QUANTITATIVE RESEARCH METHODS

DR. MEIKE MORREN

about myself

Background

- Academic experience
  - Bsc Sociology – Msc Social Sciences (University of Amsterdam)
  - PhD in methods & statistics (Tilburg University)
    - Latent class modeling
    - Cognitive interviewing
- Work experience: marketing research agency Centerdata
  - Online survey panel - non response
  - Project mobile surveys
- Teaching
  - Bachelor Marketing Research/bachelor Business Research Methods
- Interests
  - Modeling approaches & statistics
  - Survey responses

Contact data

- Office hours: by appointment

meike.morren@vu.nl

020 5982317

- Secretary:

Sandra van Arendonk (5A64)

sandra.van.arendonk@vu.nl

020-598 7145

COURSE OUTLINE

Course

- Requirements :
  - Attendence lectures (12)
  - Exercises during lecture
  - Study literature => articles & book excerpts
- Grade :
  - Assignments (6, 40%)
  - Exam (60%)

- You are allowed to miss one lecture
- Assignments are made in groups of 2 students, handed in the 11th of May via github

Literature

- Background literature on R (available in ub.vu.nl):
  - Everitt, B. S. & T. Hothorn (2011). An introduction to applied multivariate analysis with R. New York: Springer
- Articles
  - Link available via bb, search google.scholar.com
- Book excerpts
  - See https://github.com/meikemorren/BIS_QRM/issues

Lectures (1)

- Lecture 1: Introduction to R
- Lecture 2: Data cleaning, missing data analysis
- Lecture 3: Writing functions in R
- Lecture 4: Generalized linear framework
- Lecture 5: Simple and multiple regression
- Lecture 6: Categorical data analysis (logit regression)

Lectures (2)

- Lecture 7: PCA
- Lecture 8: Factor analysis
- Lecture 9: Cluster analysis
- Lecture 10: Plots
- Lecture 11: Multilevel analysis
- Lecture 12: webscraping / Q&A

Assignments

1. Analyze missing data
2. Conduct regression analysis
3. Conduct logit regression analysis
4. Analyze multi-item scale(s) using CFA
5. Apply and interpret cluster analysis
6. Compare countries or multilevel regression analysis

Assignment guidelines

- Questions to interpret the results you obtain with programming
- Deadline: 11 May (you can hand in earlier)
- Github

The assignments include:

- Programming (include your code)
- Interpretation of results (focus on argumentation)

Assignment dataset 1

Bring your own data (if suitable for assignment) or

use data that is provided for (see bb)

World Values Survey dataset includes:

- Green attitudes
- Work attitudes
- Schwarz values

Across 40+ countries

For assignment 1 (missing data), 2 (regression), 4 (factor analysis) and 6 (multi-level analysis)

Assignment dataset 2

Bring your own data (if suitable for assignment) or

use data that is provided for (see bb)

Tablet dataset which includes:

- Sales ranking
- Number of reviews
- Number of characteristics of tablets (screen size, weight, battery life etc)

For assignment 3 (logit regression) and 4 (cluster analysis)

Assignments via github

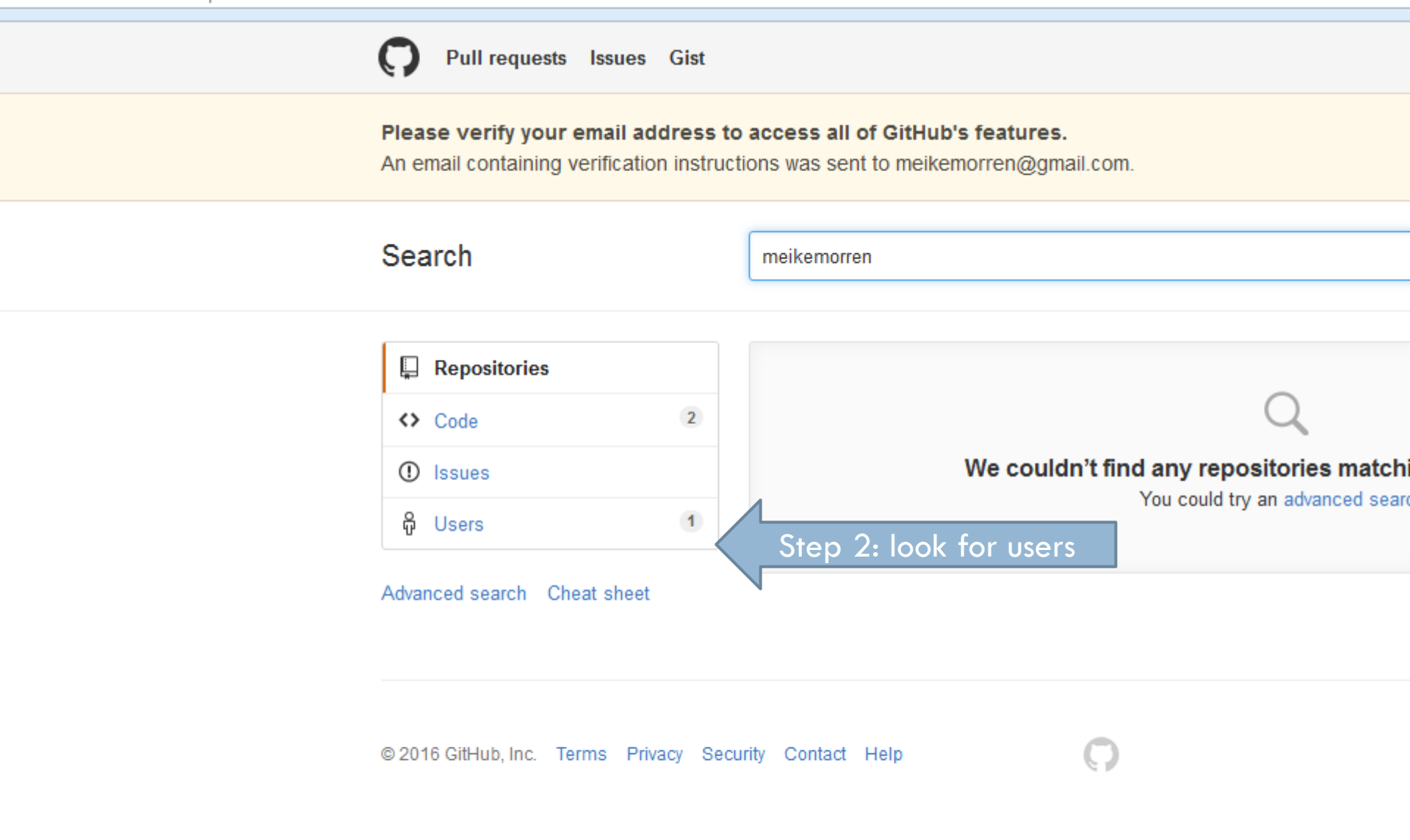
- Create an account (github.com)
  - send me your username
- Fork my github repository
  - (= copy the files to your account)
- Download the files to a location in which you will work on your exercises / assignments
- Upload your answers to github

https://github.com/meikemorren/BIS_QRM.git

GitHub

- Free
- Used for coders to work simultaneously on projects

GitHub


Pull requests
Issues
Gist
Please verify your email address to access all of GitHub's features.
An email containing verification instructions was sent to meikemorren@gmail.com.
Search
meikemorren
Repositories
Code 2
Issues
Users 1
We couldn't find any repositories matchi
You could try an advanced searc
Step 2: look for users
Advanced search
Cheat sheet
© 2016 GitHub, Inc.
Terms
Privacy
Security
Contact
Help

LECTURE 1

Contents lecture 1

- Introduction to modeling, and the concepts we will use in this lecture:
  - Introduction to R
    - Data types
    - Explore dataset
  - Scale types & distributions

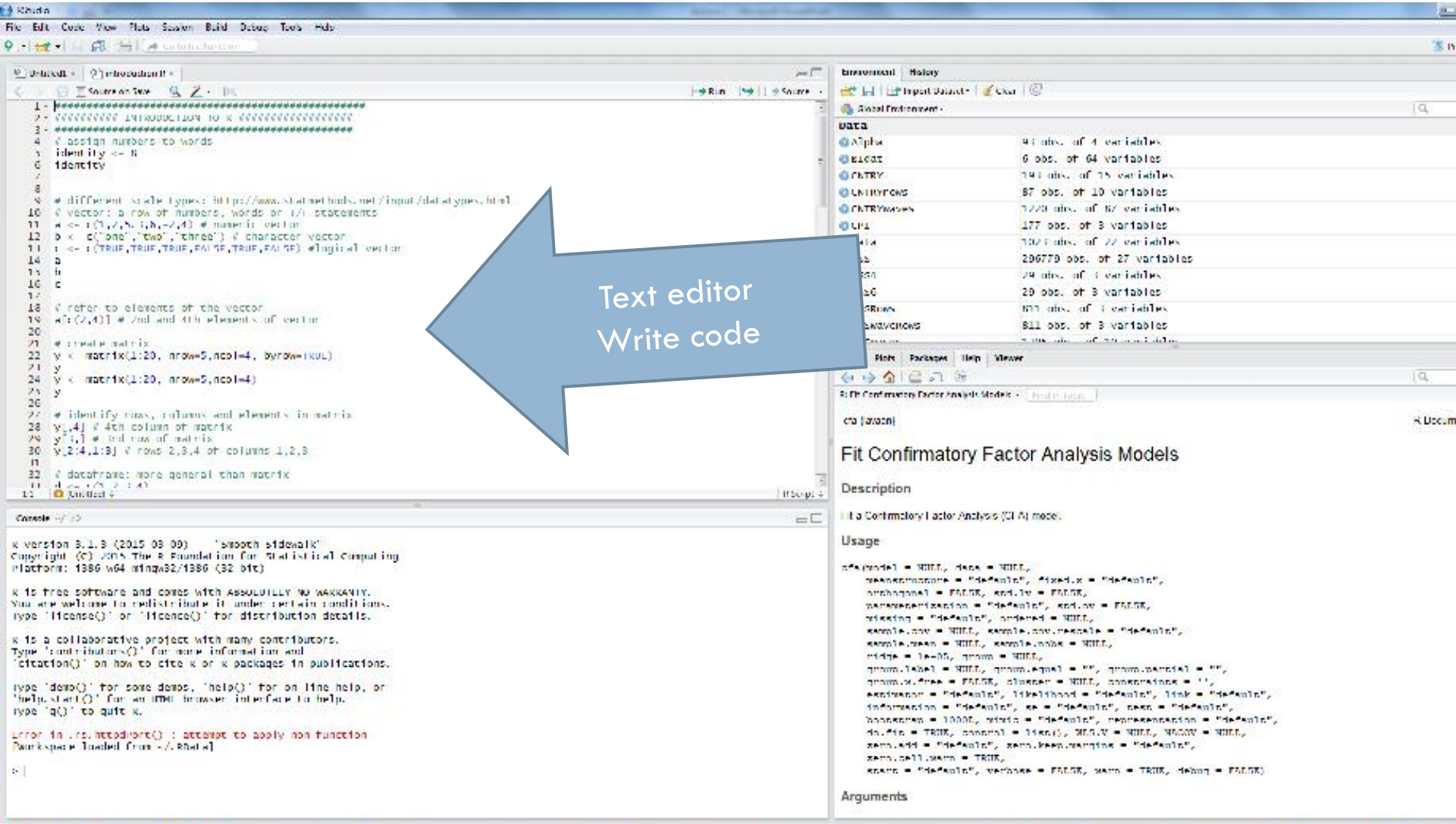
R (STUDIO)

Assignment software

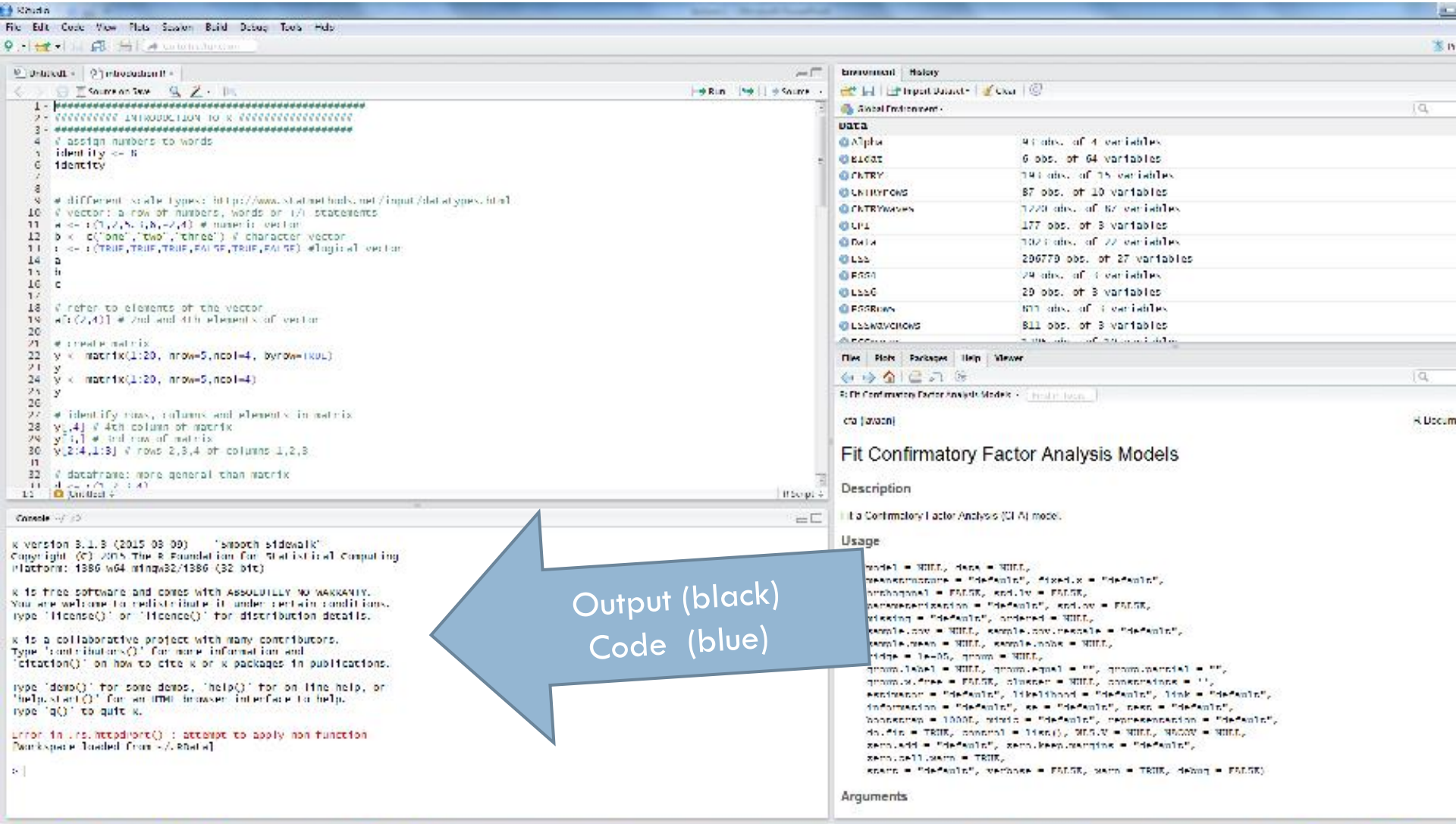
- Download R : http://cran.xl-mirror.nl/
  - Install R for the first time
  - Run exe

- R Studio : http://www.rstudio.com/products/RStudio/download/
  - Choose installers
  - Select platform

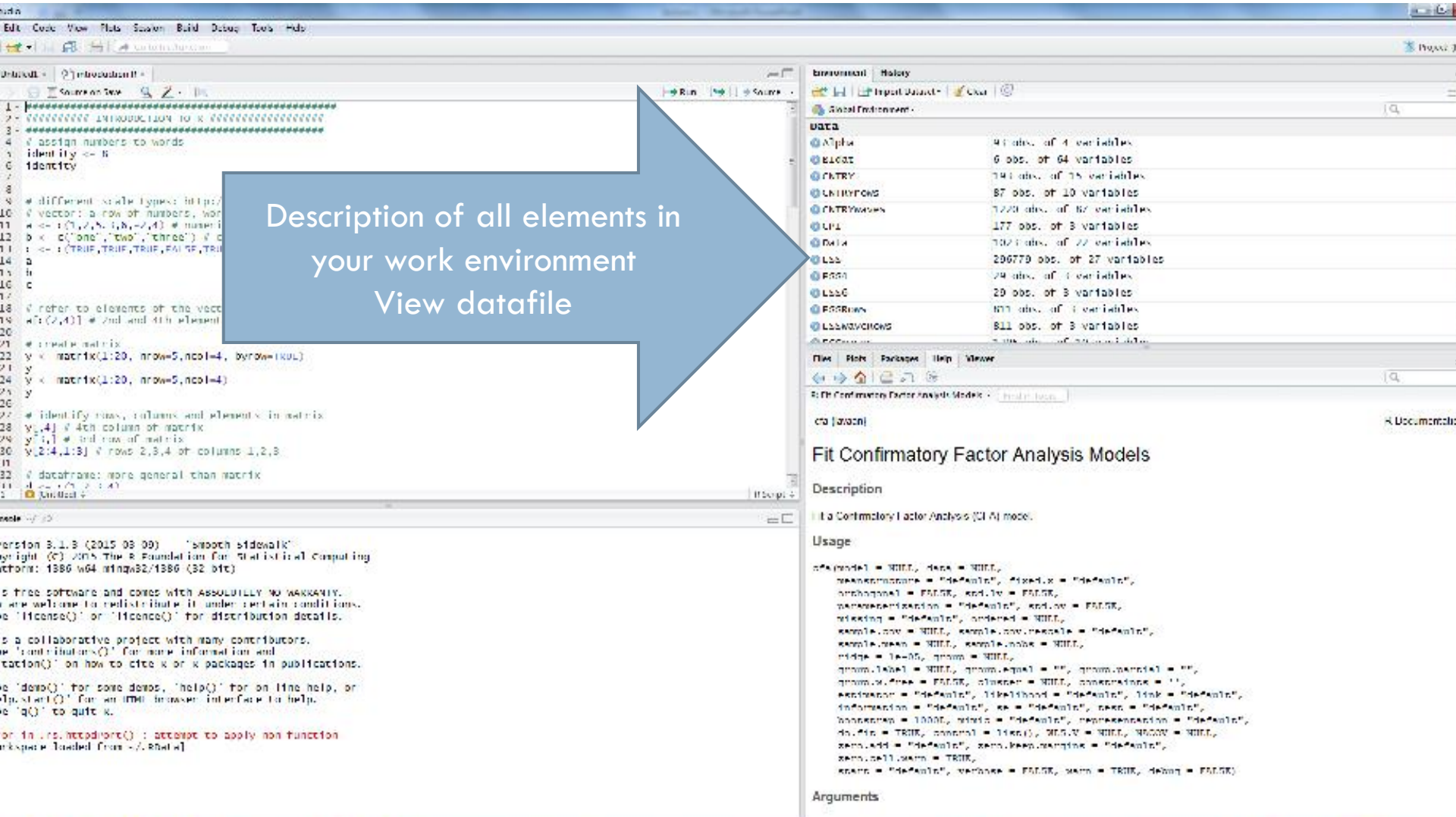
Work environment R studio



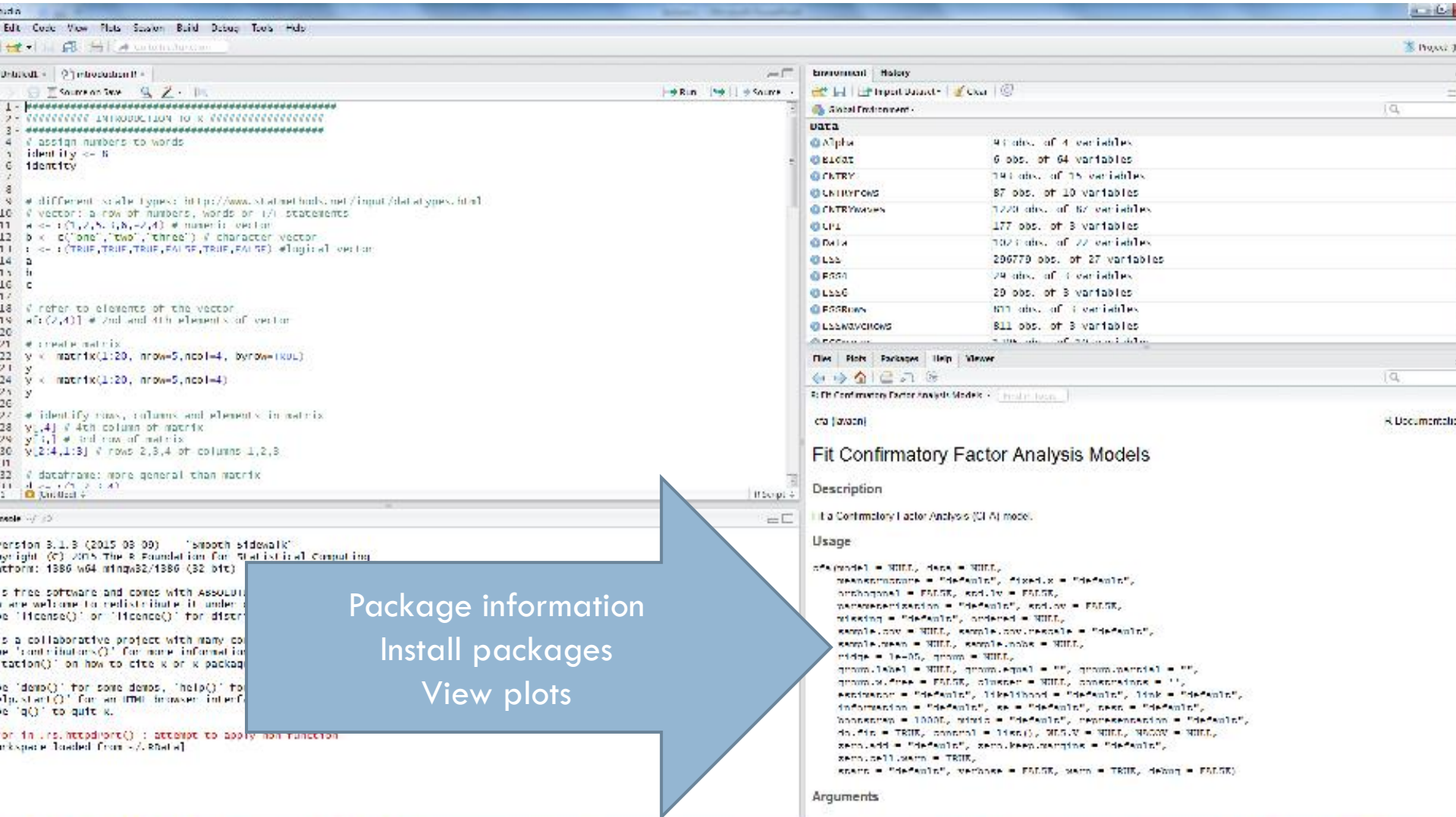
Work environment R studio



Work environment R studio



Work environment R studio



READ DATA INTO R

SPSS dataset

- Is considered to be a matrix in R:
  - Consists of variables (columns)
  - And observations (rows)
- However, variables can be of different datatypes
  - Character vectors
  - Numeric vectors
  - …
- Therefore, data is a dataframe

Useful functions

- read.table

- nrow
- ncol
- colnames

- typeof

DATA TYPES

Data types

- Integers (no decimals)
- Numeric (with decimals)
- Characters (words)
- Complex
- Logical (true, false)

DATA OBJECTS

Data objects

- Dataframes
- Factors
- Vectors
- Matrices

The base type in R is vector (not scalar)

These are the abstract data types, also called class

Mode

Data objects are stored in the memory by mode

- Numeric
- Complex
- Character
- Logical
- *List, function*

Changing the mode of an object is often called 'coercion'. The mode of an object can change without necessarily changing the class.

R MANIPULATION

Create vectors, and select from vectors

Create a vector

- Everything is an object
- You can assign a value to an object
  - `R <- 4`
- Create a vector
  - `Vec <- rep(1,4)` # of four 1s
  - `Vec <- (1:4)` # 1 thru 4
- Assign new value to third element
  - `Vec[3]<-5`

Select from vector (matrix)

- Create matrix
  - `df <- rbind(c(1,2,3),c(4,4,4))` # create matrix by row
  - `df <- cbind(c(1,2,3),c(4,4,4))` # create matrix by column
  - `df <- matrix(c(1,2,3,4,4,4), nrow=2, ncol=3, byrow=TRUE)`
- Select element, row, column
  - `df[2,3]` # select second row, third element
  - `df[,3]` # select third column
  - `df[2,]` # select second row

Select from data

- `df[df$gender==1,]` # select all data for males
- `df$job[df$gender==1]` # select only job for males
- `df$job[df$edu>3]` # select only job for edu higher than level 3

- For data levels see questionnaire

Accessing attributes

- Attributes = characteristics of a vector
  - `length(Vec)`
  - `sort(Vec)`
  - `names(Vec)`
- Assign colnames & select second
  - `colnames(df) <- c("first", "second", "third")`
  - `colnames(df)[2]`
- Get first impression
  - `summary(Vec)`

R as calculator

- Basic operations
  - `x <- rnorm(10)` # create vector of 10 random numbers from normal distribution
  - `x + 1`
  - `v <- x * 8 +1`
  - `v/5`

- Arithmetic functions
  - `mean(Vec)`
  - `sum(Vec)`
  - `var(Vec)/ sd(Vec)`

Some manipulation signs

```
x<-5; y <- 1:10
```

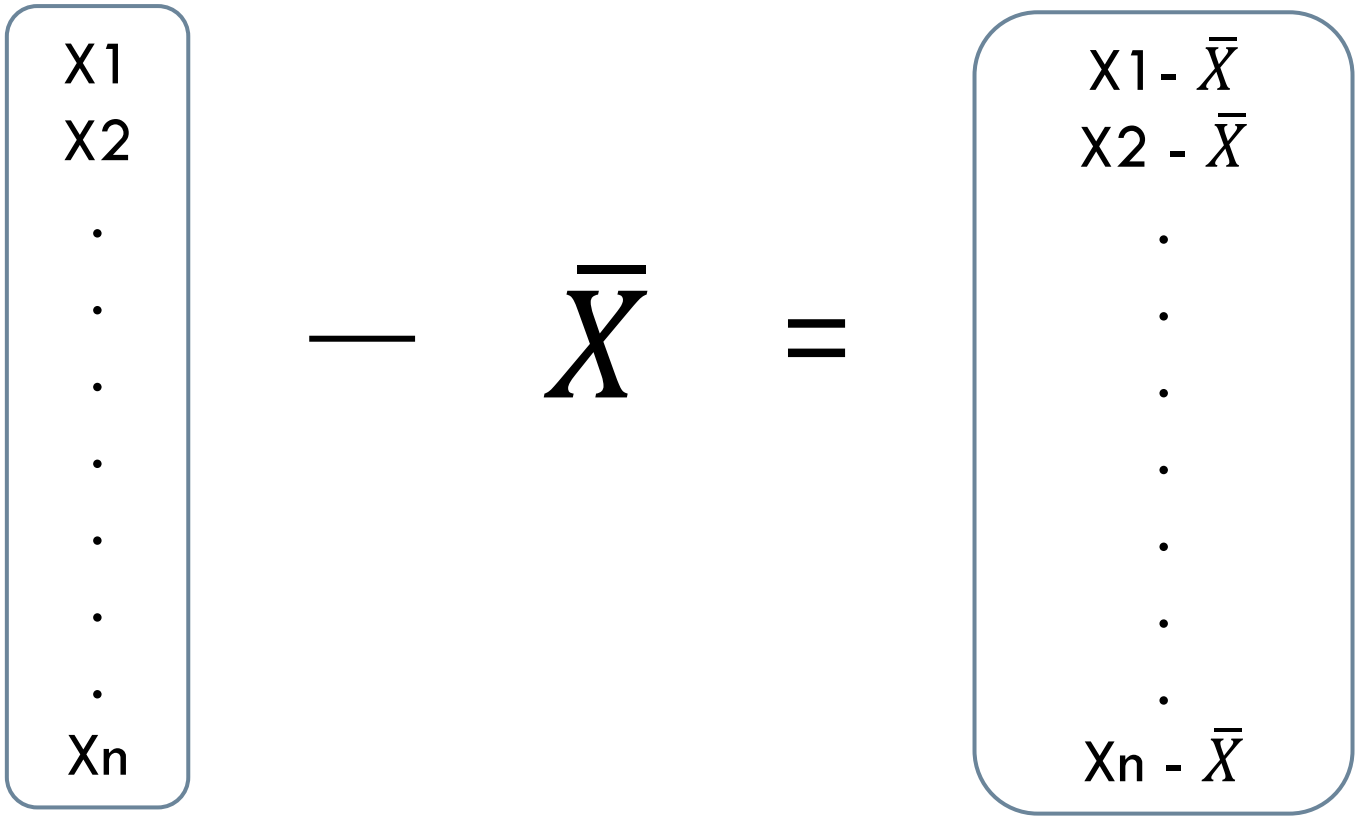
- Square
  - `x^2` or `x**2`
- Square root
  - `sqrt(x)`
- Equal to
  - `y==x`
- Not equal to
  - `y!=x`
- Not x
  - `!x`
- Smaller than
  - `x<=y`
- Greater than
  - `x>=y`
- Equal to
  - `y==x`
- x OR y
  - `x | y`
- x AND y
  - `x & y`

Simple calculations

- Calculate variance on vector
  - `var(x)`
  - `sum((x-mean(x))^2/(length(x)-1)`
- Calculate z-score
  - `mean(x)`
  - `sum(x)/length(x)`
- Calculate z-score
  - `Standard deviation: sd(x)`
  - `(x-mean(x))/sd(x)`
  - `(x-mean(x))/sqrt(var(x))`

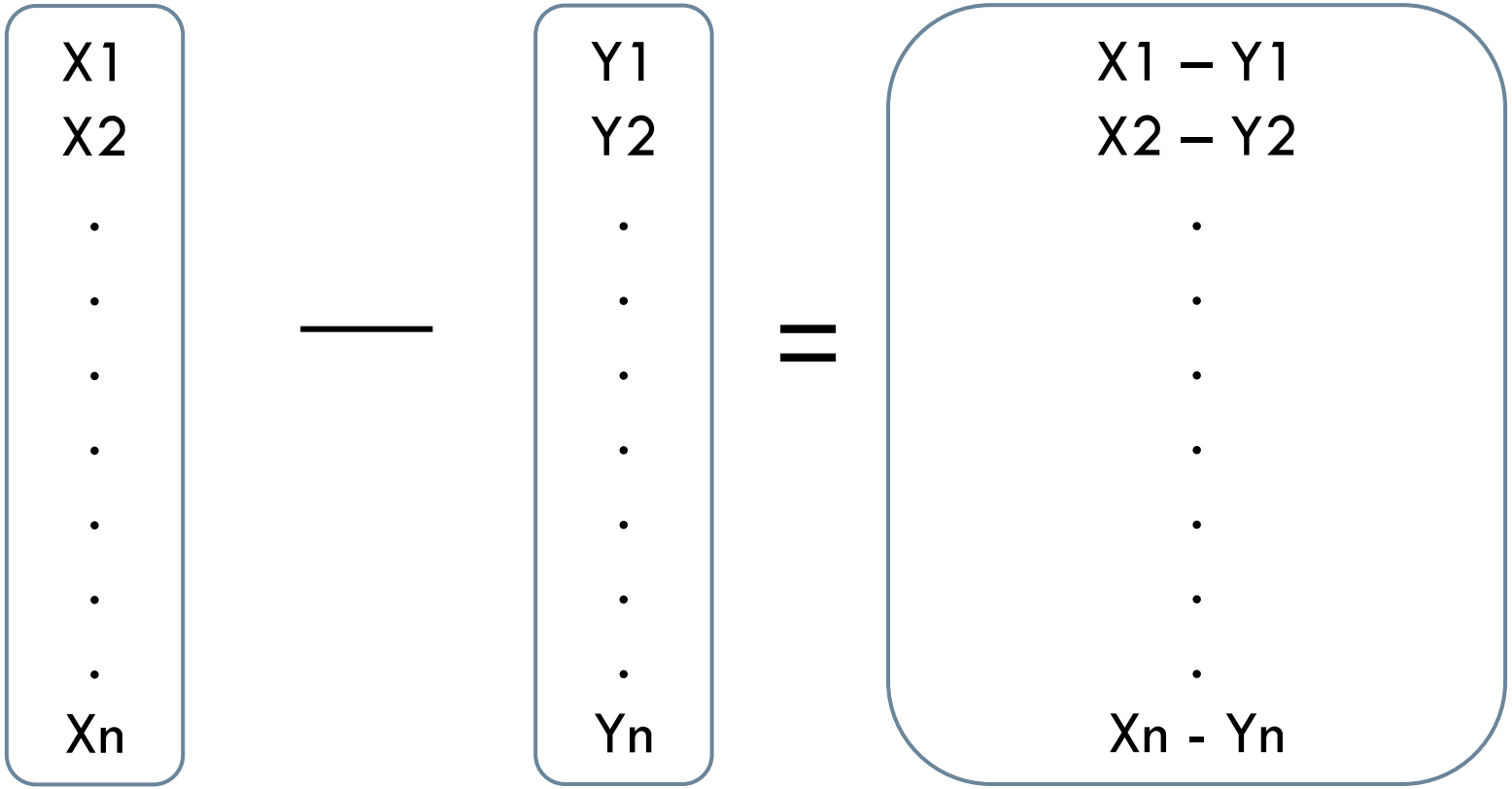
Algebra (1)

- `x-mean(x)`:



Algebra (2)

- y <- rnorm(10)
- x - y



R exercise (1)

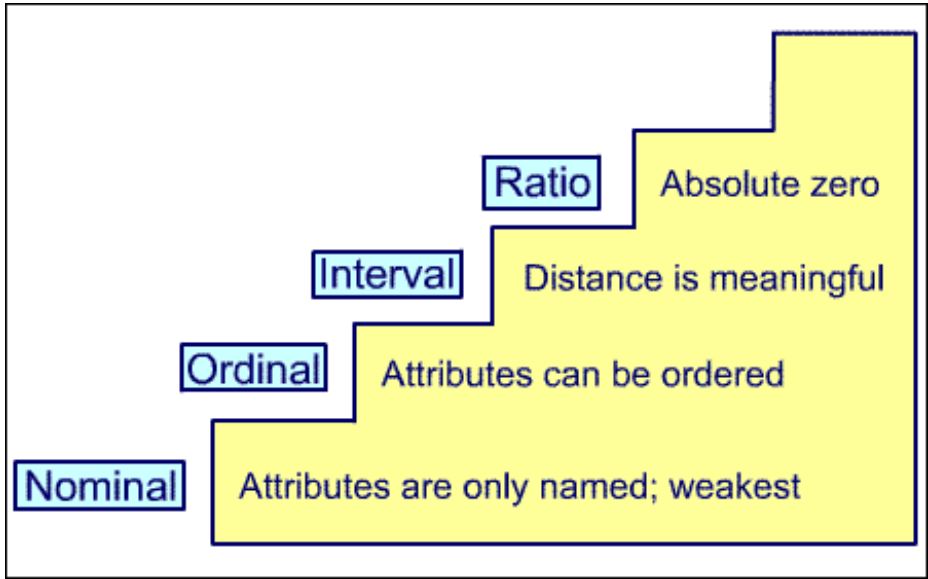
see blackboard/documents/assignment:

- Open exe1_1.r
- Download data into R
  - set directory, read.table, give a name to the dataset
- Inspect dataset
  - Explore elements, rows, columns
  - Assign names to columns
- Manipulate data
  - Calculate variance, z-score (first manually, check with R function)

SCALE TYPES

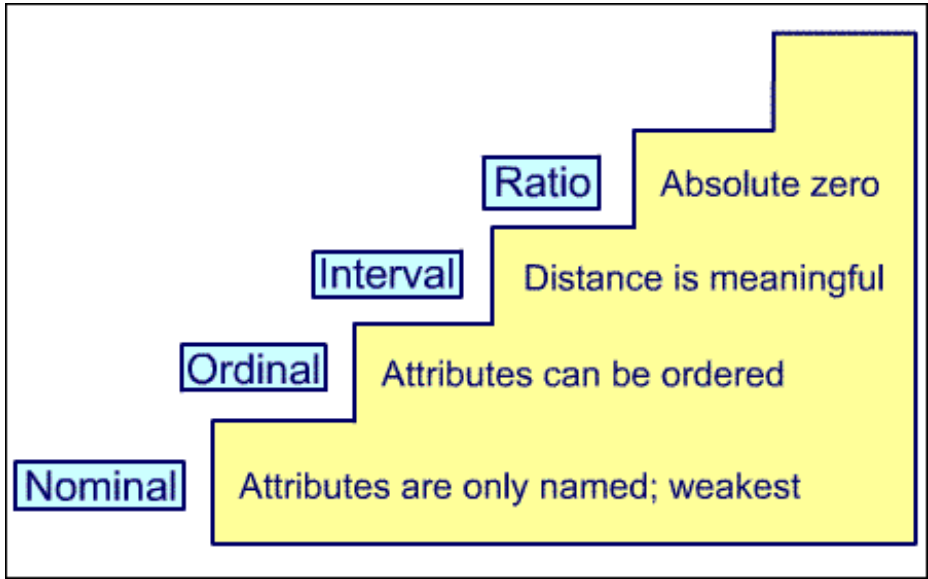
Primary scale types (1)

- Nominal
  - Categories have labels
  - Numbers no meaning
- Ordinal
  - + ordering
  - Numbers arbitrary, only order numbers has meaning
- Interval
  - + equal distances
  - Starting point arbitrary, order and distances between numbers have meaning
- Ratio
  - + starting point
  - Numbers have meaning in all mathematical senses



Primary scale types (2)

- Nominal
  - Count frequency
  - Chi square, binominal test
- Ordinal
  - rank order correlation
  - ANOVA (Kruskal-Wallis)
  - Kolmogorov-Smirnov test
- Interval
  - T-test
  - ANOVA
  - Regression
  - Factor analysis
- Ratio
  - All mathematical operations are possible



Likert Scale

Itemised scale:
Number or label to describe categories

Ordered categories

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Would you agree to the statement that muffins are more healthy than cupcakes?	○ 1	○ 2	○ 3	○ 4	○ 5

R functions to inspect variables

- Run frequencies
  - `table(df[,1])`
  - `df<-as.data.frame(df)`
  - `table(df$first)`
- Run descriptives
  - `summary(df$first)`
- Add variable labels (change into factor)
  - `dataSet$gender <-factor(dataSet$gender,levels = c(-2, 1,2), labels=c("missing","male","female"))`
  - `attributes(dataSet$gender)`

R exercise (2): Open exe1_2.r

- Run frequencies
  - `table`
- Descriptives
  - `summary, mean`
- Crosstable
  - `table(var1, var2)`
- Recode
  - `df[df==4]<-2` # recode all values of 4 to 2
  - `df[df$first==4,] <-2` # recode values of 4 in first column to 2
- Convert variables to factors
  - `factor(df$first, levels=c(1,2),`
    `labels=c("one", "two"))`

PROBABILITY DENSITY FUNCTIONS

Distributions

Used to

- describe variables
- test hypotheses
- estimate model parameters

Many distributions

- Normal
  - Continuous variables
- Logistic
  - Nominal, multi-nominal, or ordinal variables
- Binomial
  - Nominal variables
- Poisson
  - Count variables

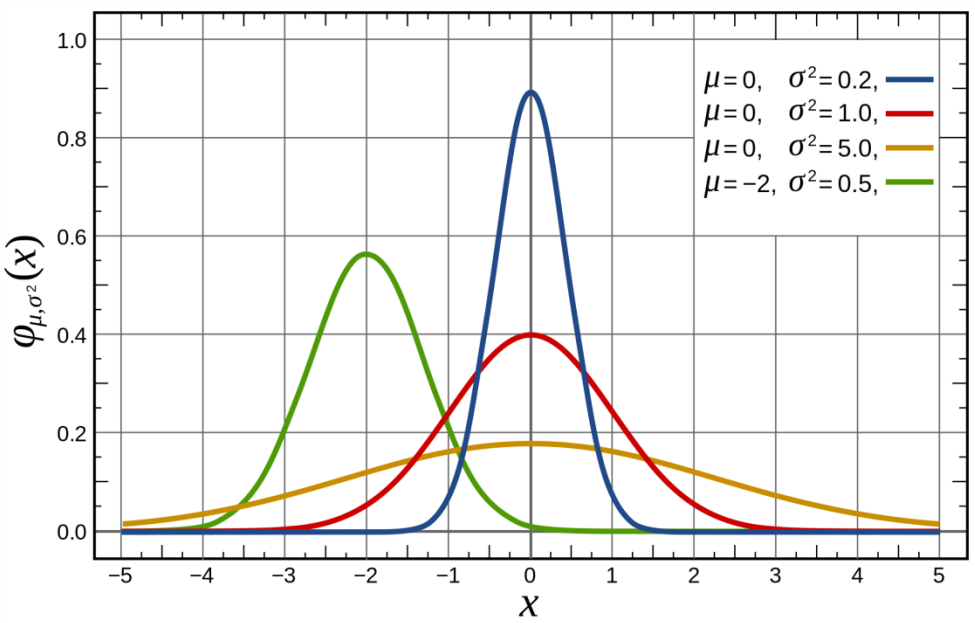
Normal distribution

- Interval / ratio variable results in normal distribution

μ= mean(location parameter)

σ = standard deviation

(scale parameter)



- PDF:

$$f(x;\sigma,\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Probability density function (PDF) describes the area under the curve

Central limit theorem

- Random variables
- Independently drawn from
- Independent distributions
- i.d.d.

- Converge to normal distribution when number of variables is large

- Mean of 0
- Standard deviation of 1

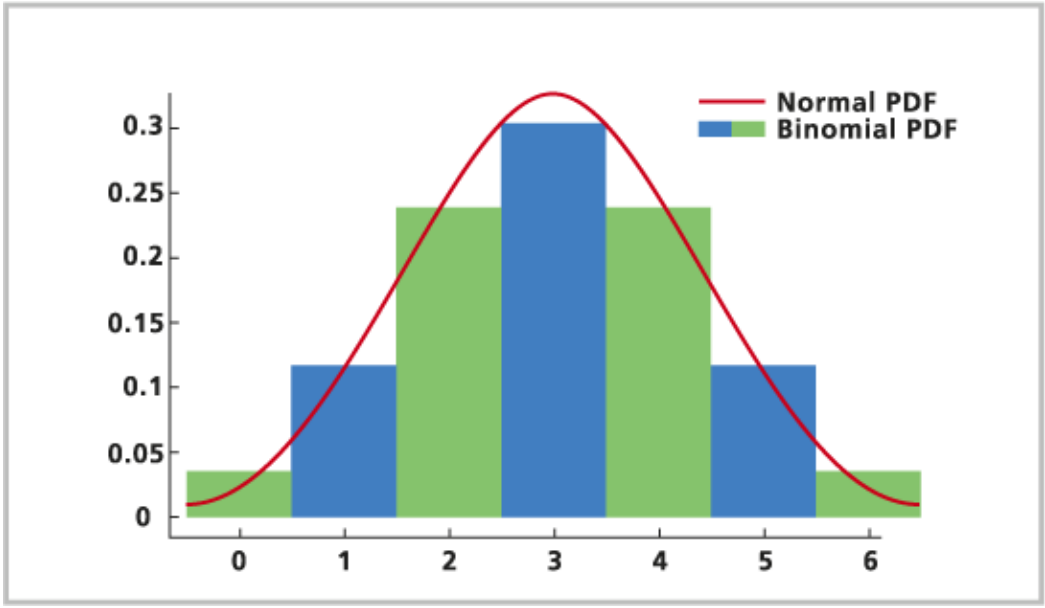
- Total area under the curve is 1

Binomial distribution

- PMF:

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for x = 1,2,…

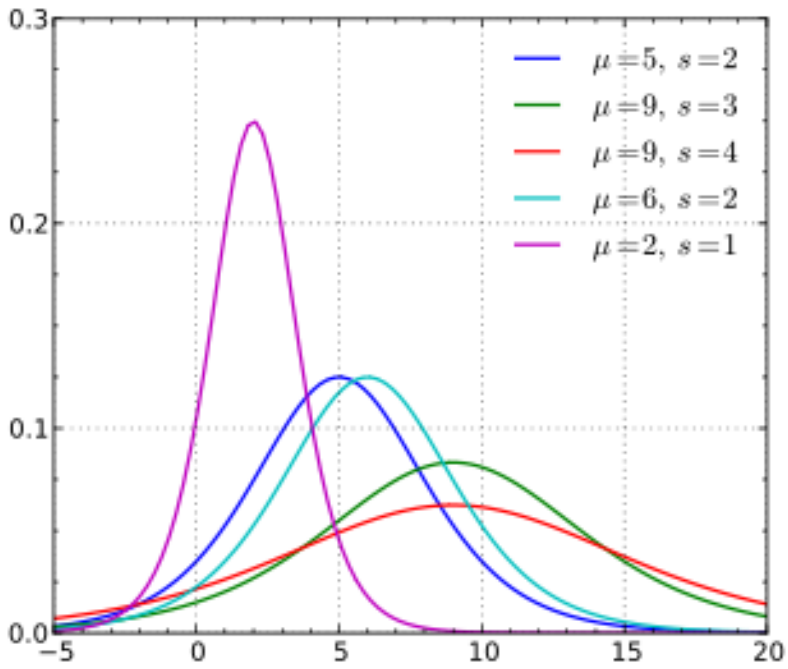


- k= # successes
- n= # trials
- p= probability
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Independent successes/failures (=with replacement)

Logistic distribution

- The logistic distribution has slightly longer tails compared to the normal distribution
  - μ = mean (location parameter)
  - s = variance (scale parameter)



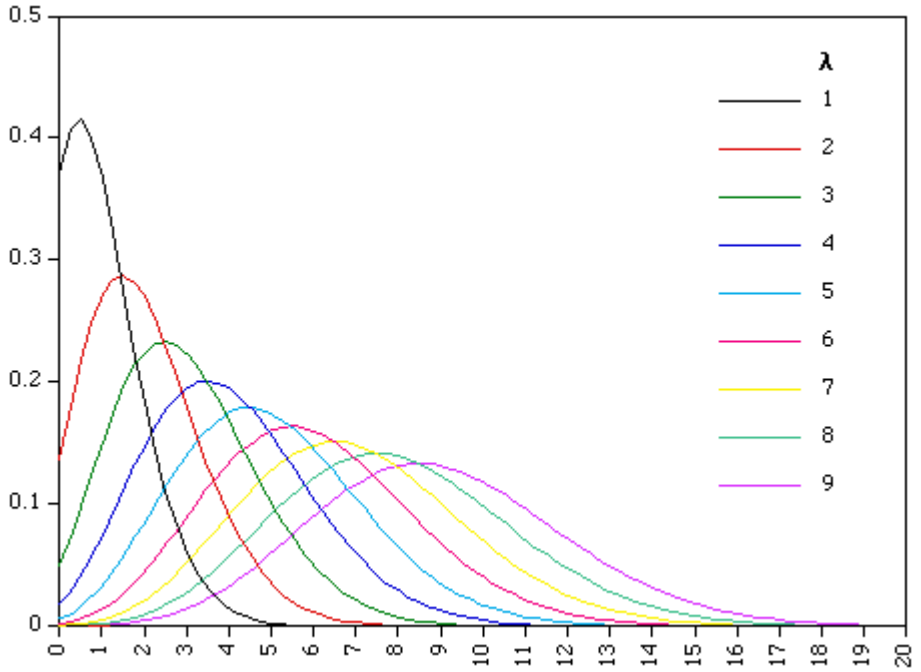
- PDF:

$$f(x;\mu,s)=\frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2}$$

Poisson distribution

- PDF:

$$f(x;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



- λ = mean
- λ = variance

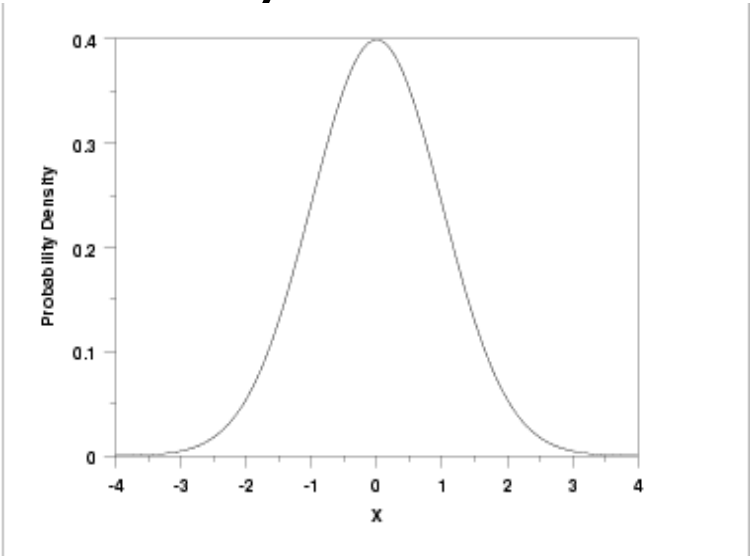
Summary

Distribution	Mean	Variance
Normal	μ	σ^2
Logistic	μ	$s^2 \frac{\pi^2}{3}$
Binomial	np	np(1-p)
Bernoulli	p	p(1-p)
Poisson	λ	λ

WHY DISTRIBUTION FUNCTIONS?

Probability density function

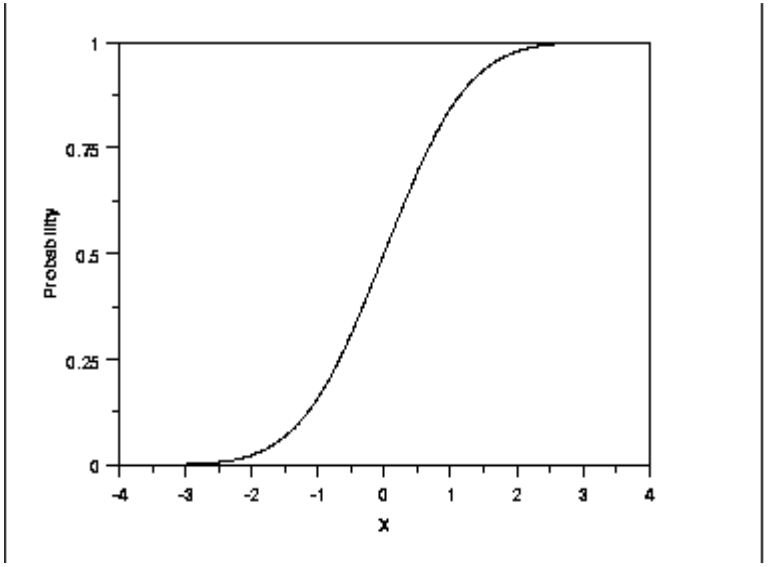
- Assumptions of a distribution allow you to estimate
- the probability of a random variable x



- A large p(x) : random variable x is close to X
- Probability of exact value is zero
- Nonnegative numbers

Cumulative distribution function

- Cumulative distribution allows you to estimate whether random variable x is lower than or equal to value X



Probability distribution in R

“d”	returns the height of the probability density function
“p”	returns the cumulative density function
“q”	returns the inverse cumulative density function (quantiles)
“r”	returns randomly generated numbers

Various distributions in R

- Create normally distributed variable :
  - `y<-rnorm(n=100)`
- Create binomially distributed variable:
  - `ybi <-rbinom(n=100,size=10,prob=.2)`
- Create chisquared distributed variable:
  - `ychi <- rchisq(n=100, df=2)`
- Create f distributed variable:
  - `yf <- rf(n=100, df1=2, df2=4)`

PLOT YOUR DATA

First data inspection

- Basic plots available in R
- More advanced functions in ggplot2 (we'll discuss later)

Types of plots

- Barplot
  - a barplot is especially useful for nominal or ordinal variables and shows the frequencies of the various categories
- Histogram
  - a histogram is useful to show the distribution of a interval or continuous variable. This illustrates whether the variable has a normal or other distribution
- Boxplot
  - a boxplot is useful to visually detect outliers
  - especially appropriate with interval or continuous variables which is split by groups

R exercise (exe1_3.r)

- Plot (first create variable)
  - Plot one variable
  - Plot two variables
- Barplot
  - Nominal / ordinal/ multinomial variables
- Histogram
  - Distribution of interval/continuous variables
- Boxplot (first: create variables with certain distribution)
  - `new <-rnorm(nrow(df))` # create normal variable

Next lecture

- Writing functions in R