Exploratory Data Analysis
Introduction to R

Raquel Menezes
rmenezes@math.uminho.pt

Department of Mathematics and Applications

University of Minho, Portugal

September, 2018

Syllabus

- **Introduction**
  - Statistics
  - The R environment
- **Exploratory data analysis**
  - Univariate data (categorical/numerical)
  - Bivariate data and correlation
  - Multivariate data

Statistics is the science that guide us in decision making under uncertainty

Main concerns:

1. **obtain information/data** (sampling design, design of experiments, surveys)
2. **initial data processing** (derive main sample characteristics, grouping into classes, tables and graphical representations)
3. **make inference from a sample to a larger population** (decision on the assumptions, estimation of the population parameters)
4. **predict the future evolution of a phenomenon** (prediction).

Some definitions

- **Population** denotes the set of elements whose characteristics (attributes) are the subject of a specific study.

  Examples: 1- postgraduate students in Portugal; 2- steel bars produced by a given company.

- **Census** are conducted to acquire complete knowledge about the population.

  Example: In Portugal the population census takes place every 10 years, and the last has been made in the end of 2011.

- **Sample** – the study of the population characteristics can be done over a finite and representative subset of the population, denoted as **sample**.

  Examples: 1- students from Minho University; 2- twenty steel bars.

- **Variables** need to be defined to allow the study of the characteristics of interest, based on observed values from the sample (or the whole population). One variable is defined per each characteristic.

  Examples:

  1- X="final mark of a PhD student"

  2- X="tensile strength of steel bars"

Given a sample of **dimension** n, and a **variable** X, one has

$$x_1, x_2, \dots, x_n$$

where x_i $(i = 1, \dots, n)$ represents the i^{th} value of the observed characteristic.

Main type of variables

- **Qualitative or categorical**: assume a set of categories.
  - **Nominal**
    Example: X="gender of student"
  - **Ordinal**
    Example: X="age group"
- **Quantitative**: assume a set of numerical values according to a range of intensities or values.
  - **Discrete**: can take a finite number or a countable infinity of values.
    Example: X="age of student"
  - **Continuous**: can take any value within an interval of real numbers.
    Example: X="height of a person"

ToDo: Identify the type of variables from previous slide.

R an open-source (GPL) **statistical environment**

Why R ?

- R is free and it runs on UNIX, Windows and Macintosh.
- R has an excellent built-in help system.
- R has excellent graphing capabilities.
- R's language has a powerful, easy to learn syntax with **many built-in statistical functions**. It is easy to extend with user-written functions.

Some drawbacks:

- Limited graphical interface
- No commercial support (although one can argue the international mailing list is even better)
- The command language is a programming language so user must learn to appreciate syntax issues etc.

How to manipulate data in R ?

- R is most easily used in an interactive manner.
- You ask it a question and R gives you an answer. **Questions are asked and answered on the command line** (prompt ">").
- Typing less! Use arrow keys to retrieve your previous commands.

Statistics is the study of data !

- To quickly enter a small data set use the c function. This function combines, or concatenates terms together.
- EXAMPLE: suppose that the number of typos per page are
  2 3 0 3 1 0 0 1

```
> typos <- c(2,3,0,3,1,0,0,1)
> typos
[1] 2 3 0 3 1 0 0 1
```

Some built-in functions

```
> is.vector(typos)
[1] TRUE
> class(typos)
[1] "numeric"
> mean(typos)
[1] 1.25
> sd(typos)
[1] 1.642857
> summary(typos)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    0.00    1.00    1.25    2.25    3.00
> length(typos)
[1] 8
```

The so useful **help** function: > help(length) or > ?length

Exploratory analysis of categorical data

R knows about some of the differences between types of data in statistics: if categorical, if discrete numeric or if continuous numeric.

Methods for viewing and summarising the data depend on the type (e.g. it doesn't make sense to derive the mean of a categorical variable).

Example: Suppose we collect a sample of 100 students from Minho University, and we record two variables for each - **month of birth** (1-12) and **gender** (m or f). Note that both variables are qualitative, although the former is labeled with numbers 1 to 12 (it could also be J F M A . . .).

```
> month <- c(9, 10, 4, 8, 7, 9, 3, 8, 4, 1, 5, 2, 12, 4, 10, 3, 2, 2, 12, 4, 6, 1, 6, 6,
12, 4, 12, 1, 8, 2, 2, 1, 6, 12, 6, 5, 10, 11, 8, 5, 4, 8, 3, 11, 11, 9, 8, 9, 8, 6, 5, 7,
12, 2, 2, 6, 4, 4, 7, 8, 4, 2, 6, 1, 4, 6, 2, 8, 11, 10, 6, 5, 4, 1, 8, 3, 7, 8, 9, 7, 8, 9,
8, 8, 5, 1, 1, 3, 7, 7, 12, 3, 1, 7, 11, 9, 4, 5, 2, 9)
```


Using tables

Alternatively, data may be read from a external file (excel, csv, text, etc). Suppose, we have a text file, then

```
> aux <- read.table(file="GenderBirthMonth.txt", header=T)
> names(aux)
> month <- aux$birthmonth
> gender <- aux$gender
> summary(month)    #WRONG
> summary(gender)   #OK
```

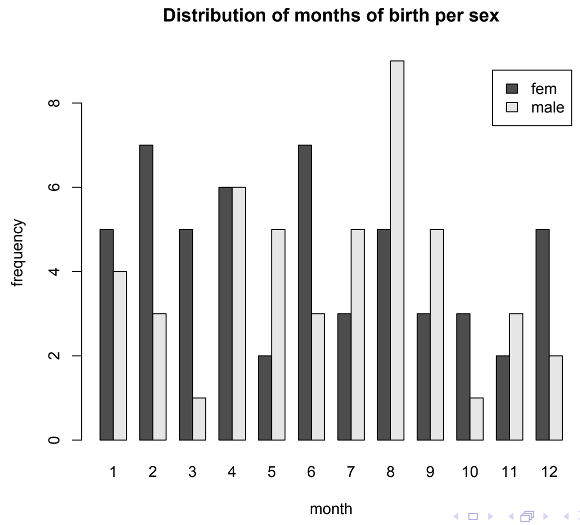
R function table is quite important for the analysis of categorical variables, as it allows to cross-classify factors to build a contingency table of the counts at each combination of factor levels.


```
> table(gender)
gender
 f  m
53 47
> table(month)
month
 1  2  3  4  5  6  7  8  9 10 11 12
 9 10  6 12  7 10  8 14  8  4  5  7
> table(gender, month)
      month
gender  1  2  3  4  5  6  7  8  9 10 11 12
     f  5  7  5  6  2  7  3  5  3  3  2  5
     m  4  3  1  6  5  3  5  9  5  1  3  2
```

ToDo: Apply function `barplot` to `table(gender, month)`

```
> barplot(table(gender,month), beside=T, legend.text=c("fem","male"), main=
"Distribution of birth months per gender", xlab="month", ylab="frequency")
```

Example of graphical representation for bivariate data



Graphical representations for univariate data

A **bar chart** draws a bar with a height proportional to the count in the table. The height can be given by **absolute frequency** or **relative frequency** (proportion).

Example: Suppose, a group of 25 people are surveyed about their **beer-drinking preference**. The categories were (1) Domestic can, (2) Domestic bottle, (3) Microbrew and (4) import

```
> beer <- c(3, 4, 1, 1, 3, 4, 3, 3, 1, 3, 2, 1, 2, 1, 2, 3, 2, 3, 1, 1, 1, 1, 4, 3, 1)
> barplot(beer)                          # NO, this isn't correct
> barplot(table(beer))                   # Yes, call with summarized data
> barplot(table(beer)/length(beer))      # divide by n for proportion
```

Try also a **pie chart**:

```
> factor(beer);
beer.counts <- table(beer);
names(beer.counts) <- c("d can", "d bottle", "microbrew", "import");
pie(beer.counts, col=c("blue", "green", "red", "yellow"))
```


Exploratory analysis of numerical data

Descriptive statistics are used to summarize a set of numerical observations, $(x_1, x_2, \ldots, x_n)$, in order to communicate the largest amount of information as simply as possible. Normally classified as

1. **Measures of central tendency**, such as
   - mean: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
   - **median**: middle value that separates the higher half from the lower half of the data
   - **mode**: most frequent value in the data set (to be used with nominal data)
   - **quantile**: quartiles Q_1, Q_2 (median), Q_3; some percentile $P_{10}, P_{20}, \ldots, P_{90}$
   - **truncated mean**: mean after a certain number or proportion of the highest and lowest data values been discarded

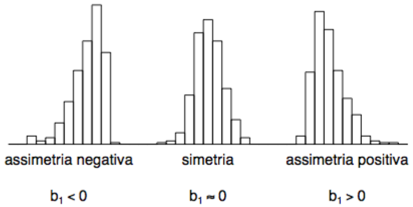
Note: The median is more robust to outliers than the mean.

2. **Measures of dispersion or variability**, such as
   - **sample range**: size of the smallest interval which contains all the data
   - **interquartile range**: $Q_3 - Q_1$
   - standard deviation (or variance): $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$
   - **minimum** and **maximum** values of the data
   - **coefficient of variation** (CV): a normalized measure of dispersion, given as $\frac{s}{\overline{x}}$. The absolute value of CV is sometimes known as **relative standard deviation** (RSD), which is expressed as a percentage

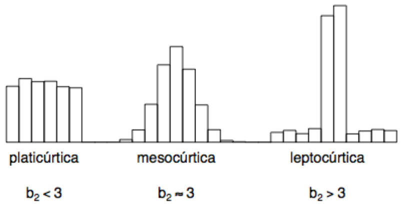
Note: The CV and RSD are useful to compare different samples, for example to compare the variability of weights of rats and elephants.

3. **Measures of shape**:

- skewness: measure of asymmetry of the distribution shape of a real-valued variable about its mean, it can be positive, negative, or even undefined.

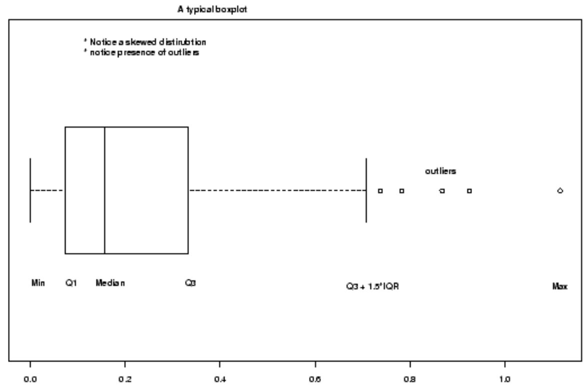


- kurtosis: measure of "peakedness" of the distribution shape of a real-valued variable

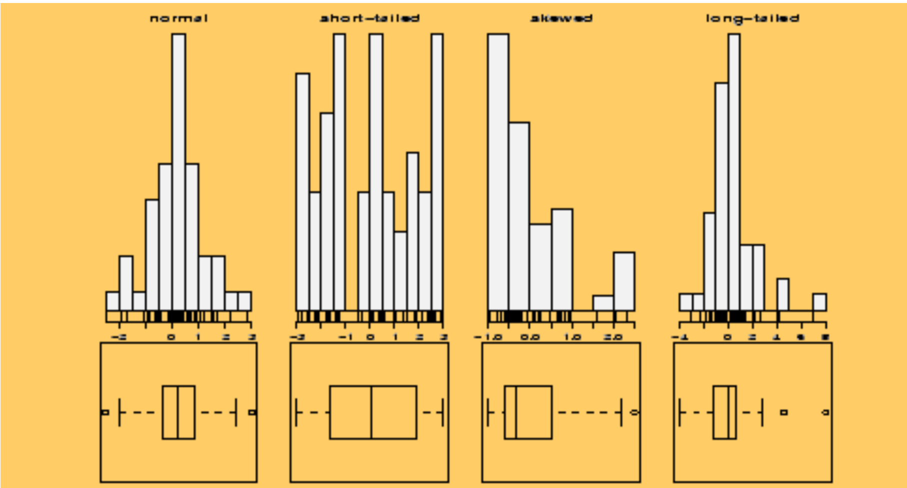


Graphical representations of numerical data

- Stem-and-leaf chart: mainly useful for relatively small datasets, for seeing the shape of the distribution and the values. > ?stem
- Histogram: the most common, very useful for large datasets, being similar to the bar plot (the height can be frequencies or proportions). > ?hist
- Boxplot: very useful to summarize data succinctly, quickly displaying if the data is symmetric or has suspected outliers. > ?boxplot



Histogram vs. Boxplot



Random distributions with both a histogram and the boxplot.

Example: Old Faithful geyser data in Yellowstone National Park, USA

Consider the data set *faithful* available in R environment, with the duration of the eruption and waiting time between eruptions in minutes for this geyser. A data set with 272 observations on 2 numerical variables: **eruptions** and **waiting**.

```
> ?faithful
> names(faithful); attach(faithful)
> summary(eruptions)
 Min. 1st Qu.  Median  Mean 3rd Qu.  Max.
1.600   2.163   4.000   3.488  4.454   5.100
> stem(eruptions, scale=0.5)
```

The decimal point is at the |

```
1 | 6777888888888888888888888888899999999999999999
2 | 00000000000000001111111222222222222333333344444444
2 | 566899
3 | 133344
3 | 555666666677788888888889999999
4 | 000000000000111111111111112222222222222223333333333333333444444444444
4 | 555555555555555555556666666666666666777777777777888888888888889999999
5 | 0011
```

Try a histogram and a boxplot for eruptions data

> hist(eruptions, breaks=seq(from=1.05, to=6, by=1/4), freq=F, main = "Histogram for eruptions", xlab = "duration (min)", ylab = "frequency")

> boxplot(eruptions, main="Boxplot for all eruptions", ylab="duration of eruption (min)")

Now, try a boxplot restricted to those eruptions with duration smaller than 3 mins

```
> er.small <- eruptions[eruptions<3]
> summary(er.small)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
1.600   1.833   1.983   2.038   2.200   2.900
> boxplot(er.small, main="Boxplot for small eruptions", ylim=range(1.5,5.1), ylab="duration (min)")
```

Hint: Use "> par(mfrow=c(1,2))" to plot the boxplots side by side.

Handling bivariate data: categorical vs. numerical

A simple example might be in a drug test, where you have data (in suitable units) for an experimental group and for a control group.

```
> experimental <- c(5, 5, 5, 13, 7, 11, 11, 9, 8, 9)
> control <- c(11, 8, 4, 5, 9, 5, 10, 5, 4, 10)
> boxplot(experimental, control)
```

Most of the times, one has a unique data set

```
> data <- c(experimental, control)
> type <- c(rep("experim.",10), rep("control",10))
```

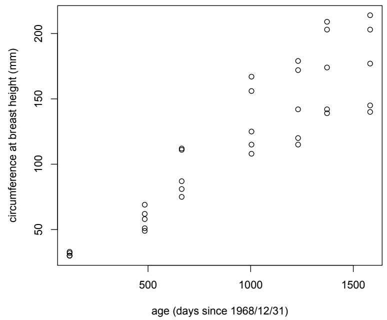
then you should

> boxplot(data ∼ type) # Pay attention to symbol ∼

Handling bivariate data: numerical vs. numerical

If you expect a relationship between two numerical variables, you might like to look for that by plotting pairs of points, i.e. a **scatterplot**. If independence is expected, you might like to compare their distributions in some manner.

```
> ?Orange
> names(Orange); attach(Orange)
> plot(age, circumference, main="scatterplot")
```



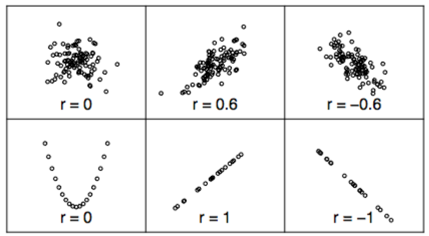
The graph seems to illustrate a strong linear trend, which should be investigated.

Bivariate data. Correlation

A valuable numeric summary of the **strength of the linear relationship** is the **Pearson correlation coefficient**, r, defined as

$$r = \frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i - \overline{x}}{s_x}\frac{y_i - \overline{y}}{s_y}$$

where s_x and s_y are the sample standard deviations for X and Y, respectively.



Note: It is possible to prove that $-1 \leq r \leq 1$.

```
> cor(Orange$age, Orange$circumference)
[1] 0.9135189
```

Correlation

Please note that a **strong correlation does not necessarily imply a cause and effect relationship** between the variables.

One common example is the **positive correlation** between the number of cases of dehydration and agricultural production, when there is **no direct cause and effect** between the two variables, but rather a **common cause to influence them** - the atmospheric temperature.

In fact, both the variables are positively correlated with the atmospheric temperature.

Spearman rank correlation

The Spearman rank correlation can be applied to the ranks of the data, i.e. to categorical ordinal variables. Suppose that d_i are the differences between the ranks of x_i and y_i

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \qquad \text{with} \quad -1 \leq r_S \leq 1$$

Example: To study the association between SO_2 and quality of white wine, 10 brands were analyzed. The wine experts assigned a rate from 1 (best) to 10 (worst) to each brand. The SO_2 content (ppm) in each brand was also given.

brand	A	B	C	D	E	F	G	H	I	J
quality	1	2	3	4	5	6	7	8	9	10
SO2 content	0.9	2.7	1.8	2.9	3.5	3.1	3.7	3.3	4.9	4.7

```
>quality <- 1:10
>SO2 <- c(0.9,2.7,1.8,2.9,3.5,3.1,3.7,3.3,4.9,4.7)
>cor(quality, SO2, method="spearman")
[1] 0.9151515
```


Multivariate data

Often in statistics, **data is presented in a tabular format similar to a spreadsheet**. The columns are for different variables, and each row is a different measurement or variable for same person or thing. R uses data frames to store these variables together.

Example: Suppose 4 people are asked three questions, their weight, height and gender and the data is entered into R as separate variables as follows

```
> weight = c(150, 135, 210, 140)
> height = c(65, 61, 70, 65)
> gender = c("Fe","Fe","M","Fe")
> study = data.frame(weight,height,gender)    # make the data frame
> study
      weight height gender
Mary   150    65      Fe
Alice  135    61      Fe
Bob    210    70      M
Judy   140    65      Fe
```


Handling data frames

You can give the rows names as well.

```
> row.names(study)<-c("Mary","Alice","Bob","Judy")
```

Different ways to get the weight variable

```
> study$weight
> study[,"weight"]
> study[,1]
```

Different ways to get the data for Alice

```
> study["Alice",]
> study[2,]
```

To get just the females information

```
> study[study$gender == "Fe", ]    # use $ to access gender via a list
```

n-way contingency tables

```
> # library MASS
> library(MASS); data(Cars93); attach(Cars93)

> ## make some categorical variables using cut
> priceCat <- cut(Price,c(0,12,20,max(Price)))
> levels(priceCat) <- c("cheap","okay","expensive")
> mpgCat <- cut(MPG.highway,c(0,20,30,max(MPG.highway)))
> levels(mpgCat) <- c("miser","okay","gas guzzler")

> ## now look at the relationships
> table(Type)
> table(priceCat,Type)
> table(priceCat,Type,mpgCat)
```

Or, do a **boxplot** with the numerical variable “Price”

```
> boxplot(Cars93$Price ~ Cars93$Type)
```