Laboratorio: Análisis exploratorio de una variable

Objetivos: El alumno al finalizar la práctica ha de ser capaz de:

- Navegar por la hoja de datos y las diferentes ventanas del programa.
- Utilizar la ayuda de Minitab.
- Crear una nueva variable realizando operaciones sobre las variables ya existentes.
- Manejar los conjuntos de datos con operaciones básicas.
- Obtener la distribución de frecuencias de variables cualitativas y cuantitativas con pocos valores.
- Describir las características de una variable usando los estadísticos básicos.
- Utilizar gráficos para representar la información que proporcionan los datos.

En las sesiones prácticas se va a utilizar el **software estadístico Minitab** en la versión 17 (14,15) (http://www.minitab.com). La ejecución del programa se desarrolla en un entorno Windows y se facilita la selección de procedimientos a través de cuadros de diálogo. Los datos aparecen en **Worksheet** en forma de matriz. Cada fila es un individuo sobre el que se ha medido diferentes características y cada columna es una variable que se mide para cada individuo. Otra ventana se denomina **Session** y en ella aparecen los resultados numéricos solicitados y refleja todos los movimientos realizados. Por último, cuando se ejecuta un procedimiento gráfico se abren las ventanas con los gráficos requeridos. Es posible guardar todo el análisis sobre una colección de datos en un fichero que Minitab denomina **Project**.

El **software R** se distribuye de manera gratuita a través de Internet en CRAN (http://www.r-project.org). R es un entorno en el que se han implementado muchas técnicas estadísticas, tanto clásicas como modernas. El programa también se puede descargar en la página de la oficina de software libre de la Universidad de Zaragoza, así como manuales para su utilización. El manejo de este software puede ser en lenguaje de comandos o bien se puede instalar el paquete R-Commander. Una vez cargado este paquete se puede manejar el software en un entorno de ventanas.

Dada una colección de datos, la primera tarea a realizar es organizar y resumir la información contenida en los mismos para conocer su distribución. El primer análisis en un conjunto de datos consiste en explorar las características más importantes de cada variable. Además, las técnicas básicas para la descripción de datos resultan de gran utilidad en la depuración de muestras o identificación de valores anómalos y errores. El análisis exploratorio de cualquier variable incluye calcular las medidas numéricas adecuadas y realizar los gráficos necesarios.

Los datos pueden ser de naturaleza numérica o cualitativa. La consideración de una variable como cualitativa, numérica discreta o continua condiciona su tratamiento descriptivo. Son **cualitativos** cuando clasifican a los

individuos en diferentes categorías que se distinguen por alguna característica no numérica, como el sexo. Aunque estos datos estén codificados con valores numéricos su tratamiento se realiza como variable cualitativa. Son **cuantitativos** cuando proceden de una medición numérica, como el peso, la altura o el número de hermanos. En este caso se distingue entre numéricas discretas como el número de hermanos y numéricas continuas, como el peso y la altura. Las variables numéricas discretas pueden tomar un número pequeño de valores, como el número de hermanos, o un número muy grande de valores, como la población de la ciudad origen del alumno.

En la siguiente sección se introduce el tratamiento de variables numéricas continuas y en la segunda sección, el tratamiento de variables cualitativas. En la siguiente tabla se indica en qué sección se ha explicado el correspondiente análisis exploratorio de acuerdo con el tipo de variable:

	Cualitativa	Numérica		
		Continua	Discreta	
			Escasos valores	Elevado número de valores
Uso de gráficos estadísticos	Sección 2	Sección 1	Sección 2	Sección 1
Cálculo de medidas numéricas	Sección 2	Sección 1	Sección 1 y 2	Sección 1

En este guión se hace referencia a los datos contenidos en el fichero ALTURA_PESO_TELECO.mtw. Este archivo contiene el peso, en kilogramos, y la altura, en metros, de los alumnos de Ingeniería de Telecomunicaciones, indicando también su sexo y el año de inicio de los estudios.

1. Análisis descriptivo de variables numéricas

En primer lugar, para obtener una idea rápida sobre la distribución de los datos se considera la realización de gráficos. En este caso se va a estudiar la distribución del *Peso*. Cuando el número de datos a representar no es muy elevado se puede utilizar un gráfico de puntos **Graph/Dotplot** que conserva los valores exactos de cada una de las observaciones. Si se realiza un histograma **Graph/Histogram**, se agrupan los datos en clases y se pierde el valor exacto de los mismos, a cambio suaviza la forma de su distribución. El software determina por defecto el número de clases definidas. Es habitual que en el histograma se superponga la curva que describe el patrón de un modelo teórico, por ejemplo la campana de Gauss del modelo normal. La observación de cualquier gráfico de la variable *Peso* identifica la presencia de datos "raros" que hay que comprobar con el tomador de datos. Notemos que, el histograma proporciona una primera idea sobre la densidad de probabilidad de la variables representada: los valores más probables, la simetría o no de la distribución y la concentración o no de los valores en torno a algún valor. Estos gráficos resultan adecuados para analizar variables numéricas continuas y discretas con muchos valores.

La interpretación de un gráfico tiene cierta componente subjetiva. Por este motivo, cualquier análisis exploratorio se acompaña del cálculo de un conjunto de medidas estadísticas. Las medidas que aportan información sobre la posición donde se sitúan los datos son medidas de localización, y las que aportan información sobre la variabilidad de los datos son medidas de dispersión. Además hay otras medidas que informan sobre la forma de la distribución. En este caso se van a calcular estas medidas para estudiar la distribución del *Peso*. En Minitab, con el procedimiento **Stat/Basic Statistics/Display Descriptive Statistics** se obtienen, entre otras, las siguientes medidas:

N Número de datos.

Mean Media, el peso medio de todos ellos:

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Median Mediana, el 50 % de los estudiantes pesan menos que la mediana.

TrMean Media truncada al 5 %, es la media de los pesos, una vez se han eliminado el 5 % de los estudiantes que menos pesan y el 5 % de los estudiantes que más pesan.

StDev Desviación típica muestral es un medida de la dispersión de los datos, $\hat{S}$. La expresión de la cuasivarianza muestral es:

$$\hat{S}^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{X})^2}{N-1}$$

SE Mean Desviación típica de la media muestral, es una medida de la precisión de la estimación dada por la media muestral:

$$\frac{\hat{S}}{\sqrt{N}}$$

Minimum El valor mínimo.

Maximum El valor máximo.

Q1 Primer cuartil, el 25 % de los estudiantes pesan menos.

Q3 Tercer cuartil, el 75 % de los estudiantes pesan menos.

La **media truncada** es una medida de localización o tendencia central menos sensible a valores extremos que la media. Si toman valores muy diferentes indica que existen valores extremos. En el caso de la variable *Peso*, como existen valores extremos por debajo y por arriba, ambas medias toman valores próximos. El hecho de que la media y la mediana no sean próximas es un indicio de que la distribución no es simétrica. Para valorar la variabilidad entre los datos de la muestra son de gran interés las medidas de dispersión. La desviación típica y la diferencia entre Q1 y Q3 (rango intercuártilico) dan una idea de la dispersión de los datos. El coeficiente de asimetría (skewness) informa sobre la simetría de la distribución. Otra medida de forma es la kurtosis, para ello cuando se estudia una distribución de carácter simétrico se compara con la campana de Gauss (distribución normal), si es más plana la kurtosis es negativa y si es más puntiaguda la kurtosis es positiva.

Otro gráfico interesante es el diagrama de caja **Graph/Boxplot**. En la figura 1, los lados de la caja del gráfico boxplot pasan por el primer y tercer cuartil y la línea central se sitúa a nivel de la mediana. Por tanto, la amplitud de la caja coincide con el rango intercuartílico. Los segmentos van desde el extremo de la caja hasta los datos más extremos que se encuentran a una distancia menor de 1.5 veces el rango intercuartílico desde el borde de la caja. Los valores más alejados de esos límites se representan con *. El diagrama de caja

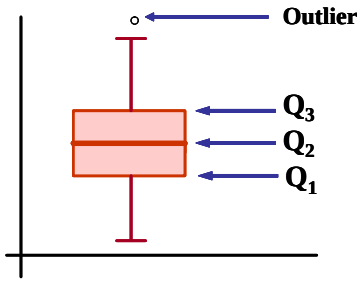

Figura 1: Diagrama de caja o Boxplot

además de informar sobre la forma de la distribución, identifica el valor de los cuartiles, la mediana y el rango

intercuártilico y los valores atípicos. Estos valores requieren un análisis especial. Si se reconocen como errores en la captura de datos se eliminan y se realiza de nuevo el análisis, ya que tanto los gráficos como las medidas numéricas están muy influidos por la presencia de datos atípicos o "outliers". En otro caso, son motivo de un análisis más exhaustivo. En las figuras 2 y 3 se muestra el efecto de los datos atípicos en las medidas numéricas y en los gráficos de dos colecciones de 15 datos que sólo difieren en uno.

Descriptive Statistics: Peso1; Peso2

Variable	Total Count	Mean	SE Mean	TrMean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness	Kurtosis
Peso1	15	69,79	1,49	69,76	5,77	60,55	66,41	68,27	77,12	79,45	0,62	-0,62
Peso2	15	73,50	3,63	70,92	14,04	60,55	66,73	69,36	77,40	120,00	2,87	9,58

Figura 2: Efecto de los datos atípicos en las medidas numéricas

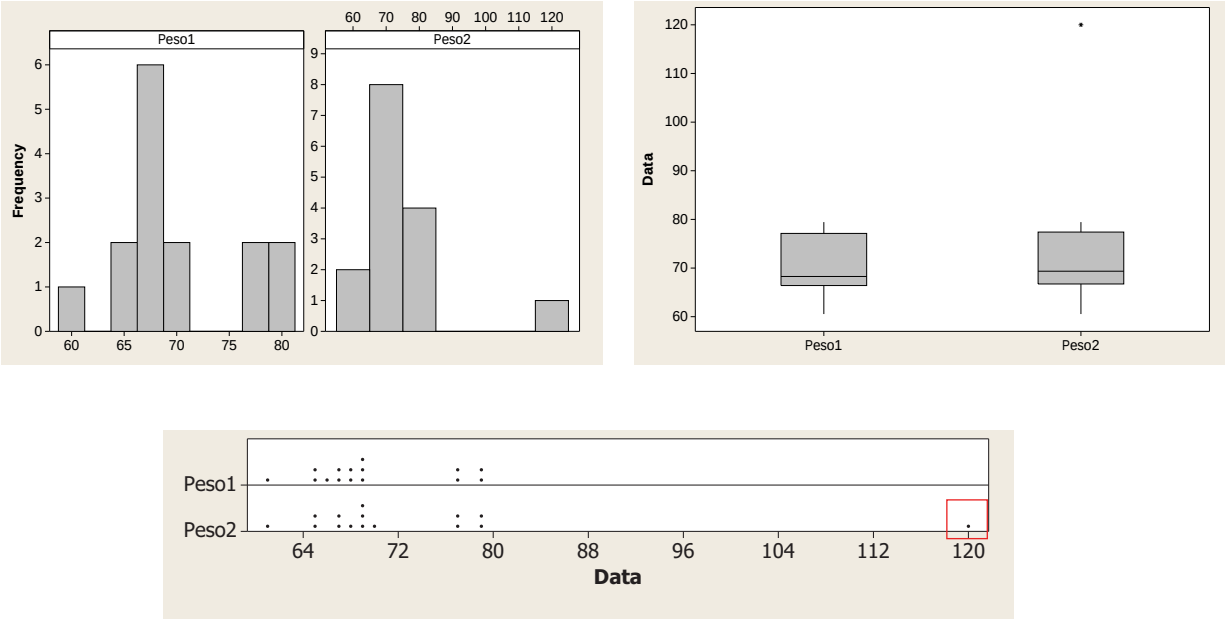

Figura 3: Efecto de los datos atípicos en los gráficos

La opción de resumen gráfico **Stat/Basic Statistics/Graphical Summary** proporciona una descripción muy exhaustiva de los datos. El resumen gráfico (figura 4) incluye un análisis sobre el comportamiento normal de los datos. La hipótesis de normalidad es básica en la aplicación de otras técnicas estadísticas como el diseño de experimentos, regresión y control de calidad. Esto sugiere la gran importancia de evaluar dada una colección de datos si su distribución teórica asociada puede ser o no una distribución normal. En el histograma se superpone la campana de Gauss correspondiente a una distribución normal de la misma media y varianza que los datos, que permite observar el ajuste al perfil del histograma. Si los datos fueran normales, en el histograma se ha de apreciar un comportamiento simétrico de la distribución de los datos así como una acumulación alrededor de los valores centrales. En la parte derecha del gráfico aparece un resumen con las principales medidas numéricas descriptivas. El primer bloque hace referencia a un contraste de normalidad de los datos. En el segundo bloque medidas de localización y sus correspondientes de dispersión. En el tercer bloque, medidas basadas en el orden de los datos. En el último bloque, intervalos de confianza para la media, mediana y desviación típica.

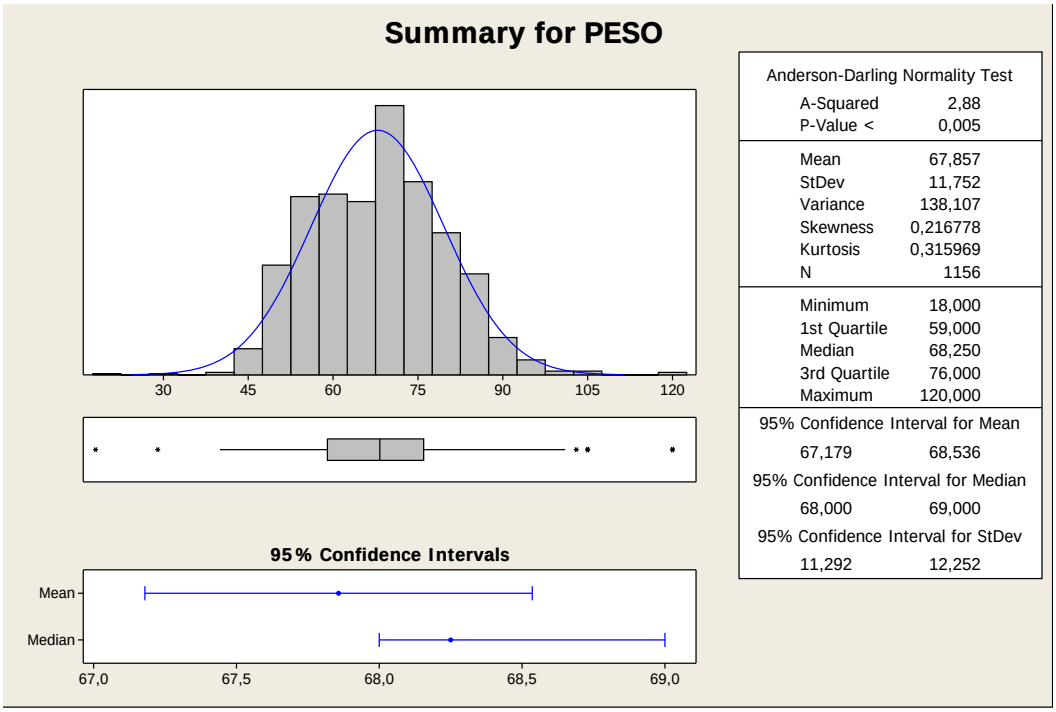

Figura 4: Resumen Gráfico

2. Análisis descriptivo de variables cualitativas

La distribución de una variable cualitativa, o de una variable discreta con pocos valores, se describe mediante su tabla de frecuencias y frecuencias relativas (**Stat/Tables/Tally**) y su representación gráfica habitual es el diagrama de sectores (**Graph/Pie chart**) o el diagrama de barras (**Graph/Bar char**). Para expresar la mayor o menor frecuencia de las categorías de datos cualitativos se construye un gráfico Pareto **Stat/Quality Tools/Pareto Chart**. En este gráfico, las barras se organizan según la frecuencia.

La presencia de variables cualitativas en un conjunto de datos permiten realizar análisis del resto de variables dependiendo de su valor. Por ejemplo, se puede analizar la variable altura según el sexo del estudiante (**Stat/Basic Statistics/Display Descriptive Statistics**) con la opción **By variable**. Hay que ser cuidadoso al segmentar con más de una variable cualitativa a la vez, ya que puede ocurrir que algún grupo se quede con escasos datos que los resultados ya no sean informativos. Los gráficos, como el diagrama de caja o el diagrama de puntos, pueden representar una variable numérica según los valores de una cualitativa. Estos gráficos proporcionan una idea visual muy efectiva sobre diferencias en la distribución de la variable numérica para cada una de las clases dadas por la variable cualitativa. La comparación se realiza tanto en valores medios como en la variabilidad de los datos.

Cuando los datos se han recogido según su orden de ocurrencia es de interés representarlos en un gráfico en el que el eje de abcisas se corresponde con el tiempo, de forma que se analice su evolución temporal. Por ejemplo, el peso de los chicos y las chicas (**Graph/Individual Value Plot**) a lo largo del tiempo.

En ocasiones los datos se han de transformar. En Minitab, los menús **Manip** y **Calc** ofrecen diferentes posibilidades de operar con los datos (dividir el archivo de datos **Manip/Split Worksheet**, establecer un subconjunto de los datos **Manip/Subset Worksheet**, codificar variables **Manip/Code**) y calcular nuevas

variables (como función de variables conocidas **Calc/Calculator**).

	Variable cualitativa	Variable numérica	
		Continua	Discreta
Gráficos	Sectores circulares Diagrama de barras	Dot plot Histograma Boxplot Probability plot	Dot plot Histograma Boxplot
Medidas numéricas	Frecuencia relativa Moda	Frec. relativa Frec. relativa acumulada Percentiles Mediana Cuartiles Rango intercuartílico Media Media truncada Desviación típica Coef. de variación Coef. de asimetría (skewness) Coef. de curtósis	Frec. relativa Moda Frec. relativa acumulada Percentiles Mediana Cuartiles Rango intercuartílico Media Media truncada Desviación típica Coeficiente de variación Coef. de asimetría

3. Consejos y errores comunes en este laboratorio

- El primer paso en el análisis consiste en identificar qué variables son categóricas y cuáles son numéricas. En el caso de ser numérica, si se trata de una variable discreta o de una variable continua. De acuerdo con el tipo de variable se realiza el análisis con medidas numéricas y gráficos adecuados. El análisis es diferente para cada tipo de variable.
- En el análisis de una variable categórica debe incluir un diagrama de sectores circulares, sólo cuando el número de categorías no es elevado, o un diagrama de barras que represente la frecuencia relativa, o el porcentaje, de casos de cada categoría.
- Para las variables categóricas o cualitativas no tiene sentido elaborar una descriptiva con medias, medianas, percentiles, desviación típicas, medidas de forma o representar un histograma.
- En variable numérica debe analizarse la posición central (media, media recortada, mediana) y la variabilidad (desviación típica, rango intercuartílico, coeficiente de variación).
- Además de las medidas de localización hay que establecer conclusiones también sobre la variabilidad.
- El coeficiente de asimetría o skewness expresa el grado de simetría en que se disponen los datos de la muestra alrededor de su media.
- Los gráficos para variables numéricas más adecuados son el histograma, el dotplot y el boxplot. El gráfico de barras no es adecuado, porque tiene un eje X en el que no existe una escala numérica y hay que utilizarlo para variables categóricas o numéricas discretas.
- Cuidado al elaborar conclusiones el uso de la palabra "significativo", en Estadística tiene una connotación especial.

- Un valor atípico no debe eliminarse de la muestra sistemáticamente, sino sólo si hay evidencia de que es un dato falso o que no corresponde a la población que se muestrea.
- En la comparación del comportamiento de una variable numérica según los valores de una categórica es muy útil el uso de diagramas de caja o boxplot.

4. Ejercicios propuestos

Los ficheros **Latas.mtw**, **Desperdicios.mtw**, **Hijos.mtw** y **Osos.mtw** se han obtenido del libro: Mario F. Triola. Estadística. Décima edición, Pearson Educación, México, 2009. ISBN: 978-970-26-1287-2. El fichero **Detergente.mtw** y **Coches.mtw** se ha obtenido del libro: P. Grima, L.I. Marco, J. Tort-Matorell. Estadística Práctica con Minitab. Pearson Educación, 2004. ISBN: 9788420543550.

1. El archivo **Ordenador.mtw** contiene la información dada por una revista en el año 1999 sobre 200 ordenadores divididos en cuatro categorías. Para cada ordenador se indica la relación calidad/precio en su categoría, el precio en pesetas, la velocidad en Mh, el tipo de procesador, Mb de memoria Ram, Mb de disco duro, su categoría y si es ordenador portátil, su peso.

   a) Describir el comportamiento de las variables *Precio* y *Procesador*. Utilizar medidas numéricas y los gráficos adecuados. Establecer si existen valores extremos o aspectos a destacar.

   b) Analizar el comportamiento del precio según el tipo de procesador.

   c) Estudiar la variable Precio por mega de memoria Ram.

   d) Describir el comportamiento de las variables numéricas medidas para los portátiles de gama alta.

2. La carga axial de una lata es el peso máximo que pueden soportar sus costados. Es importante tener una carga axial elevada para que la lata no se aplaste cuando la tapa superior se coloque a presión en su lugar. Las latas más delgadas tienen la ventaja de usar menos material, con lo que se reduce el coste, pero estas latas probablemente no sean tan fuertes como las más gruesas. El archivo **Latas.mtw** contiene información sobre la carga axial de latas de 0.0109 y 0.011 pulgadas de espesor. Las medidas se han tomado durante el proceso de fabricación en muestras de tamaño 7 recogidas cada cierto tiempo.

   a) Describir el comportamiento de la variable *L1*. Utiliza medidas numéricas y los gráficos adecuados. Establecer si existen valores extremos o aspectos a destacar.

   b) Analizar el comportamiento de la carga según el grosor de la lata.

   c) Estudiar la carga axial media por muestra en las 25 muestras de las latas de grosor 0.011 pulgadas.

   d) ¿Podrían usarse latas más delgadas sin disminuir significativamente la carga soportada por las mismas?

3. El archivo **Desperdicion.mtw** recoge los pesos en libras de diferentes categorías de desperdicios desechados por una muestra de 62 hogares (metales, papel, plástico, vidrio, productos alimenticios, desperdicios del jardín, textiles y otros no incluidos en las categorías anteriores).

   a) Describe el comportamiento de las variables *Orgánica* y *Tamaño*. Utiliza medidas numéricas y los gráficos adecuados. Comenta si existen valores extremos o aspectos a destacar.

   b) Analizra el comportamiento del peso en desperdicios orgánicos según el tamaño de la familia.

   c) Describir el comportamiento de las variables del problema para las familias de 2 miembros.

d) Construir un diagrama Pareto y un diagrama de sectores que ilustre la cantidad relativa de los pesos totales de cada categoría de desperdicio.

4. El archivo **Detergente.mtw** contiene el peso, en gramos, de 500 paquetes de detergente de peso nominal 4 Kg, y se indica en cuál de las dos líneas de producción disponibles se han llenado.

   a) Describir el comportamiento de la variable *Peso*. Utilizar medidas numéricas y los gráficos adecuados. Comentar si existen valores extremos o aspectos a destacar.

   b) Analizar el comportamiento del peso de los paquetes según la línea de llenado.

   c) ¿Se ajusta al valor nominal de llenado de 4 Kg la línea 1?, ¿y la línea 2?

   d) Analizar la desviación del valor nominal 4 Kg en los paquetes llenados por la línea 2.

5. El archivo **Pulso.mtw** (del software Minitab) contiene los datos sobre 92 estudiantes de una clase. Para cada estudiante se recoge su altura, peso, sexo, si fuma o no, el nivel de actividad física habitual y su pulso en reposo. Se eligió al azar un conjunto de estudiantes y corrieron 1 minuto, a continuación, todos se volvieron a tomar el pulso.

   a) Describir el comportamiento de las variables *Pulso 1* y *Actividad*. Utilizar medidas numéricas y los gráficos adecuados. Comentar si existen valores extremos o aspectos a destacar.

   b) Analizar el comportamiento del pulso en reposo según el sexo del estudiante.

   c) Analizar la variable incremento definida como diferencia entre el pulso final y el pulso en reposo.

   d) Construir un gráfico de barras para representar la frecuencia de hombres y mujeres según la actividad física que realizan.

6. El archivo **Osos.mtw** contiene los datos de 54 osos silvestres. Para cada oso se recoge su edad, el mes de medición, su sexo y medidas físicas de la cabeza y el cuerpo.

   a) Describir el comportamiento de las variables *Peso* y *Sexo*. Utilizar medidas numéricas y los gráficos adecuados. Establecer si existen valores extremos o aspectos a destacar.

   b) Analizar el peso según el sexo del animal.

   c) Definir una variable que sea la edad del oso en años y representar gráficamente la proporción de osos de cada edad.

   d) Describir el comportamiento de la variable longitud alrededor del cuello para los osos menores de 24 meses.

7. El archivo **Hijos.mtw** contiene los datos de 40 personas que proporcionan información sobre el sexo, la altura y la altura de sus padres.

   a) Describir el comportamiento de las variables *Altura* y *Sexo*. Utilizar medidas numéricas y los gráficos adecuados. Establecer si existen valores extremos o aspectos a destacar.

   b) Analizar la altura según el sexo de la persona.

   c) Definir una variable que clasifique en tres grupos a los hijos según la altura de su padre.

   d) Describir el comportamiento de la variable altura de los hijos cuyos padres se encuentran en el grupo de menor altura según la clasificación del apartado anterior.

8. El archivo **Coches.mtw** contiene información extraída de la revista *Coche Actual* de noviembre de 1994 sobre 247 coches.

a) Describir el comportamiento de las variables *Consumo* y *Marca*. Utilizar medidas numéricas y los gráficos adecuados. Establecer si existen valores extremos o aspectos a destacar.

b) Estudiar el comportamiento de la variable consumo según la marca del coche.

c) Calcular la variable volumen del coche, suponiendo que es un prisma cuadrangular, en litros (resta 50cm a la altura del coche y 100cm a la longitud) y estudiar el porcentaje de la capacidad del maletero en el volumen total.

d) Comparar el comportamiento de la variable precio para los coches de 4 y 6 cilindros.