Estadística Básica con R y R–Commander

(Versión Febrero 2008)

Autores:
A. J. Arriaza Gómez
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
S. Pérez Plaza
A. Sánchez Navas

UCA Universidad de Cádiz
Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN:

Depósito legal:

Estadística Básica con R y R-commander
(Versión Febrero 2008)
Autores: A. J. Arriaza Gómez, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, S. Pérez Plaza, A. Sánchez Navas

http://knuth.uca.es/ebrcmdr

Capítulo 2

Análisis Exploratorio de Datos Unidimensional

En este módulo, a través de una serie de medidas, gráficos y modelos descriptivos, se caracterizará a un conjunto de individuos, intentando descubrir regularidades y singularidades de los mismos y, si procede, comparar los resultados con los de otros grupos, patrones o con estudios previos. Se podría considerar que este estudio es una primera entrega de un estudio más completo o, por contra, tener un carácter finalista; en cualquier caso, se trata de un análisis calificable como de *exploratorio*, y de ahí el nombre del capítulo.

Las conclusiones obtenidas serán aplicables exclusivamente a los individuos considerados explícitamente en el estudio, sin que puedan hacerse extrapolaciones con validez científica fuera de ese contexto. Los resultados del Análisis Exploratorio de Datos (AED) sí que podrían emplearse para establecer hipótesis sobre individuos no considerados explícitamente en dicho análisis, que deberían ser posteriormente contrastadas.

Formalmente, se podría definir el AED como un conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas; aunque esta primera entrega se centrará en un análisis de tipo unidimensional.

1. La organización de la información

Al conjunto de individuos físicos considerados en un análisis se le denominará *Colectivo* o *Población*, aunque también se utilizarán esos mismos términos para referirse a la(s) característica(s) de esos individuos que son objeto de estudio. De hecho, desde un punto de vista estadístico, los individuos sólo interesan como portadores de rasgos que son susceptibles de marcar diferencias entre ellos. La obtención y materialización en formato analógico o digital de las características consideradas constituirá el conjunto de datos que será estadísticamente analizado.

Los datos constituyen pues la materia prima de la Estadística, pudiéndose establecer distintas clasificaciones en función de la forma en que éstos vengan dados. Se obtienen datos al realizar cualquier tipo de prueba, experimento, valoración, medición, observación, ..., dependiendo de la naturaleza de los mismos y del método empleado para su obtención. Una vez obtenidos los datos por los procedimientos que se consideren pertinentes, pueden generarse nuevos datos mediante transformación y/o combinación de las variables originales. Al conjunto de datos convenientemente organizados se le llamará *modelo de datos*.

1.1. La matriz de datos

En una primera instancia se supondrá que, sobre un conjunto de n individuos físicos, se obtienen una serie de k caracteres u observaciones de igual o distinta naturaleza. Es importante tener en cuenta, ya desde este momento, que la calidad del análisis que se realice, va a depender de la habilidad que se tenga a la hora de seleccionar los caracteres que se obtendrán del conjunto de individuos seleccionados.

Los datos obtenidos se organizarán en una matriz $n \times k$, donde cada fila representa a un individuo o registro y las columnas a las características observadas. Las columnas tendrán naturaleza homogénea, pudiendo tratarse de caracteres nominales, dicotómicos o politómicos, presencias–ausencias, ordenaciones, conteos, escalas de intervalo, razones,...; también se podrían tener variables compuestas como ratios, densidades,... En ocasiones se añade una columna que se suele colocar en

primer lugar y que asigna un nombre a cada individuo; dicha columna recibe el nombre de *variable etiqueta*.

Físicamente, la estructura de una matriz de datos se corresponde con el esquema de una base de datos o una hoja de cálculo. Al igual que pasa con los editores de los programas de tratamiento de datos, las dos dimensiones de una pantalla se acomodan perfectamente al tanden individuo–variable. Si se consideran los individuos identificados por los términos $I_1, I_2, \ldots, I_n$ y los caracteres por $C_1, C_2, \ldots, C_k$, la casilla x_{ij} representa el comportamiento del individuo I_i respecto al carácter C_j. En la figura se muestra la matriz de datos del fichero `Iris` del paquete `datasets` de **R**.

Editor de datos

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa
14	4.3	3	1.1	0.1	setosa
15	5.8	4	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa

R se refiere a este tipo de estructura de datos como *data.frame*. Este es el formato que requiere el programa para aplicar la mayoría de los procedimientos estadísticos.

1.1.1. Anomalías de la matriz de datos

Hay veces en que por distintos motivos la matriz de datos presenta casillas vacías, ello se debe a que no se ha podido medir un dato o a que se ha perdido la observación. En otras ocasiones un dato presente en la matriz ha sido *depurado* por presentar algún tipo de anomalía, como haber sido mal medido, mal transcrito a la matriz de datos, pertenecer a un colectivo distinto del que se está analizando, etc... La identificación de estos elementos anómalos se realiza mediante un proceso de detección de inconsistencias o de evaluación de valores extremos, muy grandes o muy pequeños, que determinará si razonablemente pueden pertenecer al colectivo bajo estudio. A veces se sustituye el valor depurado de un

individuo por uno que sea congruente con el resto de caracteres del mismo, mediante técnicas que se conocen como de *imputación*. Los huecos que definitivamente queden en la matriz se referirán como *valores omitidos* o, más comunmente, como *valores missing*. En **R** estos valores se representan con `NA` (Not Available). En función del tipo de análisis que se esté realizando, el procedimiento desestimará sólo el dato o todo el registro completo.

En este módulo se analizarán –salvo excepciones que se indicarán con antelación– de forma independiente cada uno de los caracteres de la matriz de datos, de forma que cada carácter describirá parcialmente al conjunto de individuos. La integración de todos los análisis deberá dar una cierta visión general de la población. En cualquier caso, este enfoque está muy lejos de ser eficiente, entre otras cosas porque habitualmente las variables individuales comparten información y dicha redundancia distorsionaría las conclusiones del estudio, siendo en general preferible decantarse por un análisis global en vez del secuencial. Por tanto, la pretensión de este capítulo es tratar algunos conceptos básicos y adquirir destreza en el manejo de medidas estadísticas que serán empleadas masivamente cuando se aborden, más adelante, modelos más sofisticados.

2. Naturaleza de los caracteres: Atributos y Variables

Respecto a la *cantidad de información* que porta cada tipo de carácter, se puede considerar que los caracteres nominales son los más "pobres", puesto que ni siquiera poseen orden, mientras que los más ricos serían las escalas de intervalos y las razones, que tienen orden, son cuantitativas y en el caso de las razones el cero lo es en términos absolutos, es decir, el 0 representa la ausencia de la característica. En posiciones intermedias se situarían el resto en el orden en que se han introducido en la figura 2.1.

Ejemplo 2.1

El caso más evidente para apreciar las diferencias entre las escalas de intervalo y las razones o escalas de cociente, lo ofrece el termómetro. Un termómetro genera una variable de escala de intervalo, porque la

Figura 2.1: Esquema de cantidad de información

diferencia real entre 2 y 3 grados es la misma que entre 40 y 41 grados, pero no se puede decir que cuando el termómetro marca 30 grados hace el doble de calor que cuando marca 15.

Por otra parte, muchas magnitudes físicas, como el peso, la longitud o la intensidad de corriente, son razones porque, por ejemplo en el caso del peso, un objeto de 20 kilogramos pesa el doble que otro de 10 kilogramos. Es decir existe el cero absoluto.

Como ya se ha comentado, la naturaleza del carácter condicionará su tratamiento, aunque en ningún caso hay que confundir la cantidad de información que porta con su valor intrínseco para analizar a los individuos del colectivo.

En una primera instancia, se distinguirá entre los caracteres que no están ordenados y los que sí lo están, los primeros jugarán en general un rol de *atributos* mientras que los segundos habitualmente actuarán como *variables*. Los atributos tendrán la misión de establecer clases, dividiendo el colectivo global en subgrupos o categorías; por su parte, las variables caracterizarán a dichos subgrupos e intentarán establecer diferencias entre unos y otros, para lo que necesariamente se debe considerar algun tipo de métrica. Pero ello es una regla general que tiene muchas excepciones y así, en ocasiones, un carácter llamado a adoptar el papel de variable podría, mediante una operación de *punto de corte*, actuar como atributo, mientras que es factible definir una medida de asociación sobre caracteres intrínsecamente de clase que permita caracterizar a los individuos del colectivo en base a una serie de atributos.

Ejemplo 2.2

Es habitual que la edad, que es intrínsecamente una variable –medida en un soporte temporal– se emplee para dividir la población en clases dando cortes en el intervalo de tiempo, obteniéndose por ejemplo grupos de alevines, adultos y maduros de una comunidad de peces y adoptando por tanto la variable un rol de atributo.

En el extremo opuesto, hay investigaciones médicas que relacionan el tipo de patología con el sexo del paciente y con el desenlace de la enfermedad, caracteres todos ellos intrínsecamente atributos.

Las variables pueden clasificarse según su conjunto soporte. El soporte de una variable es el conjunto de todos los posibles valores que toma. Cuando el conjunto soporte es finito o numerable se habla de variable discreta. Por el contrario, cuando el conjunto soporte es no numerable, se habla de variable continua. Si la variable continua no toma valores en puntos aislados se dice absolutamente continua. Esta diferencia tendrá relevancia cuando se planteen, más adelante, estructuras de probabilidad para modelizar la población bajo estudio.

Ejemplo 2.3

El número de lunares en la piel de pacientes aquejados de una cierta patología, el número de hijos de las familias de una comunidad o el número de meteoritos que surcan una cierta región estelar en periodos de tiempo determinados son variables discretas. La distancia por carretera entre las capitales de provincia peninsulares españolas, el tiempo de reacción de los corredores de una carrera de 100 metros o las longitudes de los cabellos de una persona son variables continuas.

Una vez identificadas, recolectadas y organizadas, las variables serán tratadas estadísticamente combinando un análisis numérico, a través de una serie de medidas estadísticas, con representaciones gráficas. El software estadístico **R** ofrece una amplia gama de ambos elementos: numéricos y gráficos, aunque conviene ser selectivos y tomar aquellos

Figura 2.2: Ventana de selección de datos en paquetes adjuntos

que verdaderamente aportan información relevante. A tal efecto, se proponen las siguientes opciones:

Escala de Medida	**Medidas centrales**	**Medidas de dispersión**	**Representaciones gráficas**
Atributo	Moda Porcentajes		Diagrama de sectores
Ordenación	Mediana Percentiles	Recorrido Intercuartílico	Diagrama de barras
Recuento	Media	Desviación típica	Diagramas de barras
Intervalo	Media	Desviación típica	Histograma
Razón	Media geométrica	Coeficiente de variación	Histograma Diagrama de dispersión Diagrama de cajas

Tabla 2.1: Medidas y gráficos según tipo de variable

En última instancia corresponde al investigador el tomar las decisiones correctas en cada momento, de forma que sin transgredir los principios básicos, den como resultado un análisis eficiente de los datos.

3. Análisis de atributos

Los atributos son susceptibles de ser tratados de forma individual o en grupo, para obtener los porcentajes de cada subgrupo en el colectivo global. De hecho, cada carácter o conjunto de ellos establece una partición o catálogo de la población bajo estudio. Por otra parte, el

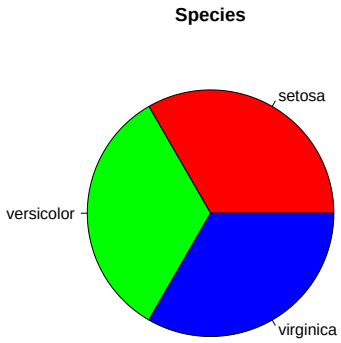


Figura 2.3: Diagrama de sectores del fichero `iris`

tratamiento gráfico más usual que se le daría a un atributo individual sería a través de un *diagrama de sectores* o *diagrama de tarta*.

Ejemplo 2.4

Se consideran ahora los datos del ejemplo `iris` *del paquete* `datasets` *de* **R** *que se describe en el apéndice A. Se carga el fichero en* **Rcmdr** *mediante la selección de las opciones del menú* `Datos`→ `Datos en paquetes`→`Leer datos desde paquete adjunto...`, *en el cuadro de diálogo se elige el paquete* `datasets` *y dentro de éste el juego de datos* `iris`,*figura 2.2. Del conjunto de variables de la matriz se considera la denominada* `Species`, *que es un atributo con los tres tipos de flores de Iris:* `Setosa, Virginica y Versicolor`.

Análisis numérico: *Se selecciona* `Estadísticos`→`Resúmenes`→ `Distribuciones de frecuencias...` *y en el cuadro de diálogo se elige el único atributo,* `Species`. *Se observa que los 150 individuos se reparten a partes iguales entre las tres variedades de flores,* 50 *para cada una, y que por tanto los* porcentajes *son iguales a* $33,33$. *No tiene sentido hablar de moda, puesto que las tres clases lo son.*

```
> .Table <- table(iris$Species)
> .Table # counts for Species
setosa     versicolor     virginica
50            50            50
> 100*.Table/sum(.Table) # percentages for Species
setosa     versicolor     virginica
33.33333    33.33333      33.33333
```

Análisis gráfico: *A continuación se selecciona el diagrama de sectores mediante* `Gráficas→Gráfica de sectores...`

Si el fichero de datos activo tiene más de una variable de clase se permite seleccionar la que se quiera. En este caso, la única variable elegible es `Species`*, que el programa da por defecto. Si se pulsa el botón* `Aceptar` *el programa dibuja el gráfico de sectores que se muestra en la figura 2.3. Como era de esperar, la tarta se divide en tres trozos exactamente iguales.*

4. Análisis de variables ordenadas

Las diferencias que se establecen entre variables de clase pura y ordenada se concretan desde el punto de vista del análisis numérico en que el grupo de medidas recomendables son las de posición, es decir los cuantiles en sus distintas versiones. Como medidas de representación, pensando que en general se dispondrá de pocas clases, se recurrirá a los cuartiles y como medida de dispersión al recorrido intercuartílico. En cuanto al análisis gráfico, se recomienda el uso del diagrama de barras.

Este tipo de variables ordenadas suele venir dada en forma de tabla de frecuencias. Por ello, en el ejemplo que ilustra el tratamiento de este tipo de variables, se comenzará explicando como transformar una tabla de frecuencias en una matriz de datos, al objeto de que puedan ser tratadas por **R** como un `data.frame`.

Ejemplo 2.5

Un caso de variable ordenada es la correspondiente a un estudio estadístico sobre el nivel académico de la población gaditana en el año 2001 (Fuente: Instituto Estadístico de Andalucía).

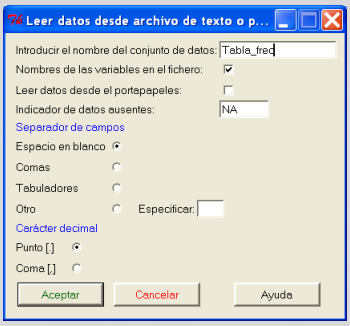
Los valores que toma la variable son: `Sin estudios`*,* `Elementales` *(primaria),* `Medios` *(secundaria, bachillerato y fp grado medio) y* `Superiores` *(fp superior, diplomatura, licenciatura y doctorado).*

Los datos se recogen en la tabla:

	NIVEL DE ESTUDIOS			
SEXO	*Sin estudios*	*Elementales*	*Medios*	*Superiores*
Hombre	*79309*	*107156*	*183488*	*70594*
Mujer	*108051*	*109591*	*174961*	*64858*

Debido al gran número de individuos que forman esta muestra puede ser útil almacenar la variable estudiada a partir de su tabla de frecuencias, transformándola en base de datos en el momento de realizar los análisis. El fichero en cuestión se ha guardado bajo el nombre de `tabla_freq_niv_estudios.dat`*, conteniendo tres variables:* `sexo`*,* `nivel` *y* `frec`*. En total consta de 8 filas que se correponden con los cruces de las clases sexo y nivel.*

Para cargar en **Rcmdr** *la tabla de frecuencias se selecciona* `Datos`→ `Importar datos desde archivo de texto o portapapeles...`*, en este ejemplo se ha elegido el nombre* `Tabla_frec` *para denominar al fichero que contendrá los datos de la tabla de frecuencias, como se muestra en la ventana de diálogo. A continuación se elige el archivo* `tabla_freq_niv_estudios.dat`*.*



Ahora se tendrá que transformar esta tabla de frecuencias en un conjunto de datos, `data.frame`*, con el que* **R** *pueda trabajar. Para conseguir esto se procede de la siguiente manera:*

```
>nivel<-rep(Tabla_frec$nivel,Tabla_frec$frec)
>sexo<-rep(Tabla_frec$sexo,Tabla_frec$frec)
>niv_estudios_cadiz< −data.frame(nivel,sexo)
```

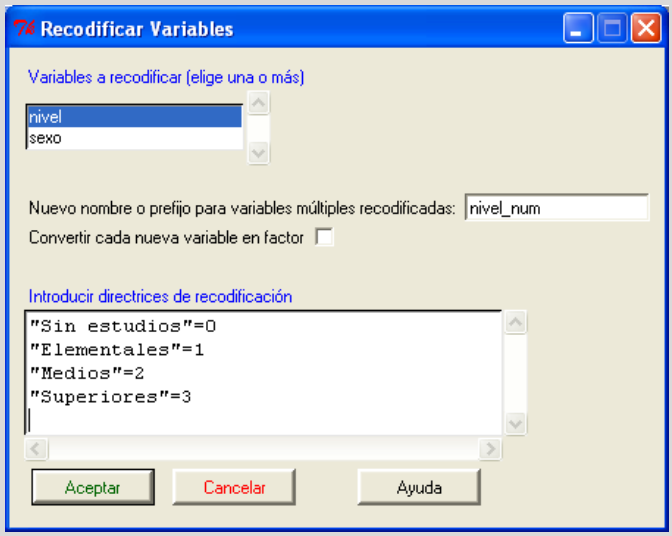
Es decir, se crean las variables `nivel` *y* `sexo` *a partir de la repetición de cada una de las clases de las respectivas variables, tantas veces como indique su frecuencia. A partir de ahí, se construye el* `data.frame` `niv_estudios_cadiz` *con las dos variables creadas.*

Este `data.frame` *se encuentra entre los datos que se facilitan en este libro y se puede cargar directamente sin realizar las operaciones anteriores. Para ello, basta con seleccionar* `Datos`→`Importar datos`→ `desde archivo de texto o portapapeles...`*, eligiendo ahora el ar-*

chivo `niv_estudios_cadiz.dat`.

Análisis numérico: *En variables de tipo ordenado es aconsejable utilizar, como medida de posición, los cuartiles.*

Para realizar este análisis la variable `nivel` *debe ser codificada numéricamente. Se creará una nueva variable en la base de datos, que se llamará* `nivel_num` *y que representará los valores numéricos de la variable* `nivel`. *Los valores* `Sin estudios`, `Elementales`, `Medios` *y* `Superiores` *han sido codificados mediante los valores 0, 1, 2 y 3, respectivamente. En* **Rcmdr** *esto se realizará seleccionando* `Datos`→`Modificar variables de los datos activos`→`Recodificar variables...` , *desmarcando la pestaña* `Convertir cada nueva variable en factor`.



Para realizar el análisis numérico de la variable `nivel_num` *se selecciona:* `Estadísticos`→`Resúmenes`→`Resúmenes numéricos...`, *eligiendo en la ventana emergente la variable* `nivel_num` *y marcando la opción de* **cuantiles**. *Se puede observar entre los cuartiles que la mediana recae sobre el valor 2.*

```
> numSummary(Niv_estudios[,''niv_num''],
statistics=c(''quantiles''))
0%   25%   50%   75%   100%
0     1     2     2     3
```

Desde **Rcmdr** *existe otra forma de realizar el análisis numérico de una variable ordenada.*

Para ello, se reordenan los niveles de la variable factor usando las opciones del menú `Datos→Modificar variables del conjunto de datos activo→Reordenar niveles de factor...`, *almacenando la variable* `nivel` *como factor de tipo ordenado. A la nueva variable se le ha llamado* `nivel_ord`*. A continuación se almacena ésta como variable de tipo numérico, escribiéndo en la ventana de instrucciones:*

```
Datos$nivel_num<-as.numeric(Datos$nivel_ord)
```

siendo ya posible calcular los cuantiles, para la variable numérica `Datos$nivel_num`*.*

Como medida de dispersión se ha recomendado el recorrido intercuartílico relativo, definido como el cociente entre la diferencia de los cuartiles tercero y primero, y la mediana. **Rcmdr** *no proporciona directamente este estadístico, pero se puede implementar fácilmente en la ventana de instrucciones, mediante las órdenes siguientes:*

```
>Q1<-quantile(niv_estudios_cadiz$nivel_num, 0.25)
>Q2<-quantile(niv_estudios_cadiz$nivel_num, 0.5)
>Q3<-quantile(niv_estudios_cadiz$nivel_num, 0.75)
>RIR<-as.numeric((Q3-Q1)/Q2)
>RIR
[1] 0.5
```

Análisis gráfico: *Para realizar el análisis gráfico de la variable se utiliza el diagrama de barras. En* **Rcmdr** *se selecciona:* `Gráficas→Gráfica de barras...` *y se elige en la ventana de diálogo, la variable* `nivel_ord`*.*

En **R** *existe una gran variedad de opciones que ayudan a mejorar el aspecto de los gráficos. Se puede acceder a ellas escribiéndolas en la ventana de instrucciones. En este ejemplo se ha optado por modificar el*

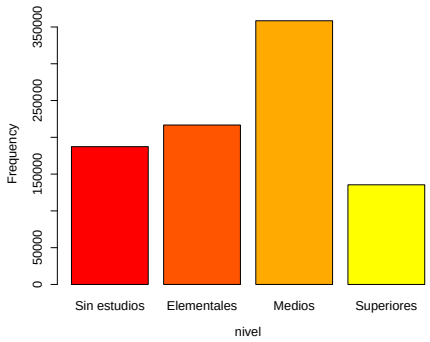


Figura 2.4: Diagrama de barras de la variable nivel de estudios

color, siguiendo una escala de colores cálidos. Esto se consigue agregando `col=heat.colors(5)` *a las opciones de* `barGraph` *(figura 2.4).*

5. Análisis de variables de escala

Ejemplo 2.6

Se estudiará ahora el tratamiento de una variable continua. Para ello se considera la base de datos `chickwts`*, del paquete* `datasets` *de* **R***. En ella se recogen los pesos finales, en gramos, de 71 polluelos, según el tipo de dieta seguida durante un periodo de 6 semanas.*

Análisis numérico: *Para la variable que da el peso de los polluelos las medidas básicas recomendadas son la media y la desviación típica. Estas medidas se calculan desde* `Estadísticos`→`Resúmenes`→ `Resúmenes numéricos...`*, seleccionando para la variable* `weight` *las opciones deseadas.*

```
> numSummary(chickwts[,''weight''], statistics=c(''mean'',
''sd''))
mean      sd       n
261.3099 78.0737 71
```

Aunque se está hablando de la desviación típica, la función `sd` *calcula en realidad la cuasidesviación típica. Cabe la posibilidad de que*

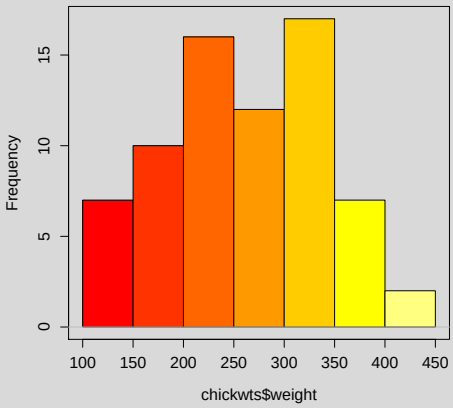
se necesiten otro tipo de medidas que completen el estudio, como la simetría, el apuntamiento, ... Para ello, en el apéndice B, se incluye una tabla de medidas estadísticas. Por ejemplo, si se deseara calcular la simetría y la curtosis de la variable `weight`*, habría en primer lugar que instalar y cargar en* ***R****, si no lo está ya, el paquete* `fBasics`*. Y a continuación:*

```
> kurtosis(chickwts$weight)
 -0.9651994
attr(,''method'')
''excess''
```

```
> skewness(chickwts$weight)
 -0.01136593
attr(,''method'')
''moment''
```

Ambos coeficientes están calculados a partir de los momentos y, en el caso de la curtosis, se le ha restado 3. Se podría concluir que la distribución es bastante simétrica y algo aplastada.

Análisis gráfico: *Para analizar gráficamente la variable peso se comienza con la realización del histograma que se muestra al margen mediante las instrucciones* `Gráficas→Histograma...` *En el histograma se observa un comportamiento bastante simétrico y la posibilidad de que existan dos modas.*



A continuación, se construye el diagrama de caja (figura 2.5). Se puede observar en el gráfico que la variable no posee valores atípicos, es simétrica y está relativamente dispersa.

El `data.frame` *que se está utilizando incluye un factor,* `Feed`*, que se corresponde con las diferentes dietas sumimistradas a los pollos. Ello permite la realización de un análisis por grupo, tanto numérico como gráfico, que permita evaluar las diferencias de peso en función del tipo de alimentación seguida. Los valores que toma la variable* `Feed` *son:*

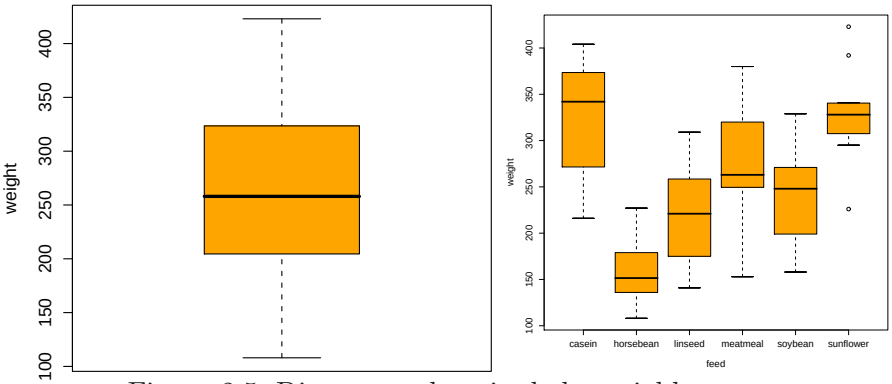


Figura 2.5: Diagramas de caja de la variable peso

`horsebean` (habas), `linseed` (linaza), `soybean` (soja), `sunflower` (girasoles), `meatmeal` (carne) y `casein` (caseína).

Es interesante la representación del diagrama de caja de la variable peso, según el tipo de alimentación (figura 2.5). Se observa que los valores de la variable peso están más concentrados para la dieta `sunflower`. También éste es el único grupo en el que se dan valores atípicos. Por contra la mayor dispersión de los datos se produce con la dieta `casein`. Una evaluación inicial, parece indicar que la dieta que produce pollos de mayor peso es `sunflower`, ya que los pesos que consigue están más concentrados en torno a uno de los valores más altos.

El análisis numérico ofrece los siguientes resultados:

```
> numSummary(chickwts[,''weight''], groups=chickwts$feed,
statistics=c(''mean''))


           mean       sd         n
casein     323.5833   64.43384   12
horsebeen  160.2000   38.62584   10
lindseed   218.7500   52.23570   12
meatmeal   276.9091   64.90062   11
soybean    246.4286   54.12907   14
sunflower  328.9167   48.83638   12
```

6. Ejercicios

2.1 Al comenzar el curso se pasó una encuesta a los alumnos del primer curso de un colegio, preguntándoles, entre otras cuestiones, por el número de hermanos que tenían. Se obtuvieron los siguientes resultados:

3, 3, 2, 2, 8, 5, 2, 4, 3, 1, 4, 5, 3, 3, 3, 3, 3, 2, 5
1, 3, 3, 2, 2, 4, 3, 3, 2, 2, 4, 4, 3, 6, 3, 3, 2, 2, 4
3, 4, 3, 2, 2, 4, 4, 3, 3, 4, 2, 5, 4, 1, 2, 8, 2 ,3, 3, 4

a) Represente este conjunto de datos con un diagrama de barras.

b) Calcule media, moda y mediana.

c) Estudie la dispersión de los datos.

d) Analice la simetría de la distribución.

2.2 Los pesos de un colectivo de niños son:

60, 56, 54, 48, 99, 65, 58, 55, 74, 52, 53, 58, 67, 62, 65
76, 85, 92, 66, 62, 73, 66, 59, 57, 54, 53, 58, 57, 55, 60
65, 65, 74, 55, 73, 97, 82, 80, 64, 70, 101, 72, 96, 73, 55
59, 67, 49, 90, 58, 63, 96, 100, 70, 53, 67, 60, 54

Obtenga:

a) La distribución de frecuencias agrupando por intervalos.

b) La mediana de la distribución.

c) La media de la distribución, indicando su nivel de representatividad.

d) Utilizando la agrupación en intervalos, el porcentaje de alumnos que tienen un peso menor de 65 kg y el número de alumnos con un peso mayor de 60 kg dentro del grupo de los que pesan menos de 80 kg.

2.3 En el Consejo de Apuestas del Estado se han ido anotando, durante una temporada, el número de premiados de quinielas según la cantidad de aciertos. Los resultados se recogen en la siguiente tabla:

Nº de aciertos	11	12	13	14	15
Nº de personas (miles)	52	820	572	215	41

Calcule:

a) La mediana, la moda y los cuartiles de la distribución.

b) La simetría de la distribución.

2.4 En un puerto se controla diariamente la entrada de pesqueros según su tonelaje, resultando para un cierto día los siguientes datos:

Peso(Tm.)	0-25	25-50	50-70	70-100	100-500
Nº de barcos	5	17	30	25	3

Se pide:

a) El peso medio de los barcos que entran en el puerto diariamente, indicando la representatividad de dicha medida.

b) El intervalo donde se encuentra el 60 % central de la distribución.

c) El grado de apuntamiento.

d) El tonelaje más frecuente en este puerto.