP?*

   ***Pista**: Prueba a utilizar la opción de* Tablas de entrada múltiples *para responder a la pregunta.*

5. Representación gráfica para dos variables cualitativas

Existen dos tipos de gráficos para la representación de dos variables cualitativas: *Gráfico de barras apiladas* y *Gráfico de barras lado a lado*. Para conseguir estos gráficos basta con seguir los siguientes pasos:

Gráficas → Gráfica de barras

En la pestaña de *Datos* en la opción *Gráfica por grupos* seleccionar la variable por la que queremos agrupar. En la pestaña *Opciones*, en *Estilo de grupo barras* debemos seleccionar *lado a lado* si queremos un gráfico de barras lado a lado y *Dividido* si queremos un gráfico de barras apilado.

Si por el contrario queremos representar una variable cuantitativa en función de una variable cualitativa la opción que tenemos es realizar un Histograma y utilizar la opción de gráfica por grupos para seleccionar la variable cualitativa que realizará los grupos.

Ejercicio 13. *Con los datos del archivo submuestra_pisa.RData:*

1. *Dibuja el gráfico de barras apiladas y el gráfico de barras lado a lado para las variables Ordenador y Género. Trata de visualizar ambos gráficos en la misma ventana. ¿Se observa dependencia entre las variables?*

   ***Pista**: Utiliza la opción par(mfrow = c(2,2)) en la ventana R Script*

2. *Representa el histograma de las variables Matemáticas y Ciencias en función del género. Interpreta los resultados.*

3. *¿Puedes responder a las preguntas del apartado 3 del ejercicio anterior realizando un gráfico? En caso afirmativo realiza el gráfico correspondiente.*

6. Ejercicios

Ejercicio 14. *El archivo Seat2013 contiene los datos sobre el consumo (l/100km), emisiones (g de CO_2/km), clasificación energética (letras de la A = Muy eficiente a la G = Poco Eficiente, S = Sin clasificar) y tipo de combustible (D = Diésel, E = Eléctrico, G = Gasolina) de los vehículos de la marca Seat comercializados en España en 2013* [1].

1. *Define la población objeto de estudio. ¿Cuales son los individuos del estudio? Los datos disponibles, ¿corresponden a una población o a una muestra?*

2. *Abre el archivo, visualízalo y clasifica sus variables.*

3. *Obtén la tabla de frecuencias y la representación gráfica más adecuada para las variables tipo de combustible y clasificación energética. ¿Qué porcentaje de vehículos está clasificado como muy eficiente? ¿Qué vehículos son más frecuentes los que usan diésel o los que usan gasolina?*

4. *Representa gráficamente las variables consumo y emisiones. Interpreta los resultados.*

5. *¿Existe alguna medida de posición, dispersión o forma que puedas calcular para las variables combustible y clasificación energética? Si la respuesta es afirmativa, calcúlalas e interpreta el resultado.*

6. *Calcula las medidas de posición, dispersión y forma de las variables consumo y emisiones. Interpreta cada una de ellas. Ayúdate de las representaciones gráficas realizadas anteriormente. ¿Qué variable es más homogénea?*

7. *¿Cuál es el consumo máximo del 30 % de los vehículos que menos consumen? ¿Y la cantidad de CO_2/km mínima del 30 % de los vehículos que más Co_2/km emiten?*

8. *Compara el consumo de los vehículos según si utilizan gasolina o diesel. Para ello calcula las medidas de posición, dispersión y forma del consumo para cada uno de los grupos y utiliza el histograma y el diagrama de caja correspondiente. ¿Depende el consumo de los vehículos del tipo de combustible que utilizan?*

9. *Repite el apartado anterior para comparar la cantidad de CO_2/km que emiten los vehículos según el tipo de combustible que utilizan.*

10. *Responde a las preguntas del apartado 7 distinguiendo según el tipo de combustible que utilizan.*

[1] Fuente: Instituto para la Diversificación y Ahorro de la Energía, `www.idae.es`

Ejercicio 15. *El archivo EncuestaAguaAseoEstudiantes contiene los datos recogidos en una encuesta sobre consumo de agua en el aseo de los estudiantes de primero de Ciencias Ambientales en los años 2014 y 2015.*

1. *Obtén la tabla de frecuencias y la represetnación gráfica de la variable residencia habitual. ¿Qué porcentaje de encuestados vive en una residencia de estudiantes?*

2. *Analiza las variables número de lavabos, número de inodoros y número de duchas. Para ello:*

   a) *Dibuja el diagrama de barras de la variable número de lavabos y comprueba que toma muchos valores distintos. Observa que algunos encuestados han respondido que no disponen de lavabo en su residencial habitual. Lo mismo ocurre con las variables número de inodoros y número de duchas.*

   b) *Recodifica los valores* 0 *de las tres variables anteriores como datos faltantes.*
   ***AYUDA**: Recuerda que puedes recodificar las tres variables simultáneamente. Después basta que indiques en las directrices de recodificación tan solo la orden* 0 =*NA, y quites la marca en la opción* Convertir cada nueva variable en factor. *De esta forma la nueva variable sigue teniendo carácter numérico.*

   c) *Recodifica los valores de las tres nuevas variables que has creado con los valores* 1, 2 *y* 3 *o más.*

   d) *Dibuja en la misma ventana gráfica los diagramas de barras de las tres variables anteriores de forma que los ejes verticales representen la misma escala de valores.*
   ***AYUDA**: Antes de hacer el gráfico ejecuta la orden* par(mfrow=c(1,3)) *para que aparezcan los tres gráficos en una misma ventana. Si deseas volver a utilizar la pantalla con un único espacio, basta volver a ejecutar* par(mfrow=c(1,1)).

   e) *Interpreta los resultados obtenidos.*

3. *La variable veces.manos.residencia y veces.manos.fuera responden a la pregunta número de veces al día que te lavas las manos en tu residencia habitual y número de veces a la semana que te lavas las manos fuera de ella, respectivamente. Elige la representación gráfica más adecuada para comparar el número de veces al día que se lavan las manos los estudiantes dentro y fuera de la residencia habitual. Interpreta los resultados.*

   ***AYUDA**: Observa que primero debes cambiar las unidades de medida de la variable veces.manos.fuera.*

4. *Compara el gasto en agua semanal de los estudiantes cuando se duchan dentro y fuera de la residencia habitual. Para ello:*

   a) *Teniendo en cuenta que si una ducha tiene dispositivo de ahorro gasta* 6.5 *litros por minuto y si no tiene gasta* 10 *litros por minuto, crear una nueva variable que indique lo que gasta la ducha de cada estudiante por minuto.*
   ***AYUDA**: Recodifica la variable dispo.ahorro.ducha asignando* 10 *al valor No,* 6.5 *al valor Sí y el promedio* 8.25 *al valor No se, y manteniendo la nueva variable como variable cuantitativa.*

b) *Calcula una variable que represente el agua que gasta semanalmente cada estudiante cuando se ducha en su residencia habitual.*

***AYUDA**: Debes tener en cuenta lo que gasta su ducha por minuto, el número de veces por semana que se ducha en su residencia habitual (veces.ducha.residencia) y el tiempo que tarda en hacerlo (tiempo.ducha.residencia).*

c) *Calcula una variable que represente el agua que gasta semanalmente cada estudiante cuando se ducha fuera de su residencia habitual.*

***AYUDA**: Debes tener en cuenta el número de veces a la semana que se ducha fuera de su residencia habitual (veces.ducha.fuera) y el tiempo que tarda en hacerlo (tiempo.ducha.fuera). Considera que fuera de la residencia habitual una ducha gasta 8.25 litros por minuto (se utiliza el valor promedio puesto que no sabemos si dispone o no de dispositivo de ahorro).*

d) *Representa gráficamente las nuevas variables creadas en la misma ventana.*

e) *Interpreta los resultados.*

5. *Compara el número de veces que se duchan los chicos y las chicas teniendo en cuenta que la variable veces.ducha.residencia y veces.ducha.fuera responde a las preguntas cuántas veces a la semana te duchas dentro y fuera de tu residencia habitual, respectivamente.*

a) *Calcula la variable número de veces que se ducha un estudiante a la semana.*

b) *Representa gráficamente la variable que has calculado en el apartado anterior. ¿Hay datos atípicos? Si la respuesta es afirmativa, elimina los datos correspondiente a ducharse más de 14 veces a la semana.*

***AYUDA**: Puedes filtrar los datos para quitar aquellos que se duchan 14 o más veces a la semana.*

c) *Calcula las medidas de posición, dispersión y forma del número de veces que se duchan los estudiantes a la semana. Interpreta los resultados. Ayúdate de las representaciones gráficas más adecuadas. ¿Qué grupo presenta menos dispersión?*

d) *¿Cuántas veces se duchan como mínimo el 50 % de las chicas que más veces se duchan a la semana? ¿Y de chicos? ¿Qué medida has utilizado para responder a esta pregunta?*

e) *¿Qué valor o valores son los más frecuentes en los chicos? ¿Y en las chicas? ¿Qué medida has utilizado para responder a esta pregunta?*

f) *Repite los apartados anteriores para comparar el número de veces que se duchan los estudiantes distinguiendo además si se duchan dentro y fuera de la residencia habitual.*

6. *Compara la existencia de dispositivos de ahorro en las duchas según la antigüedad de la vivienda.*

a) *Calcula la tabla de doble entrada de las variables existencia de dispositivo de ahorro en las duchas y la antigüedad de la vivienda eligiendo en la pestaña Opciones porcentajes por filas o porcentajes por columnas según como hayas elegido las variables en la pestaña Datos.*

b) *Observa que las respuestas No sé deben ser eliminadas para hacer la comparación. Recodifica la variable existencia de dispositivo de ahorro en la ducha para eliminar estos valores.*

c) *Vuelve a calcular la tabla de doble entrada con la variable recodificada. Interpreta los resultados obtenidos.*

d) *Dibuja el diagrama de barras asociado a la tabla de contingencia del apartado c).*

e) *¿Depende la existencia de dispositivos de ahorro en la ducha de la antigüedad de la vivienda?*

7. *Calcula la distribución conjunta de frecuencias absolutas y de frecuencias porcentuales de las variables tipo de grifo en el lavabo y antigüedad de la vivienda.*

   a) *¿Cuántos estudiantes viven en viviendas anteriores a* 1990*? ¿Y en viviendas de* 1990 *o posterior?*

   b) *¿Cuántos estudiantes disponen de grifos monomando? ¿Y de grifos bimando?*

   c) *¿Cuántos estudiantes viven en viviendas posteriores a* 1990 *y disponen de un grifo monomando en el lavabo?*

   d) *¿Qué porcentaje de estudiantes dispone de un grifo bimando en el lavabo y vive en una vivienda anterior a* 1990*?*

8. *Dibuja la distribución conjunta de frecuencias absolutas de las variables tipo de grifo en el lavabo y antigüedad de la vivienda.*

   a) *Dibuja en la misma ventana un diagrama de barras apilado y otro diagrama de barras lado a lado.*

   b) *Comprueba que las alturas de las barras coinciden con los valores de la tabla de frecuencias absolutas obtenidas en el apartado anterior.*

   c) *Cambia la forma de elegir las variables en la pestaña Datos y observa la diferencia con los diagrama de barras anteriores. Para compararlos puedes dibujar los cuatro diagramas de barras en la misma ventana.*

Ejercicio 16. *En el fichero Informacion.RData, que recoge información sobre un conjunto de personas entrevistadas por un departamento de recursos humanos, se encuentran, entre otras, las variables Sexo (varon o mujer), Lugar (si el entrevistado prefiere trabajar en Madrid o no le importaría trabajar en otro lugar de España), y Calif1 y Calif2 (puntuaciones otorgadas a cada entrevistado por dos personas del departamento de recursos humanos).*

1. *Compara la media y la mediana de la variable Calif1 e interpretalas. ¿Crees que existe algún tipo de asimetría en los datos?*

2. *Representa gráficamente la variable Calif1 utilizando un diagrama de caja y comenta toda la información que observes en él.*

3. *Representa gráficamente la variable Calif1 utilizando un histograma y calcula su coeficiente de asimetría.*

4. *¿Que nota (Calif1) han sacado, como mínimo, el 10 % de los entrevistados que han sacado mayor nota?*

5. *Compara la variabilidad en las calificaciones de cada uno de los entrevistadores, ¿cuál de los dos es más homogéneo?*

6. *Calcula una nueva variable CalifM que sea la media aritmética de las calificaciones Calif1 y Calif2.*

7. *Calcula, a partir de la variable CalifM, una nueva variable CalifCuali que contenga el valor cualitativo de las calificaciones (suspenso, aprobado, notable y sobresaliente).*

8. *Representa gráficamente, utilizando un diagrama de sectores, la variable CalifCuali (no olvides ordenar los valores de la variable, ya que es cualitativa ordinal). A la vista del diagrama de sectores, ¿puedes decir cuál es el valor del primer cuartil, la mediana y el tercer cuartil?.*

9. *Calcula el primer cuartil, la mediana y el tercer cuartil de la variable CalifM y observa si coinciden sus valores con tus respuestas al apartado anterior.*

10. *Obtén el diagrama de caja de la variable CalifM. ¿Observas algún valor atípico? En caso afirmativo, comenta qué casos son y qué calificaciones han obtenido por parte de cada uno de los entrevistadores.*

Ejercicio 17. *En un seguimiento realizado a un grupo de 284 clientes de una empresa, se ha anotado el número de veces que los clientes han realizado reclamaciones a la empresa en el último año, para resolver alguna incidencia. Así, un cliente puede que no haya tenido que reclamar nunca a lo largo del año, y otro puede que haya tenido que reclamar 4 veces. Los datos se encuentran en el archivo Veces.RData.*

1. *Calcula la mediana e interprétala.*

2. *Representa gráficamente el número de reclamaciones utilizando un diagrama de caja. Explica el porqué de la forma de dicho diagrama.*

3. *Utiliza un diagrama de barras para representar gráficamente el número de reclamaciones.*

4. *Si preguntamos a 213 de los 284 clientes (un 75 %), elegidos al azar, seguro que alguno de ellos ha reclamado, al menos, ___ veces.*