

Práctica 2. Estadística Descriptiva

Estadística

Grado en Ingeniería Informática

Índice

1. Tablas de frecuencias	2
2. Representaciones gráficas	5
3. Medidas características	9
4. Tabla resumen de procedimientos y funciones	12

1. Tablas de frecuencias

Una tabla de frecuencias está compuesta por varias columnas: las frecuencias absolutas, las frecuencias relativas y sus correspondientes acumuladas, cuando tengan sentido. Comenzaremos por obtener las frecuencias absolutas, para lo que se utiliza la función `table`. En el caso de las frecuencias relativas es tan sencillo como dividir las frecuencias absolutas previamente calculadas entre el tamaño muestral, que se puede obtener fácilmente con la función `length`.

Consideremos una muestra de 20 observaciones de la variable $X :=$ “Satisfacción con el servicio de atención al cliente” que introducimos en R del siguiente modo:

```
> satisfaccion<-c("buena", "buena", "regular", "muy buena", "mala",
+                "buena", "buena", "regular", "muy mala", "buena",
+                "muy buena", "muy buena", "regular", "buena", "mala",
+                "buena", "muy buena", "muy buena", "buena", "regular")
```

Esta variable así definida es almacenada en R como una variable de tipo carácter. Para convertirla en tipo factor (que es el tipo de variables que R emplea para las variables cualitativas) y que permite comprender que existen unas determinadas categorías que se van repitiendo en nuestra muestra haremos:

```
> satisfaccion<-as.factor(satisfaccion)
> levels(satisfaccion)

[1] "buena"      "mala"       "muy buena"  "muy mala"   "regular"
```

Por defecto, y tal y como se observa en la salida, R ordena las categorías alfabéticamente. Si queremos que aparezcan en un orden “lógico” debemos de permutarlas según ese criterio:

```
> satisfaccion<-factor(satisfaccion, levels=levels(satisfaccion)[c(4,2,5,1,3)])
> levels(satisfaccion)

[1] "muy mala"    "mala"        "regular"      "buena"       "muy buena"
```

En un primer lugar obtendremos las frecuencias absolutas y relativas:

```
> Fabs<-table(satisfaccion)
> Fabs

satisfaccion
muy mala      mala    regular    buena    muy buena
           1         2         4         8         5
```

```
> Frel<-table(satisfaccion)/length(satisfaccion)
> Frel
```

satisfaccion	muy mala	mala	regular	buena	muy buena
	0.05	0.10	0.20	0.40	0.25

Para obtener las **frecuencias acumuladas** (tanto absolutas como relativas) que en este caso tiene sentido calcular por tratarse de una variable cualitativa ordinal, se utiliza la función `cumsum`:

```
> FabsAcum<-cumsum(Fabs)
> FabsAcum
```

	muy mala	mala	regular	buena	muy buena
	1	3	7	15	20

```
> FrelAcum<-cumsum(Frel)
> FrelAcum
```

	muy mala	mala	regular	buena	muy buena
	0.05	0.15	0.35	0.75	1.00

Para ver la tabla de frecuencias completa, podemos “pegar por columnas” los vectores calculados previamente con la función `cbind`:

```
> Tabla<-cbind(Fabs,Frel,FabsAcum,FrelAcum)
> Tabla
```

	Fabs	Frel	FabsAcum	FrelAcum
muy mala	1	0.05	1	0.05
mala	2	0.10	3	0.15
regular	4	0.20	7	0.35
buena	8	0.40	15	0.75
muy buena	5	0.25	20	1.00

Veamos ahora un **ejemplo con datos reales**. En el archivo `quine`¹ de la librería MASS se recogen datos sobre el absentismo escolar de 146 escolares en Nueva Gales (Australia). Entre las variables recogidas están el sexo (`Sex`) y la edad (`Age`) medida en grupos. Para construir las correspondientes tablas de frecuencias emplearemos el siguiente código (hay que tener en cuenta que la variable `Age` está categorizada y es cualitativa ordinal, mientras que `Sex` es cualitativa nominal).

¹La función `attach` permite acceder a las variables de la base de datos directamente por su nombre, es decir, que cuando no usamos `attach` para acceder a los datos de la variable `Age` tendríamos que emplear el código `quine$Age`, mientras que una vez ejecutada la función `attach` podemos acceder directamente con `Age`.

```
> library(MASS)
> attach(quine)
> ni<-table(Age)
> fi<-ni/length(Age)
> Ni<-cumsum(ni)
> Fi<-cumsum(fi)
> cbind(ni,fi,Ni,Fi)

      ni      fi  Ni      Fi
F0 27 0.1849315  27 0.1849315
F1 46 0.3150685  73 0.5000000
F2 40 0.2739726 113 0.7739726
F3 33 0.2260274 146 1.0000000

> ni<-table(Sex)
> fi<-ni/length(Sex)
> cbind(ni,fi)

      ni      fi
F 80 0.5479452
M 66 0.4520548
```

La función `table` también permite construir **tablas de contingencia** en las cuales se representan datos de dos (o más) variables. Para el ejemplo de los datos de absentismo escolar, podemos representar, de manera conjunta, a los individuos según sexo y edad, así como obtener las distribuciones marginales con la función `margin.table` y completar la tabla con la función `addmargins` de la siguiente forma:

```
> Tabla<-table(Age,Sex)
> margin.table(Tabla,1)
> margin.table(Tabla,2)

> addmargins(Tabla)

      Sex
Age   F   M Sum
F0   10  17  27
F1   32  14  46
```

```
F2  19  21  40
F3  19  14  33
Sum 80  66 146
```

Nota 1. El cálculo de las distribuciones marginales también puede programarse a mano (en lugar de usar las funciones `margin.table` y `addmargins`) usando las funciones `colSums`, `cbind`, `rowSums` y `rbind`.

2. Representaciones gráficas

Las representaciones gráficas más simples para variables cualitativas o cuantitativas discretas son el **diagrama de barras** y el **diagrama de sectores**, como vimos en el Tema 1. Sobre los datos del ejemplo inicial sobre la variable $X :=$ “Satisfacción con el servicio de atención al cliente”, podríamos construir un diagrama de barras ejecutando:

```
> barplot(table(satisfaccion),ylab="Proporción",col=4)
```

y para dibujar el diagrama de sectores haríamos:

```
> pie(table(satisfaccion),main="Satisfacción",
+      col=c("red","blue","green","orange","purple"))
```

De este modo obtendríamos las representaciones mostradas en la Figura 1. Nótese que cualquiera de las representaciones anteriores se podrían obtener para una tabla de frecuencias relativas simplemente dividiendo entre el tamaño muestral, es decir, `length(satisfaccion)`.

En el caso del **ejemplo con datos reales**, representaremos tanto Age como Sex mediante diagramas de barras, en la primera lo haremos con las frecuencias absolutas y en la segunda con las relativas:

```
> barplot(table(Age),ylab="Frecuencias absolutas",col=3)
> barplot(table(Sex)/length(Sex),ylab="Frecuencias relativas",col=3)
```

Para variables cuantitativas continuas consideremos el ejemplo de los datos del geyser Old Faithful, disponible dentro de la librería MASS:

```
> library(MASS)
> attach(geyser)
> names(geyser)
```

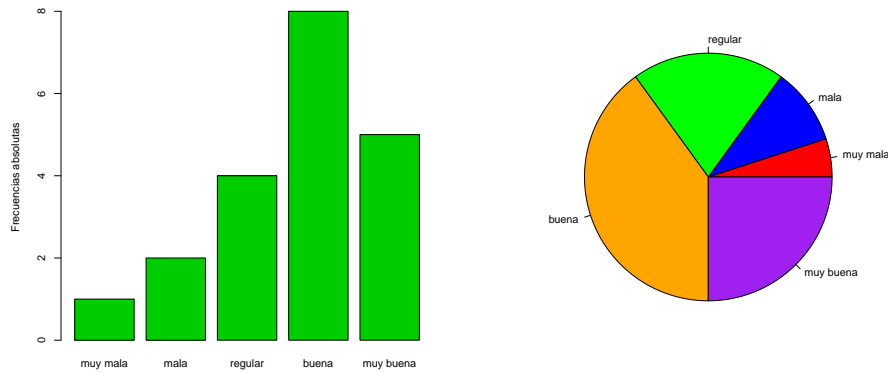


Figura 1: Diagramas de barras y sectores.

```
[1] "waiting" "duration"
```

```
> length(duration)
```

```
[1] 299
```

Si dibujamos un diagrama de barras o un diagrama de sectores para los datos de duración de erupciones:

```
> barplot(duration)
```

```
> pie(duration)
```

se pone de manifiesto que ninguno de ellos resulta informativo cuando se aplican sobre variables cuantitativas continuas (como ya sabíamos por la teoría estudiada). Una forma de poder aplicarlos es categorizando, o haciendo un resumen de los valores de la variable por intervalos, aunque ello conlleva cierta pérdida de información.

Uno de los gráficos adecuados para este tipo de datos es el histograma, que podemos dibujar en frecuencias absolutas o relativas empleando la función `hist` con el argumento `freq=FALSE` para las frecuencias relativas:

```
> hist(duration, main="Histograma del tiempo de erupción", xlab="tiempo")
> hist(duration, main="Histograma del tiempo de erupción", xlab="tiempo",
+       freq=F)
```

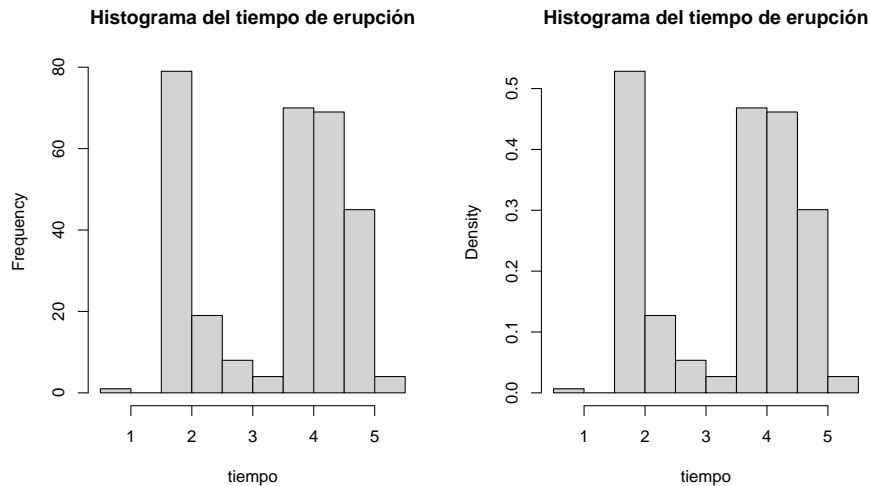



Figura 2: Histograma para la duración de erupción, con frecuencias absolutas y densidades.

El histograma es sensible al número de intervalos que se consideran. En la Figura 3 mostramos histogramas para distintos números de intervalos, que se construyen en  modificando convenientemente el argumento breaks:

```
> par(mfrow=c(2,2))
> hist(duration,breaks=4,col="grey",main="Histograma con 5 intervalos")
> hist(duration,breaks=9,col="grey",main="Histograma con 10 intervalos")
> hist(duration,breaks=24,col="grey",main="Histograma con 25 intervalos")
> hist(duration,breaks=49,col="grey",main="Histograma con 50 intervalos")
```

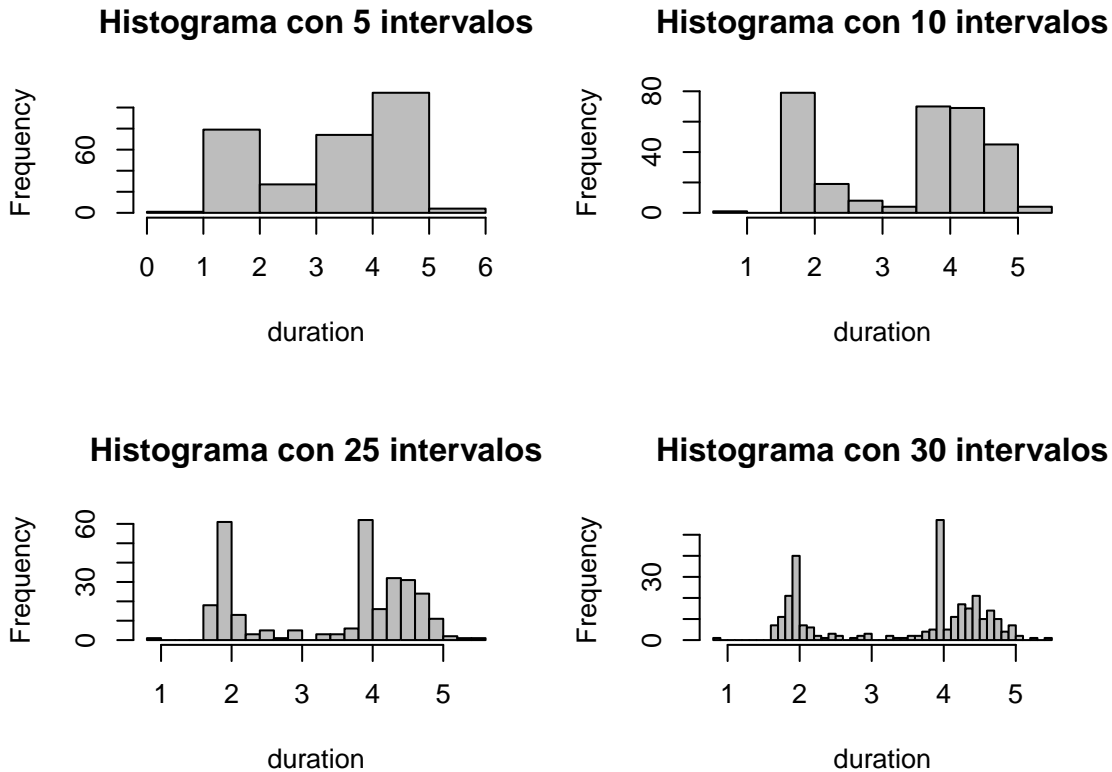



Figura 3: Histograma para la duración de erupción, con distinto número de intervalos.

Otra representación gráfica para variables cuantitativas continuas es el **diagrama de tallo y hojas**. Para obtener un diagrama de este tipo con  utilizamos la función `stem`:

```
> stem(duration)
The decimal point is 1 digit(s) to the left of the |
```

```
8 | 3
10 |
12 |
14 |
16 | 223370023357778
18 | 00022223333335557778880022333333555557778
20 | 000000000000000000000000223578023578
22 | 0278
```


[illegible]

3. Medidas características

La función `summary` nos proporciona información general sobre cualquier variable. En el caso de variables cuantitativas, lo que devuelve es el valor del mínimo, el máximo, la media y los cuartiles. Para variables cualitativas es más simple y devuelve únicamente un recuento por categorías, es decir, las frecuencias absolutas (y por tanto lo mismo que la función `table` que vimos anteriormente).

Para calcular la **media** muestral utilizamos la función `mean` con la que también se pueden obtener **medias truncadas** mediante el argumento `trim`, que nos permite introducir una fracción (de 0 a 0.5) que indica el porcentaje de datos más altos y más bajos que se eliminan. De este modo, se obtiene una media *robustificada* en el sentido de que ya no es tan sensible a la presencia de datos atípicos.

No existe una función en que calcule la **varianza** muestral. La función `var` devuelve la **cuasi-varianza** muestral, por lo que debemos transformar el resultado para obtener la varianza:

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \implies S^2 = \frac{n-1}{n} S_c^2.$$

Al igual que ocurre con la función `var`, la función `sd` proporciona la cuasi-desviación típica muestral, si queremos calcular la desviación típica muestral podemos obtenerla empleando su definición como raíz cuadrada de la varianza.

Otras funciones para obtener medidas características que se han visto en el Tema 1 son:

- `median` para calcular la **mediana** muestral.
- `min` para calcular el **mínimo** de la muestra.
- `max` para calcular el **máximo** de la muestra
- `diff(range)` para calcular el **rango muestral**.
- `quantile` para calcular diferentes **cuantiles muestrales** adaptando el argumento `probs`.
- `IQR` para calcular el **rango intercuartílico** de la muestra.

Retomemos el ejemplo de los datos de `geyser` de la librería `MASS` y la variable `duration` para la que calcularemos algunas de estas medidas:

```
> mean(duration) #media
[1] 3.460814


> median(duration) #mediana
[1] 4

> quantile(duration,probs=0.7) #cuantil 0.7
70%
4.266667

> quantile(duration, probs = c(0.25,0.5,0.75)) #cuartiles
      25%      50%      75%
2.000000 4.000000 4.383333

> var(duration) #cuasi-varianza
[1] 1.317683

> diff(range(duration)) #rango
[1] 4.616667
```

Como también vimos, una forma de resumir la información de las medidas características es representando un **diagrama de caja** o Box-Plot. Obtendríamos esta representación en  de la siguiente forma:

```
> boxplot(duration)
```

El diagrama de caja o Box-Plot resume en un único gráfico algunas de las medidas de centralización y dispersión más usadas en estadística. La barra negra central muestra donde se sitúa la mediana, mientras que los extremos de la caja son el primer cuartil, Q_1 , (inferior) y el tercer cuartil, Q_3 , (superior). De este modo, la altura de la caja es el rango intercuantílico $Q_3 - Q_1$. En caso de que la variable sea simétrica, la mediana ha de situarse en un punto equidistante de ambos cuartiles y coincidirá con la media. Los extremos de las barras en el Box-Plot representan los límites superior e inferior. Las observaciones que caigan fuera de estos límites son datos atípicos. Véase Figura 4.

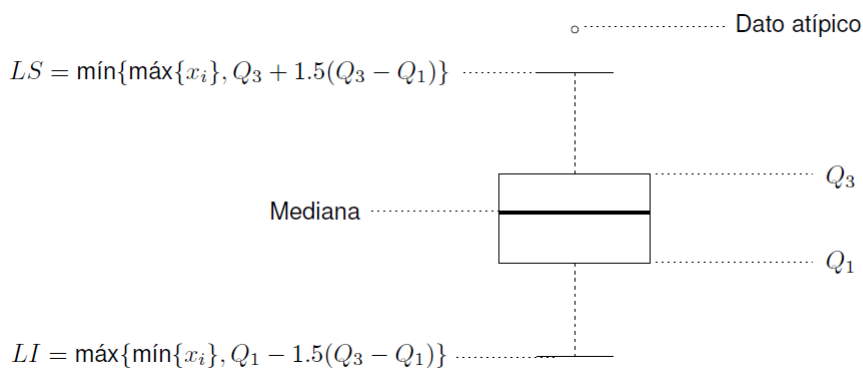


Figura 4: Diagrama de cajas.

Veamos ahora un nuevo **ejemplo con datos de calidad de aire**. El conjunto de datos `airquality` contiene información sobre calidad del aire. Entre otras variables, se recoge la velocidad de viento registrada (`Wind`), en millas por hora, que se trata de una variable cuantitativa continua. Para representar gráficamente esta variable necesitamos:

```
> attach(airquality)
> names(airquality)
[1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"

> hist(Wind, col="blue", main="Histograma", xlab="Velocidad del viento")
> boxplot(Wind,col="purple",main="Diagrama de caja", xlab="Velocidad
+ del viento")
```

Como ya hemos visto en el ejemplo de los datos de `geyser` se puede modificar el número de intervalos del histograma con el argumento `breaks`.

4. Tabla resumen de procedimientos y funciones

Procedimientos	Funciones
Tamaño muestral	length
Tablas	table
Marginales	addmargins
Diagrama de barras	barplot
Diagrama de sectores	pie
Histograma	hist
Diagrama de tallo y hojas	stem
Diagrama de caja	boxplot
Media	mean
Cuasivarianza	var
Mediana	median
Cuantil	quantile
Rango intercuartílico	IQR
Rango	diff y range
Asimetría	skewness
Curtosis	kurtosis

Otras funciones	
Instalar paquete	install.packages
Cargar paquete	library
Sumas acumuladas	cumsum
Redondeo	round
Resumen de medidas	summary
Adjuntar datos	attach
Nombres de objeto	names