Universidad de Cantabria
Curso 2011–2012

Práctica 1: Estadística descriptiva

En esta práctica se utilizará R como una herramienta para explorar diferentes muestras de datos con el fin de descubrir regularidades y singularidades de los mismos. Para ello calcularemos ciertas medidas estadísticas vistas en clase y procederemos a representar algunos de los tipos gráficos analizados. El objetivo principal de esta práctica se centrará en el análisis y la interpretación de los resultados obtenidos en cada caso. Nos centraremos en el análisis de una única variable.

1. Análisis numéricos

Los diferentes tipos de análisis numéricos para una variable se albergan bajo la opción *Resúmenes* del menú *Estadísticos*. De las opciones que muestra R bajo este menú se analizarán en detalle las más relevantes en nuestro caso. Para ello cargaremos el fichero *Pulso.rda* disponible en la página web de la asignatura, con el que trabajaremos en este tema. Este fichero contiene registros de 92 personas de diferentes variables (visualizar el contenido).

1.1. Resumen rápido

Pulsando *Conjunto de datos activos*, en la ventana de resultados se muestra un breve análisis descriptivo para cada una de las variables contenidas en el fichero de datos. Cabe destacar que en el caso de variables cuantitativas se indican los valores correspondientes al máximo, mínimo, cuartiles y media mientras que para variables cualitativas, se da la frecuencia absoluta de las modalidades más frecuentes (y la de los valores ausentes, si hay alguno).

Si hay más de diez variables en el conjunto de datos, R pide confirmación, pues la abundante información puede resultar difícil de visualizar.

1.2. Resúmenes numéricos

En *Resúmenes numéricos* podemos ampliar la información anterior obteniendo los valores de la media, desviación típica y cuantiles arbitrarios para una variable cuantitativa.

Además, este menú permite resumir esta información por grupos, seleccionando en el botón *Resumir por grupos...* la variable cuantitativa que más nos interese clasificar. Por ejemplo, utilizando el fichero *Pulso.rda* obtener los valores de la media, desviación típica, primer decil, mediana y percentil 85 de la variable *Altura*, clasificando los resultados en *fumadores* y *no fumadores*.

La orden que aparece en la ventana de instrucciones sería la siguiente:

```
> numSummary(Pulsaciones[,"Altura"], groups=Pulsaciones$Fumar,
    statistics=c("mean", "sd", "quantiles"), quantiles=c(0.1,.5,.85))
```

El comando `attach`[1] permite utilizar vectores para ejecutar esa misma orden de la siguiente manera:

[1] Con el comando `attach`, las columnas de un data.frame pasan a ser variables de tipo vector. El comando `detach` deshace esta acción.

```
> attach(Pulsaciones)
> data<-Altura[Fumar=="fuma"]
> numSummary(data,statistics=c("mean", "sd", "quantiles"),quantiles=c(0.1,.5,.85))
```

Conviene resaltar que R utiliza la cuasivarianza, es decir, cuando se le pide que calcule la varianza y la desviación típica, lo que da exactamente es el resultado de las fórmulas:

$$S^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \qquad S = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$$

que corresponden a la cuasi-varianza y la cuasi-desviación respectivamente.

Para calcular otros estadísticos descriptivos debemos escribir la orden correspondiente en la ventana de instrucciones. Vamos a indicar las funciones que calculan los estadísticos más importantes:

Media: Utilice la opción del menú vista anteriormente o bien la orden `mean`:

```
mean(Altura, na.rm=TRUE)
```

La parte "`, na.rm=TRUE`" es opcional y se utiliza para indicar que hay datos ausentes que no deben ser considerados al calcular ese estadístico. Calcular la media del siguiente vector (2,2,2,NA) considerando la opción `na.rm=TRUE` y sin considerarla.

Mediana: Utilice un cuantil de orden 0,5, como se vio arriba, o bien la orden `median`:

```
median(Altura)
```

Amplitud: Aquí es necesario combinar dos órdenes:

```
diff(range(Altura))
```

Si la va a utilizar varias veces, es mejor definir una función `amplitud`:

```
amplitud <- function(x){diff(range(x))}
amplitud(Altura)
```

Recorrido intercuartílico: Utilice la orden `IQR`:

```
IQR(Altura)
```

Coeficiente de variación: Definamos la función `CV`, bien en la forma más simple,

```
CV <- function(x) {sd(x)/mean(x)}
```

o, para obtener un valor numérico incluso en datos con valores ausentes,

```
CV <- function(x) {sd(x, na.rm=TRUE) / mean(x, na.rm=TRUE)}
```

Simetría: Utilice `skewness` así[2]:

[2]Puede ser necesario que antes cargue el paquete `fBasics`, eligiéndolo en el menu *Herramientas / Cargar paquete(s)*.

```
skewness(Altura)
```

Curtosis: La orden correspondiente es `kurtosis`:

```
kurtosis(Altura)
```

Practica tú mismo

---

1) Calcula los estadísticos anteriores para la variable *Peso* y contesta a las siguientes preguntas:

1. Unidades de la media
2. Valor del rango interquartílico
3. Valor de la varianza
4. ¿Es la distribución simétrica?
5. ¿Cuántos hombres hay en la muestra?
6. ¿Es la distribución platicúrtica, mesocúrtica o leptocúrtica?
7. ¿Cuál de las dos variables, Peso o Altura, presenta mayor dispersión?
8. Calcula la kurtosis de la variable Peso de las personas que no fuman
9. Calcula la cuasi-desviación típica de la Altura de las mujeres

---

Practica tú mismo

---

2) A lo largo de un año, los importes de las facturas mensuales de vuestro móvil han sido:

$$23, 33, 25, 45, 10, 28, 39, 27, 15, 38, 34, 29$$

1. ¿Cuanto habéis gastado en total en el año?
2. ¿Cuál ha sido el gasto mínimo?,
3. ¿y el máximo?
4. ¿Qué meses han supuesto un gasto menor que el gasto medio?
5. ¿ Existe mucha dispersión de unos meses a otros? ¿Cómo estimas esa dispersión?

1.3. Distribuciones de frecuencias

Para las variables cualitativas, puede confeccionarse con el procedimiento *Estadísticas / Resúmenes / Distribución de frecuencias* una tabla donde aparezcan los valores de la variable, sus frecuencias absolutas y las frecuencias relativas en forma de porcentajes.

Practica tú mismo

---

3) Utilizando los datos del fichero Pulso.rda, responde a las siguientes preguntas:

1. Frecuencia absoluta de las mujeres
2. Frecuencias relativas de la variable Actividad
3. Frecuencias relativas acumuladas de la variable Correr

La función `table` también puede aplicarse a variables cuantitativas. Esta orden nos muestra los valores de la frecuencia absoluta para cada una de las clases. Por ejemplo la orden:

```
> n.i <- table(Pulse1)
```

genera la tabla de frecuencias de la variable *Pulse1*. Si los datos son continuos o hay muchos valores diferentes, esta tabla no resume los datos adecuadamente y habrá que agruparlos por clases. Para esto podemos utilizar la funcion `cut`, que "corta" los datos en clases.

```
> n.i <- table( cut(Pulse1, seq(45,100,5) ))
```

A partir de esos valores se puede calcular el resto de frecuencias para completar la tabla de frecuencias utilizando los siguientes comandos:

```
> N.i <- cumsum(n.i)
> f.i <- n.i / sum(n.i)
> F.i <- cumsum(f.i)
```

Practica tú mismo

4) Utilizando los datos del fichero Pulso.rda, calcula los siguientes valores:

1. Con la variable *Altura*, calcular el número de personas que miden menos de 183 cm.
2. Con la variable *Altura*, calcular el porcentaje de personas que miden menos de 183 cm.
3. Calcular el porcentaje de mujeres que son fumadoras.
4. Determinar la frecuencia relativa de las personas que miden menos de 183 cm.
5. Determinar la frecuencia relativa de las personas que son mujeres.
6. Determinar la frecuencia relativa de las personas que son varones y su práctica deportiva habitual es *suave*.
7. Determinar la frecuencia relativa de las personas que son varones y fuman.
8. Elegida una persona al azar resulta ser mujer. Determinar la frecuencia relativa de las que miden menos de 170 cm.
9. Determinar la frecuencia relativa de las personas cuyo pulso es superior de 84.
10. Determinar la frecuencia relativa de las personas cuyo pulso es superior de 84 y cuya práctica deportiva habitual es *alta*.
11. Determinar la frecuencia relativa de las personas cuyo su pulso es superior de 84 y cuya práctica deportiva habitual no sea *mediana*.
12. Determinar la frecuencia relativa de las personas tales que su pulso es superior de 84, fuman y son mujer.
13. De entre los personas que fuman, determinar la frecuencia relativa de los que registran el Pulso1 superior a 84.

2. Representaciones gráficas unidimensionales

Las representaciones gráficas permiten captar rápidamente y sin gran esfuerzo las principales características de una distribución de frecuencias. Son un medio complementario, aunque muy importante, para realizar un análisis estadístico de los datos. Como se indicará en cada caso, existen representaciones gráficas propias de variables cualitativas (gráficos de sectores) o cuantitativas y dentro de estas últimas para variables discretas (diagrama de barras) y continuas (histograma).

Están recogidas bajo el menú *Gráficas*. Describimos sólo las opciones de interés en nuestro curso.

Si la orden ejecutada proporciona una salida gráfica, R abre una nueva ventana (*device*) que contiene el gráfico. Éste puede ser grabado en un fichero mediante la opción *Guardar gráfica* del menú *Gráficas*.

2.1. Histograma

Para representar la distribución de una variable cuantitativa continua, se puede recurrir a la opción *Histograma*. Esta opción del menú ejecuta la función `Hist` que, como veremos más adelante, no comparte exactamente los mismos argumentos que la función `hist` que también permite representar histogramas. Para hacer esta representación a través del menú, es posible introducir el número aproximado de barras o dejar la elección a un algoritmo automático. Como se observa en la ventana emergente, se puede representar la frecuencia absoluta (*Recuentos de frecuencias*), la frecuencia relativa dada en porcentaje (*Porcentajes*) y la densidad (*Densidades*) que hace que el área total del histograma sea 1. Es muy recomendable escribir un título en el gráfico así como también las etiquetas en cada uno de los ejes de manera que la figura muestre cuales son las variables representadas. La funciones de representación gráfica aceptan los atributos `main` , `xlab` e `ylab`. La orden que se genera con este menú para representar el histograma de la variable *Altura* es la siguiente:

```
Hist(Altura, scale="frequency", breaks="Sturges", col="darkgray",
     main="Histograma de frecuencias", xlab="Altura", ylab="ni")
```

Como se observa el método que usa por defecto para calcular el número de clases es el método de Sturges visto en clase. La figura 1 muestra el histograma correspondiente.

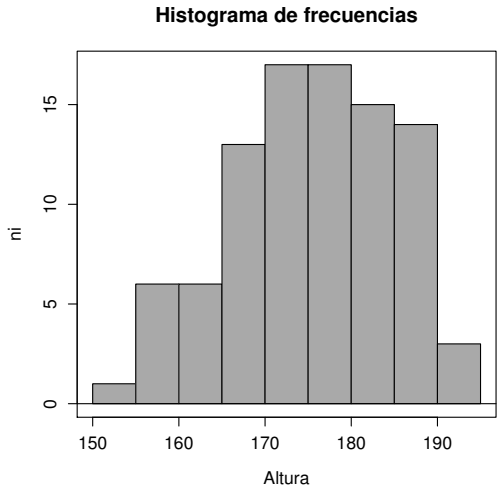


Figura 1: Histograma de frecuencias de la variable *Altura* contenida en el fichero *Pulso.rda*

Como se ha mencionado antes existe otra función en R para representar el histograma, se trata de la función `hist`. Resulta interesante consultar la ayuda de ambas funciones para ver que no todos los parámetros de entrada son los mismos. Por ejemplo, la función `hist` de R (`?hist`), permite introducir argumentos que controlan diferentes aspectos del histograma como por ejemplo la amplitud del intervalo (`breaks`), si el intervalo de clase está cerrado por la derecha o por la izquierda (`right`) o el color de las barras (`col`). Además tiene un argumento muy útil que no tiene la función `Hist`, `plot=FALSE`, que no pinta el histograma sino que saca por pantalla la lista de valores calculados. Esto nos permite poder utilizar estos valores para hacer otros cálculos. Un ejemplo de la función `hist` aplicada a la variable Altura es el siguiente:

```
> tab<-hist(Altura, breaks=seq(150,200,10), plot=F)
> tab
```

Como se observa en la figura 2, la ventana de resultados muestra todos los valores resultantes de esta función: límites de las clases, frecuencia ablsoluta, densidad, marcas de clase... Además al haber guardado el resultado en una variable, se pueden hacer otros cálculos con esos valores, como calcular la frecuencia

relativa:

```
> tab$f_relativa <- tab$counts/length(Altura)
```

En este caso la frecuencia relativa se ha incluido dentro de la propia variable *tab*.

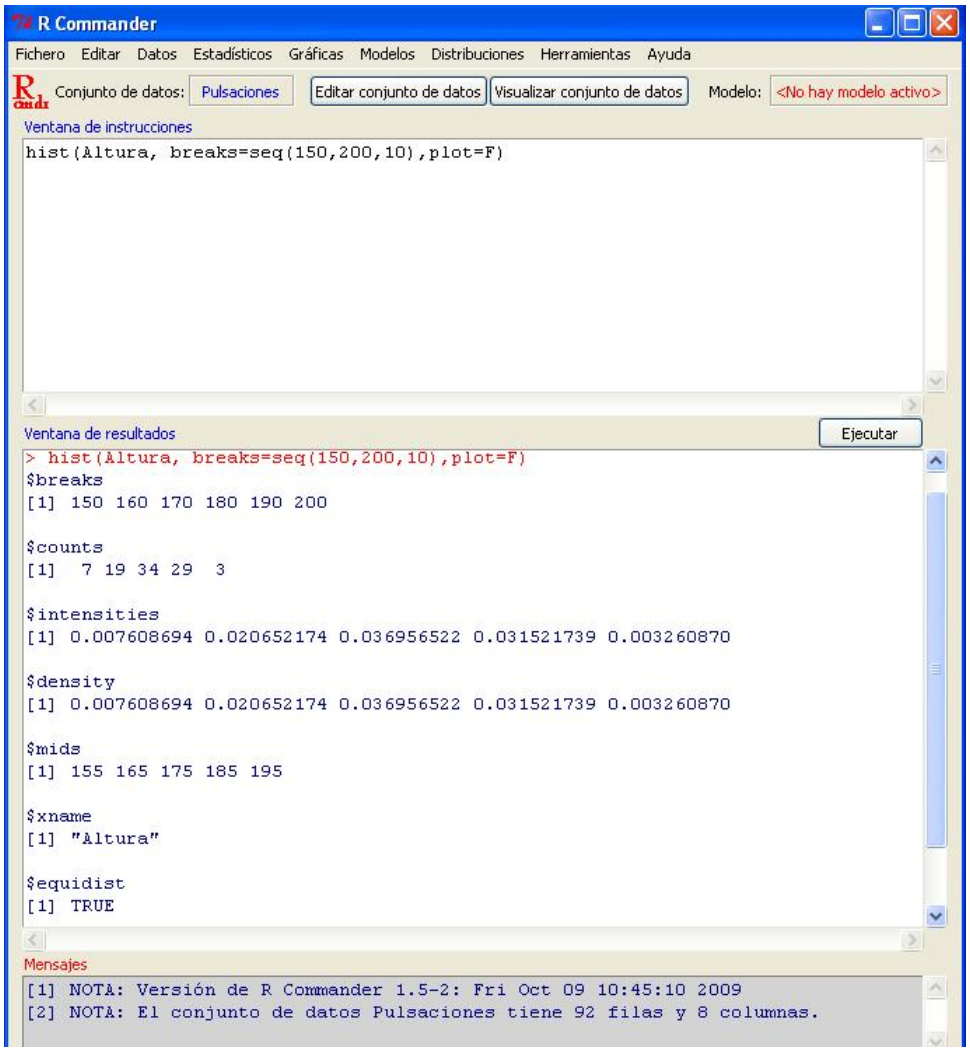


Figura 2: Resultado de la instrucción `hist` con el argumento `plot=FALSE`

Practica tú mismo

---

5) Utilizando los datos del fichero Pulso.rda, representa el histograma de la variable *Peso*. ¿Qué observas? Dibuja los histogramas de los hombres (en rojo) y de la mujeres (en azul).

2.2. Diagrama de cajas

Los diagramas de cajas son gráficos de variables cuantitativas y en R se obtienen con la opción *Diagrama de caja* del menú *Gráficas*. Si representamos el diagrama de cajas de la variable *Altura* la instrucción que obtenemos es la siguiente:

```
> boxplot(Altura, ylab="Altura (cm)", main="Box plot")
```

a la que le hemos añadido el título de la figura. Además esta instrucción permite que se pueda representar el diagrama de cajas de una determinada variable clasificada por grupos. Por ejemplo, podemos hacer el box-plot de la variable *Altura* separándola según el *Sexo*, como se muestra en la figura 3. Esta opción se puede elegir en el menú *Diagrama de caja* o bien mediante la instrucción:

```
> boxplot(Altura~Sexo, xlab="Sexo", ylab="Altura (cm)", main="Box plot")
```

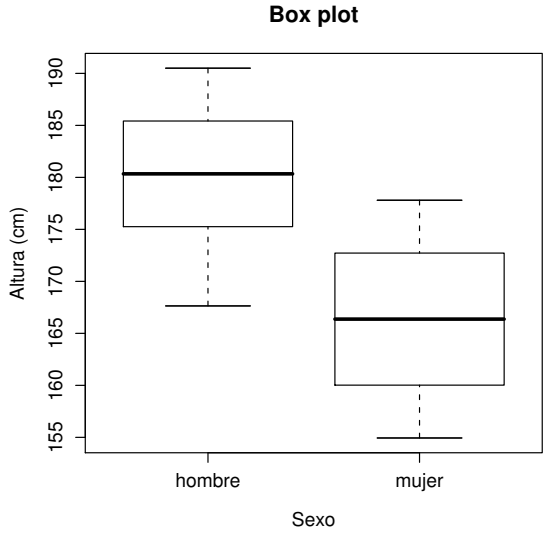


Figura 3: Diagrama de cajas de la variable *Altura* contenida en el fichero *Pulso.rda* clasificada por *Sexo*.

En algunos casos resulta interesante representar dos figuras en una única ventana para facilitar la comparación. Por ejemplo, puede interesarnos comparar el diagrama de cajas de la variable *Altura* por *Sexo* y también considerar si fuma. En este caso podemos ejecutar los siguientes comandos:

```
> split.screen(c(1,2))
> screen(1)
> boxplot(Altura~Sexo, xlab="Sexo", ylab="Altura (cm)")
> screen(2)
> boxplot(Altura~Fumar, xlab="Fumar", ylab="Altura (cm)")
```

como se observa en la figura 4, la pantalla de gráficos se ha dividido en una matriz con una fila y dos columnas como indica la orden `split.screen`.

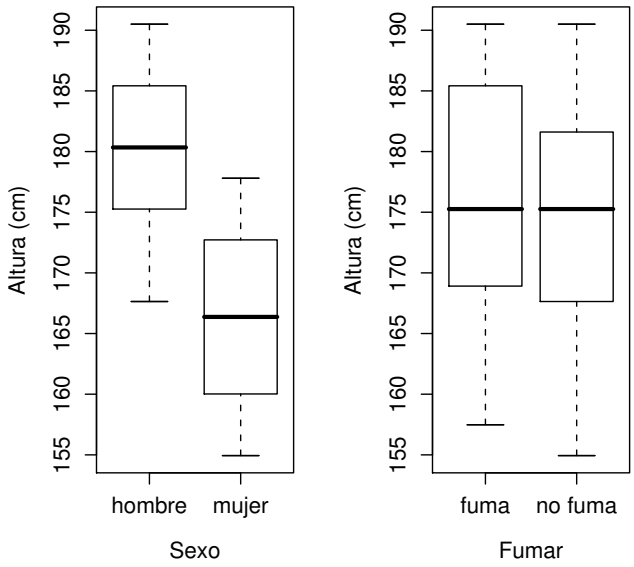


Figura 4: Ejemplo de inclusión de dos figuras en la misma ventana.

Practica tú mismo

6) Utilizando los datos del fichero Pulso.rda:

1. Representa el diagrama de cajas de la variable *Pulso2*. Observando el gráfico obtener aproximadamente el valor del pulso mínimo, del máximo, el valor del primer cuartil, de la mediana y del tercer cuartil.

2. Calcula una variable que sea la diferencia de *Pulse1* y *Pulse2* y representa un diagrama de cajas de esa nueva variable frente a la variable *Correr*. ¿Podrías sacar alguna conclusión del gráfico que obtienes?

2.3. Gráfico de barras y diagrama acumulativo

La opción *Gráfica de barras* permite representar diagrama de barras de variables cualitativas. Si elegimos hacer el diagrama de barras de la variable *Fumar* la orden que resulta es la siguiente:

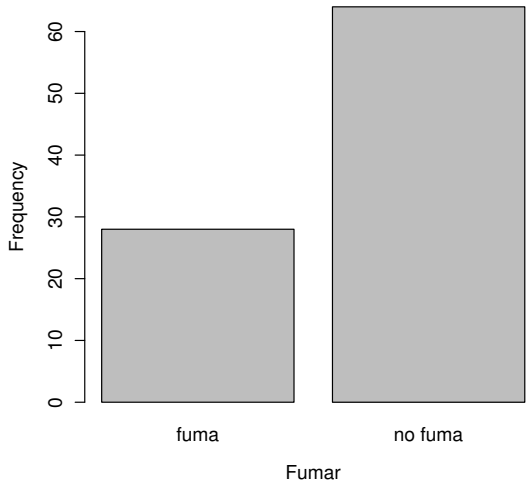
```
> barplot(table(Fumar), xlab="Fumar", ylab="Frequency")
```

que representa el siguiente gráfico.

Sin utilizar el menú, podemos usar esta misma función para realizar el diagrama de barras de variables cuantitativas discretas como es la variable *Hijos*.

```
barplot(table(Hijos), xlab="Numero de hijos", ylab="Frequency")
```

Cuando la variable observada es cuantitativa discreta tiene sentido representar el gráfico de frecuencias acumuladas (diagrama acumulativo) que tiene el aspecto de escalera con altura de cada peldaño proporcionales a la frecuencia asociada a cada valor de la variable.

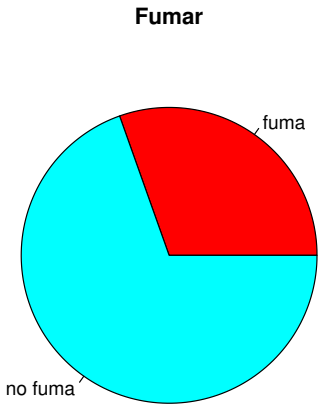



```
xval<-seq(min(Hijos),max(Hijos))
n.i<-table(Hijos)
F.i<-cumsum(n.i/length(Hijos))
plot(xval,F.i,type="s",lwd=4,xlab="Numero de Hijos",ylab="Frec. ABs. Acumuladad")
```

2.4. Gráfico de sectores

Otra opción que podemos considerar para representar variables cualitativas o cuantitativas discrestas es el gráfico de sectores que podemos ejecutar bajo la opción *Gráfica de sectores* o bien mediante la orden:

```
> pie(table(Fumar), labels=levels(Fumar), main="Fumar",
      col=rainbow(length(levels(Fumar))))
```



Practica tú mismo

7) Representa el gráfico de barras y el de sectores de la variable *Actividad*. Observa los gráficos obtenidos y responde a las siguientes preguntas:

1. ¿cuál de los dos gráficos representa mejor el la cantidad de personas que realizan cada tipo de actividad?
2. ¿cuál de los dos gráficos representa mejor el porcentaje de personas que realizan cada tipo de actividad?

Practica tú mismo

8) El fichero de datos santander.dat contiene los datos de temperatura media mensual medida en Santander desde 1950 hasta 2003. Se pide:

1. Representar las temperaturas medias registradas en febrero durante el periodo 1950-1960. Indicar en qué año se registró la temperatura más baja.
2. Representar las temperaturas medias de julio y agosto entre los años 1970 y 1980 y encontrar gráficamente en qué años ésta fue mayor en julio.
3. Calcule la mediana, media y desviación estándar de las temperaturas medias mensuales en enero y julio entre los años 1950 y 1980.
4. Determine los coeficientes de asimetría y curtosis de las temperaturas medias mensuales de junio, julio y agosto entre los años 1950 y 1980. ¿En que caso la serie presenta mayor asimetría?