A brief introduction to R

Ellen Brock

Outline

1. Brief history
2. Downloading and installing R
3. Some basic steps
4. Summary statistics
5. Visualisation
6. Resources: What next?

1. Brief history

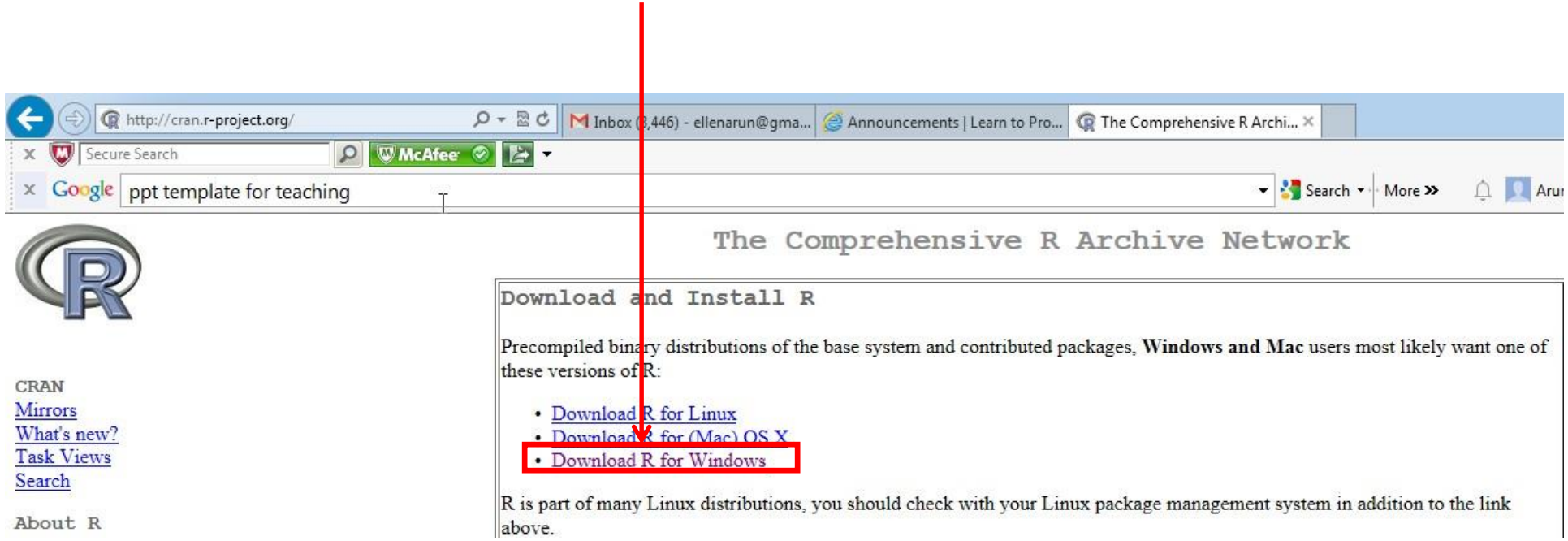
- R is a “dialect” from the S language developed at Bell Labs
- Developed by Ross Ihaka and Robert Gentleman from New Zealand in 1991
- R was first released in the public in 1993
- R version 1.0.0 is released
- Latest version 3.1.0 was released in August 2015
- Software is free of cost under the GNU General Public License

2. Downloading and installing R

- Go to the following website:

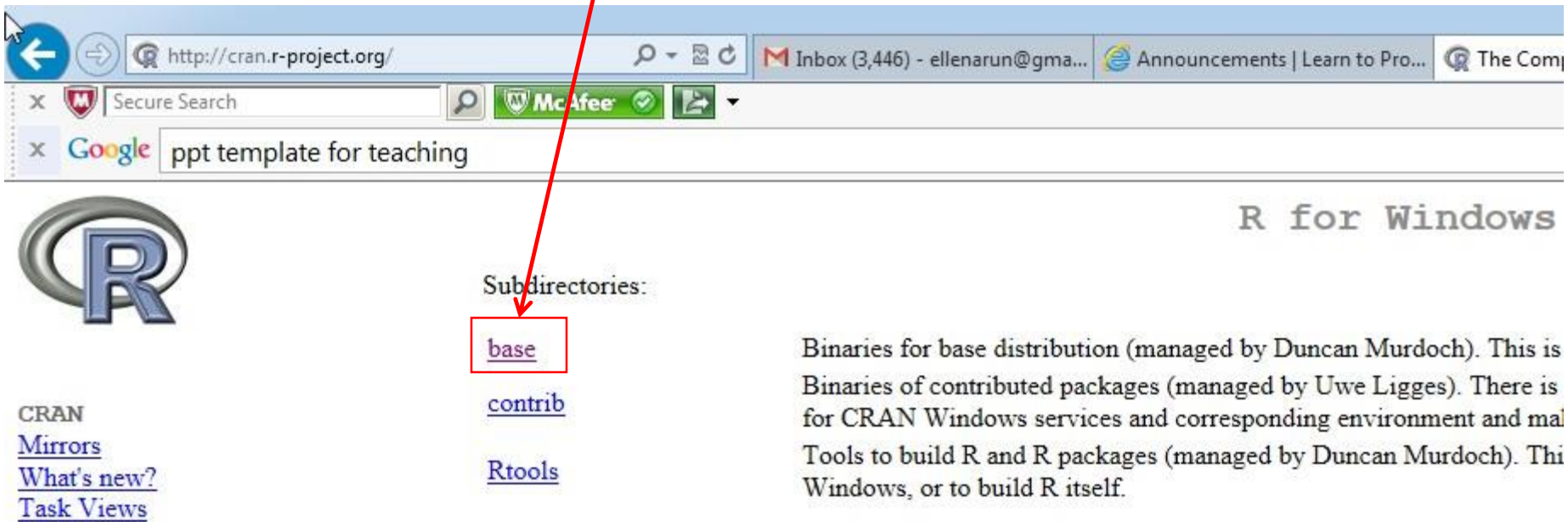
  http://cran.r-project.org/

- Click on: “Download R for Windows”



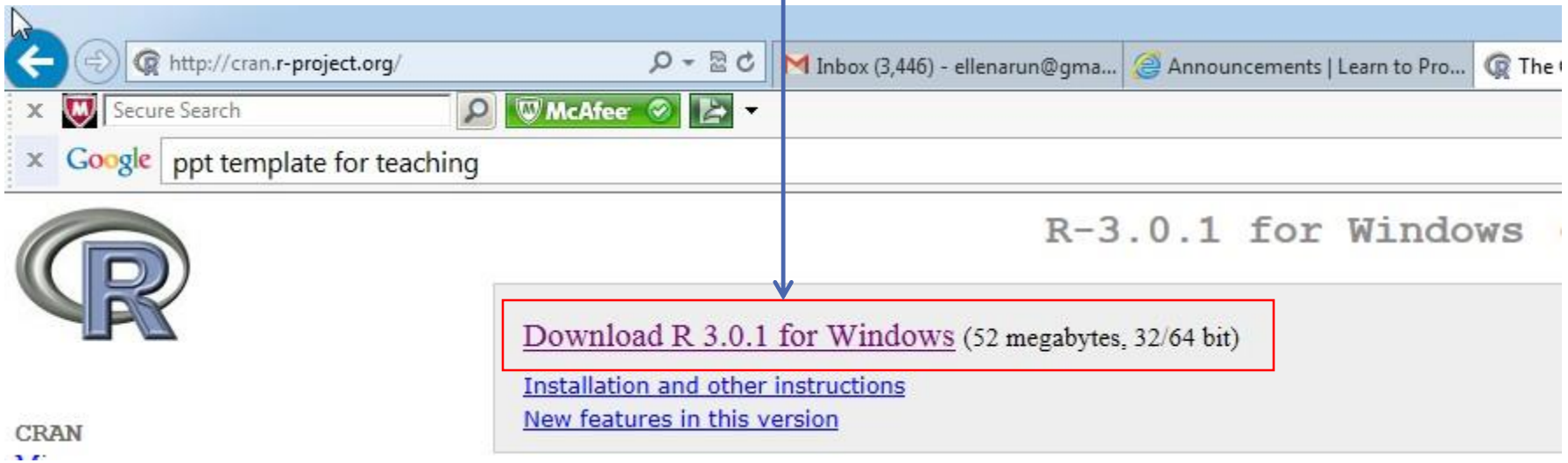
2. Downloading and installing R (contd.)

- Next, click on “base”:



2. Downloading and installing R (contd.)

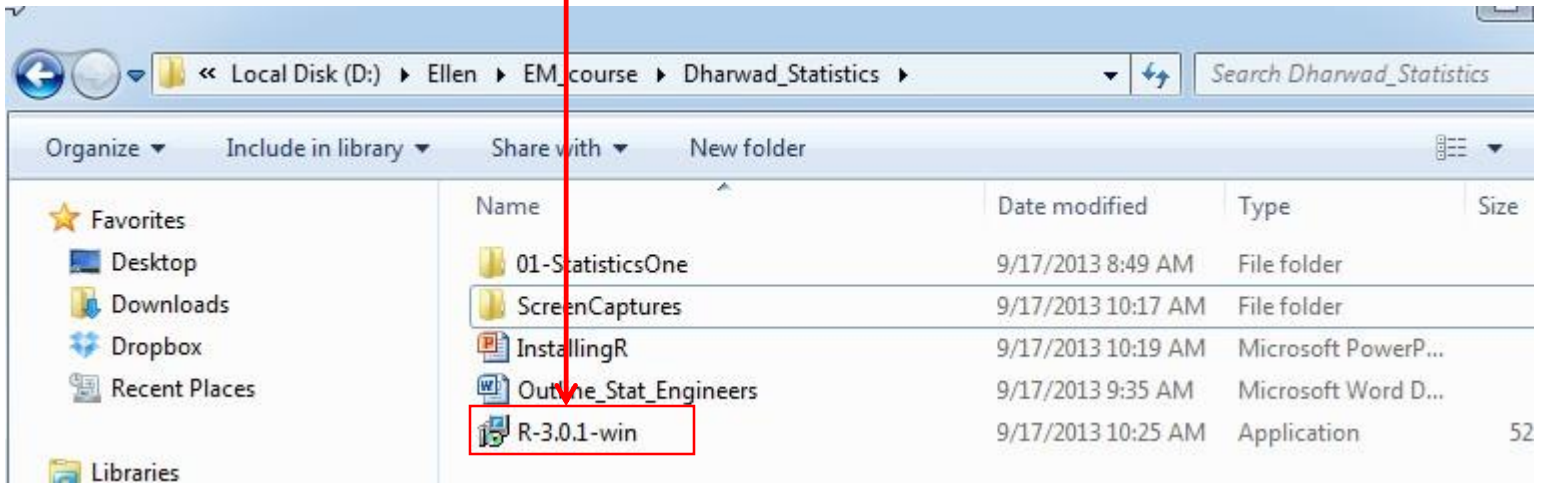
- Click on “Download R.3.0.1 for Windows”



- You will be asked to save this file somewhere on your hard disk

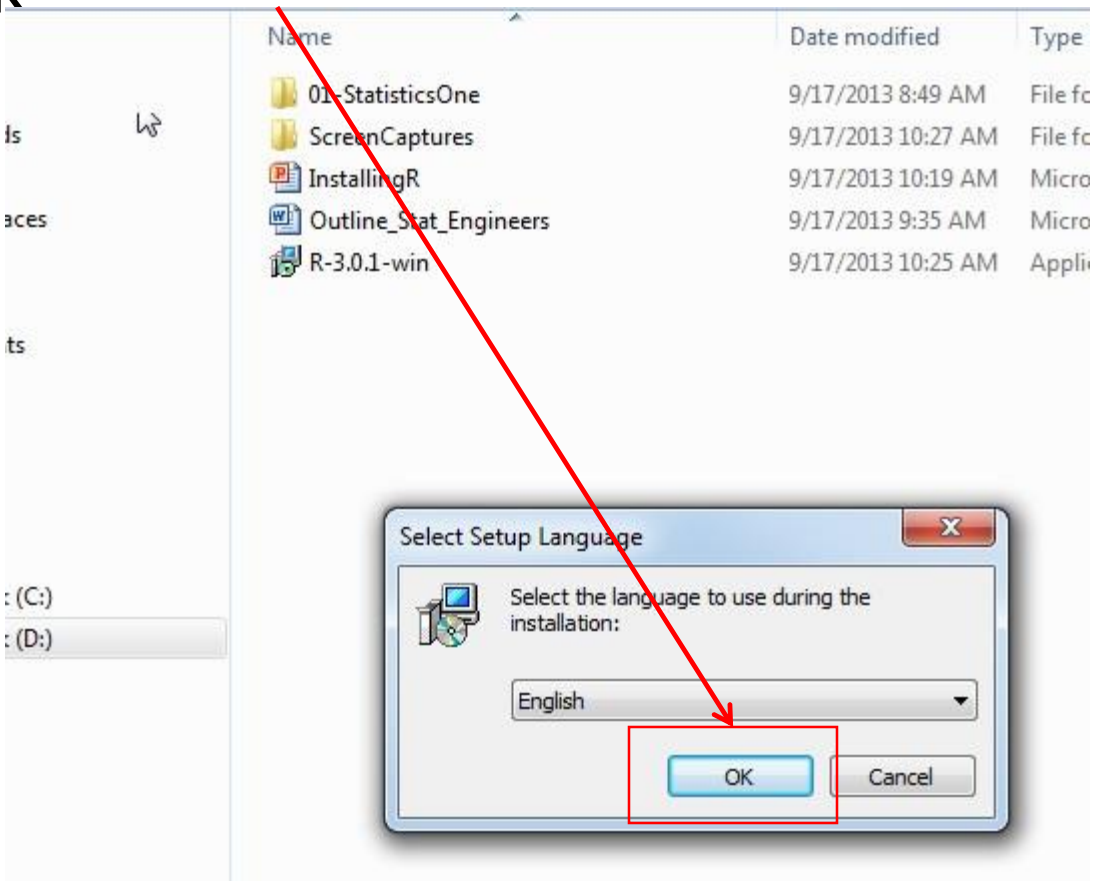
2. Downloading and installing R (contd.)

- Double click on the file you just downloaded



2. Downloading and installing R (contd.)

- Click on “OK”

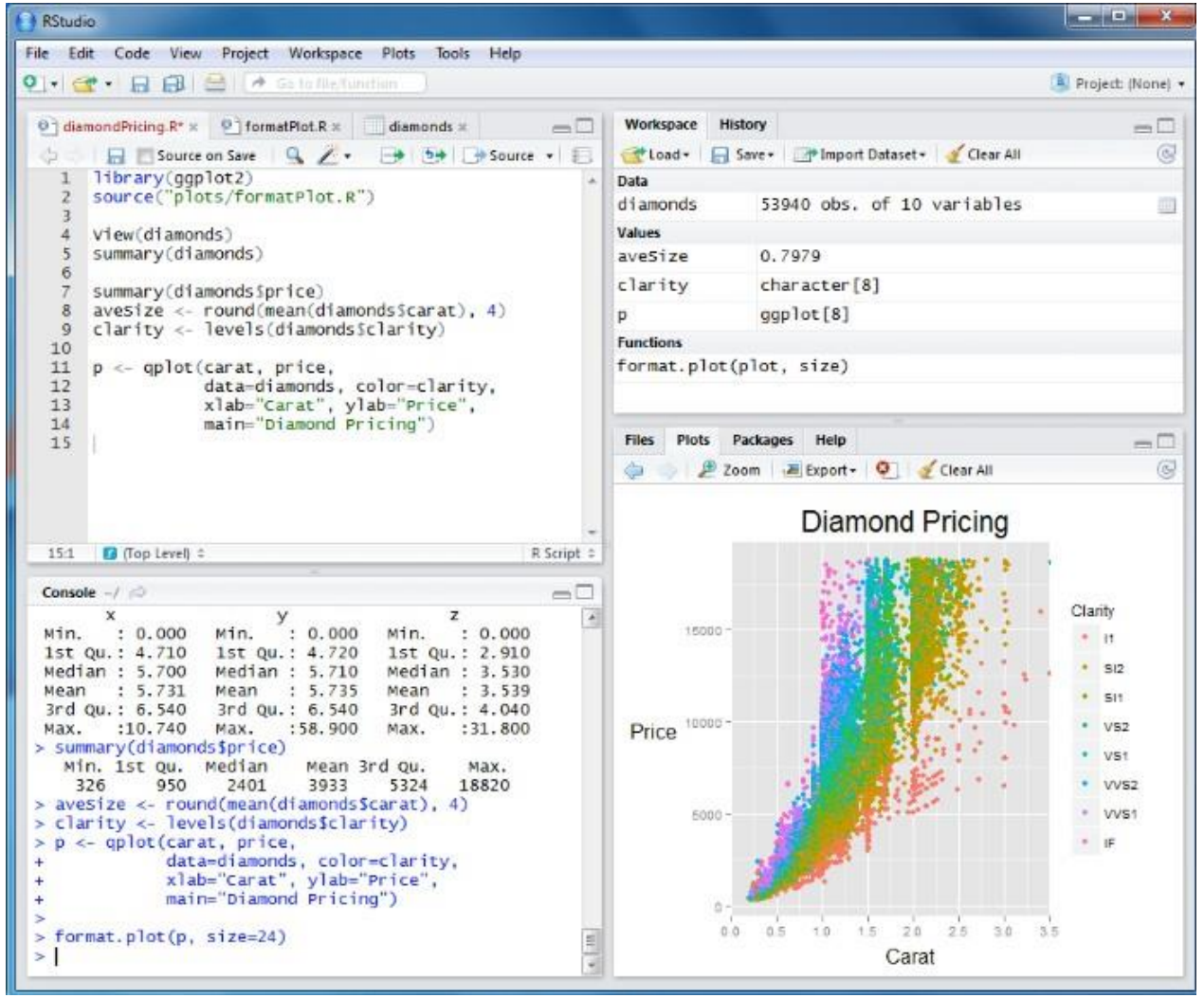


Further options

- One can use R-studio
- Nice interface that combines: the console, source editor, workspace, command history, files, plots, packages and R help
- Code, output, graphics all in one window
- Download instructions can be found here:

  http://www.rstudio.com/

RStudio


RStudio
File Edit Code View Project Workspace Plots Tools Help
diamondPricing.R* formatPlot.R diamonds
Source on Save Source
1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 View(diamonds)
5 summary(diamonds)
6
7 summary(diamonds$price)
8 aveSize <- round(mean(diamonds$carat), 4)
9 clarity <- levels(diamonds$clarity)
10
11 p <- qplot(carat, price,
12 data=diamonds, color=clarity,
13 xlab="Carat", ylab="Price",
14 main="Diamond Pricing")
15
15:1 (Top Level) R Script
Console
x y z
Min. : 0.000 Min. : 0.000 Min. : 0.000
1st Qu.: 4.710 1st Qu.: 4.720 1st Qu.: 2.910
Median : 5.700 Median : 5.710 Median : 3.530
Mean : 5.731 Mean : 5.735 Mean : 3.539
3rd Qu.: 6.540 3rd Qu.: 6.540 3rd Qu.: 4.040
Max. :10.740 Max. :58.900 Max. :31.800
> summary(diamonds$price)
Min. 1st Qu. Median Mean 3rd Qu. Max.
326 950 2401 3933 5324 18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+ data=diamonds, color=clarity,
+ xlab="Carat", ylab="Price",
+ main="Diamond Pricing")
>
> format.plot(p, size=24)
>
Workspace History
Load Save Import Dataset Clear All
Data
diamonds 53940 obs. of 10 variables
Values
aveSize 0.7979
clarity character[8]
p ggplot[8]
Functions
format.plot(plot, size)
Files Plots Packages Help
Zoom Export Clear All
Diamond Pricing
Price
Carat
Clarity
I1
SI2
SI1
VS2
VS1
VVS2
VVS1
IF

Other editors

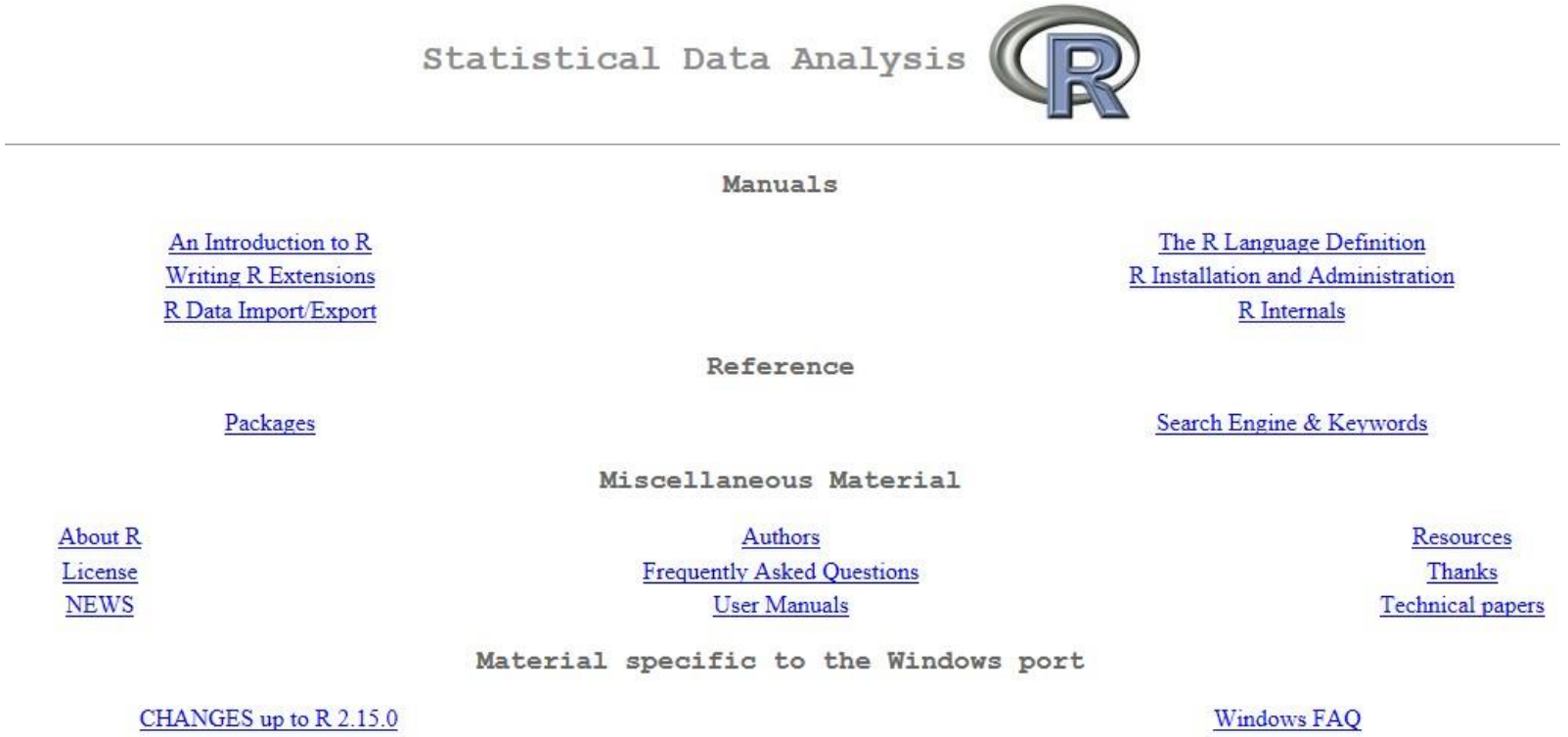
- Tinn-R: http://www.sciviews.org/Tinn-R/
- Emacs speaks Statistics: http://ess.r-project.org/
- Vim: http://www.vim.org/
- Eclipse Statet: http://www.walware.de/goto/statet
- Other editors: See the Wikipedia page on R, https://en.wikipedia.org/wiki/R_(programming_language)#Interfaces

2. Some basic steps

- R as a calculator
- Creating a simple series and compute a mean
- Loading built-in datasets

Using help

- Type in “help.start()” and you will get this in your browser:



Help for functions

- E.g. for the mean: help(mean) or ?mean

Loading packages

- R has built-in modules, called packages
- After has been started, several packages are already loaded
- To see which ones, use search()
- The first time, you need to install the package, you need to type:

  install.packages("spatial")
- Once it is installed, you need to load it and do this for every new session e.g. require(spatial)

Setting the working directory

- Working directory is the source for your file input and output:
  - Reading and writing data files
  - Opening and saving script files
  - Saving workspace image
- At the opening of a session, the know the settings of your current working directory, type (either in the console or your script): getwd()
- You will get e.g:

```
> getwd()
[1] "C:/Users/Arun/Documents"
```

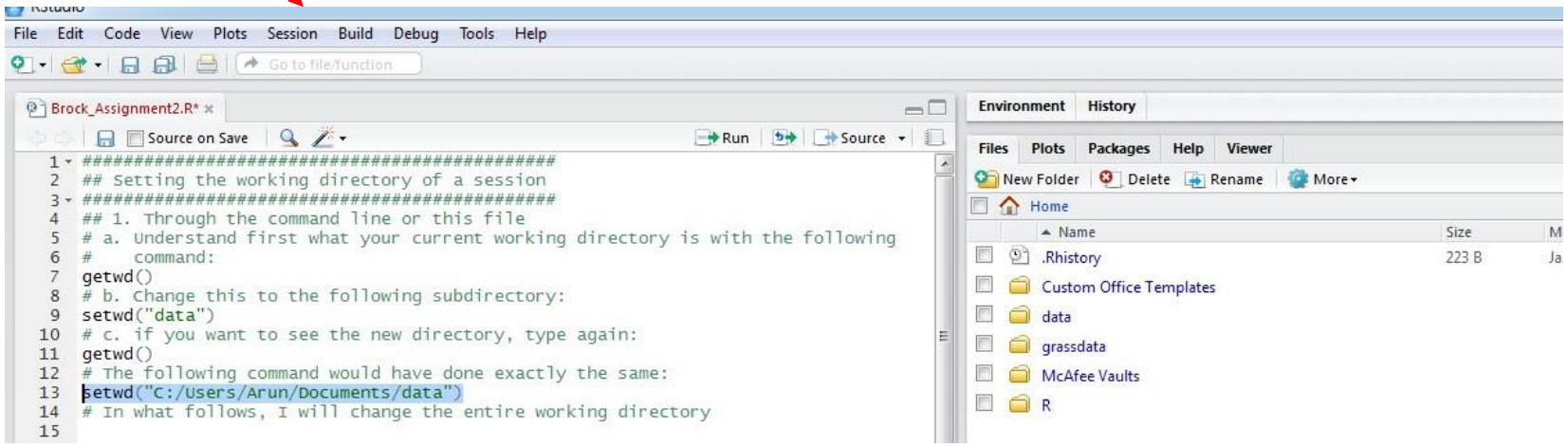
Setting the working directory

- Working directory is the source for your file input and output:
  - Reading and writing data files
  - Opening and saving script files
  - Saving workspace image
- At the opening of a session, the know the settings of your current working directory, type (either in the console or your script): getwd()
- You will get e.g:

```
> getwd()
[1] "C:/Users/Arun/Documents"
```

Changing the working directory

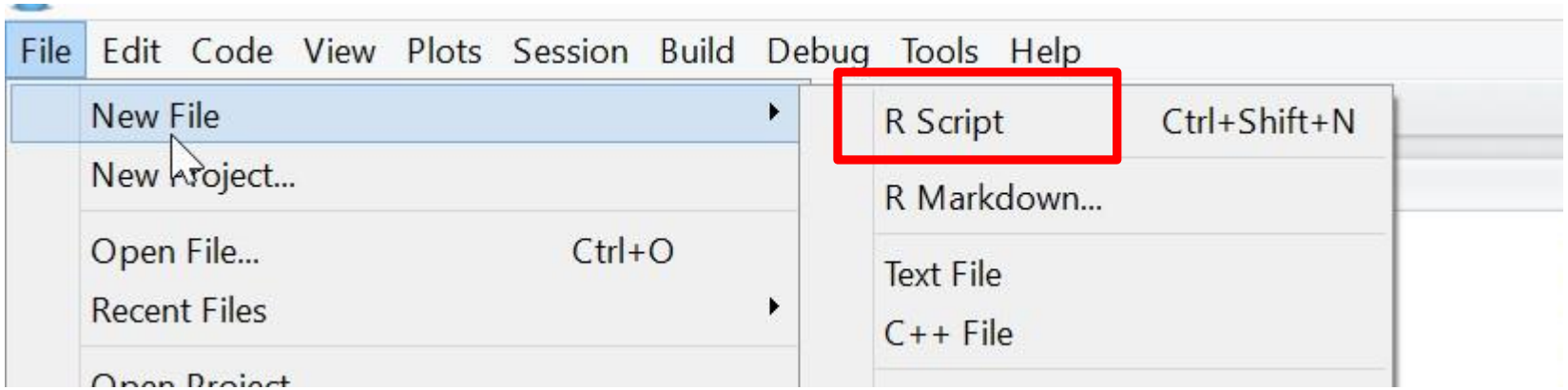
- If you want to add a subdirectory to the current working directory, you have two ways to change to this directory (see http://www.rstudio.com/ide/docs/using/workspaces):
  1. In R-studio (option 1) go to “Session/Change working directory” on the menu bar



Changing the working directory (contd)

2. In the script or command line (Option 2): setwd(“data”)
   - This is equivalent to typing:
     setwd("C:/Users/Arun/Documents/data")
   - If we want to use a different directory altogether than a subdirectory of the R default directory, we can type e.g.:
     setwd("E:/Ellen/DataMining/DS400/week2")

Creating an R script



The housing dataset

- The housing dataset among many other datasets can be found in the UCI Machine Learning Repository:
- http://archive.ics.uci.edu/ml/datasets.html
- More specifically, the housing dataset can be found in the following link: http://archive.ics.uci.edu/ml/datasets/Housing
- Housing prices in the area of Boston

The housing dataset

- **CRIM**: per capita crime rate by town
- **ZN**: proportion of residential land zoned for lots over 25,000 sq.ft.
- **INDUS:** proportion of non-retail business acres per town
- **CHAS5:** Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- **NOX**: nitric oxides concentration (parts per 10 million)
- **RM**: average number of rooms per dwelling
- **AGE:** proportion of owner-occupied units built prior to 1940
- **DIS:** weighted distances to five Boston employment centres
- **RAD:** index of accessibility to radial highways
- **TAX:** full-value property-tax rate per $10,000
- **PTRATIO:** pupil-teacher ratio by town
- **B**: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- **LSTAT:** % lower status of the population
- **MEDV:** Median value of owner-occupied homes in $1000's

Source: Harrison, D. and Rubinfeld, D.L. (1978),
'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978, see http://archive.ics.uci.edu/ml/datasets/Housing.

Accessing variables

- To access one variable of the dataset, you need to write (e.g. the variable MEDV), you have two options:
  1. Write: HousDat$MEDV
  2. Or use the function detach/attach:

     attach(HousDat)

     - Then you can access the variable directly. Assume you want to compute the mean, you can type

       mean(MEDV)

     - Once you are done, you can write again:

       detach(HousDat)

Reading in the dataset: text file and CSV file

- See demo

Data description

- See demo

Looking more deeply at the types of variables

- See Navarro (2013), p. 90 for background

  http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/
- We can see that all the variables in the data set are characterized as "numeric"
- What does this mean?
- In general, we have two types of variables:
  - Quantitative/numeric variables (displayed as "NUM" or "INT" in "str" command)
  - Qualitative variables (displayed as "FACTOR" or "Ord.FACTOR" in "str" command)
- Based on the previous slide, we see that all variables are NUMERIC (quantitative) but does that make sense?

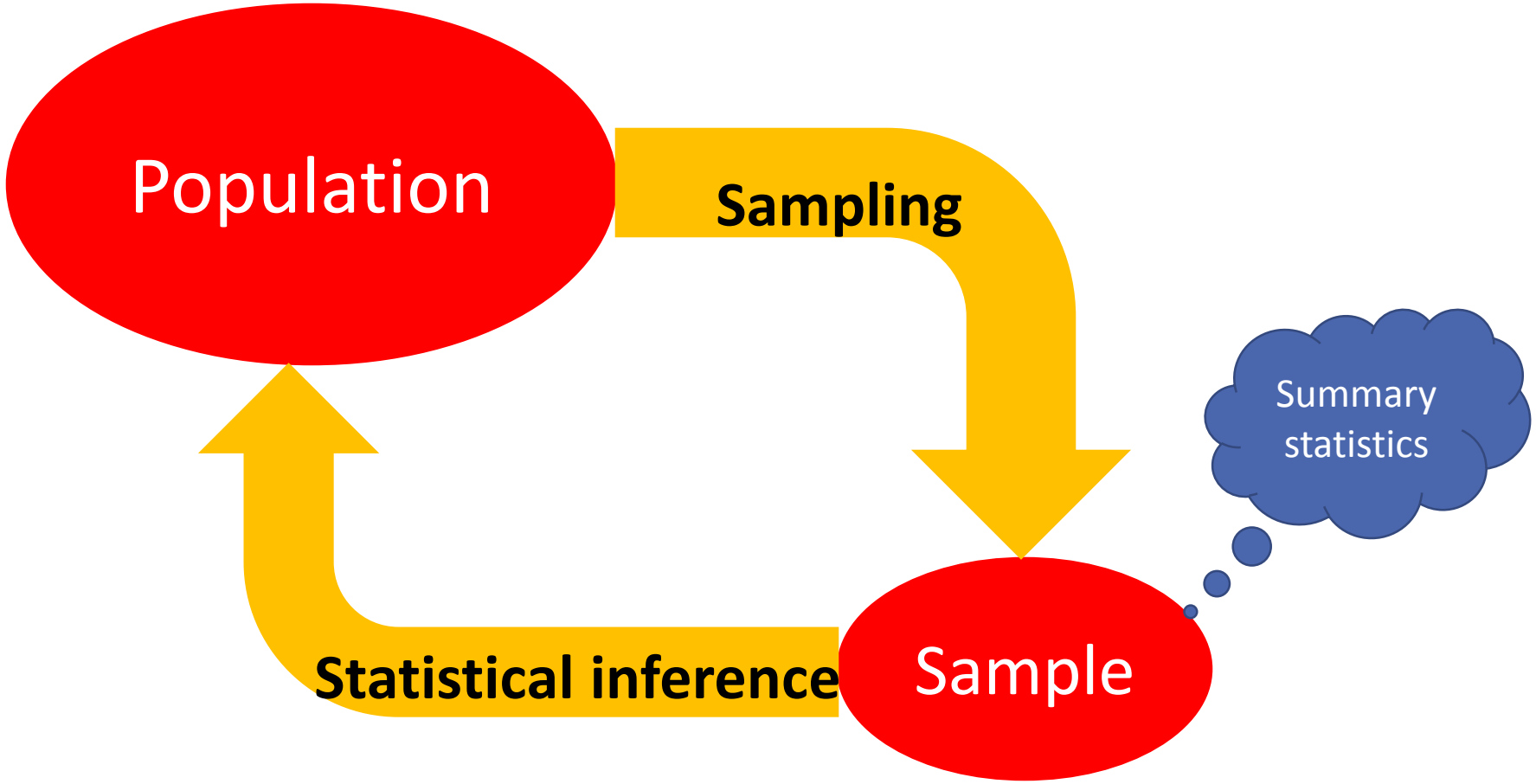
Quantitative/numeric variable

- http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+what+are+variables
- Measurable quantity and describes “how much” or “how many”
- Further distinction in:
  - **Continuous variable:** Any value between certain set of numbers and can be as small as the instrument of measurement allows (examples such as height, weight, temperature, etc)
  - **Discrete variable:** based on a count from a set of distinct whole values (examples are number of days, cars, children, etc)

Qualitative/categorical variable

- **Describe a 'quality' or 'characteristic', like “what type” or “which category”**
- Further distinction in:
  - **Ordinal variable:** There is logical ordering/ranking in the variable but the difference between categories does not necessary has meaning such as academic grades, clothing sizes, etc. See http://www.ats.ucla.edu/stat/r/modules/factor_variables.htm for more information in R.
  - **Nominal variable:** There is NO logical ordering such as male/female, eye color, etc.

Summary statistics: Statistical inference



Based on a sample drawn from the population, we will try to say something (inference) about the population. The starting point are the summary statistics of the sample.

Descriptive/summary statistics

- Consists of:
  - Summarizing the data
  - Visualising the data
- Depends on:
  - One or two (or more) variables (e.g. association between housing prices and crime rates in a certain area)
  - Type of data (see before):
    - Quantitative variable
    - Qualitative variable

Summary statistics according type of variable

Type of variable	Summary statistics
Numeric variable	Measures of central tendency (Mean, median, mode) Measures of spread (range, interquartile range, mean absolute deviation, variance, standard deviation and mean absolute deviation.
Categorical variable	Frequency table and relative frequency table

Quantitative variables: Measures of central tendency

Mean: Sum of all data / Total number of data

Quantitative variables: Measures of central tendency (contd.)

- Median: Value which divides data arranged in ascending or descending order into two equal halves
- If the number of observations is even, take the average of the two middle values
- Robust to outliers in the data

Quantitative variables: Measures of spread (contd.)

- Variance: mean square deviation meaning that you take the average of the squared deviations from the mean

$$\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

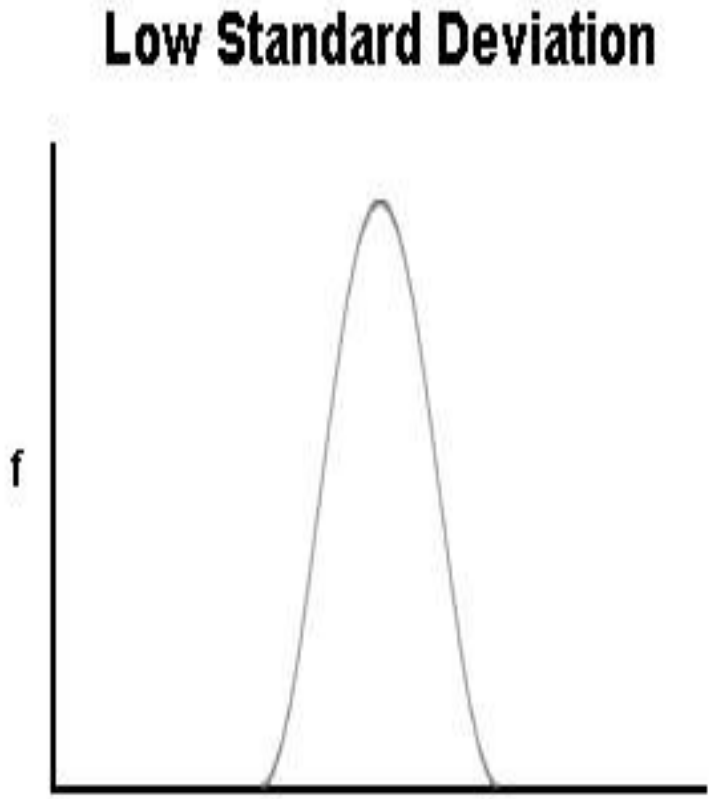
- With *N* the number of observations X_i the i^{th} observation and $\bar{X}$ the average
- Why this is divided by N-1 is beyond the scope of this presentation.
- Details can be further found in Navarro (2013, p. 117 and 290) among others why this is the case

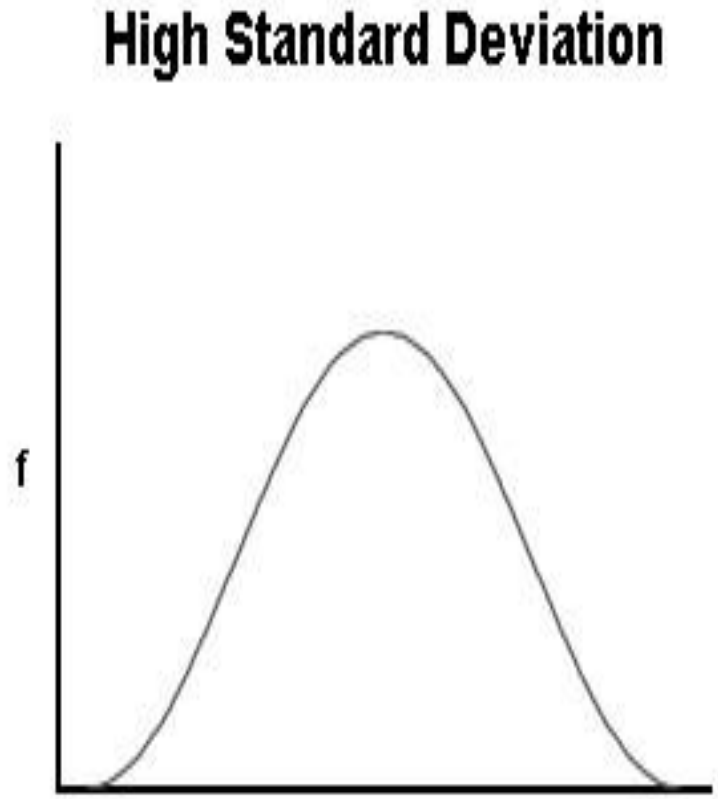
Quantitative variables: Measures of spread (contd.)

- Standard deviation:
  - Square root of the average squared deviation from mean (square root from the variance)
  - On an average how much each value is away from the mean
  - Formula:

$$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Visualising standard deviation





Obtaining summary statistics

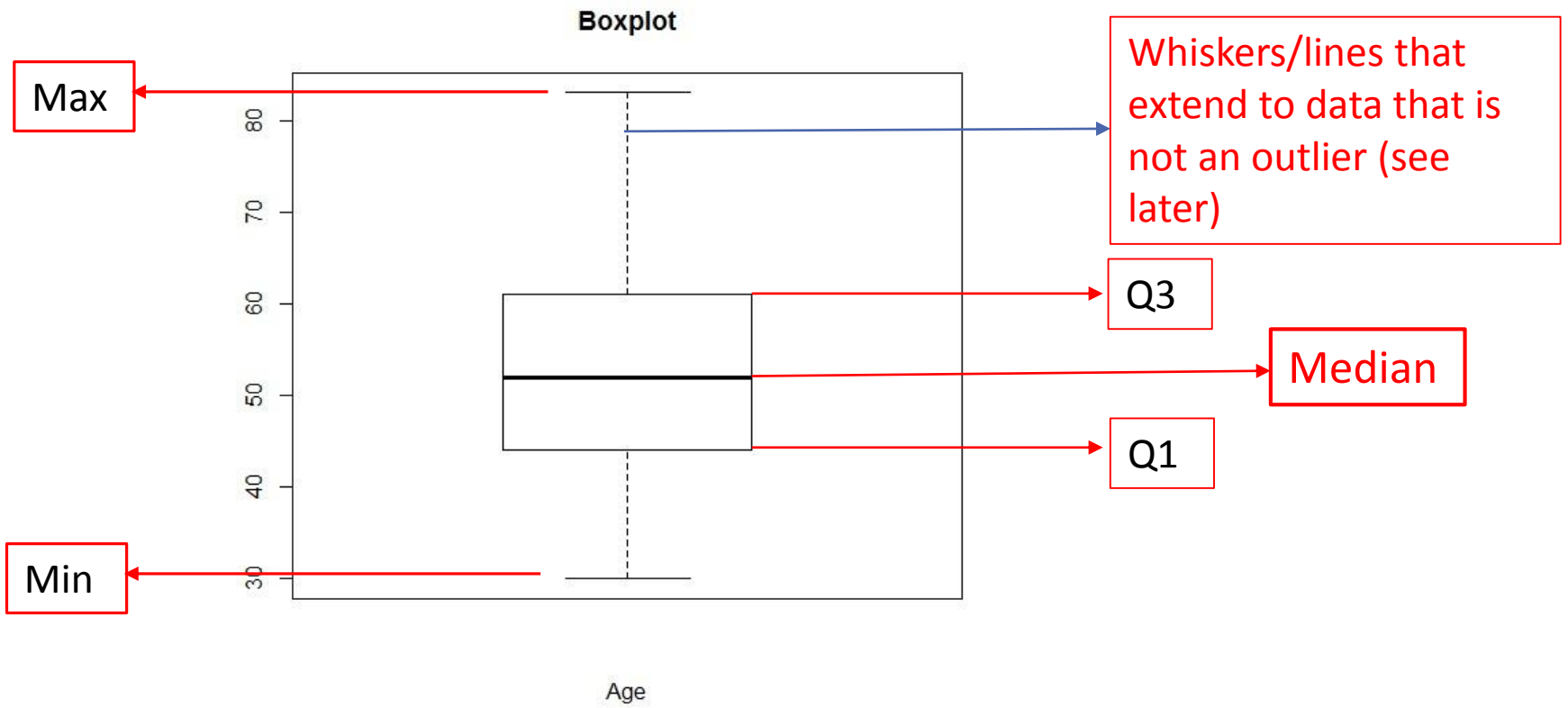
- See demo

Basic visual representation of one variable according to type of variable

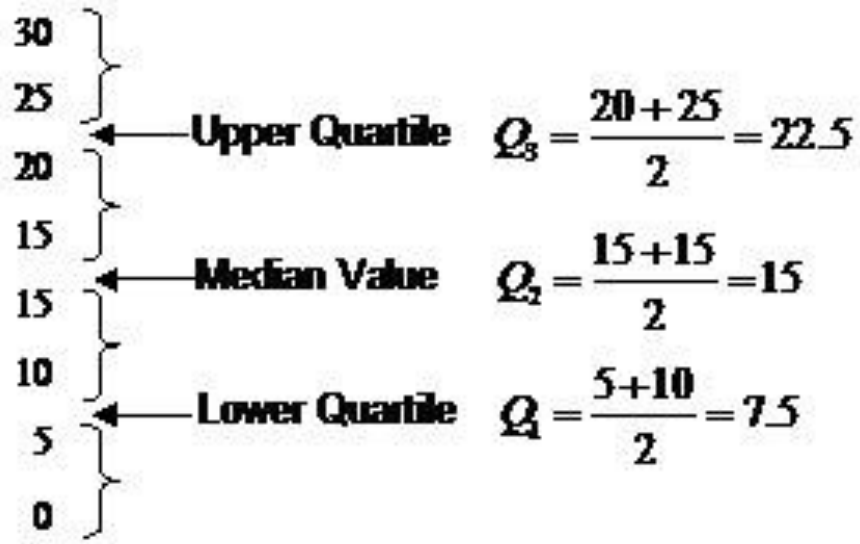
Type of variable	Graphical representation
Numeric variable	Histogram, density plot, box plot, stem and leaf display (small samples) and dot plots (small samples)
Categorical variable	Bar chart and pie chart

Quantitative variables: Graphical representation

- Boxplot: Graphical representation of the InterQuartile Range (IQR = Q3-Q1) of age of breast cancer patients who had undergone surgery



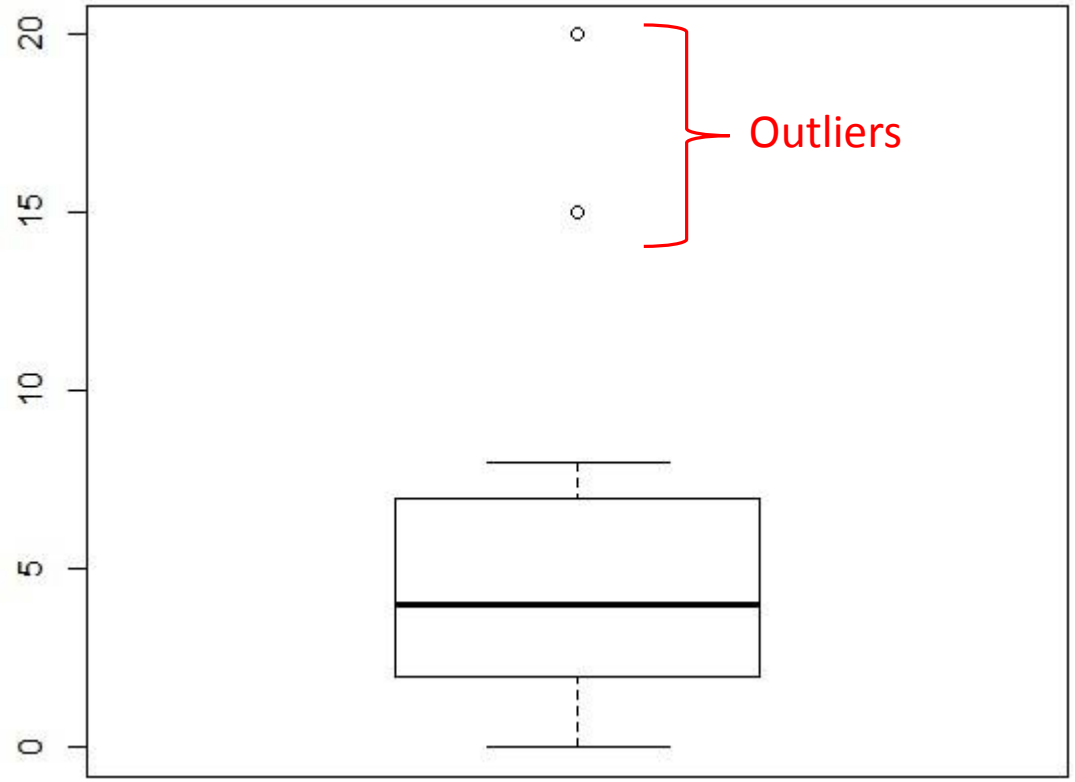
Quartiles


30
25
Upper Quartile
$Q_3 = \frac{20+25}{2} = 22.5$
20
15
Median Value
$Q_2 = \frac{15+15}{2} = 15$
15
10
Lower Quartile
$Q_1 = \frac{5+10}{2} = 7.5$
5
0

Outliers

- In general defined as more than (in absolute value) 1.5 times the IQR:
  - Less than 1.5 times Q1
  - More than 1.5 times Q3
- Consider following example:

```
x=c(0,4,15, 1, 6, 3, 20, 5, 8, 1, 3)
boxplot(x)
```



Basic visual representation of two variables according to the type of the variables

Type of first variable	Type of second variable	Graph
Numeric	Numeric	Scatterplot
Categorical	Numeric	Barplot using the mean of the continuous variable Boxplot per category of the discrete variable
Categorical	Categorical	Stacked or grouped barplot

Statistical tests/inference

- Questions such as:
  - Are housing prices different when located next to the river?
  - Is there a correlation between housing prices and the number of rooms?
  - According to origin of the car, is the mileage different?

We have no time to explore this but this is a next step

Statistical tests/inference (contd.)

- http://bama.ua.edu/~jleeper/627/choosestat.html
- http://www.ats.ucla.edu/stat/mult_pkg/whatstat/choosestat.html
- http://www.csun.edu/~amarenco/Fcs%20682/When%20to%20use%20what%20test.pdf

- Navarro, D. (2013), Learning statistics with R: A tutorial for psychology students and other beginners, http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/

Classification

- Supervised classification
- Unsupervised classification

We have no time to explore this but this is a next step and is also used in remote sensing

What next?

- Rmarkdown: http://rmarkdown.rstudio.com/
- Learn Github, sharing of R or any other code: https://github.com/
- Explore the R package ggplot2, sophisticated R plotting package: http://ggplot2.org/
- Spatial data in R (links in the material given on the pen drive)

Internet links

- R Manuals (go to “Manuals” on the left): https://cran.r-project.org/
- Contributed documents (CRAN): https://cran.r-project.org/other-docs.html
- Quick-R: http://www.statmethods.net/
- R Cookbook: http://www.cookbook-r.com/

Internet links (contd.)

- R-bloggers: http://www.r-bloggers.com/
- Twotutorials: http://www.twotorials.com/
- R tips: http://pj.freefaculty.org/R/Rtips.html
- R for beginners: http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

Internet links (contd.)

- UCLA R resources: http://www.ats.ucla.edu/stat/r/
- R tutorial: http://www.r-tutor.com/
- Short courses in R: http://courses.had.co.nz/

Blogs, discussion forums, etc

- R mailing lists: https://www.r-project.org/mail.html
- Stackoverflow R FAQ: http://stackoverflow.com/questions/tagged/r-faq
- R-bloggers: http://www.r-bloggers.com/

R reference cards

- Tom Short: Most used
  https://cran.r-project.org/doc/contrib/Short-refcard.pdf
- Jonathan Baron:
  https://cran.r-project.org/doc/contrib/refcard.pdf
- Vitto Rici: For regression analysis
  https://cran.r-project.org/doc/contrib/Ricci-refcard-regression.pdf

Overview of R packages

- CRAN Task Views: https://cran.r-project.org/web/views/
- Examples:
  - Analysis of ecological and environmental data: https://cran.r-project.org/web/views/Environmetrics.html
  - Design of experiments and analysis of experimental data: https://cran.r-project.org/web/views/ExperimentalDesign.html
  - Machine learning and statistical learning: https://cran.r-project.org/web/views/MachineLearning.html
  - Analysis of spatial data: https://cran.r-project.org/web/views/Spatial.html
  - Etc ..

Books

- Books given by the R website: https://www.r-project.org/doc/bib/R-books.html

- Books that I have:
  - Gardener, M. (2013), Beginning R, Wiley
  - Teetor, P. (2011), R Cookbook, O'Reilly Media
  - Adler, J. (2012), R in a nutshell, O'Reilly Media

Free courses in R and statistics (MOOCs)

- Coursera: https://www.coursera.org/
  - **Data science:** Good mix of theory and practice and as a series of 9 courses
  - **Data analysis and statistical inference**: Good course to understand some of the basics from a theory perspective and for doing statistical tests

  - **Other**

Free courses in R and statistics (MOOCs)

- **EdX: https://www.edx.org/**
- **Stat2.1X: Introduction to Statistics: Descriptive Statistics**
  - **Stat2.2x: Introduction to Statistics: Probability**
  - **Stat2.3x: Introduction to Statistics: Inference**
  - **The Analytics Edge**
  - **Other**

Thank you

ellenarun@gmail.com