

Course 02402 Introduction to Statistics

Lecture 1: Introduction and R

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Agenda

- 1 Practical course information
- 2 Introduction to Statistics - a primer
- 3 Statistics and Engineers
- 4 Descriptive Statistics
 - Mean and median
 - Variance and standard deviation
 - Percentiles and quantiles
 - Covariance and correlation
- 5 Software: R & RStudio

Overview

- 1 Practical course information
- 2 Introduction to Statistics - a primer
- 3 Statistics and Engineers
- 4 Descriptive Statistics
 - Mean and median
 - Variance and standard deviation
 - Percentiles and quantiles
 - Covariance and correlation
- 5 Software: R & RStudio

Practical course information

• Teaching module

- Lectures: Tuesday 13-15 at Zoom Channel introstat_F21_02402
- Exercises: Tuesday 15-17 at
 - Zoom Channels. See Email or Course Website.
 - Quizzes and Area9 Rhapsode

• Exam

- Sunday 30 May 2021 9am-1pm
- 4 hour multiple choice

• Mandatory projects

- 2 projects must be approved in order to participate in the exam.
- For each project, choose between one of four topics.

Practical course information

• Generic weekly agenda

- Before teaching: Read relevant chapters/sections in eNote/book
- Lectures: 2 hours, curriculum of the week
- Exercises: 2 hours, exercises and online quizzes
- After teaching: Online "exam quiz" (test yourself)

• Teaching material

- Available under *Material* on course website
- Optional: Use the available R script to download all the material in one go (but beware that it overwrites changes made to previously downloaded files)
- Lecture slides and R code will be updated shortly before each lecture (remember to refresh browser)

Practical Information

- Homepage: 02402.compute.dtu.dk

- Online book (website or via [lix](#))
- Syllabus
- Lecture plan / agenda
- Exercises & solutions
- Slides
- Discussion forum
- Podcasts of previous years' lectures (English and Danish)
- Quizzes

- Learn:
learn.inside.dtu.dk/d21/home/60261

- Announcements (Must subscribe to get email)
- Projects - description and submission

Overview

- 1 Practical course information
- 2 Introduction to Statistics - a primer
- 3 Statistics and Engineers
- 4 Descriptive Statistics
 - Mean and median
 - Variance and standard deviation
 - Percentiles and quantiles
 - Covariance and correlation
- 5 Software: R & RStudio

Introduction to Statistics - a primer

New England Journal of Medicine:

EDITORIAL: Looking Back on the Millennium in Medicine,
N Engl J Med, 342:42-49, January 6, 2000.

<http://www.nejm.org/doi/full/10.1056/NEJM200001063420108>

Millennium list

- Elucidation of human anatomy and physiology
- Discovery of cells and their substructures
- Elucidation of the chemistry of life
- **Application of statistics to medicine**
- Development of anesthesia
- Discovery of the relation of microbes to disease
- Elucidation of inheritance and genetics
- Knowledge of the immune system
- Development of body imaging
- Discovery of antimicrobial agents
- Development of molecular pharmacotherapy

James Lind

One of the earliest clinical trials took place in 1747, when James Lind treated 12 scorbutic ship passengers with cider, an elixir of vitriol, vinegar, sea water, oranges and lemons, or an electuary recommended by the ship's surgeon. The success of the citrus-containing treatment eventually led the British Admiralty to mandate the provision of lime juice to all sailors, thereby eliminating scurvy from the navy.

(See also http://en.wikipedia.org/wiki/James_Lind).



John Snow

The origin of modern epidemiology is often traced to 1854, when John Snow demonstrated the transmission of cholera from contaminated water by analyzing disease rates among citizens served by the Broad Street Pump in London's Golden Square. He arrested the further spread of the disease by removing the pump handle from the polluted well.

(See also [http://en.wikipedia.org/wiki/John_Snow_\(physician\)](http://en.wikipedia.org/wiki/John_Snow_(physician))).



Google - *Big Data*

A quote from the New York Times article titled *For Today's Graduate, Just One Word: Statistics* (5 August 2009)

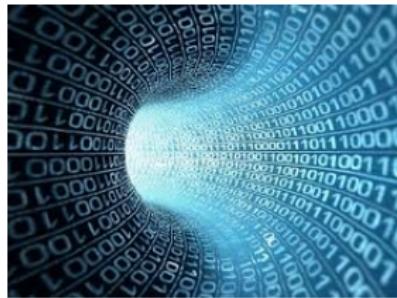
<http://www.nytimes.com/2009/08/06/technology/06stats.html>

I keep saying that the sexy job in the next 10 years will be statisticians, said Hal Varian, chief economist at Google. And I'm not kidding.



IBM - *Big Data*

The key is to let computers do what they are good at, which is trawling these massive data sets for something that is mathematically odd, said Daniel Gruhl, an I.B.M. researcher whose recent work includes mining medical data to improve treatment. And that makes it easier for humans to do what they are good at - explain those anomalies.



Intro Case stories:

IBM big data, Novo Nordisk small data, Skive fjord

- Presentation by Senior Scientist Hanne Refsgaard, Novo Nordisk A/S
- IBM Social Media podcast by Henrik H. Eliassen, IBM.
- Skive Fjord podcasts, by Jan K. Møller, DTU.

Overview

- 1 Practical course information
- 2 Introduction to Statistics - a primer
- 3 Statistics and Engineers
- 4 Descriptive Statistics
 - Mean and median
 - Variance and standard deviation
 - Percentiles and quantiles
 - Covariance and correlation
- 5 Software: R & RStudio

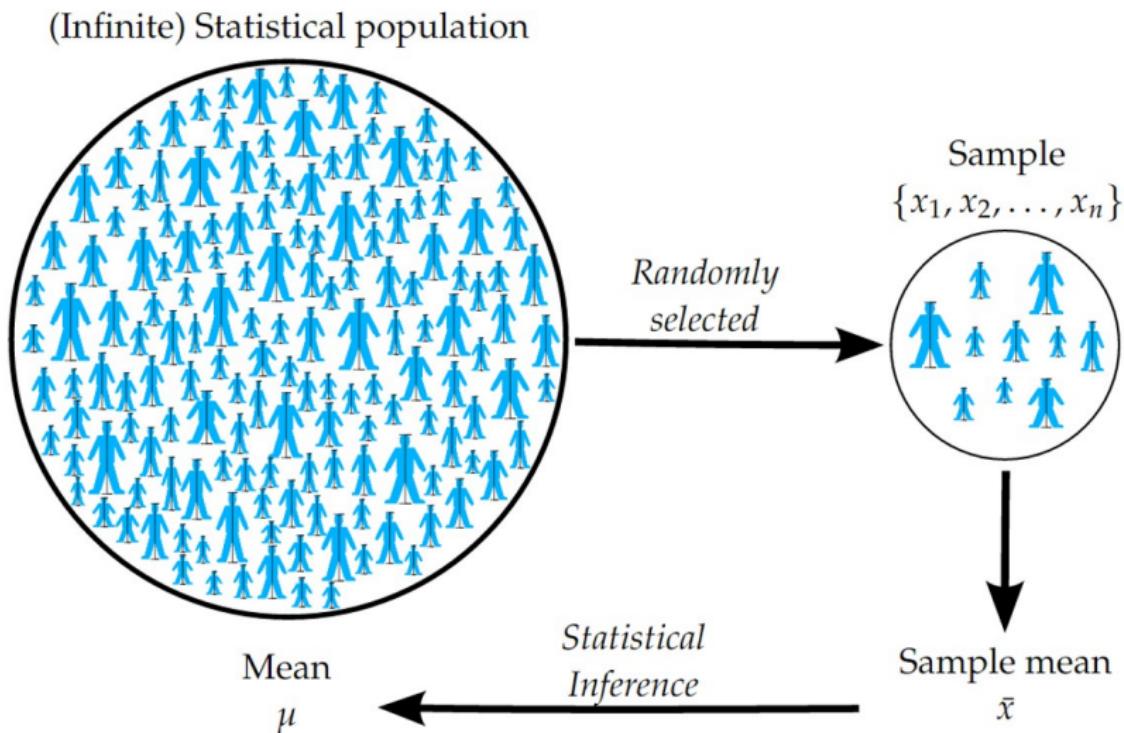
Statistics and Engineers

- Analysis of data ("both small & big")
- Understanding random variation
- Understanding the advantages (and limitations) of statistics for problem solving
- Quality improvement
- Design of experiments
- Prediction of future values
- ... and much more!

Statistics

- Descriptive statistics vs. *statistical inference*
- Statistics is often about analyzing a *sample*, taken from a *population*.
- Based on the sample, we try to generalize to the population.
- Therefore, it is important that the sample is *representative* of the population.

Statistics



Overview

- 1 Practical course information
- 2 Introduction to Statistics - a primer
- 3 Statistics and Engineers
- 4 Descriptive Statistics
 - Mean and median
 - Variance and standard deviation
 - Percentiles and quantiles
 - Covariance and correlation
- 5 Software: R & RStudio

Summary statistics

We use *summary statistics* to summarize and describe data (stochastic variables)

- Measures of centrality
 - e.g.: mean (\bar{x}) and median
- Measures of dispersion
 - e.g.: variance (s^2) and standard deviation (s)
- Measures of relation
 - e.g.: covariance and correlation

Note the difference between, e.g., the (*sample*) mean \bar{x} and the (*population*) mean μ .

Mean, Definition 1.4

The mean value is a key number which indicates the centre of gravity or centering of the data.

The sample mean (average):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

We say that \bar{x} is an *estimate* of the population mean.

Median, Definition 1.5

The median is also a key number indicating the center of the data.

In some cases, for example in the case of extreme values, the median is preferable to the mean.

Sample median:

The observation in the middle (in sorted order).

Example: Student heights

- **Sample:** Student heights in cm, $n = 5$.

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Mean:**

$$\bar{x} = \frac{1}{5}(185 + 184 + 194 + 180 + 182) = 185$$

- **Median:**

- First order the data: 180, 182, 184, 185, 194.
- Then choose the third/middle number (n uneven): 184
- If a person with height 235 cm is added to the data:
 - *Mean:* 193
 - *Median:* 184.5

Variance and standard deviation, Definition 1.10

The variance and the standard deviation indicate the dispersion ("spread") of the data:

- Variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Example: Student heights

- **Sample:** Student heights in cm, $n = 5$.

$$(x_1, x_2, x_3, x_4, x_5) = (185, 184, 194, 180, 182)$$

- **Variance:**

$$s^2 = \frac{1}{4}((185 - 185)^2 + (184 - 185)^2 + \cdots + (182 - 185)^2) = 29$$

- **Standard deviation:**

$$s = \sqrt{29} = 5.385$$

The coefficient of variation, Definition 1.12

The standard deviation and the variance are key numbers for absolute variation.

If it is of interest to compare variation between different data sets, it might be a good idea to use a *relative* key number.

Coefficient of variation:

$$V = \frac{s}{\bar{x}} \quad (1)$$

Percentiles and quantiles

The median is the value that divides the data into two halves.

More generally, we may compute *percentiles*, e.g.:

- 0, 25, 50, 75, 100 % percentiles and/or
- 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 % percentiles

Note:

- The median is the 50% percentile.
- The 25, 50, 75 % percentiles are often referred to as the first, second and third quartiles, and denoted Q_1 , Q_2 , and Q_3 , respectively.
- Inter Quartile Range (IQR): $Q_3 - Q_1$

Quantiles, Definition 1.7

The p 'th *quantile*, also named the $100p$ 'th *percentile*, can be defined by the following procedure:

- ① Order the n observations from smallest to largest: $x_{(1)}, \dots, x_{(n)}$.
- ② Compute pn .
- ③ If pn is an integer: Average the pn 'th and $(pn + 1)$ 'th ordered observations:

$$\text{The } p\text{'th quantile} = (x_{(np)} + x_{(np+1)}) / 2$$

- ④ If pn is a non-integer, take the next ordered observation:

$$\text{The } p\text{'th quantile} = x_{(\lceil np \rceil)}$$

where $\lceil np \rceil$ is the *ceiling* of np , that is, the smallest integer larger than np .

Example: Student heights

- **Sample:** *Ordered* student heights in cm.

$$(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}) = (180, 182, 184, 185, 194)$$

- **Lower quartile, Q1:**

- Establish that $np = 1.25$, as $p = 0.25$ and $n = 5$.
- The smallest integer larger than np is 2.
- $Q1 = x_{(2)} = 182$.

- **Upper quartile, Q3:**

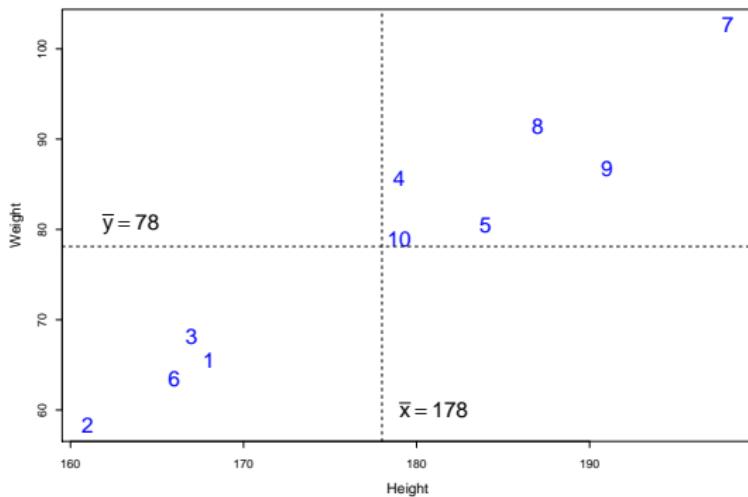
- Establish that $np = 3.75$, as $p = 0.75$ and $n = 5$.
- The smallest integer larger than np is 4.
- $Q3 = x_{(4)} = 185$.

- **IQR:**

- $Q3 - Q1 = 3$

Covariance and correlation

Height (x_i)	168	161	167	179	184	166	198	187	191	179
Weight (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Covariance and correlation, Definitions 1.18 and 1.19

The sample covariance is given by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

The sample correlation coefficient is given by

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{s_{xy}}{s_x \cdot s_y}$$

where s_x and s_y are the sample standard deviations for x and y respectively.

Covariance and correlation

Student	1	2	3	4	5	6	7	8	9	10
Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9
$(x_i - \bar{x})$	-10	-17	-11	1	6	-12	20	9	13	1
$(y_i - \bar{y})$	-12.6	-19.8	-10	7.6	2.4	-14.7	24.5	13.3	8.6	0.8
$(x_i - \bar{x})(y_i - \bar{y})$	126.1	336.8	110.1	7.6	14.3	176.5	489.8	119.6	111.7	0.8

$$\begin{aligned}
 s_{xy} &= \frac{1}{9} (126.1 + 336.8 + 110.1 + 7.6 + 14.3 + 176.5 + 489.8 \\
 &\quad + 119.6 + 111.7 + 0.8) \\
 &= \frac{1}{9} \cdot 1493.3 \\
 &= 165.9
 \end{aligned}$$

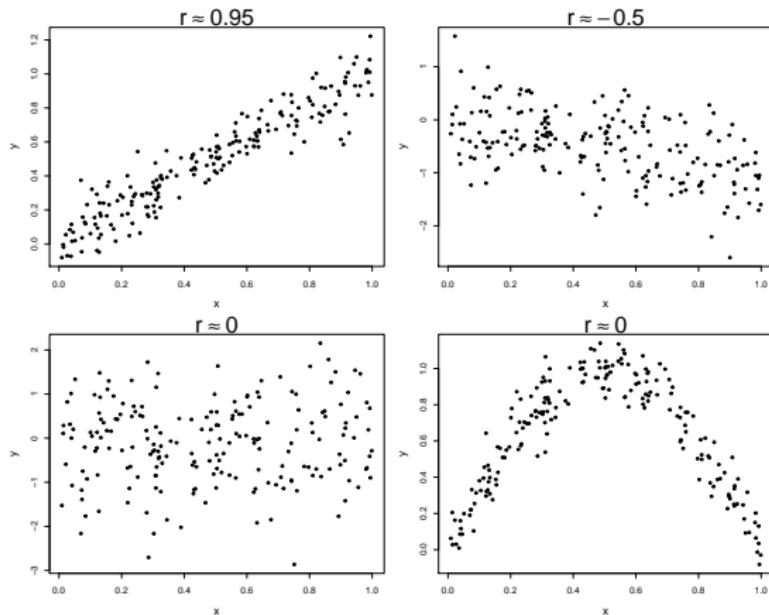
$$s_x = 12.21, \text{ and } s_y = 14.07$$

$$r = \frac{165.9}{12.21 \cdot 14.07} = 0.97$$

Correlation - properties

- r is always between -1 and 1 : $-1 \leq r \leq 1$.
- r measures the degree of linear relation between x and y .
- $r = \pm 1$ if and only if all points in the scatterplot are exactly on a line.
- $r > 0$ if and only if the general trend in the scatterplot is positive.
- $r < 0$ if and only if the general trend in the scatterplot is negative.

Correlation



Figures/Tables

- Quantitative data

- Scatter plot (xy plot)
- Histogram
- Cumulative distribution
- Box plot

- Count data

- Bar chart
- Pie chart

Overview

- 1 Practical course information
- 2 Introduction to Statistics - a primer
- 3 Statistics and Engineers
- 4 Descriptive Statistics
 - Mean and median
 - Variance and standard deviation
 - Percentiles and quantiles
 - Covariance and correlation
- 5 Software: R & RStudio

Software: R & RStudio

- R: Software/language for statistical analysis and visualization of data.
- R & RStudio: Free to download, can be installed on Linux, Mac, Windows.
- R, RStudio, extra packages under continuous development.
- Introduction in the book.
- Integrated in the course material and teaching.
- Learning by doing. Also: use Google!

Software: R

```
> # Adding numbers in the console
> 2 + 3

## [1] 5
```

```
> # Assigning a number to a variable
> x <- 3
> x

## [1] 3
```

```
> # Assigning a vector to a variable
> x <- c(1, 4, 6, 2); x

## [1] 1 4 6 2
```

```
> # A vector of integers from 1 to 10
> ( x <- 1:10 )

## [1] 1 2 3 4 5 6 7 8 9 10
```

Software: R

```
# Height data from before
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
```

```
# Sample mean
mean(x)

## [1] 178
```

```
# Sample median
median(x)

## [1] 179
```

```
# Sample variance
var(x)

## [1] 149
```

Software: R

```
# Sample standard deviation
sd(x)
```

```
## [1] 12
```

```
# Sample quartiles
quantile(x, type = 2)
```

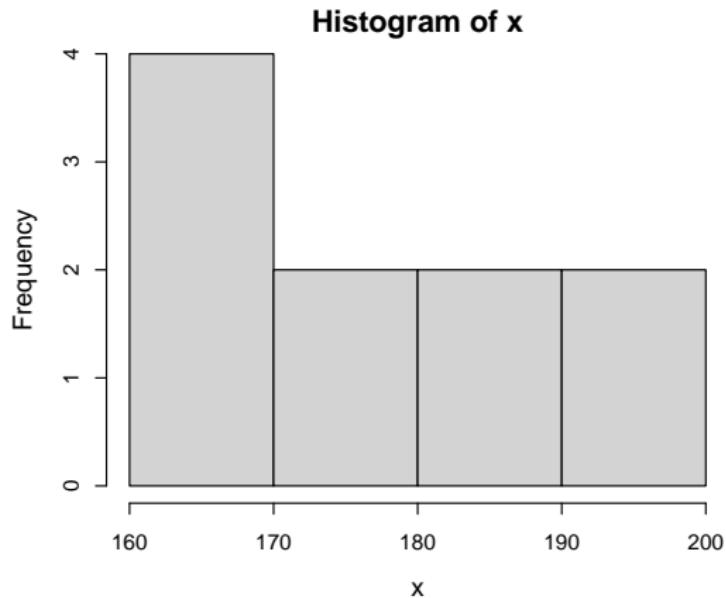
```
##    0%   25%   50%   75% 100%
## 161 167 179 187 198
```

```
# Sample quantiles 0%, 10%,..,90%, 100%
quantile(x, probs = seq(0, 1, by = 0.10), type = 2)
```

```
##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90% 100%
## 161 164 166 168 174 179 184 187 189 194 198
```

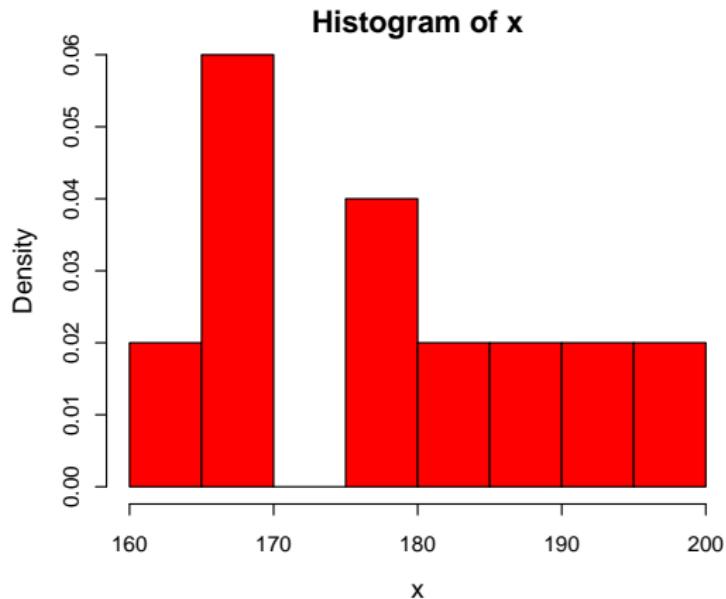
Software: R

```
# A histogram of the heights  
hist(x)
```



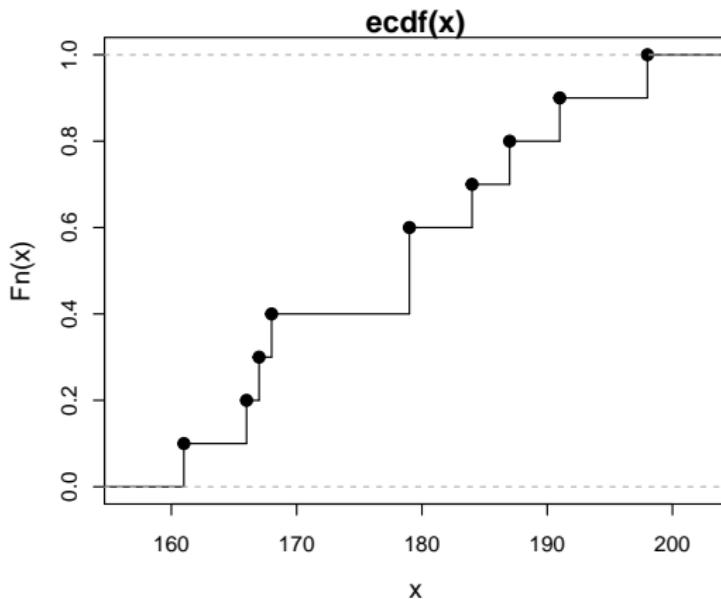
Software: R

```
# A density histogram of the heights  
hist(x, prob = TRUE, col = "red", nclass = 8)
```



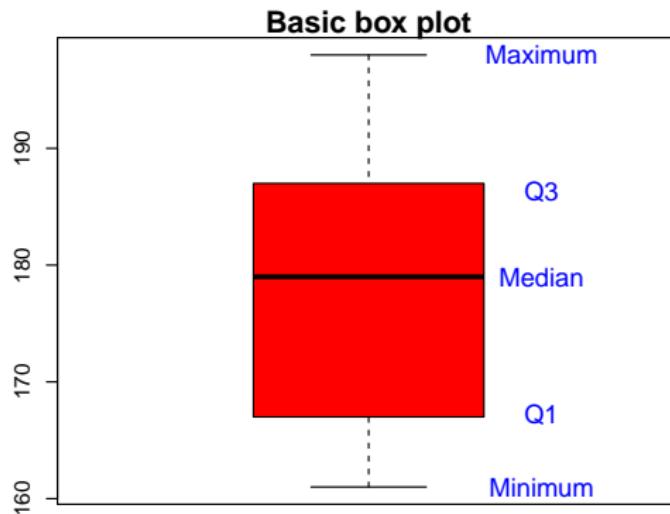
Software: R

```
# Empirical cumulative distribution function of the heights  
plot(ecdf(x), verticals = TRUE)
```



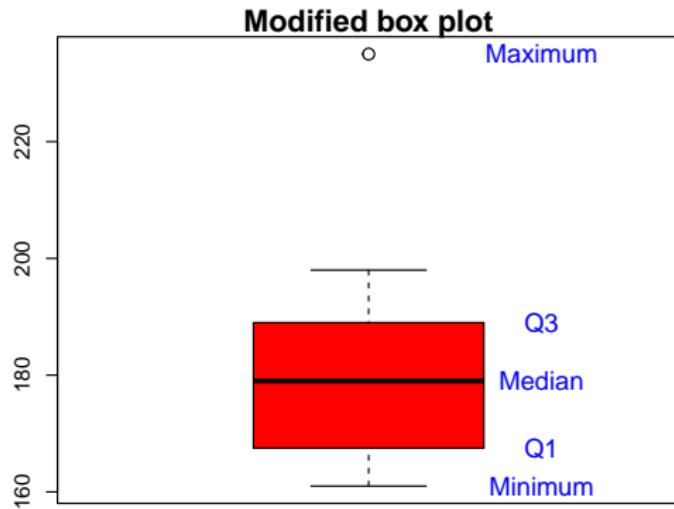
Software: R

```
# Basic box plot of the heights ('range = 0' makes it "basic")
boxplot(x, range = 0, col = "red", main = "Basic box plot")
text(1.3, quantile(x), c("Minimum", "Q1", "Median", "Q3", "Maximum"), col = "blue")
```



Software: R

```
# Modified box plot of heights with an additional extreme observation (235 cm).  
# The modified version is the default.  
boxplot(c(x, 235), col = "red", main = "Modified box plot")  
text(1.3, quantile(c(x, 235)), c("Minimum", "Q1", "Median", "Q3", "Maximum"), col = "blue")
```



Next week:

- Probability, part 1 - eNote/book chapter 2.

Agenda

- 1 Practical course information
- 2 Introduction to Statistics - a primer
- 3 Statistics and Engineers
- 4 Descriptive Statistics
 - Mean and median
 - Variance and standard deviation
 - Percentiles and quantiles
 - Covariance and correlation
- 5 Software: R & RStudio

Course 02402 Introduction to Statistics

Lecture 2: Random variables and discrete distributions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Agenda

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

Overview

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

Random variables

A random variable represents a value of an outcome *before* the corresponding *experiment* is carried out.

Random variables

A random variable represents a value of an outcome *before* the corresponding *experiment* is carried out.

- A throw of a dice.
- The number of six'es in ten dice throws.
- Fuel consumption of a car.
- Measurement of glucose level in blood sample.
- ...

Discrete and continuous random variables

- We distinguish between *discrete* and *continuous* random variables.
- Discrete:
 - Number of people in this room who wear glasses.
 - Number of planes departing from CPH within the next hour.
- Continuous:
 - Wind speed measurement.
 - Transport time to DTU.
- Today: Discrete. Next week: Continuous.

Random variable

Before the experiment is carried out, we have a random variable

X (or X_1, \dots, X_n)

indicated with capital letters.

Random variable

Before the experiment is carried out, we have a random variable

X (or X_1, \dots, X_n)

indicated with capital letters.

After the experiment is carried out, we have a *realization* or *observation*

x (or x_1, \dots, x_n)

indicated with lowercase letters.

Simulate rolling a dice in R

```
# One random draw from (1,2,3,4,5,6)
# with equal probability for each outcome
sample(1:6, size = 1)
```

```
[1] 1
```

Discrete distributions

- Random variables are used to describe an experiment before it is carried out.
- How to do this without yet knowing the outcome?

Discrete distributions

- Random variables are used to describe an experiment before it is carried out.
- How to do this without yet knowing the outcome?
- Solution: Use a *density function*.

Density function, discrete random variable: Definition 2.6

The *density function* (probability density function, pdf) of a discrete random variable:

Definition

$$f(x) = P(X = x)$$

Describes the probability that X takes the value x when the experiment is carried out.

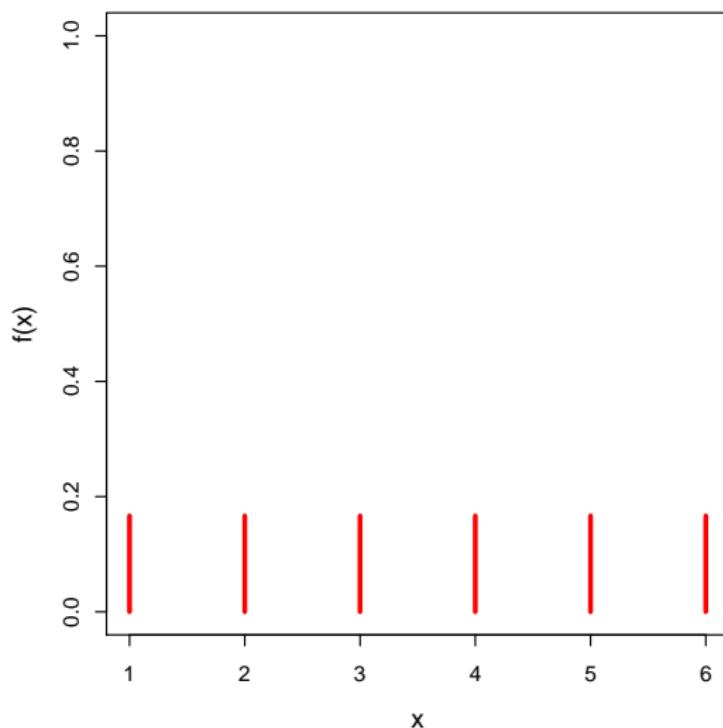
Density function, discrete random variable, Definition 2.6

The density function of a discrete random variable satisfies two properties:

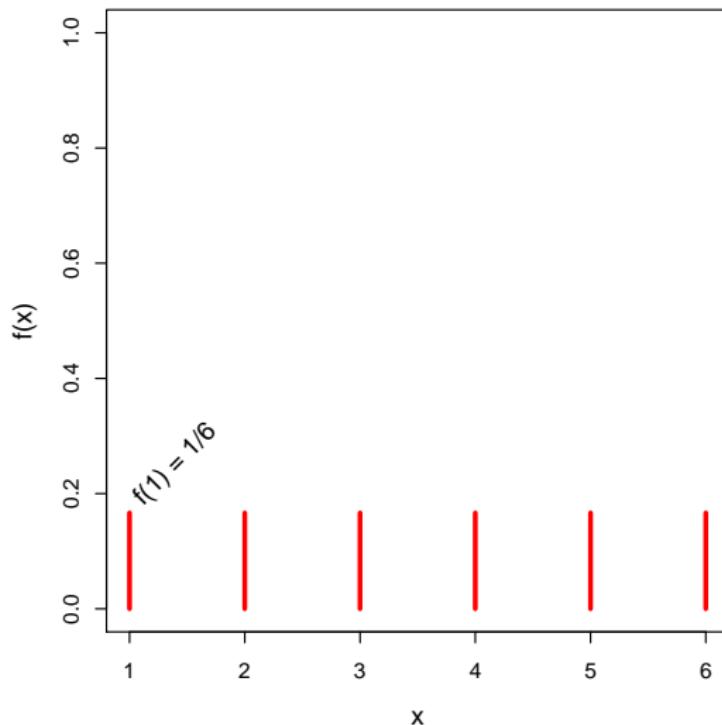
Definition

$$f(x) \geq 0 \text{ for all } x \quad \text{and} \quad \sum_{\text{all } x} f(x) = 1$$

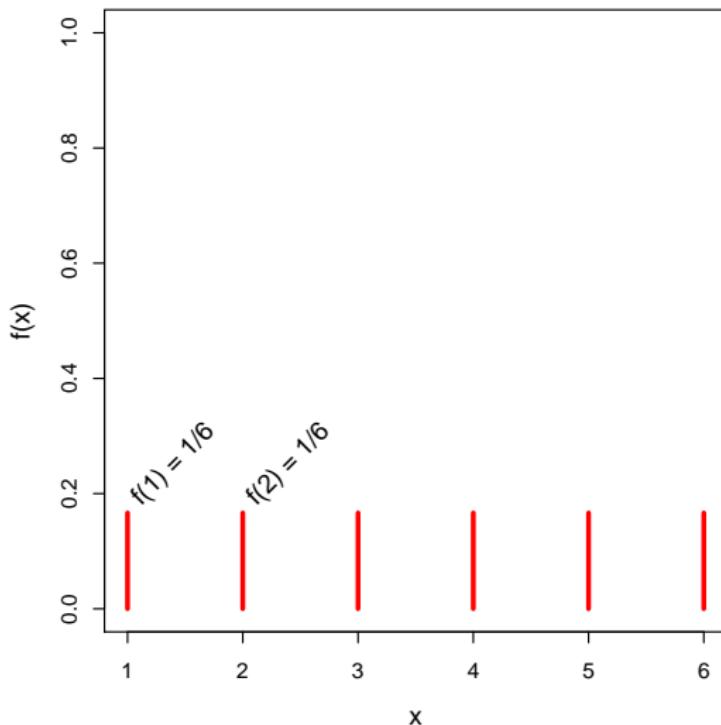
Density function for a fair dice



Density function for a fair dice



Density function for a fair dice



Sample

If we only have a single observation, can we see the distribution?

Sample

If we only have a single observation, can we see the distribution? **No!**

But if we have n observations, then we have a *sample*

$$\{x_1, x_2, \dots, x_n\}$$

and we can begin to get an idea of the distribution.

Simulation of n rolls with a fair dice

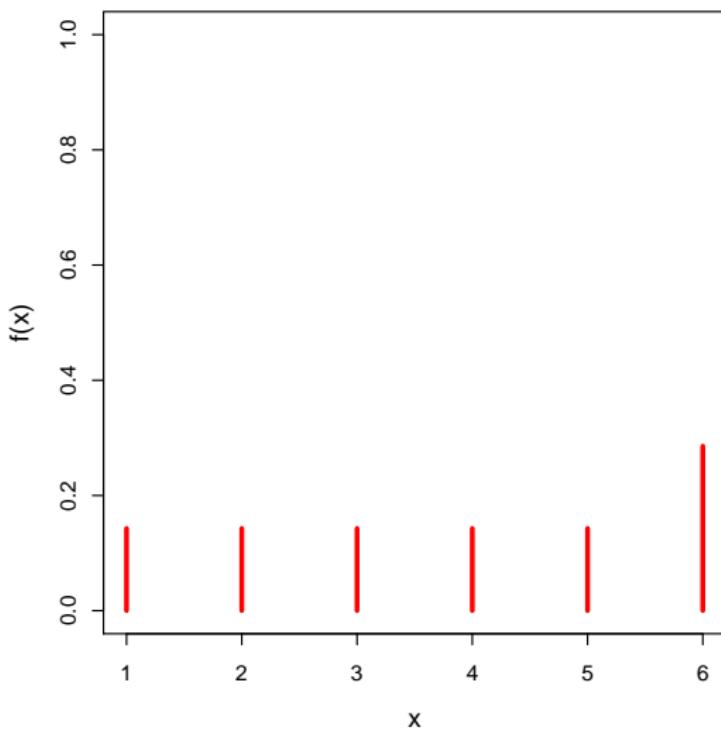
```
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability of each outcome
xFair <- sample(1:6, size = n, replace = TRUE)
xFair

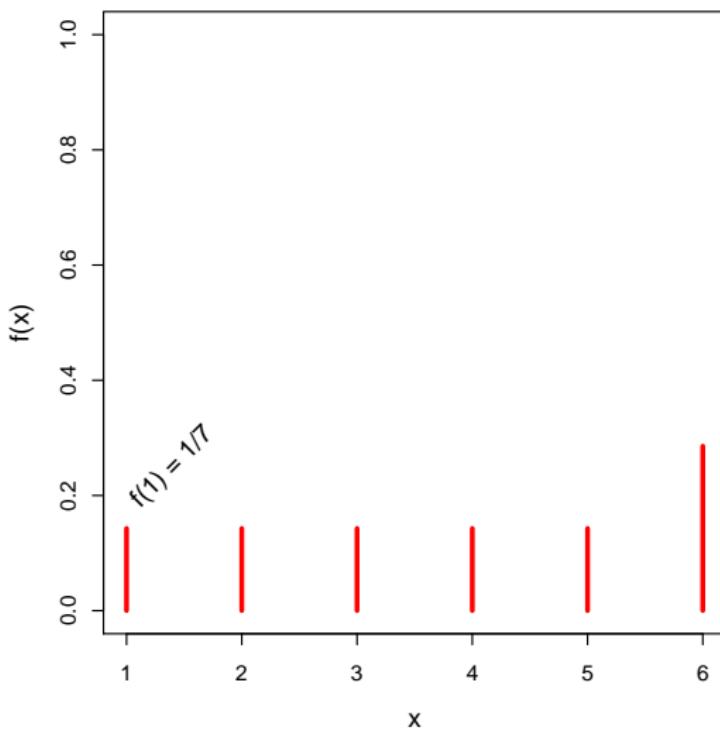
# Count number of each outcome using the 'table' function
table(xFair)

# Plot the empirical pdf
plot(table(xFair)/n, lwd = 10, ylim = c(0,1), xlab = "x",
      ylab = "Density f(x)")
# Add the true pdf to the plot
lines(rep(1/6,6), lwd = 4, type = "h", col = 2)
# Add a legend to the plot
legend("topright", c("Empirical pdf","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```

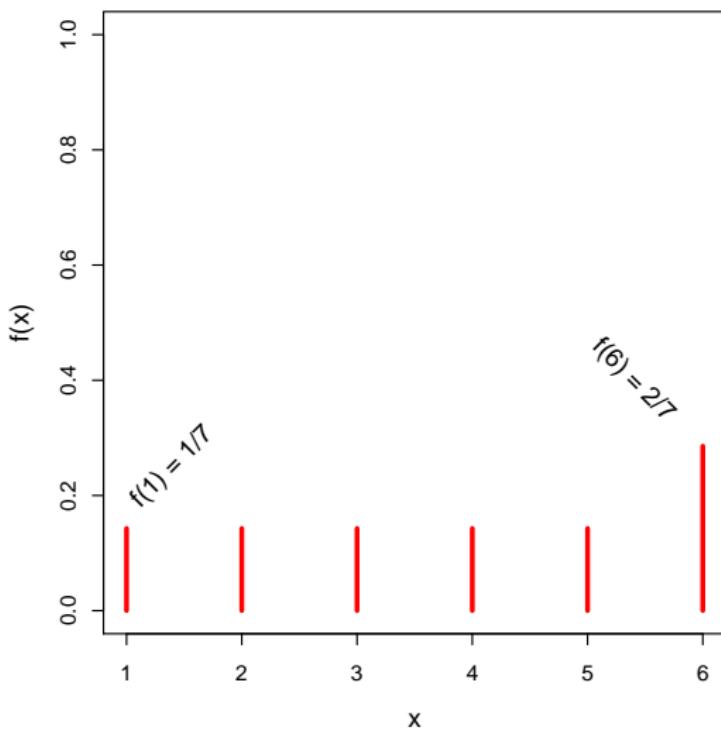
Density function for an unfair dice



Density function for an unfair dice



Density function for an unfair dice



Simulation of n rolls with an unfair dice

```
# Number of simulated realizations (sample size)
n <- 30

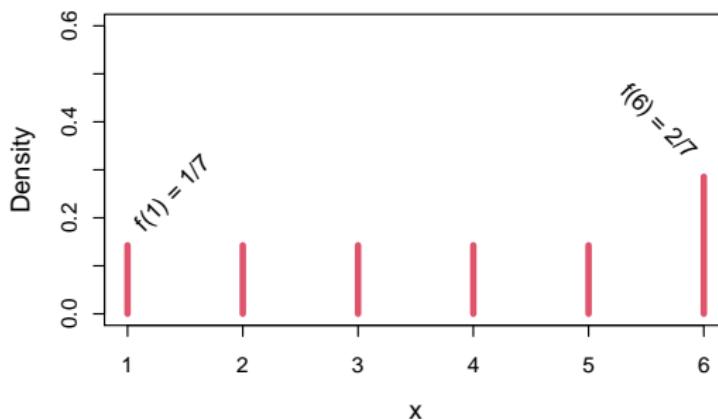
# n independent random draws from the set (1,2,3,4,5,6)
# with higher probability of getting a six
xUnfair <- sample(1:6, size = n, replace = TRUE, prob = c(rep(1/7,5),2/7))
xUnfair

# Plot the empirical pdf
plot(table(xUnfair)/n, lwd = 10, ylim = c(0,1), xlab = "x",
      ylab = "Density f(x)")
# Add the true pdf to the plot
lines(c(rep(1/7,5),2/7), lwd = 4, type = "h", col = 2)
# Add a legend to the plot
legend("topright", c("Empirical pdf","True pdf"), lty = 1, col = c(1,2),
       lwd = c(5, 2), cex = 0.8)
```

Some questions

Let X describe one throw with the *unfair* dice. What is:

- The probability of getting a 4?
- The probability of getting a 5 or a 6?
- The probability of getting less than 3?



Overview

- 1 Random variables and density functions
- 2 **Distribution functions**
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

Distribution function, discrete random variable: Definition 2.9

The *distribution function* (cumulative distribution function, cdf) of a discrete random variable:

Definition

$$F(x) = P(X \leq x) = \sum_{j \text{ where } x_j \leq x} f(x_j)$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$P(X < 3)$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$P(X < 3) = P(X \leq 2)$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \text{ *the distribution function*} \end{aligned}$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \text{ *the distribution function*} \\ &= P(X = 1) + P(X = 2) \end{aligned}$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \text{ *the distribution function*} \\ &= P(X = 1) + P(X = 2) \\ &= f(1) + f(2) \text{ *the density function*} \end{aligned}$$

Fair dice example

Let X represent one throw with a fair dice.

Find the probability of throwing less than 3:

$$\begin{aligned} P(X < 3) &= P(X \leq 2) \\ &= F(2) \text{ *the distribution function*} \\ &= P(X = 1) + P(X = 2) \\ &= f(1) + f(2) \text{ *the density function*} \\ &= \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \end{aligned}$$

Fair dice example

Find the probability of throwing greater than or equal to 3:

Fair dice example

Find the probability of throwing greater than or equal to 3:

$$P(X \geq 3)$$

Fair dice example

Find the probability of throwing greater than or equal to 3:

$$P(X \geq 3) = 1 - P(X \leq 2)$$

Fair dice example

Find the probability of throwing greater than or equal to 3:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - F(2) \text{ *the distribution function*} \end{aligned}$$

Fair dice example

Find the probability of throwing greater than or equal to 3:

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) \\ &= 1 - F(2) \text{ *the distribution function*} \\ &= 1 - \frac{1}{3} = \frac{2}{3} \end{aligned}$$

Overview

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

Specific discrete distributions

- A number of different statistical distributions exist, which may be used to describe and analyse different types of problems.
- Today, we consider only discrete distributions:
 - The binomial distribution
 - The hypergeometric distribution
 - The Poisson distribution

The Binomial distribution

- An experiment with two outcomes, "success" or "failure", is repeated (independent repetitions).
- X is the number of successes after n repetitions.

The Binomial distribution

- An experiment with two outcomes, "success" or "failure", is repeated (independent repetitions).
- X is the number of successes after n repetitions.
- Then X follows a binomial distribution:

$$X \sim B(n, p)$$

- n : number of repetitions
- p : probability of success in each repetition

The density function of the binomial distribution

The probability of x successes:

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

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Example: Binomial distribution

Suppose that $X \sim B(4, p)$, i.e. $n = 4$. Find the probability of 3 successes.

Example: Binomial distribution

Suppose that $X \sim B(4, p)$, i.e. $n = 4$. Find the probability of 3 successes.

- Probability of 3 successes: $P(X = 3)$.

Example: Binomial distribution

Suppose that $X \sim B(4, p)$, i.e. $n = 4$. Find the probability of 3 successes.

- Probability of 3 successes: $P(X = 3)$.
- Three successes can be obtained in four "ways":
SSSF, SSFS, SFSS, FSSS.

Example: Binomial distribution

Suppose that $X \sim B(4, p)$, i.e. $n = 4$. Find the probability of 3 successes.

- Probability of 3 successes: $P(X = 3)$.
- Three successes can be obtained in four "ways": SSSF, SSFS, SFSS, FSSS.
- Thus,

$$\binom{n}{x} = \binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4,$$

and

$$P(X = 3) = 4p^3(1 - p).$$

Simulation from a binomial distribution

```
## Probability of success
p <- 0.1

## Number of repetitions
nRepeat <- 30

## Simulate Bernoulli experiment 'nRepeat' times
tmp <- sample(c(0,1), size = nRepeat, prob = c(1-p,p), replace = TRUE)

# Compute 'x'
sum(tmp)

## Or: Use the binomial distribution simulation function
rbinom(1, size = 30, prob = p)
```

Example: Fair dice

```
# Number of simulated realizations (sample size)
n <- 30

# n independent random draws from the set (1,2,3,4,5,6)
# with equal probability for each outcome
xFair <- sample(1:6, size = n, replace = TRUE)

# Count the number of six'es
sum(xFair == 6)

## Do the same using 'rbinom()' instead
rbinom(n = 1, size = 30, prob = 1/6)
```

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- **Step 1)** What should be described by a random variable X ?

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- **Step 1)** What should be described by a random variable X ?

The number of corrected errors.

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- **Step 1)** What should be described by a random variable X ?

The number of corrected errors.

- **Step 2)** What is the distribution of X ?

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- **Step 1)** What should be described by a random variable X ?

The number of corrected errors.

- **Step 2)** What is the distribution of X ?

A binomial distribution with $n = 6$ and $p = 0.7$.

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- **Step 3)** Which probability should be computed?

Example 1

In the call center of a phone company, customer satisfaction is an issue. It is especially important that when errors/faults occur, they are corrected within the same day.

Assume that six errors occur, and that the probability of any error being corrected within the same day is 70%. **What is the probability that all six errors are corrected within the same day that they occurred?**

- **Step 3)** Which probability should be computed?

$$\underline{P(X = 6) = f(6; 6, 0.7)}$$

Overview

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

The hypergeometric distribution

- Again, X is the the number of successes, but now *without* replacement when repeating.

The hypergeometric distribution

- Again, X is the the number of successes, but now *without* replacement when repeating.
- X follows the hypergeometric distribution

$$X \sim H(n, a, N)$$

- n is the number of draws (repetitions)
- a is the number of successes in the population
- N is the number of elements in the (entire) population

The hypergeometric distribution

- The probability of getting x successes is

$$f(x; n, a, N) = P(X = x) = \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}}$$

- n is the number of draws (repetitions)
- a is the number of successes in the population
- N is the number of elements in the (entire) population

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

A random sample of 3 harddisks is taken. **What is the probability that at least 1 of them has scratches?**

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

A random sample of 3 harddisks is taken. **What is the probability that at least 1 of them has scratches?**

- **Step 1)** What should be described by a random variable X ?

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

A random sample of 3 harddisks is taken. **What is the probability that at least 1 of them has scratches?**

- **Step 1)** What should be described by a random variable X ?

Number of harddisks with scratches in the random sample.

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

A random sample of 3 harddisks is taken. **What is the probability that at least 1 of them has scratches?**

- **Step 1)** What should be described by a random variable X ?

Number of harddisks with scratches in the random sample.

- **Step 2)** What is the distribution of X ?

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

A random sample of 3 harddisks is taken. **What is the probability that at least 1 of them has scratches?**

- **Step 1)** What should be described by a random variable X ?

Number of harddisks with scratches in the random sample.

- **Step 2)** What is the distribution of X ?

A hypergeometric distribution with $n = 3$, $a = 2$, $N = 10$.

- **Step 3)** Which probability should be computed?

Example 2

In a shipment of 10 harddisks, 2 of them have small scratches.

A random sample of 3 harddisks is taken. **What is the probability that at least 1 of them has scratches?**

- **Step 1)** What should be described by a random variable X ?

Number of harddisks with scratches in the random sample.

- **Step 2)** What is the distribution of X ?

A hypergeometric distribution with $n = 3$, $a = 2$, $N = 10$.

- **Step 3)** Which probability should be computed?

$P(X \geq 1) = 1 - P(X = 0) = 1 - f(0; 3, 2, 10)$

Binomial vs. hypergeometric

- The binomial distribution is used to analyse samples with replacement.
- The hypergeometric distribution is used to analyse samples without replacement.

Overview

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

The Poisson distribution

- The Poisson distribution is often used as a distribution (model) for counts, which do not have a natural upper bound.
- The Poisson distribution is often characterized by its *intensity*, which is on the form "number/unit", and often denoted λ .

The Poisson distribution

$$X \sim Po(\lambda)$$

The density function:

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

The distribution function:

$$F(x) = P(X \leq x)$$

Example 3

Assume that, on average, 0.3 patients per day are hospitalized in Copenhagen due to air pollution.

What is the probability that at most two patients are hospitalized in Copenhagen due to air pollution on any given day?

Example 3

Assume that, on average, 0.3 patients per day are hospitalized in Copenhagen due to air pollution.

What is the probability that at most two patients are hospitalized in Copenhagen due to air pollution on any given day?

- **Step 1)** What should be described by a random variable X ?

Example 3

Assume that, on average, 0.3 patients per day are hospitalized in Copenhagen due to air pollution.

What is the probability that at most two patients are hospitalized in Copenhagen due to air pollution on any given day?

- **Step 1)** What should be described by a random variable X ?

The number of patients on a given day.

Example 3

Assume that, on average, 0.3 patients per day are hospitalized in Copenhagen due to air pollution.

What is the probability that at most two patients are hospitalized in Copenhagen due to air pollution on any given day?

- **Step 1)** What should be described by a random variable X ?

The number of patients on a given day.

- **Step 2)** What is the distribution of X ?

Example 3

Assume that, on average, 0.3 patients per day are hospitalized in Copenhagen due to air pollution.

What is the probability that at most two patients are hospitalized in Copenhagen due to air pollution on any given day?

- **Step 1)** What should be described by a random variable X ?

The number of patients on a given day.

- **Step 2)** What is the distribution of X ?

A Poisson distribution with $\lambda = 0.3$.

Example 3

Assume that, on average, 0.3 patients per day are hospitalized in Copenhagen due to air pollution.

What is the probability that at most two patients are hospitalized in Copenhagen due to air pollution on any given day?

- **Step 1)** What should be described by a random variable X ?

The number of patients on a given day.

- **Step 2)** What is the distribution of X ?

A Poisson distribution with $\lambda = 0.3$.

- **Step 3)** Which probability should be computed?

Example 3

Assume that, on average, 0.3 patients per day are hospitalized in Copenhagen due to air pollution.

What is the probability that at most two patients are hospitalized in Copenhagen due to air pollution on any given day?

- **Step 1)** What should be described by a random variable X ?

The number of patients on a given day.

- **Step 2)** What is the distribution of X ?

A Poisson distribution with $\lambda = 0.3$.

- **Step 3)** Which probability should be computed?

$$\underline{P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)}$$

Overview

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 **Distributions in R**
- 7 Mean and variance of discrete distributions

R	Name
<code>binom</code>	Binomial
<code>hyper</code>	Hypergeometric
<code>pois</code>	Poisson

- d** $f(x)$, probability density function
- p** $F(x)$, cumulative distribution function
- r** random numbers from the distribution
- q** quantiles of the distribution ("inverse" of $F(x)$)

Example: The binomial distribution, $P(X \leq 5) = F(5; 10, 0.6)$

```
pbinom(q = 5, size = 10, prob = 0.6)
```

```
[1] 0.37
```

```
# Get help with:  
?pbinom
```

Overview

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

Mean (expectation, expected value)

Mean of a discrete random variable, Definition 2.13:

Definition

$$\mu = E(X) = \sum_{\text{all } x} xf(x)$$

- The “*true mean*” of X (as opposed to the sample mean).
- Expresses the “center” of the distribution of X .

Example: Mean of a throw with a fair dice

$$\mu = E(X)$$

$$\begin{aligned} &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5 \end{aligned}$$

Link to sample mean - learning from simulations

```
# Number of simulated realizations (sample size)
n <- 30

# Sample independently from the set (1,2,3,4,5,6)
# with equal probability of outcomes
xFair <- sample(1:6, size = n, replace = TRUE)

# Compute the sample mean
mean(xFair)
```

```
[1] 3.3
```

Asymptotics, increasing the sample size

The more observations (the larger the sample size), the closer you get to the true mean:

$$\lim_{n \rightarrow \infty} \hat{\mu} = \mu$$

- Try increasing n in the simulations in R.

Variance

Variance of a discrete random variable, Definition 2.16:

Definition

$$\sigma^2 = \text{Var}(X) = \sum_{\text{all } x} (x - \mu)^2 f(x)$$

- Measures average dispersion/spread.
- The “true variance” of X (as opposed to the sample variance).

Example: Variance of a throw with a fair dice

$$\begin{aligned}\sigma^2 &= E[(X - \mu)^2] \\ &= (1 - 3.5)^2 \cdot \frac{1}{6} + (2 - 3.5)^2 \cdot \frac{1}{6} + (3 - 3.5)^2 \cdot \frac{1}{6} \\ &\quad + (4 - 3.5)^2 \cdot \frac{1}{6} + (5 - 3.5)^2 \cdot \frac{1}{6} + (6 - 3.5)^2 \cdot \frac{1}{6} \\ &\approx 2.92\end{aligned}$$

Link to sample variance - learning from simulations

```
# Number of simulated realizations (sample size)
n <- 30

# Sample independently from the set (1,2,3,4,5,6)
# with equal probability of outcomes
xFair <- sample(1:6, size = n, replace = TRUE)

# Compute the sample variance
var(xFair)
```

```
[1] 2.4
```

Mean and variance of specific discrete distributions

The binomial distribution

- Mean:

$$\mu = n \cdot p$$

- Variance:

$$\sigma^2 = n \cdot p \cdot (1 - p)$$

Mean and variance of specific discrete distributions

The hypergeometric distribution

- Mean:

$$\mu = n \cdot \frac{a}{N}$$

- Variance:

$$\sigma^2 = \frac{n \cdot a \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$$

Mean and variance of specific discrete distributions

The poisson distribution

- Mean:

$$\mu = \lambda$$

- Variance:

$$\sigma^2 = \lambda$$

Agenda

- 1 Random variables and density functions
- 2 Distribution functions
- 3 Specific (discrete) distributions I: The binomial
 - Example 1
- 4 Specific distributions II: The hypergeometric
 - Example 2
- 5 Specific distributions III: The Poisson
 - Example 3
- 6 Distributions in R
- 7 Mean and variance of discrete distributions

Course 02402 Introduction to Statistics

Lecture 3: Random variables and continuous distributions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Continuous random variables and distributions
 - Density and distribution functions
 - Mean, variance, and covariance
- 2 Specific continuous distributions
 - The uniform distribution
 - The normal distribution
 - The log-normal distribution
 - The exponential distribution
- 3 Calculation rules for random variables

Overview

- 1 Continuous random variables and distributions
 - Density and distribution functions
 - Mean, variance, and covariance
- 2 Specific continuous distributions
 - The uniform distribution
 - The normal distribution
 - The log-normal distribution
 - The exponential distribution
- 3 Calculation rules for random variables

The density function, Definition 2.32

- The density function (probability density function, pdf) for a random variable is denoted by $f(x)$.

The density function, Definition 2.32

- The density function (probability density function, pdf) for a random variable is denoted by $f(x)$.
- The density function says something about the frequency of the outcome x for the random variable X .

The density function, Definition 2.32

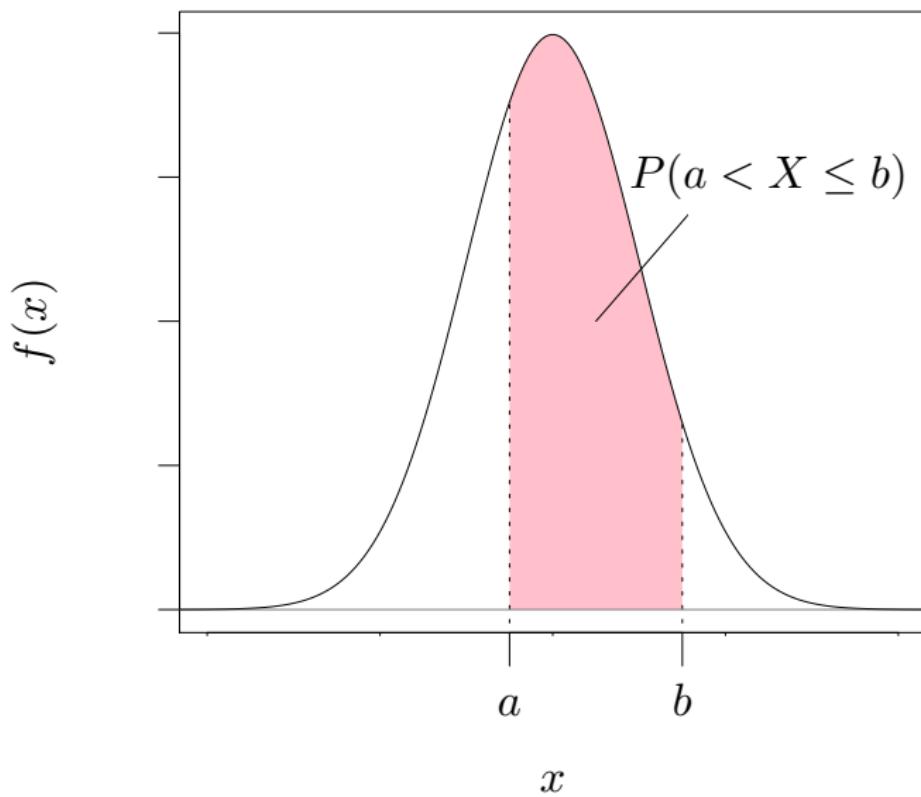
- The density function (probability density function, pdf) for a random variable is denoted by $f(x)$.
- The density function says something about the frequency of the outcome x for the random variable X .
- The density function for a continuous random variable does *not* correspond directly to a probability. In fact, $f(x) \neq P(X = x)$ and $P(X = x) = 0$ for all x .

The density function, Definition 2.32

- The density function (probability density function, pdf) for a random variable is denoted by $f(x)$.
- The density function says something about the frequency of the outcome x for the random variable X .
- The density function for a continuous random variable does *not* correspond directly to a probability. In fact, $f(x) \neq P(X = x)$ and $P(X = x) = 0$ for all x .
- The density function $f(x)$ for the distribution of a continuous random variable satisfies that

$$f(x) \geq 0 \text{ for all } x \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

The density function



The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.

The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.
- The distribution function is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.
- The distribution function is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

- Note that as a consequence of this definition,

$$f(x) = F'(x).$$

The distribution function, Definition 2.33

- The distribution function (cumulative density function, cdf) for a continuous random variable is denoted by $F(x)$.
- The distribution function is defined by

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt.$$

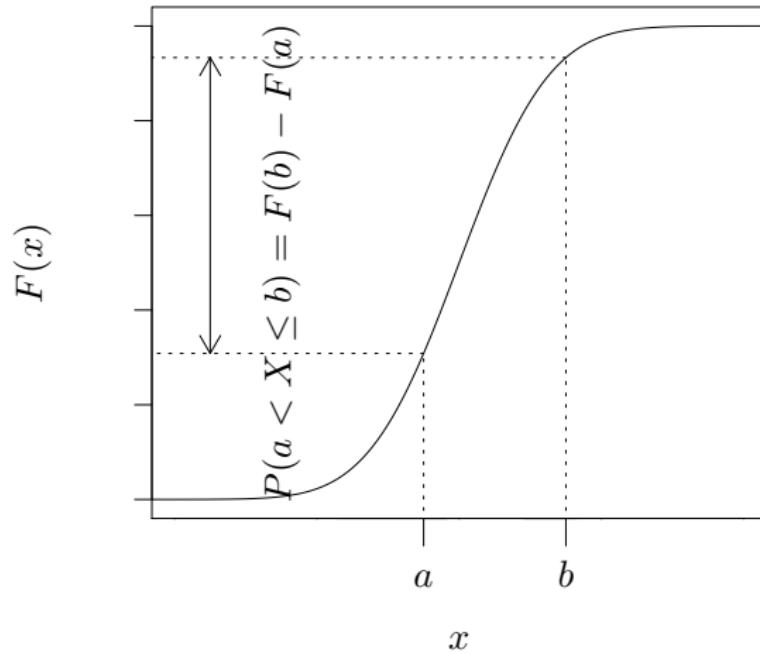
- Note that as a consequence of this definition,

$$f(x) = F'(x).$$

- It's particularly useful to note that

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx.$$

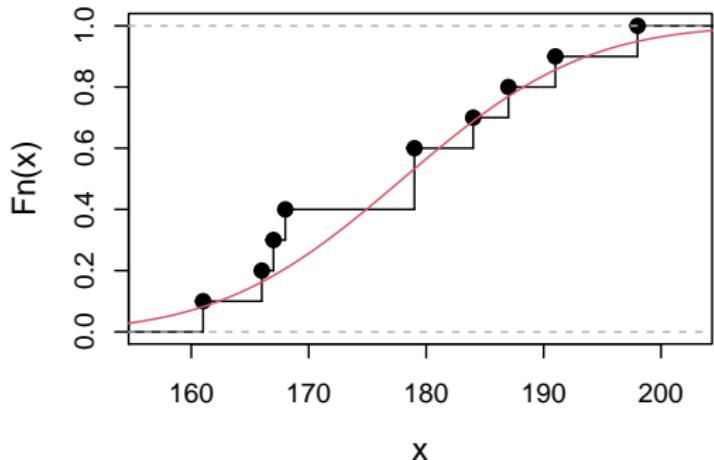
The distribution function



The empirical cumulative distribution function (ecdf)

```
# Empirical cdf for sample of height data from Chapter 1
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
plot(ecdf(x), verticals = TRUE, main = "")

# 'True cdf' for normal distribution (with sample mean and variance)
xp <- seq(0.9*min(x), 1.1*max(x), length = 100)
lines(xp, pnorm(xp, mean(x), sd(x)), col = 2)
```



Mean, continuous random variable, Definition 2.34

The mean/expected value of a continuous random variable:

$$\mu = \int_{-\infty}^{\infty} xf(x) dx$$

Mean, continuous random variable, Definition 2.34

The mean/expected value of a continuous random variable:

$$\mu = \int_{-\infty}^{\infty} xf(x) dx$$

Compare with the mean of a discrete random variable:

$$\mu = \sum_{\text{all } x} xf(x)$$

Variance, continuous random variable, Definition 2.34

The variance of a continuous random variable:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Variance, continuous random variable, Definition 2.34

The variance of a continuous random variable:

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Compare with the variance of a discrete random variable:

$$\sigma^2 = \sum_{\text{all } x} (x - \mu)^2 f(x)$$

Covariance, Definition 2.58

The covariance between two random variables:

Let X and Y be two random variables. Then, the covariance between X and Y is

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Relationship between covariance and independence:

If two random variables are *independent* their covariance is 0. *The reverse is not necessarily true!*

Overview

- 1 Continuous random variables and distributions
 - Density and distribution functions
 - Mean, variance, and covariance
- 2 Specific continuous distributions
 - The uniform distribution
 - The normal distribution
 - The log-normal distribution
 - The exponential distribution
- 3 Calculation rules for random variables

Specific continuous distributions

A number of statistical distributions exist (both continuous and discrete) that can be used to describe and analyze different types of problems.

Today, we'll take a closer look at the following continuous distributions:

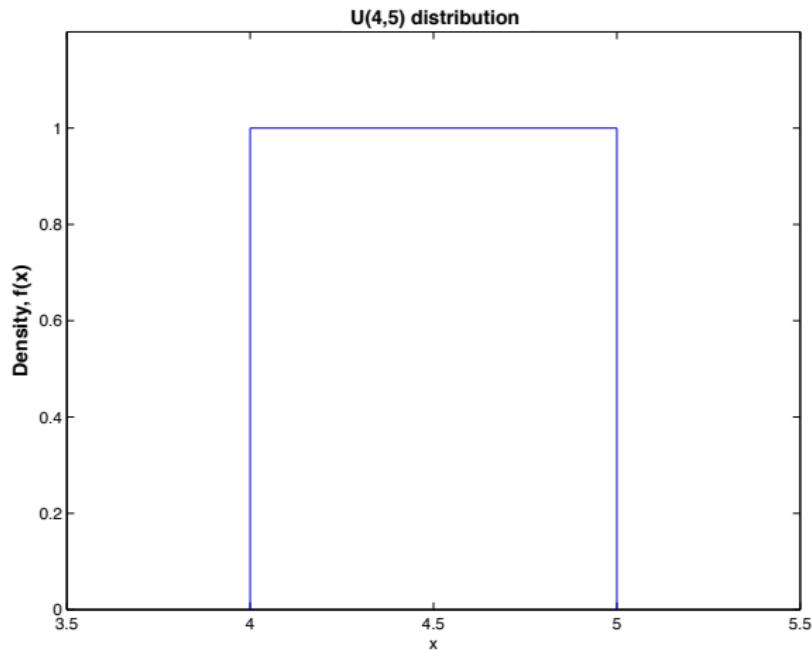
- The uniform distribution
- The normal distribution
- The log-normal distribution
- The exponential distribution

Continuous distributions in R

R	Distribution
<code>norm</code>	The normal distribution
<code>unif</code>	The uniform distribution
<code>lnorm</code>	The log-normal distribution
<code>exp</code>	The exponential distribution

- d** Probability density function, $f(x)$.
- p** Cumulative distribution function, $F(x)$.
- q** Quantile function.
- r** Random numbers from the distribution.

Density of a uniform distribution (example)



The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$$X \sim U(\alpha, \beta)$$

The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$$X \sim U(\alpha, \beta)$$

Density function:

$$f(x) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq x \leq \beta$$

The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$$X \sim U(\alpha, \beta)$$

Density function:

$$f(x) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq x \leq \beta$$

Mean:

$$\mu = \frac{\alpha + \beta}{2}$$

The uniform distribution, Def. 2.35 & Theo. 2.36

Syntax:

$$X \sim U(\alpha, \beta)$$

Density function:

$$f(x) = \frac{1}{\beta - \alpha} \text{ for } \alpha \leq x \leq \beta$$

Mean:

$$\mu = \frac{\alpha + \beta}{2}$$

Variance:

$$\sigma^2 = \frac{1}{12}(\beta - \alpha)^2$$

Example 1

Students attending a stats course arrive at a lecture between 8.00 and 8.30. It is assumed that the arrival times can be described by a uniform distribution.

Question:

What is the probability that a randomly selected student arrives between 8.20 and 8.30?

Example 1

Students attending a stats course arrive at a lecture between 8.00 and 8.30. It is assumed that the arrival times can be described by a uniform distribution.

Question:

What is the probability that a randomly selected student arrives between 8.20 and 8.30?

Answer:

$$10/30 = 1/3$$

Let $X \sim U(0, 30)$ represent arrival time. Then:

$$P(20 \leq X \leq 30) = P(X \leq 30) - P(X \leq 20) = 1 - 2/3 = 1/3$$

```
punif(30, 0, 30) - punif(20, 0, 30)
```

[1] 0.33

Example 1 (continued)

Question:

What is the probability that a randomly selected student arrives after 8.30?

Example 1 (continued)

Question:

What is the probability that a randomly selected student arrives after 8.30?

Answer:

0

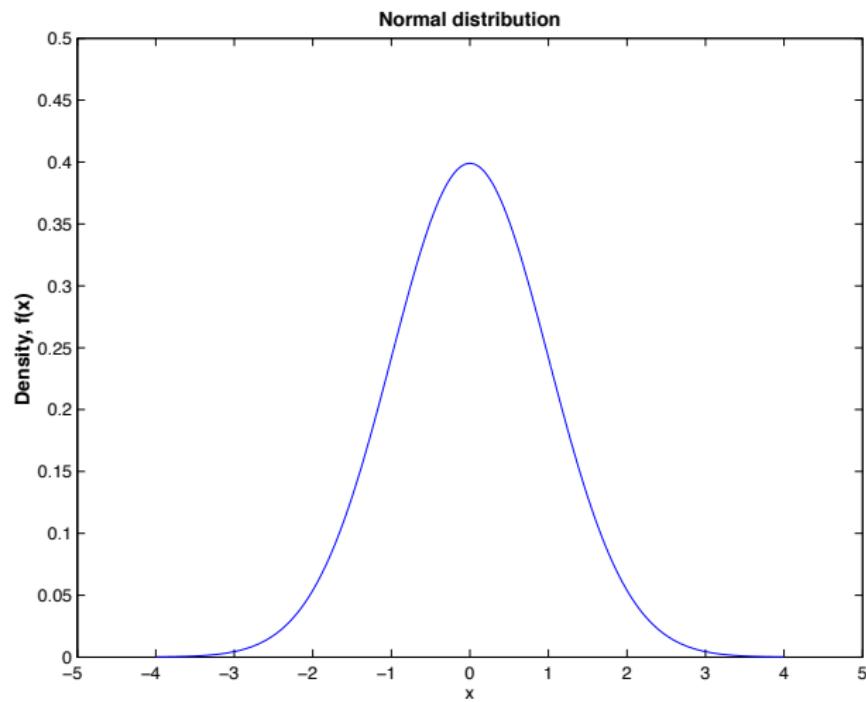
Let $X \sim U(0, 30)$ represent arrival time. Then:

$$P(X > 30) = 1 - P(X \leq 30) = 1 - 1 = 0$$

```
1 - punif(30, 0, 30)
```

```
[1] 0
```

Density of a normal distribution (example)



The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$$X \sim N(\mu, \sigma^2)$$

The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$$X \sim N(\mu, \sigma^2)$$

Density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$$X \sim N(\mu, \sigma^2)$$

Density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

Mean:

$$\mu = \mu$$

The normal distribution, Def. 2.37 & Theo. 2.38

Syntax:

$$X \sim N(\mu, \sigma^2)$$

Density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

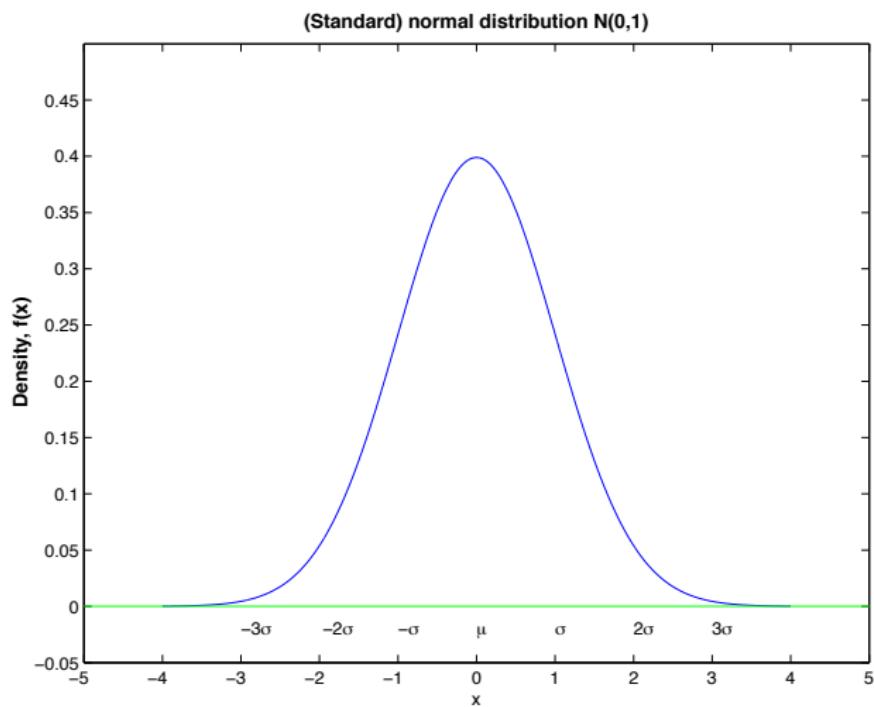
Mean:

$$\mu = \mu$$

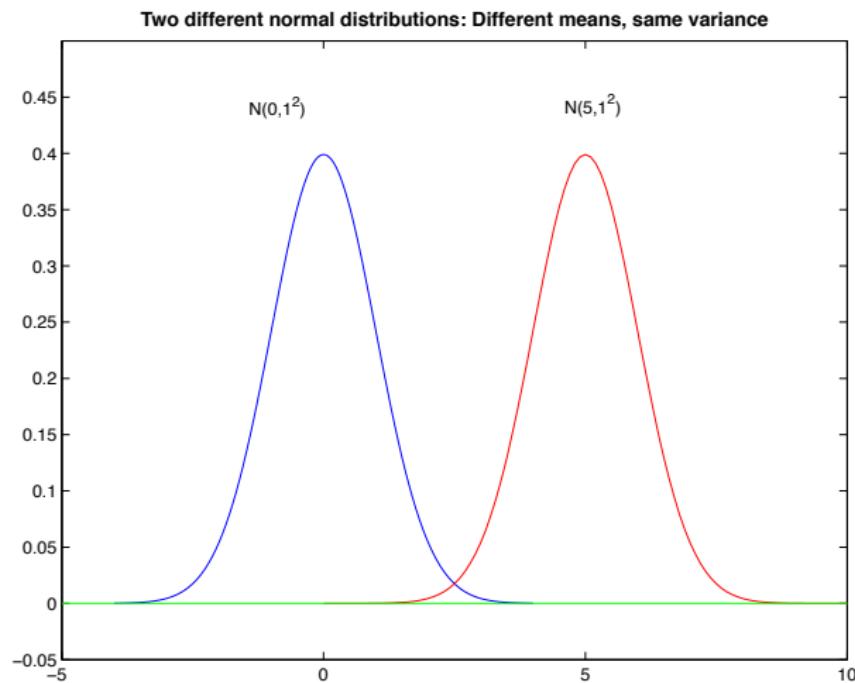
Variance:

$$\sigma^2 = \sigma^2$$

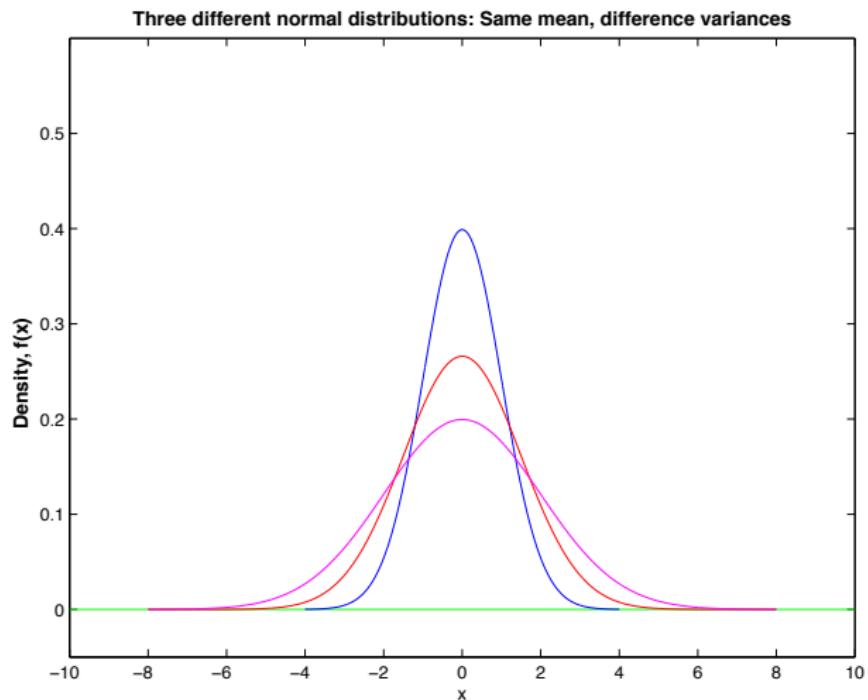
Density of a standard normal distribution



Density of two normal distributions (example)



Density of three normal distributions (example)



The standard normal distribution

The standard normal distribution:

$$Z \sim N(0, 1^2)$$

The normal distribution with mean 0 and variance 1.

The standard normal distribution

The standard normal distribution:

$$Z \sim N(0, 1^2)$$

The normal distribution with mean 0 and variance 1.

Standardization:

An arbitrary normal distributed variable $X \sim N(\mu, \sigma^2)$ can be *standardized* by

$$Z = \frac{X - \mu}{\sigma}$$

Example 2

Measurement error:

A scale has a measurement error, Z , that can be described by the standard normal distribution, i.e.

$$Z \sim N(0, 1^2).$$

That is, the mean measurement error is $\mu = 0$ with standard deviation $\sigma = 1$ gram. The scale is used to measure the weight of a product.

Example 2

Measurement error:

A scale has a measurement error, Z , that can be described by the standard normal distribution, i.e.

$$Z \sim N(0, 1^2).$$

That is, the mean measurement error is $\mu = 0$ with standard deviation $\sigma = 1$ gram. The scale is used to measure the weight of a product.

Question a):

What is the probability that the scale yields a measurement which is at least 2 grams smaller than the true weight of the product?

Example 2

Measurement error:

A scale has a measurement error, Z , that can be described by the standard normal distribution, i.e.

$$Z \sim N(0, 1^2).$$

That is, the mean measurement error is $\mu = 0$ with standard deviation $\sigma = 1$ gram. The scale is used to measure the weight of a product.

Question a):

What is the probability that the scale yields a measurement which is at least 2 grams smaller than the true weight of the product?

Answer:

$$P(Z \leq -2) = 0.02275$$

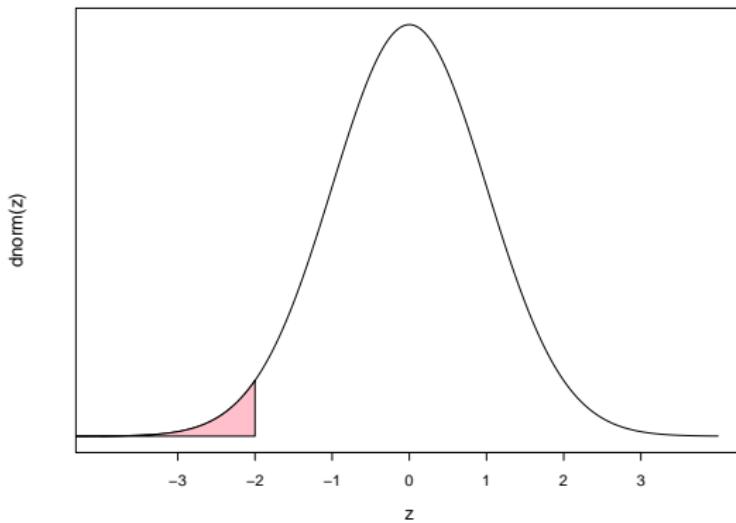
```
pnorm(-2)
```

Example 2

Answer:

```
pnorm(-2)
```

```
[1] 0.023
```



Example 2

Question b):

What is the probability that the scale yields a measurement which is at least 2 grams larger than the true weight of the product?

Example 2

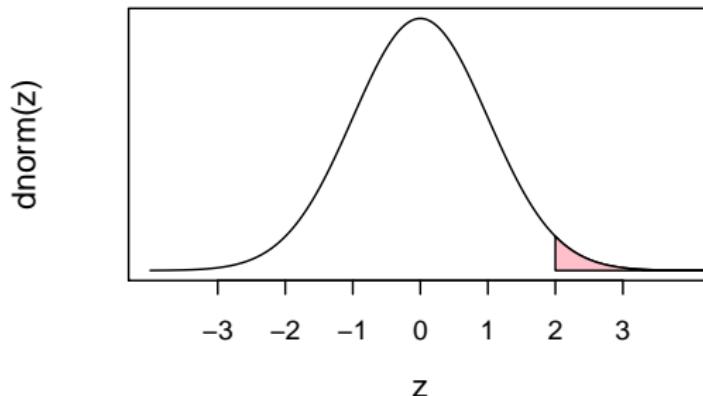
Question b):

What is the probability that the scale yields a measurement which is at least 2 grams larger than the true weight of the product?

Answer:

$$P(Z \geq 2) = 0.02275$$

```
1 - pnorm(2)
```



Example 2

Question c):

What is the probability that the scale is off by at most ± 1 gram?

Example 2

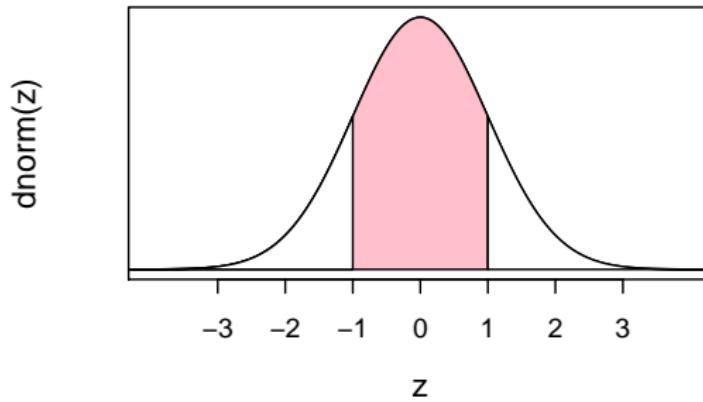
Question c):

What is the probability that the scale is off by at most ± 1 gram?

Answer:

$$P(|Z| \leq 1) = P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq -1) = 0.683$$

```
pnorm(1) - pnorm(-1)
```



Example 3

Income distribution:

It is assumed that the annual salary distribution of elementary school teachers can be described using a normal distribution with mean $\mu = 290$ (in DKK thousand) and standard deviation $\sigma = 4$ (DKK thousand).

Example 3

Income distribution:

It is assumed that the annual salary distribution of elementary school teachers can be described using a normal distribution with mean $\mu = 290$ (in DKK thousand) and standard deviation $\sigma = 4$ (DKK thousand).

Question a):

What is the probability that a randomly selected teacher earns more than DKK 300.000?

Example 3

Question a):

What is the probability that a randomly selected teacher earns more than DKK 300.000?

Example 3

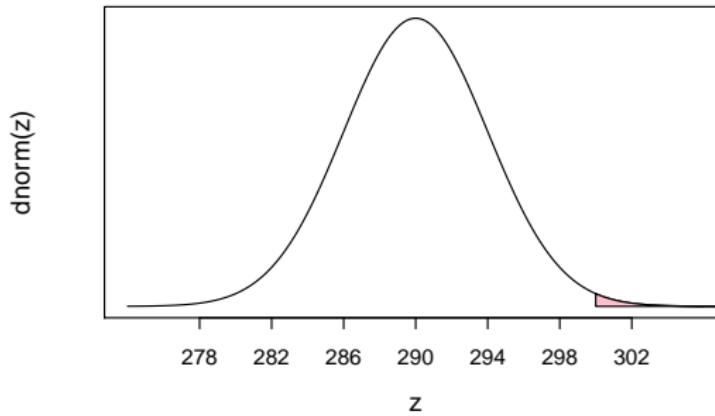
Question a):

What is the probability that a randomly selected teacher earns more than DKK 300.000?

Answer:

```
1 - pnorm(300, m = 290, s = 4)
```

```
[1] 0.0062
```



Example 4

(Same income distribution):

It is assumed that the annual salary distribution of elementary school teachers can be described using a normal distribution with mean $\mu = 290$ (DKK thousand) and standard deviation $\sigma = 4$ (DKK thousand).

Example 4

(Same income distribution):

It is assumed that the annual salary distribution of elementary school teachers can be described using a normal distribution with mean $\mu = 290$ (DKK thousand) and standard deviation $\sigma = 4$ (DKK thousand).

"Opposite question"

Give a salary interval (symmetric around the mean) which covers 95% of all teachers' salary.

Example 4

(Same income distribution):

It is assumed that the annual salary distribution of elementary school teachers can be described using a normal distribution with mean $\mu = 290$ (DKK thousand) and standard deviation $\sigma = 4$ (DKK thousand).

"Opposite question"

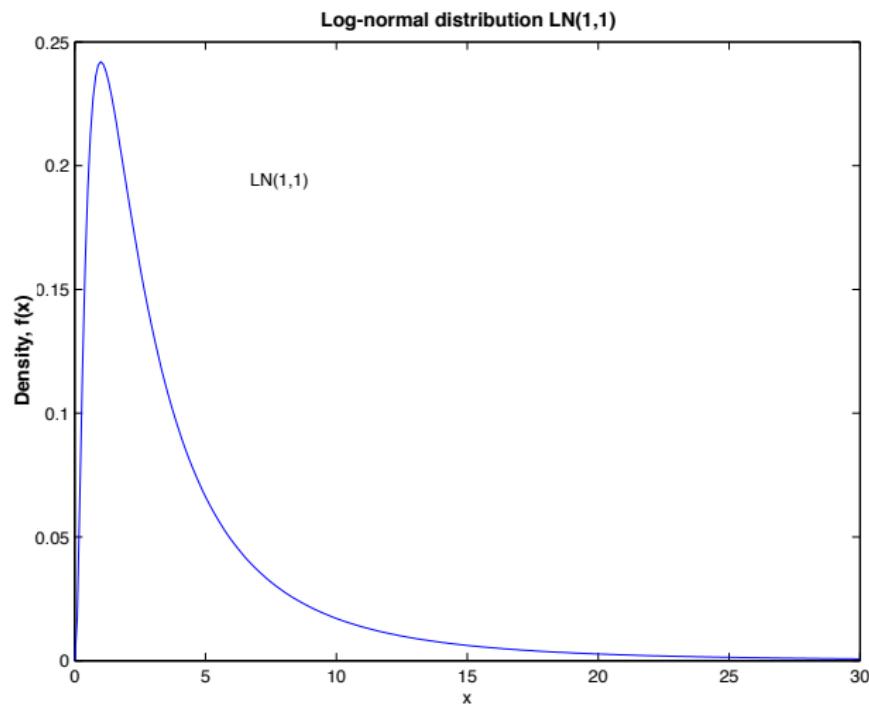
Give a salary interval (symmetric around the mean) which covers 95% of all teachers' salary.

Answer:

```
qnorm(c(0.025, 0.975), m = 290, s = 4)
```

```
[1] 282 298
```

The log-normal distribution



The log-normal distribution, Def. 2.46 & Theo. 2.47

Syntax:

$$X \sim LN(\alpha, \beta^2) \text{ (with } \beta > 0)$$

The log-normal distribution, Def. 2.46 & Theo. 2.47

Syntax:

$$X \sim LN(\alpha, \beta^2) \text{ (with } \beta > 0)$$

Density function:

$$f(x) = \begin{cases} \frac{1}{\beta\sqrt{2\pi}}x^{-1}e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The log-normal distribution, Def. 2.46 & Theo. 2.47

Syntax:

$$X \sim LN(\alpha, \beta^2) \text{ (with } \beta > 0)$$

Density function:

$$f(x) = \begin{cases} \frac{1}{\beta\sqrt{2\pi}}x^{-1}e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean:

$$\mu = e^{\alpha + \beta^2/2}$$

The log-normal distribution, Def. 2.46 & Theo. 2.47

Syntax:

$$X \sim LN(\alpha, \beta^2) \text{ (with } \beta > 0)$$

Density function:

$$f(x) = \begin{cases} \frac{1}{\beta\sqrt{2\pi}}x^{-1}e^{-(\ln(x)-\alpha)^2/2\beta^2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean:

$$\mu = e^{\alpha + \beta^2/2}$$

Variance:

$$\sigma^2 = e^{2\alpha + \beta^2}(e^{\beta^2} - 1)$$

The log-normal distribution

Log-normal and normal distributions:

A log-normal distributed variable $Y \sim LN(\alpha, \beta^2)$ can be transformed into a normal distributed variable:

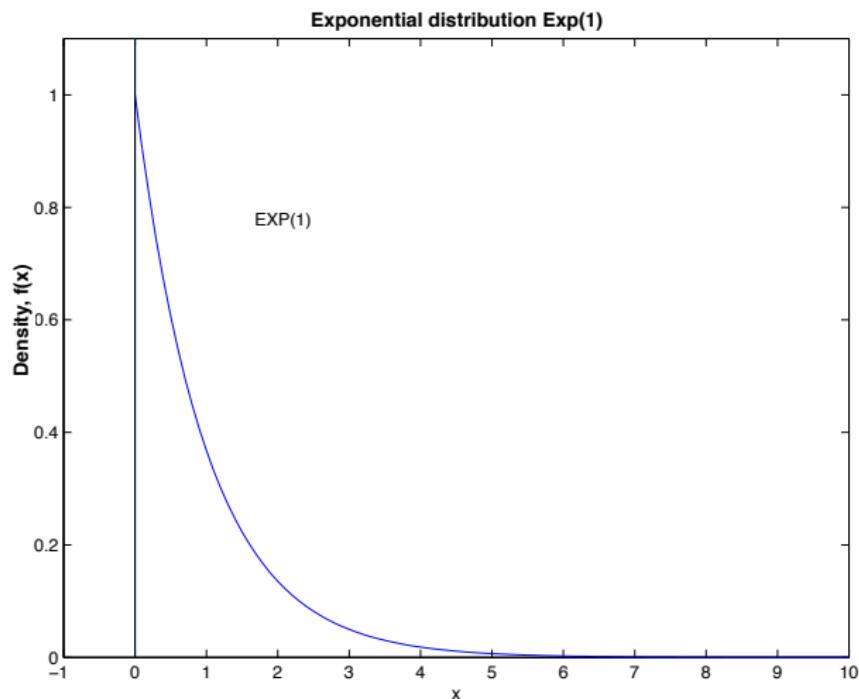
$$X = \ln(Y)$$

is normal distributed with mean α and variance β^2 , i.e. $X \sim N(\alpha, \beta^2)$.

$$Z = \frac{\ln(Y) - \alpha}{\beta}$$

is standard normal distributed, i.e. $Z \sim N(0, 1)$.

The exponential distribution



The exponential distribution, Def. 2.48 & Theo. 2.49

Syntax:

$$X \sim \text{Exp}(\lambda)$$

with $\lambda > 0$.

Density function:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Mean:

$$\mu = \frac{1}{\lambda}$$

Variance:

$$\sigma^2 = \frac{1}{\lambda^2}$$

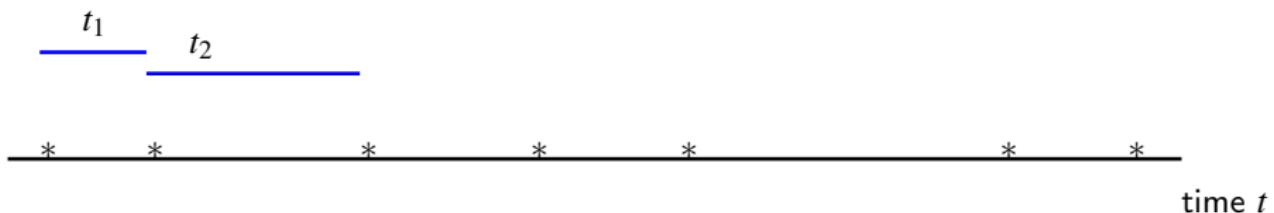
The exponential distribution

- The exponential distribution is a special case of the gamma distribution.
- The exponential distribution is used to describe lifespan and waiting times.
- The exponential distribution can be used to describe (waiting) time between Poisson events.

Connection between the exponential and Poisson distributions

Poisson: Discrete events per unit

Exponential: Continuous distance between events



Example 5

Queuing model – Poisson process

The time between customer arrivals at a post office is exponentially distributed with mean $\mu = 2$ minutes.

Example 5

Queuing model – Poisson process

The time between customer arrivals at a post office is exponentially distributed with mean $\mu = 2$ minutes.

Question:

One customer has just arrived. What is the probability that no other customers will arrive during the next 2 minutes?

Example 5

Queuing model – Poisson process

The time between customer arrivals at a post office is exponentially distributed with mean $\mu = 2$ minutes.

Question:

One customer has just arrived. What is the probability that no other customers will arrive during the next 2 minutes?

Answer:

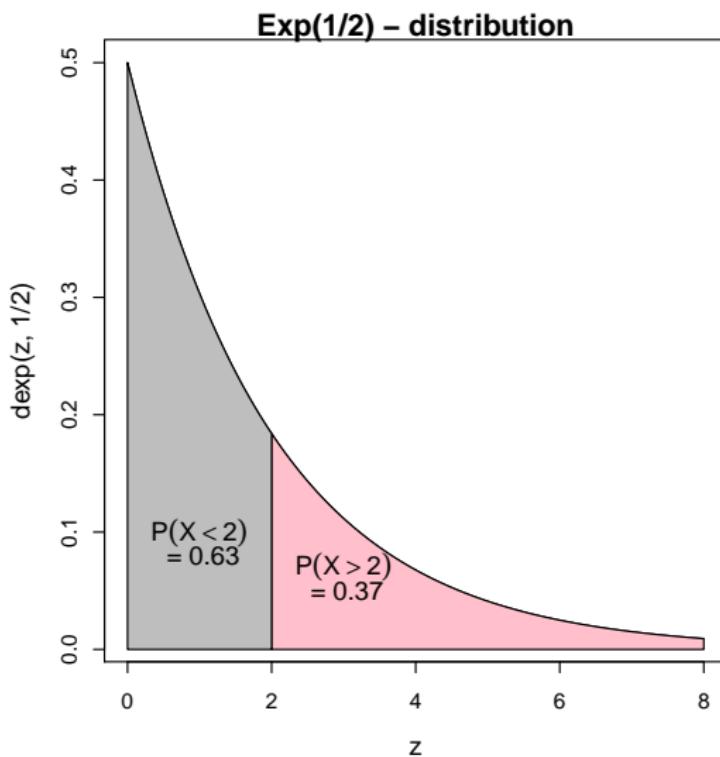
$X \sim \text{Exp}(1/2)$ represents waiting time until next customer.

$$P(X > 2) = 1 - P(X \leq 2)$$

```
1 - pexp(2, rate = 1/2)
```

```
[1] 0.37
```

Example 5



Example 6

Question:

One customer has just arrived. Use the Poisson distribution to calculate the probability that no other customers will arrive during the next two minutes.

Example 6

Question:

One customer has just arrived. Use the Poisson distribution to calculate the probability that no other customers will arrive during the next two minutes.

Answer:

$$\lambda_{2\text{min}} = 1, P(X = 0) = \frac{e^{-1}}{1!} 1^0 = e^{-1}$$

```
dpois(0, 1)
```

```
[1] 0.37
```

```
exp(-1)
```

```
[1] 0.37
```

Overview

- 1 Continuous random variables and distributions
 - Density and distribution functions
 - Mean, variance, and covariance
- 2 Specific continuous distributions
 - The uniform distribution
 - The normal distribution
 - The log-normal distribution
 - The exponential distribution
- 3 Calculation rules for random variables

Calculation rules for random variables

These rules work for both continuous and discrete random variables!

X is a random variable, a and b are constants.

Calculation rules for random variables

These rules work for both continuous and discrete random variables!

X is a random variable, a and b are constants.

Mean rule:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Calculation rules for random variables

These rules work for both continuous and discrete random variables!

X is a random variable, a and b are constants.

Mean rule:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Variance rule:

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

Example 7

X is a random variable with mean 4 and variance 6.

Question:

Calculate the mean and variance of $Y = -3X + 2$

Example 7

X is a random variable with mean 4 and variance 6.

Question:

Calculate the mean and variance of $Y = -3X + 2$

Answer:

$$\mathbb{E}(Y) = -3\mathbb{E}(X) + 2 = -3 \cdot 4 + 2 = -10$$

$$\text{Var}(Y) = (-3)^2 \text{Var}(X) = 9 \cdot 6 = 54$$

Calculation rules for random variables

X_1, \dots, X_n are *independent* random variables.

Calculation rules for random variables

X_1, \dots, X_n are *independent* random variables.

Mean rule:

$$\begin{aligned} & \mathbb{E}(a_1X_1 + a_2X_2 + \dots + a_nX_n) \\ &= a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \dots + a_n\mathbb{E}(X_n) \end{aligned}$$

Calculation rules for random variables

X_1, \dots, X_n are *independent* random variables.

Mean rule:

$$\begin{aligned} & \mathbb{E}(a_1X_1 + a_2X_2 + \dots + a_nX_n) \\ &= a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \dots + a_n\mathbb{E}(X_n) \end{aligned}$$

Variance rule:

$$\begin{aligned} & \text{Var}(a_1X_1 + a_2X_2 + \dots + a_nX_n) \\ &= a_1^2\text{Var}(X_1) + \dots + a_n^2\text{Var}(X_n) \end{aligned}$$

Example 8

Airline Planning

The weight of each passenger on a flight is assumed to be normal distributed $X \sim N(70, 10^2)$.

A plane, which can take 55 passengers, may not have a load exceeding 4000 kg (only the weight of the passengers is considered load).

Example 8

Airline Planning

The weight of each passenger on a flight is assumed to be normal distributed $X \sim N(70, 10^2)$.

A plane, which can take 55 passengers, may not have a load exceeding 4000 kg (only the weight of the passengers is considered load).

Question:

Calculate the probability that the plain is overloaded

Example 8

Airline Planning

The weight of each passenger on a flight is assumed to be normal distributed $X \sim N(70, 10^2)$.

A plane, which can take 55 passengers, may not have a load exceeding 4000 kg (only the weight of the passengers is considered load).

Question:

Calculate the probability that the plain is overloaded

What is $Y = \text{Total passenger weight}$?

Example 8

Airline Planning

The weight of each passenger on a flight is assumed to be normal distributed
 $X \sim N(70, 10^2)$.

A plane, which can take 55 passengers, may not have a load exceeding 4000 kg (only the weight of the passengers is considered load).

Question:

Calculate the probability that the plain is overloaded

What is $Y = \text{Total passenger weight}$?

What is Y ?

Definitely NOT: $Y = 55 \cdot X$

Example 8

What is $Y = \text{Total passenger weight}$?

$Y = \sum_{i=1}^{55} X_i$, where $X_i \sim N(70, 10^2)$ (and assumed to be independent)

Example 8

What is $Y = \text{Total passenger weight}$?

$Y = \sum_{i=1}^{55} X_i$, where $X_i \sim N(70, 10^2)$ (and assumed to be independent)

Mean and variance of Y :

$$\mathbb{E}(Y) = \sum_{i=1}^{55} \mathbb{E}(X_i) = \sum_{i=1}^{55} 70 = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = \sum_{i=1}^{55} \text{Var}(X_i) = \sum_{i=1}^{55} 100 = 55 \cdot 100 = 5500$$

Example 8

What is $Y = \text{Total passenger weight}$?

$Y = \sum_{i=1}^{55} X_i$, where $X_i \sim N(70, 10^2)$ (and assumed to be independent)

Mean and variance of Y :

$$\mathbb{E}(Y) = \sum_{i=1}^{55} \mathbb{E}(X_i) = \sum_{i=1}^{55} 70 = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = \sum_{i=1}^{55} \text{Var}(X_i) = \sum_{i=1}^{55} 100 = 55 \cdot 100 = 5500$$

Y is normal distributed, so we may find $P(Y > 4000)$ using:

```
1-pnorm(4000, mean = 3850, sd = sqrt(5500))
```

[1] 0.022

Example 8 - WRONG ANALYSIS

What is Y ?

Definitely NOT: $Y = 55 \cdot X$

Example 8 - WRONG ANALYSIS

What is Y ?

Definitely NOT: $Y = 55 \cdot X$

Mean and variance of WRONG Y :

$$\mathbb{E}(Y) = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = 55^2 \text{Var}(X) = 55^2 \cdot 100 = 550^2$$

Example 8 - WRONG ANALYSIS

What is Y ?

Definitely NOT: $Y = 55 \cdot X$

Mean and variance of WRONG Y :

$$\mathbb{E}(Y) = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = 55^2 \text{Var}(X) = 55^2 \cdot 100 = 550^2$$

Wrong Y is also normal distributed. Finding $P(Y > 4000)$ using WRONG Y :

```
1 - pnorm(4000, mean = 3850, sd = 550)
```

```
[1] 0.39
```

Example 8 - WRONG ANALYSIS

What is Y ?

Definitely NOT: $Y = 55 \cdot X$

Mean and variance of WRONG Y :

$$\mathbb{E}(Y) = 55 \cdot 70 = 3850$$

$$\text{Var}(Y) = 55^2 \text{Var}(X) = 55^2 \cdot 100 = 550^2$$

Wrong Y is also normal distributed. Finding $P(Y > 4000)$ using WRONG Y :

```
1 - pnorm(4000, mean = 3850, sd = 550)
```

```
[1] 0.39
```

Consequence of wrong calculation:

A LOT of wasted money for the airline company!!!

Overview

- 1 Continuous random variables and distributions
 - Density and distribution functions
 - Mean, variance, and covariance
- 2 Specific continuous distributions
 - The uniform distribution
 - The normal distribution
 - The log-normal distribution
 - The exponential distribution
- 3 Calculation rules for random variables

Course 02402 Introduction to Statistics

Lecture 4: Confidence intervals

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

Example: Heights

Sample, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Example: Heights

Sample, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean and standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Example: Heights

Sample, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean and standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimate population mean and standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

Example: Heights

Sample, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean and standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimate population mean and standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

NEW: Confidence interval for μ :

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} = [169.3; 186.7]$$

NEW: Confidence interval for σ :

$$[8.4; 22.3]$$

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

(Empirical) distribution of sample mean

```
# 'True' mean and standard deviation
mu <- 178
sigma <- 12

# Sample size
n <- 10

# Simulate normal distributed X_i for n = 10
x <- rnorm(n = n, mean = mu, sd = sigma)
x

# Empirical density
hist(x, prob = TRUE, col = 'blue')
# Compute sample mean
mean(x)

# Repeat the simulated sampling many times (100 samples)
mat <- replicate(100, rnorm(n = n, mean = mu, sd = sigma))

# Compute the sample mean for each sample
xbar <- apply(mat, 2, mean)
xbar

# See the distribution of the sample means
hist(xbar, prob = TRUE, col = 'blue')
# Empirical mean and variance of sample means
mean(xbar)
var(xbar)
```

Theorem 3.3: Distribution of the sample mean of i.i.d. normal random variables

The distribution of \bar{X}

Assume that X_1, \dots, X_n are independent and identically distributed (*i.i.d.*) normal random variables, $X_i \sim N(\mu, \sigma^2), i = 1, \dots, n$, then:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Mean, variance and 'normality' follow from 'rules':

The mean of \bar{X} (Theorem 2.56):

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu$$

Mean, variance and 'normality' follow from 'rules':

The mean of \bar{X} (Theorem 2.56):

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu$$

The variance of \bar{X} (Theorem 2.56):

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Mean, variance and 'normality' follow from 'rules':

The mean of \bar{X} (Theorem 2.56):

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu$$

The variance of \bar{X} (Theorem 2.56):

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

The 'normality' of \bar{X} (Theorem 2.40):

By this theorem, the distribution of \bar{X} is a normal distribution with mean μ and variance σ^2/n as specified above.

Distribution of the error $\bar{X} - \mu$

The standard deviation of \bar{X} :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Distribution of the error $\bar{X} - \mu$

The standard deviation of \bar{X} :

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of $\bar{X} - \mu$:

$$\sigma_{(\bar{X} - \mu)} = \frac{\sigma}{\sqrt{n}}$$

Standardized version of the above, Theorem 3.4

Distribution of the standardized sample mean (or standardized error):

Assume that X_1, \dots, X_n are i.i.d. normal random variables, $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$, then:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

That is, the standardized sample mean Z follows a standard normal distribution.

Practical problem (and solution)

How do we use the results from the previous slides to say something about μ ...

... when the 'true', unknown, population standard deviation σ enters into all the formulas?

Practical problem (and solution)

How do we use the results from the previous slides to say something about μ ...

... when the 'true', unknown, population standard deviation σ enters into all the formulas?

Obvious solution:

Use the estimate s instead of σ in formulas.

Practical problem (and solution)

How do we use the results from the previous slides to say something about μ ...

... when the 'true', unknown, population standard deviation σ enters into all the formulas?

Obvious solution:

Use the estimate s instead of σ in formulas.

BUT:

Then, we need new theory! (There is also uncertainty linked to s .)

Theorem 3.5, a more applicable extension of the above

The t -distribution takes the uncertainty of s into account:

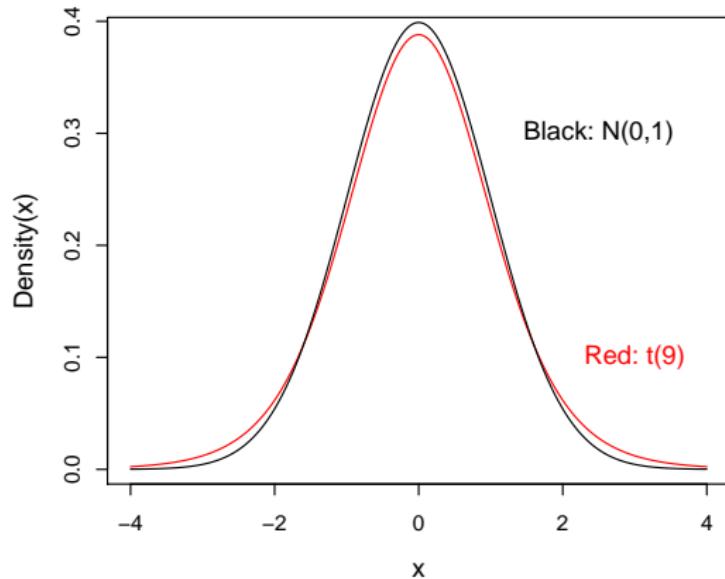
Assume that X_1, \dots, X_n are i.i.d. normal distributed random variables, where $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$, then:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

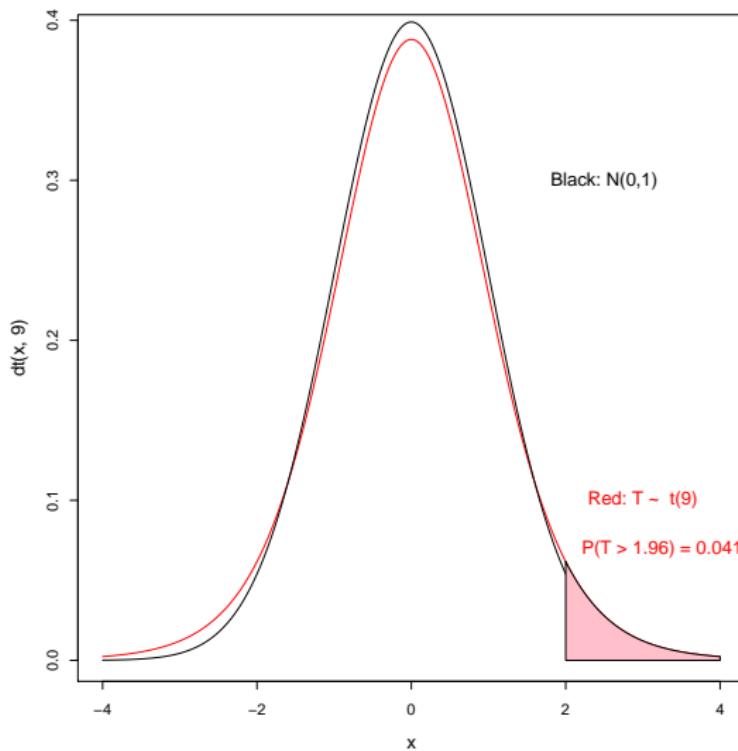
where $t(n-1)$ is the t -distribution with $n-1$ degrees of freedom.

The *t*-distribution with 9 degrees of freedom ($n = 10$)

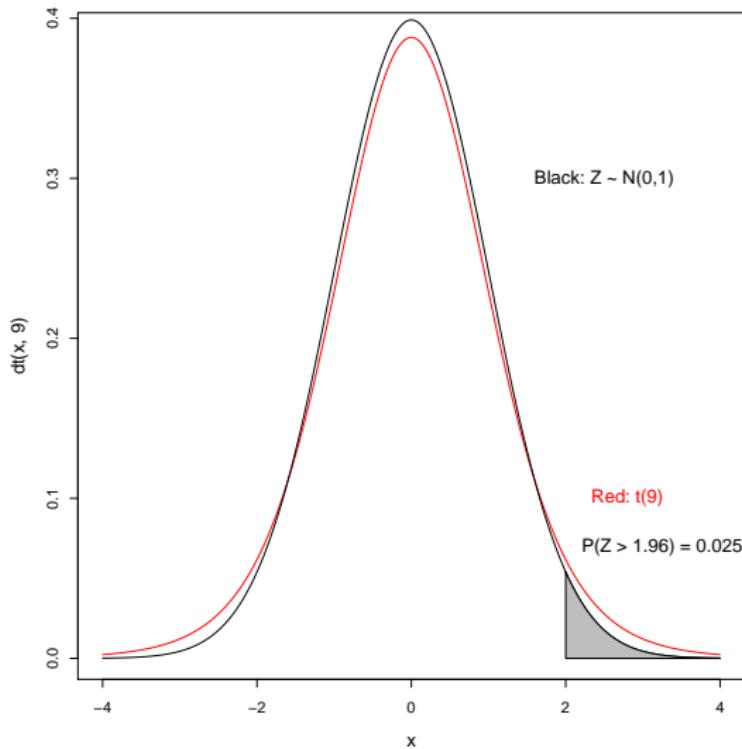
```
x <- seq(-4, 4, by = 0.01)
plot(x, dt(x, df = 9), type = "l", col = "red", ylab = "Density(x)")
lines(x, dnorm(x), type = "l")
text(2.5, 0.3, "Black: N(0,1)")
text(3, 0.1, "Red: t(9)", col = "red")
```



The t -distribution with 9 degrees of freedom and standard normal distribution



The t -distribution with 9 degrees of freedom and standard normal distribution



Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

Method 3.9: One-sample Confidence Interval (CI) for μ

Use the correct t -distribution to construct the confidence interval:

For a sample x_1, \dots, x_n the $100(1 - \alpha)\%$ confidence interval is given by:

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where $t_{1-\alpha/2}$ is the $100(1 - \alpha/2)\%$ quantile from the t -distribution with $n - 1$ degrees of freedom.

Method 3.9: One-sample Confidence Interval (CI) for μ

Use the correct t -distribution to construct the confidence interval:

For a sample x_1, \dots, x_n the $100(1 - \alpha)\%$ confidence interval is given by:

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where $t_{1-\alpha/2}$ is the $100(1 - \alpha/2)\%$ quantile from the t -distribution with $n - 1$ degrees of freedom.

Most commonly using $\alpha = 0.05$:

The most commonly used is the 95% confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

Example: Heights, 95% CI

```
# 0.975 quantile for the t(9) distribution (n = 10):  
qt(0.975, df = 9)
```

Gives the result $t_{0.975} = 2.26$.

Now, we can recognize the already given result

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}}$$

which is

$$178 \pm 8.74 = [169.3, 186.7].$$

Example: Heights, 99% CI

```
# 0.995 quantile for the t(9) distribution (n = 10):  
qt(0.995, df = 9)
```

Gives the result $t_{0.995} = 3.25$.

In this case,

$$178 \pm 3.25 \cdot \frac{12.21}{\sqrt{10}}$$

giving

$$178 \pm 12.55 = [165.5; 190.5]$$

An R function for computing these CI (and more):

```
# Data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)

# 99% CI for mu
t.test(x, conf.level = 0.99)

##
##  One Sample t-test
##
## data: x
## t = 46, df = 9, p-value = 5e-12
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
## 165.5 190.5
## sample estimates:
## mean of x
## 178
```

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

The formal framework for *statistical inference*

From eNote, Chapter 1:

- An *observational unit* is the single entity/level at which information is sought (e.g. a person). (**Observationsenhed**)
- The *statistical population* consists of all possible “measurements” on each possible *observational unit*. (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Stikprøve**)

The formal framework for *statistical inference*

From eNote, Chapter 1:

- An *observational unit* is the single entity/level at which information is sought (e.g. a person). (**Observationsenhed**)
- The *statistical population* consists of all possible “measurements” on each possible *observational unit*. (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Stikprøve**)

Language and concepts:

- μ and σ are parameters describing the population.
- \bar{x} is the *estimate* of μ (specific realization).
- \bar{X} is the *estimator* of μ (now seen as a random variable).
- The word '*statistic(s)*' is used for both.

The formal framework for *statistical inference* - Example

From eNote, Chapter 1. Example: Heights

We measured the heights of 10 randomly selected students.

The formal framework for *statistical inference* - Example

From eNote, Chapter 1. Example: Heights

We measured the heights of 10 randomly selected students.

The sample:

The 10 specific numbers (heights): x_1, \dots, x_{10} .

The formal framework for *statistical inference* - Example

From eNote, Chapter 1. Example: Heights

We measured the heights of 10 randomly selected students.

The sample:

The 10 specific numbers (heights): x_1, \dots, x_{10} .

The population:

The heights for all people in Denmark.

The formal framework for *statistical inference* - Example

From eNote, Chapter 1. Example: Heights

We measured the heights of 10 randomly selected students.

The sample:

The 10 specific numbers (heights): x_1, \dots, x_{10} .

The population:

The heights for all people in Denmark.

Observational unit:

A person.

Statistical inference = Learning from data

Learning from data:

Learning about parameters of distributions that describe populations.

Statistical inference = Learning from data

Learning from data:

Learning about parameters of distributions that describe populations.

Important:

The sample must, in a meaningful way, represent some well defined population.

Statistical inference = Learning from data

Learning from data:

Learning about parameters of distributions that describe populations.

Important:

The sample must, in a meaningful way, represent some well defined population.

How to ensure this:

For example, by making sure that the sample is taken completely at random.

Random Sampling

Definition 3.12:

- A random sample from an (infinite) population: A set of observations X_1, X_2, \dots, X_n constitutes a random sample of size n from the infinite population $f(x)$ if:
 - ① Each X_i is a random variable whose distribution is given by $f(x)$.
 - ② These n random variables are independent.

Random Sampling

Definition 3.12:

- A random sample from an (infinite) population: A set of observations X_1, X_2, \dots, X_n constitutes a random sample of size n from the infinite population $f(x)$ if:
 - ① Each X_i is a random variable whose distribution is given by $f(x)$.
 - ② These n random variables are independent.

What does that mean?

- ① All observations must come from the same population.
- ② They cannot share any information with each other (e.g., shouldn't be related).

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

Theorem 3.14: The Central Limit Theorem (CLT)

"No matter the distribution of X_i ", the distribution of the mean of i.i.d. random variables approaches a normal distribution:

Let \bar{X} be the mean of a random sample of size n taken from a population with mean μ and variance σ^2 . Then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a random variable whose distribution function approaches that of the standard normal distribution, $N(0, 1^2)$, as $n \rightarrow \infty$.

Theorem 3.14: The Central Limit Theorem (CLT)

"No matter the distribution of X_i ", the distribution of the mean of i.i.d. random variables approaches a normal distribution:

Let \bar{X} be the mean of a random sample of size n taken from a population with mean μ and variance σ^2 . Then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is a random variable whose distribution function approaches that of the standard normal distribution, $N(0, 1^2)$, as $n \rightarrow \infty$.

Hence, if n is large enough, we can assume (approximately) that:

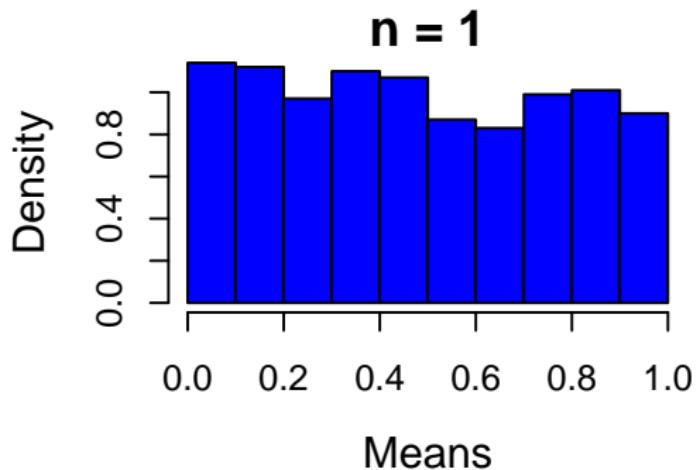
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

CLT example: Mean of uniformly distributed observations

```
n <- 1 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 1)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 1", xlab = "Means", prob = TRUE)
```

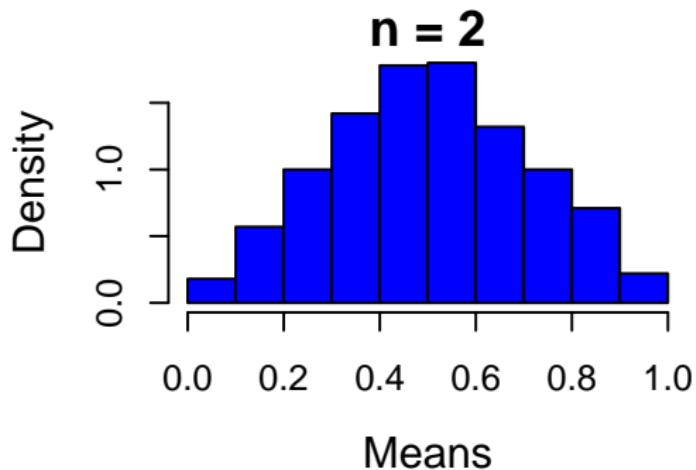


CLT example: Mean of uniformly distributed observations

```
n <- 2 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 2)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 2", xlab = "Means", xlim = c(0,1), prob = TRUE)
```

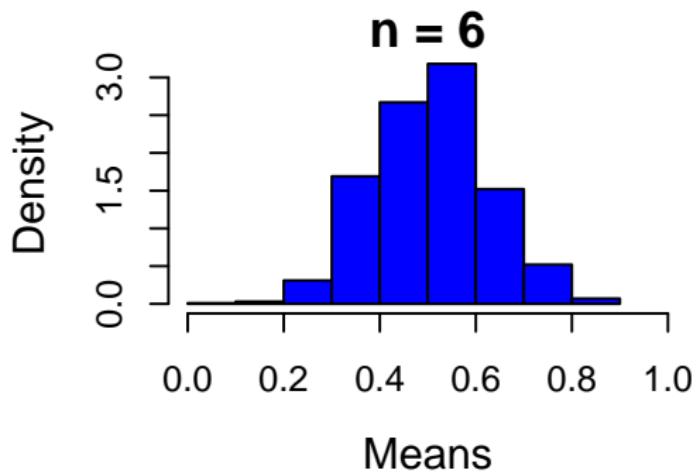


CLT example: Mean of uniformly distributed observations

```
n <- 6 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 6)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 6", xlab = "Means", xlim = c(0,1), prob = TRUE)
```

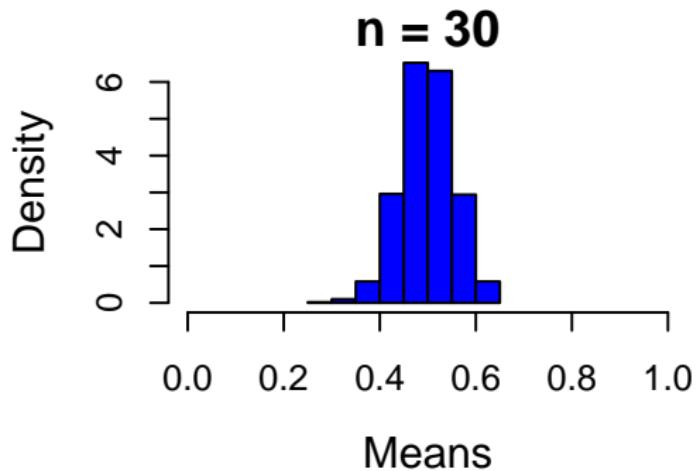


CLT example: Mean of uniformly distributed observations

```
n <- 30 # Sample size
k <- 1000 # No. of samples (i.e. no. of means to be computed)

# Simulations from U(0,1)-distribution (k = 1000 samples, each of size n = 30)
u <- matrix(runif(k*n), ncol = n)

# Empirical density of means
hist(apply(u, 1, mean), col = "blue", main = "n = 30", xlab = "Means", xlim = c(0,1), prob = TRUE)
```



Consequence of the CLT:

Our CI-method also works for non-normal data:

We can use the confidence-interval based on the t -distribution in basically any situation, as long as n is large enough.

Consequence of the CLT:

Our CI-method also works for non-normal data:

We can use the confidence-interval based on the t -distribution in basically any situation, as long as n is large enough.

When is n "large enough"?

Actually difficult to say exactly, BUT:

- Rule of thumb: $n \geq 30$
- Even for smaller n the approach can be (almost) valid for non-normal data.

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 **Formal interpretation of the CI**
- 7 CI for variance σ^2 and standard deviation σ

'Repeated sampling' interpretation

In the long run, we catch the true value in 95% of cases (95% CI):

The confidence interval will vary in both width (s) and position (\bar{x}) if the study is repeated.

'Repeated sampling' interpretation

In the long run, we catch the true value in 95% of cases (95% CI):

The confidence interval will vary in both width (s) and position (\bar{x}) if the study is repeated.

More formally expressed (Theorem 3.5):

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < t_{0.975}\right) = 0.95$$

'Repeated sampling' interpretation

In the long run, we catch the true value in 95% of cases (95% CI):

The confidence interval will vary in both width (s) and position (\bar{x}) if the study is repeated.

More formally expressed (Theorem 3.5):

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} < t_{0.975}\right) = 0.95$$

Which is equivalent to:

$$P\left(\bar{X} - t_{0.975} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0.975} \frac{S}{\sqrt{n}}\right) = 0.95$$

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

Motivating Example

Production of tablets

In the production of tablets, an active matter is mixed with a powder and then the mixture is formed to tablets. It is important that the mixture is homogenous, so that each tablet has the same strength.

We consider a mixture (of the active matter and powder) from where a large amount of tablets is to be produced.

We seek to produce the mixtures (and the final tablets) so that the mean content of the active matter is 1 mg/g with the smallest variance as possible. A random sample is collected where the amount of active matter is measured. It is assumed that all the measurements follow a normal distribution with the unit mg/g.

The sampling distribution of the variance estimator, Theorem 2.81

Assume i.i.d. normal distributed variables, $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$.

Variance estimators behaves like a χ^2 -distribution:

Let

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

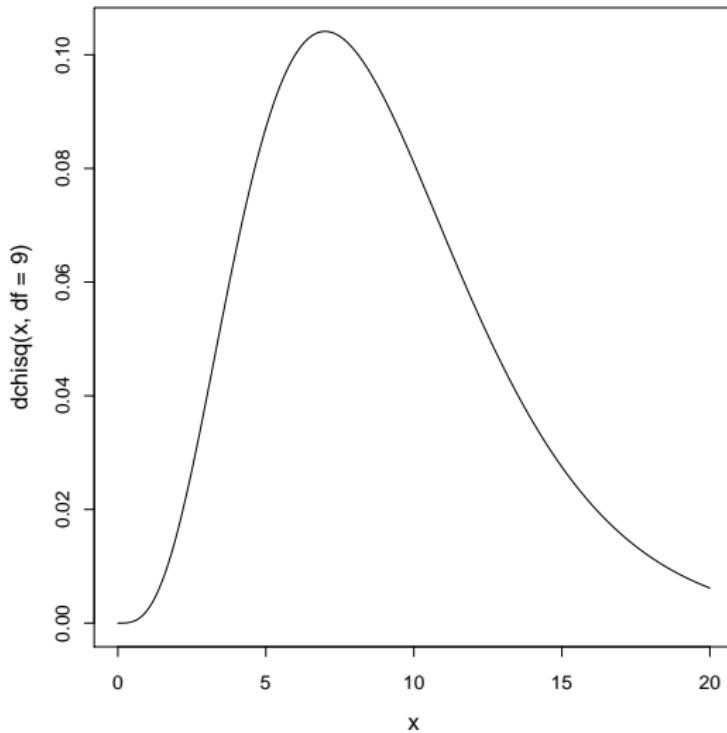
then:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is a stochastic variable following the χ^2 -distribution with $v = n - 1$ degrees of freedom.

χ^2 -distribution with $v = 9$ degrees of freedom

```
x <- seq(0, 20, by = 0.1)
plot(x, dchisq(x, df = 9), type = "l")
```



Method 3.19: Confidence interval for the variance and standard deviation

Assume i.i.d. normal distributed variables, $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$.

The variance:

A $100(1 - \alpha)\%$ confidence interval for the variance σ^2 is:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \right]$$

where the quantiles come from a χ^2 -distribution with $v = n - 1$ degrees of freedom.

Method 3.19: Confidence interval for the variance and standard deviation

Assume i.i.d. normal distributed variables, $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots, n$.

The variance:

A $100(1 - \alpha)\%$ confidence interval for the variance σ^2 is:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \right]$$

where the quantiles come from a χ^2 -distribution with $v = n - 1$ degrees of freedom.

The standard deviation:

A $100(1 - \alpha)\%$ confidence interval for the standard deviation σ is:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} \right]$$

Example

Data:

A random sample with $n = 20$ tablets is taken and from this we get:

$$\hat{\mu} = \bar{x} = 1.01, \hat{\sigma}^2 = s^2 = 0.07^2$$

Example

Data:

A random sample with $n = 20$ tablets is taken and from this we get:

$$\hat{\mu} = \bar{x} = 1.01, \hat{\sigma}^2 = s^2 = 0.07^2$$

95% confidence interval for the variance - we need the χ^2 -quantiles (19 degrees of freedom):

$$\chi^2_{0.025} = 8.9065, \chi^2_{0.975} = 32.8523$$

```
qchisq(c(0.025, 0.975), df = 19)
```

```
[1] 8.907 32.852
```

Example

So the confidence interval for the variance σ^2 becomes:

$$\left[\frac{19 \cdot 0.07^2}{32.85}; \frac{19 \cdot 0.07^2}{8.907} \right] = [0.002834; 0.01045]$$

Example

So the confidence interval for the variance σ^2 becomes:

$$\left[\frac{19 \cdot 0.07^2}{32.85}; \frac{19 \cdot 0.07^2}{8.907} \right] = [0.002834; 0.01045]$$

and the confidence interval for the standard deviation σ becomes:

$$\left[\sqrt{0.002834}; \sqrt{0.01045} \right] = [0.053; 0.102]$$

Example: Heights

We need the χ^2 -quantiles with $v = 9$ degrees of freedom:

$$\chi^2_{0.025} = 2.700389, \chi^2_{0.975} = 19.022768$$

```
qchisq(c(0.025, 0.975), df = 9)
```

```
[1] 2.70 19.02
```

Example: Heights

We need the χ^2 -quantiles with $v = 9$ degrees of freedom:

$$\chi^2_{0.025} = 2.700389, \chi^2_{0.975} = 19.022768$$

```
qchisq(c(0.025, 0.975), df = 9)
```

```
[1] 2.70 19.02
```

So the confidence interval for the height standard deviation σ becomes:

$$\left[\sqrt{\frac{9 \cdot 12.21^2}{19.022768}}; \sqrt{\frac{9 \cdot 12.21^2}{2.700389}} \right] = [8.4; 22.3]$$

Example: Heights

Sample, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean and standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimate population mean and standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

NEW: Confidence interval, μ :

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} = [169.3; 186.7]$$

NEW: Confidence interval, σ :

$$[8.4; 22.3]$$

Overview

- 1 Intro and example
- 2 Distribution of the sample mean
 - The t -distribution
- 3 Confidence interval (CI) for μ
 - Example: Heights
- 4 The language of statistics and the formal framework
- 5 Non-normal data, the Central Limit Theorem (CLT)
- 6 Formal interpretation of the CI
- 7 CI for variance σ^2 and standard deviation σ

Course 02402 Introduction to Statistics

Lecture 5: Hypothesis testing

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Motivating example - sleep medicine
- 2 One-sample t -test and p -value
- 3 Critical value and relation to the confidence interval
- 4 Hypothesis tests in general
 - The alternative hypothesis
 - The general method
 - Errors in hypothesis testing
- 5 Checking the normality assumption
 - The normal q-q plot
 - Transformation towards normality

Overview

- 1 Motivating example - sleep medicine
- 2 One-sample t -test and p -value
- 3 Critical value and relation to the confidence interval
- 4 Hypothesis tests in general
 - The alternative hypothesis
 - The general method
 - Errors in hypothesis testing
- 5 Checking the normality assumption
 - The normal q-q plot
 - Transformation towards normality

Motivating example - sleep medicine

Difference between sleep medicines?

In a study, the aim is to compare two kinds of sleep medicine, A and B. 10 test persons tried both kinds of medicine, and the following 10 *differences* between the two types of medicine were measured:

(For person 1, sleep medicine B was 1.2 sleep hours better than medicine A, etc.):

Sample, $n = 10$:

person	$x = B \text{ effect} - A \text{ effect}$
1	1.2
2	2.4
3	1.3
4	1.3
5	0.9
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

Example - sleep medicine

The hypothesis of no difference:

$$H_0: \mu = 0$$

where μ represents mean difference in sleep length.

Example - sleep medicine

The hypothesis of no difference:

$$H_0: \mu = 0$$

where μ represents mean difference in sleep length.

Sample mean and std. deviation:

$$\bar{x} = 1.670 = \hat{\mu}$$

$$s = 1.13 = \hat{\sigma}$$

Example - sleep medicine

The hypothesis of no difference:

$$H_0: \mu = 0$$

where μ represents mean difference in sleep length.

Sample mean and std. deviation:

$$\bar{x} = 1.670 = \hat{\mu}$$

$$s = 1.13 = \hat{\sigma}$$

NEW p -value:

$$p\text{-value} = 0.00117$$

(Computed under the scenario that H_0 is true).

Is data in accordance with the null hypothesis H_0 ?

Data: $\bar{x} = 1.67$, $H_0: \mu = 0$

NEW Conclusion:

As the data is far away from H_0 (unlikely under H_0), we **reject** H_0 . There is a **significant effect** of medicine B compared to A.

Overview

- 1 Motivating example - sleep medicine
- 2 One-sample t -test and p -value
- 3 Critical value and relation to the confidence interval
- 4 Hypothesis tests in general
 - The alternative hypothesis
 - The general method
 - Errors in hypothesis testing
- 5 Checking the normality assumption
 - The normal q-q plot
 - Transformation towards normality

Method 3.23: One-sample *t*-test and *p*-value

How to compute the *p*-value?

For a (quantitative) one sample situation, the (non-directional) *p*-value is given by:

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|)$$

where T follows a *t*-distribution with $(n - 1)$ degrees of freedom.
The observed value of the test statistics to be computed is

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where μ_0 is the value of μ under the null hypothesis:

$$H_0 : \mu = \mu_0$$

The definition and interpretation of the *p*-value (completely general)

The *p*-value expresses evidence against the null hypothesis – Table 3.1:

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

Definition 3.22 of the *p*-value:

The *p*-value is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

Example - sleep medicine

The hypothesis of no difference:

$$H_0: \mu = 0$$

where μ represents mean difference in sleep length.

Example - sleep medicine

The hypothesis of no difference:

$$H_0: \mu = 0$$

where μ represents mean difference in sleep length.

Compute the test-statistic:

$$t_{\text{obs}} = \frac{1.67 - 0}{1.13/\sqrt{10}} = 4.67$$

Example - sleep medicine

The hypothesis of no difference:

$$H_0: \mu = 0$$

where μ represents mean difference in sleep length.

Compute the *p*-value:

Compute the test-statistic:

$$t_{\text{obs}} = \frac{1.67 - 0}{1.13/\sqrt{10}} = 4.67$$

$$2P(T > 4.67) = 0.00117$$

```
2 * (1 - pt(4.67, df = 9))
```

Interpretation of the *p*-value in light of Table 3.1:

There is strong evidence against the null hypothesis.

Example - sleep medicine - in R, manually

```
# Enter data
x <- c(1.2, 2.4, 1.3, 1.3, 0.9, 1.0, 1.8, 0.8, 4.6, 1.4)
n <- length(x) # sample size

# Compute 'tobs' - the observed test statistic
tobs <- (mean(x) - 0) / (sd(x) / sqrt(n))

# Compute the p-value as a tail-probability
# in the relevant t-distribution:
2 * (1 - pt(abs(tobs), df = n-1))

## [1] 0.001166
```

Example - sleeping medicine - in R, with built-in function

```
t.test(x)

##
##  One Sample t-test
##
## data: x
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of x
##      1.67
```

Definition of a hypothesis test and significance (general)

Definition 3.24. Hypothesis test:

We say that we *carry out a hypothesis test* when we decide against a null hypothesis or not, using the data.

A null hypothesis is *rejected* if the *p*-value, calculated after the data has been observed, is less than some α , that is if the p -value $< \alpha$, where α is some pre-specified (so-called) *significance level*.

Otherwise, the null hypothesis is said to be '*accepted*'.

Definition 3.29. Statistical significance:

An *effect* is said to be (*statistically*) *significant* if the *p*-value is less than the significance level α .

Often, we use $\alpha = 0.05$.

Example - sleep medicine

With $\alpha = 0.05$:

Since the p -value is less than α , we **reject** the null hypothesis.

Example - sleep medicine

With $\alpha = 0.05$:

Since the *p*-value is less than α , we **reject** the null hypothesis.

In conclusion:

We have found a **significant effect** of medicine B when compared to A (and B works better than A).

Overview

- 1 Motivating example - sleep medicine
- 2 One-sample t -test and p -value
- 3 Critical value and relation to the confidence interval
- 4 Hypothesis tests in general
 - The alternative hypothesis
 - The general method
 - Errors in hypothesis testing
- 5 Checking the normality assumption
 - The normal q-q plot
 - Transformation towards normality

Critical value

Definition 3.31 - the critical values of the t -test:

The $(1 - \alpha)100\%$ critical values for the (non-directional) one-sample t -test are the $(\alpha/2)100\%$ and $(1 - \alpha/2)100\%$ quantiles of the t -distribution with $n - 1$ degrees of freedom:

$$t_{\alpha/2} \text{ and } t_{1-\alpha/2}$$

Critical value

Definition 3.31 - the critical values of the t -test:

The $(1 - \alpha)100\%$ critical values for the (non-directional) one-sample t -test are the $(\alpha/2)100\%$ and $(1 - \alpha/2)100\%$ quantiles of the t -distribution with $n - 1$ degrees of freedom:

$$t_{\alpha/2} \text{ and } t_{1-\alpha/2}$$

Method 3.32: One-sample t -test by critical value:

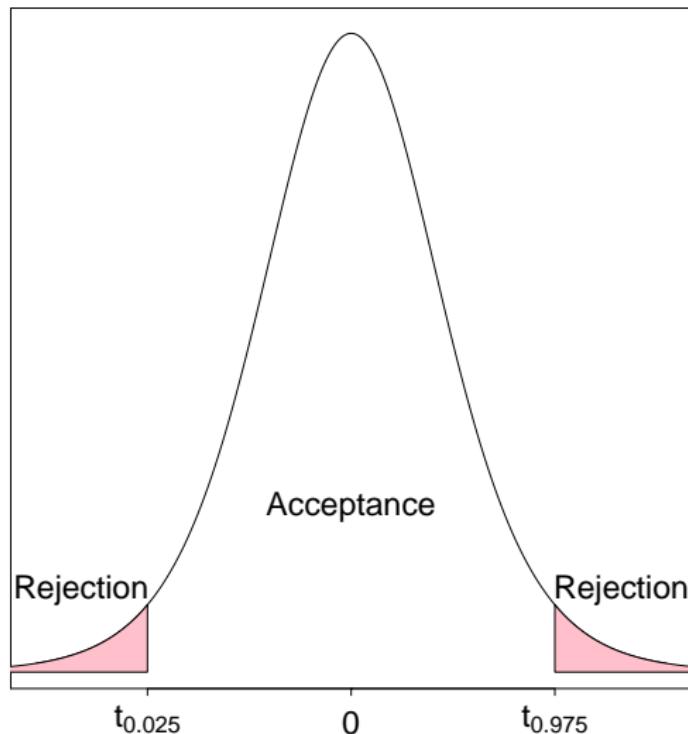
A null hypothesis is *rejected* if the observed test-statistic is more extreme than the critical values:

If $|t_{\text{obs}}| > t_{1-\alpha/2}$ then *reject*

otherwise *accept*.

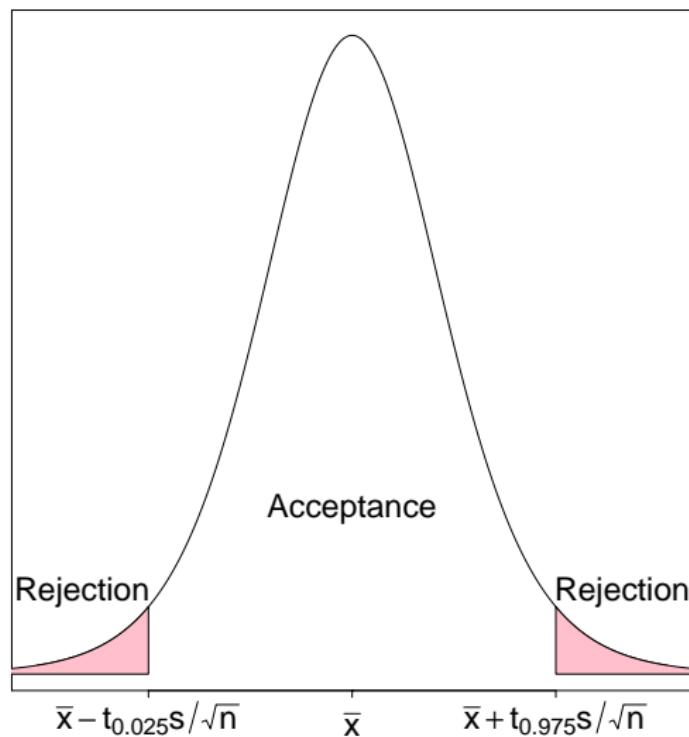
Critical value and hypothesis test

The acceptance region consists of the values of μ which are not too far away from the sample mean - here on the standardized scale:



Critical value and hypothesis test

The acceptance region consists of the values of μ which are not too far away from the sample mean - now on the original scale:



Critical value, confidence interval and hypothesis test

Theorem 3.33: Critical value method = Confidence interval method

We consider a $(1 - \alpha) \cdot 100\%$ confidence interval for μ :

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The confidence interval corresponds to the acceptance region for H_0 when testing the (non-directional) hypothesis

$$H_0 : \mu = \mu_0$$

Critical value, confidence interval and hypothesis test

Theorem 3.33: Critical value method = Confidence interval method

We consider a $(1 - \alpha) \cdot 100\%$ confidence interval for μ :

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The confidence interval corresponds to the acceptance region for H_0 when testing the (non-directional) hypothesis

$$H_0 : \mu = \mu_0$$

(New) interpretation of the confidence interval:

The confidence interval covers those values of the parameter that we believe in given the data.

(Those values that we accept by the corresponding hypothesis test.)

Proof:

Remark 3.34

A μ_0 inside the confidence interval satisfies that

$$|\bar{x} - \mu_0| < t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

which is equivalent to

$$\frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} < t_{1-\alpha/2}$$

and again to

$$|t_{\text{obs}}| < t_{1-\alpha/2}$$

which then exactly states that μ_0 is accepted, since the t_{obs} is within the critical values.

Overview

- 1 Motivating example - sleep medicine
- 2 One-sample t -test and p -value
- 3 Critical value and relation to the confidence interval
- 4 Hypothesis tests in general
 - The alternative hypothesis
 - The general method
 - Errors in hypothesis testing
- 5 Checking the normality assumption
 - The normal q-q plot
 - Transformation towards normality

The alternative hypothesis

So far - implied: (= non-directional)

The alternative to $H_0 : \mu = \mu_0$ is $H_1 : \mu \neq \mu_0$.

The alternative hypothesis

So far - implied: (= non-directional)

The alternative to $H_0 : \mu = \mu_0$ is $H_1 : \mu \neq \mu_0$.

BUT there are other possible settings, e.g. one-sided (= directional), "less":

The alternative to $H_0 : \mu = \mu_0$ is $H_1 : \mu < \mu_0$.

The alternative hypothesis

So far - implied: (= non-directional)

The alternative to $H_0 : \mu = \mu_0$ is $H_1 : \mu \neq \mu_0$.

BUT there are other possible settings, e.g. one-sided (= directional), "less":

The alternative to $H_0 : \mu = \mu_0$ is $H_1 : \mu < \mu_0$.

We stick to the "non-directional" in this course!

Steps of a hypothesis test - an overview

Generally, a hypothesis test consists of the following steps:

- ① Formulate the hypothesis and choose the level of significance α (choose the "risk-level").
- ② Calculate, using the data, the value of the test statistic.
- ③ Calculate the p-value using the test statistic and the relevant distribution. Compare the p -value to the significance level α and make a conclusion.

OR:

Alternatively, make a conclusion based on the relevant critical value(s).

The one-sample t-test again

Method 3.36 The level α one-sample t-test:

- ① Compute t_{obs} as before.
- ② Compute evidence against the *null hypothesis* $H_0: \mu = \mu_0$ vs. the *alternative hypothesis* $H_1: \mu \neq \mu_0$ by the

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|),$$

where the t -distribution with $n - 1$ degrees of freedom is used.

- ③ If $p\text{-value} < \alpha$, we reject H_0 . Otherwise, we accept H_0 .

OR:

The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value(s) $\pm t_{1-\alpha/2}$:

If $|t_{\text{obs}}| > t_{1-\alpha/2}$ we reject H_0 , otherwise we accept H_0 .

Errors in hypothesis testing

Two kind of errors can occur (but only one at a time!):

Type I: Rejection of H_0 when H_0 is true.

Type II: Non-rejection (acceptance) of H_0 when H_1 is true.

The risks of the two types of errors are:

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

Court of law analogy

A man is standing in a court of law:

A man is standing in a court of law, accused of criminal activity.

The null- and the alternative hypotheses are:

H_0 : The man is not guilty.

H_1 : The man is guilty.

Court of law analogy

A man is standing in a court of law:

A man is standing in a court of law, accused of criminal activity.

The null- and the alternative hypotheses are:

H_0 : The man is not guilty.

H_1 : The man is guilty.

Not being able to prove that the man is guilty is not the same as *proving* that he is innocent.

Or, put differently:

Accepting a null hypothesis is NOT a statistical proof of the null hypothesis being true!

Errors in hypothesis testing

Theorem 3.39: Significance level = The risk of a Type I error

The significance level α in hypothesis testing is the overall Type I risk:

$$P(\text{Type I error}) = P(\text{Rejection of } H_0 \text{ when } H_0 \text{ is true}) = \alpha$$

Errors in hypothesis testing

Theorem 3.39: Significance level = The risk of a Type I error

The significance level α in hypothesis testing is the overall Type I risk:

$$P(\text{Type I error}) = P(\text{Rejection of } H_0 \text{ when } H_0 \text{ is true}) = \alpha$$

Two possible truths vs. two possible conclusions:

	Reject H_0	Fail to reject H_0
H_0 is true	Type I error (α)	Correct acceptance of H_0
H_0 is false	Correct rejection of H_0 (Power)	Type II error (β)

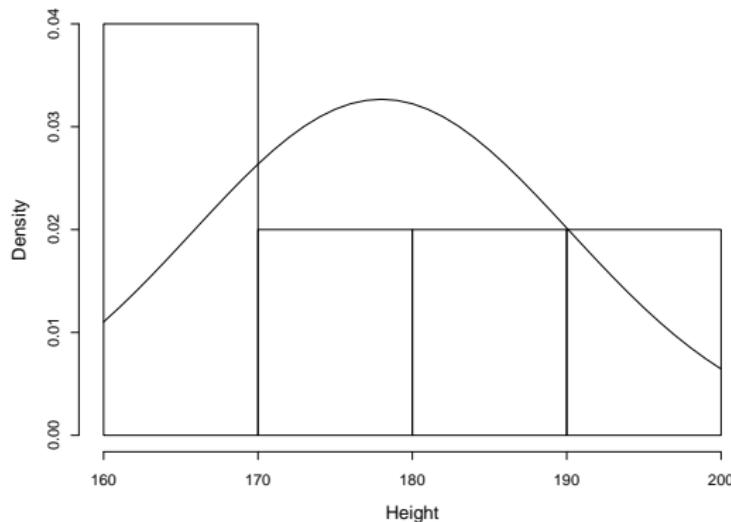
Overview

- 1 Motivating example - sleep medicine
- 2 One-sample t -test and p -value
- 3 Critical value and relation to the confidence interval
- 4 Hypothesis tests in general
 - The alternative hypothesis
 - The general method
 - Errors in hypothesis testing
- 5 Checking the normality assumption
 - The normal q-q plot
 - Transformation towards normality

Example - student heights

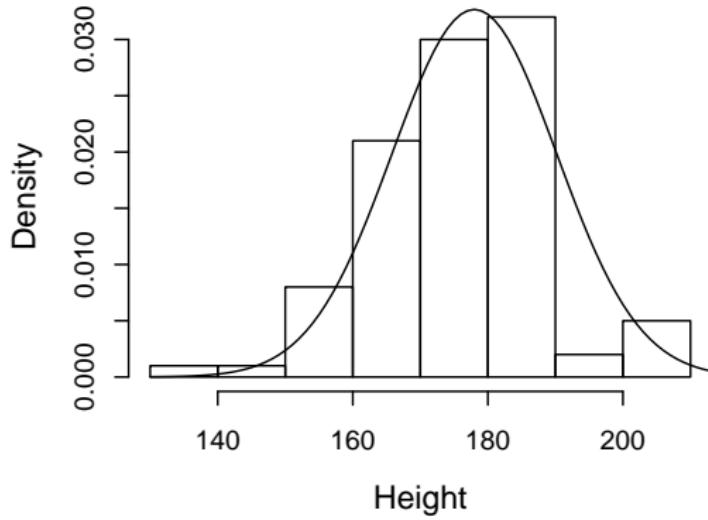
```
# Student heights data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)

# Density histogram of student height data together with normal pdf
hist(x, xlab = "Height", main = "", freq = FALSE)
lines(seq(160, 200, 1), dnorm(seq(160, 200, 1), mean(x), sd(x)))
```



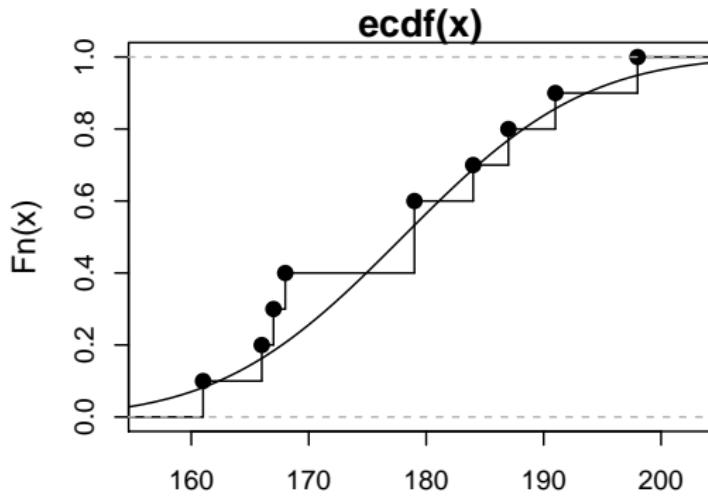
Example - 100 observations from a normal distribution

```
# Density histogram of simulated data from normal distribution
# (n = 100) together with normal pdf
xr <- rnorm(100, mean(x), sd(x))
hist(xr, xlab = "Height", main = "", freq = FALSE, ylim = c(0, 0.032))
lines(seq(130, 230, 1), dnorm(seq(130, 230, 1), mean(x), sd(x)))
```



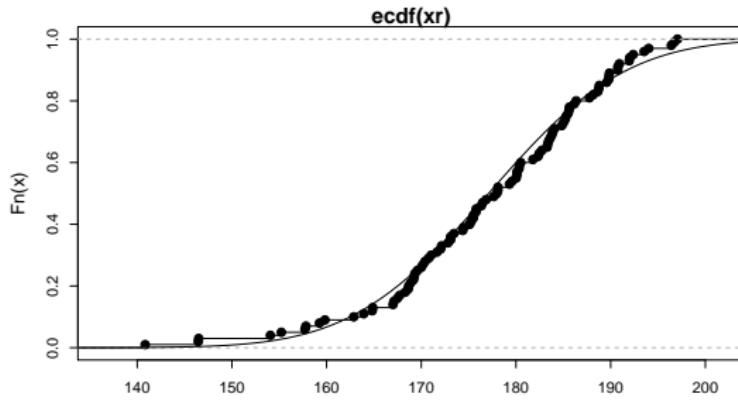
Example - student heights - ecdf

```
# Empirical cdf for student height data together  
# with normal cdf  
plot(ecdf(x), verticals = TRUE)  
xp <- seq(0.9*min(x), 1.1*max(x), length.out = 100)  
lines(xp, pnorm(xp, mean(x), sd(x)))
```



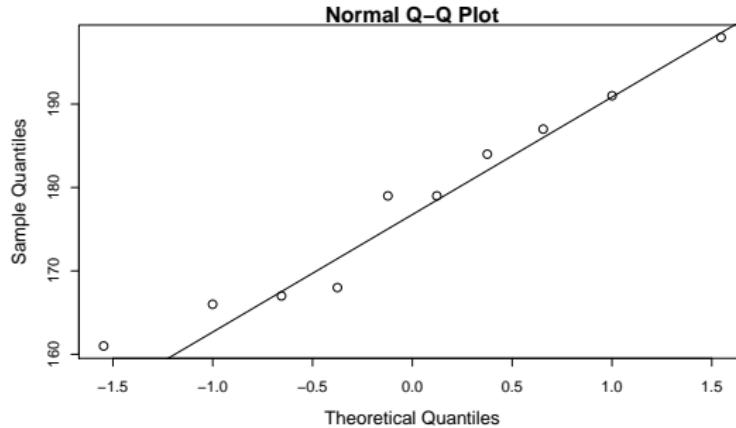
Example - 100 observations from a normal distribution - ecdf

```
# Empirical cdf of simulated data from normal distribution
# (n = 100) together with normal cdf
xr <- rnorm(100, mean(x), sd(x))
plot(ecdf(xr), verticals = TRUE)
xp <- seq(0.9*min(xr), 1.1*max(xr), length.out = 100)
lines(xp, pnorm(xp, mean(xr), sd(xr)))
```



Example - student heights - normal q-q plot

```
# Normal q-q plot of student heights  
qqnorm(x)  
qqline(x)
```



Normal q-q plot

Method 3.42- The formal definition

The ordered observations $x_{(1)}, \dots, x_{(n)}$ are plotted versus a set of expected normal quantiles z_{p_1}, \dots, z_{p_n} . Different definitions of p_1, \dots, p_n exist:

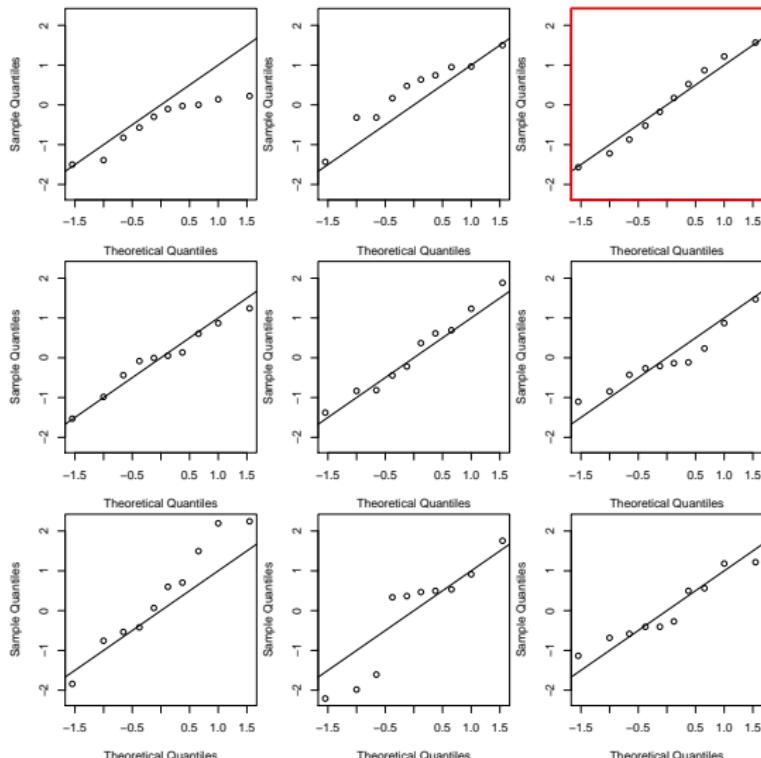
- In R, when $n > 10$:

$$p_i = \frac{i - 0.5}{n}, \quad i = 1, \dots, n$$

- In R, when $n \leq 10$:

$$p_i = \frac{i - 3/8}{n + 1/4}, \quad i = 1, \dots, n$$

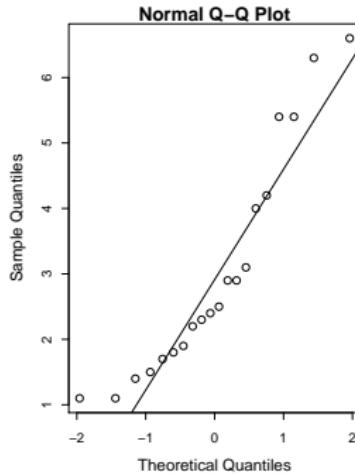
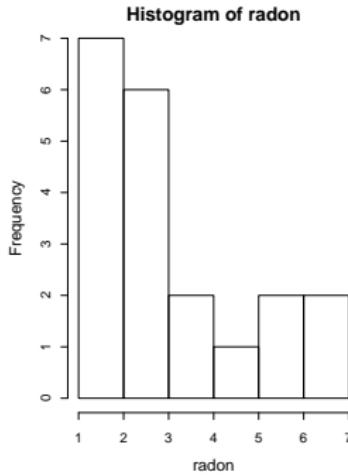
Example - student heights - compare with other simulated normal distributed data



Example - Radon data

```
## Reading in the data
radon <- c(2.4, 4.2, 1.8, 2.5, 5.4, 2.2, 4.0, 1.1, 1.5, 5.4, 6.3,
         1.9, 1.7, 1.1, 6.6, 3.1, 2.3, 1.4, 2.9, 2.9)

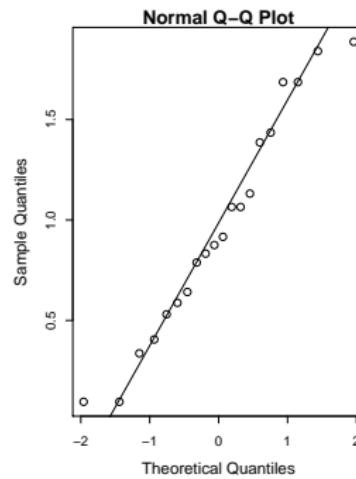
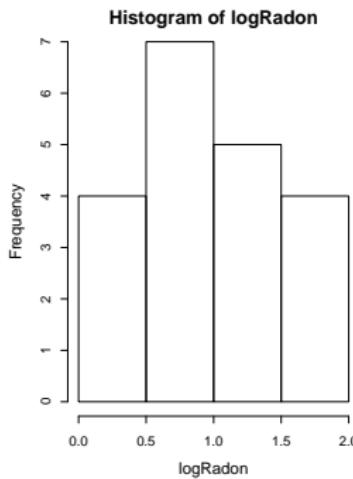
## Histogram and q-q plot of data
par(mfrow = c(1,2))
hist(radon)
qqnorm(radon)
qqline(radon)
```



Example - Radon data - log-transformed data are closer to a normal distribution

```
# Transform data using the natural logarithm
logRadon<-log(radon)

## Histogram and q-q plot of transformed data
par(mfrow = c(1,2))
hist(logRadon)
qqnorm(logRadon)
qqline(logRadon)
```



Agenda

- 1 Motivating example - sleep medicine
- 2 One-sample t -test and p -value
- 3 Critical value and relation to the confidence interval
- 4 Hypothesis tests in general
 - The alternative hypothesis
 - The general method
 - Errors in hypothesis testing
- 5 Checking the normality assumption
 - The normal q-q plot
 - Transformation towards normality

Course 02402 Introduction to Statistics

Lecture 6: Two-sample comparisons and power/sample size

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

Overview

- ➊ Motivating example: Nutrition study
- ➋ Repetition: p -values and hypothesis tests
- ➌ Two-sample t -test and p -value
- ➍ Confidence interval for the mean difference
- ➎ Overlapping confidence intervals?
- ➏ The paired setup
- ➐ Checking the normality assumptions
- ➑ Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- ➒ The pooled t-test - a possible alternative

Motivating example - nutrition study

Nutrition study

In a nutrition study, the aim is to investigate whether there is a difference in the energy usage for two different types of (moderately physically demanding) work.

In the study, the energy usage of 9 nurses from hospital *A* and 9 (other) nurses from hospital *B* have been measured. The measurements are given in the following table in mega Joule (MJ):

	Hospital A	Hospital B
Sample from each hospital,	7.53	9.21
$n_1 = n_2 = 9$:	7.48	11.51
	8.08	12.79
	8.09	11.85
	10.15	9.97
	8.40	8.79
	10.88	9.69
	6.13	9.68
	7.90	9.19

Example - nutrition study

The hypothesis of no difference in mean energy usage is in focus:

$$H_0 : \mu_A = \mu_B$$

Example - nutrition study

The hypothesis of no difference in mean energy usage is in focus:

$$H_0 : \mu_A = \mu_B$$

Sample means and standard deviations:

$$\hat{\mu}_A = \bar{x}_A = 8.293, (s_A = 1.428)$$

$$\hat{\mu}_B = \bar{x}_B = 10.298, (s_B = 1.398)$$

Example - nutrition study

The hypothesis of no difference in mean energy usage is in focus:

$$H_0 : \mu_A = \mu_B$$

Sample means and standard deviations:

$$\hat{\mu}_A = \bar{x}_A = 8.293, (s_A = 1.428)$$

$$\hat{\mu}_B = \bar{x}_B = 10.298, (s_B = 1.398)$$

Is data in accordance with the null hypothesis H_0 ?

Data: $\bar{x}_B - \bar{x}_A = 2.005$

Null hypothesis: $H_0 : \mu_B - \mu_A = 0$

Example - nutrition study

The hypothesis of no difference in mean energy usage is in focus:

$$H_0 : \mu_A = \mu_B$$

Sample means and standard deviations:

$$\hat{\mu}_A = \bar{x}_A = 8.293, (s_A = 1.428)$$

$$\hat{\mu}_B = \bar{x}_B = 10.298, (s_B = 1.398)$$

Is data in accordance with the null hypothesis H_0 ?

$$\text{Data: } \bar{x}_B - \bar{x}_A = 2.005$$

$$\text{Null hypothesis: } H_0 : \mu_B - \mu_A = 0$$

NEW: ***p*-value for difference:**

$$p\text{-value} = 0.0083$$

(Found under the scenario that H_0 is true.)

NEW: **Confidence interval for difference:**

$$2.005 \pm 1.412 = [0.59; 3.42]$$

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

Definition of a hypothesis test and significance (repetition)

Definition 3.24. Hypothesis test:

When we *carry out a hypothesis test*, we decide against a null hypothesis or not, using the data.

A null hypothesis is *rejected* if the *p*-value, calculated after the data has been observed, is less than some α – that is, if the p -value $< \alpha$, where α is some pre-specified (so-called) *significance level*.

If we do not reject the null hypothesis, it is said to be *accepted*.

Definition 3.29. Statistical significance:

An *effect* is said to be *(statistically) significant* if the *p*-value is less than the significance level α .

Often (and unless otherwise mentioned) we use $\alpha = 0.05$.

Steps of a hypothesis test - overview (repetition)

Generally, a hypothesis test consists of the following steps:

- ① Formulate the hypothesis and choose the level of significance α (the "risk-level").
- ② Calculate, using the data, the value of the test statistic.
- ③ Calculate the p -value using the test statistic and the relevant distribution. Compare the p -value to the significance level α and make a conclusion.

OR:

Alternatively, make a conclusion based on the relevant critical value(s).

Definition and interpretation of the *p*-value (repetition)

The *p*-value expresses the *evidence* against the null hypothesis – Table 3.1:

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

Definition 3.22 of the *p*-value:

The *p*-value is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

Critical value, confidence interval and hypothesis test (repetition)

Theorem 3.33: Critical value method \approx Confidence interval method

We consider a $(1 - \alpha) \cdot 100\%$ confidence interval for μ :

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The confidence interval corresponds to the acceptance region for H_0 when testing the (non-directional) hypothesis

$$H_0 : \mu = \mu_0$$

Critical value, confidence interval and hypothesis test (repetition)

Theorem 3.33: Critical value method \approx Confidence interval method

We consider a $(1 - \alpha) \cdot 100\%$ confidence interval for μ :

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The confidence interval corresponds to the acceptance region for H_0 when testing the (non-directional) hypothesis

$$H_0: \mu = \mu_0$$

(New) interpretation of the confidence interval:

The confidence interval covers those values of the parameter that we believe in given the data.

The confidence interval contains those values of the parameter that we would accept by the corresponding hypothesis test.

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 **Two-sample t -test and p -value**
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t -test - a possible alternative

Method 3.49: Two-sample *t*-test

Computing the test statistic:

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0 : \delta = \delta_0$$

the (Welch) two-sample *t*-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Theorem 3.50: Distribution of the (Welch) *t*-test statistic

The Welch *t*-test statistic is (approximately) *t*-distributed:

Under the null hypothesis, the (Welch) two-sample statistic, seen as a random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

approximately follows a *t*-distribution with v degrees of freedom, where

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

if the two population distributions are normal or if the two sample sizes are large enough.

Method 3.51: Two-sample t -test

The level α test is:

- ① Compute t_{obs} and v as given above.
- ② Compute the evidence against the *null hypothesis*^a $H_0: \mu_1 - \mu_2 = \delta_0$ vs. the *alternative hypothesis* $H_1: \mu_1 - \mu_2 \neq \delta_0$ by the

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|)$$

where the t -distribution with v degrees of freedom is used.

- ③ If $p\text{-value} < \alpha$: We reject H_0 , otherwise we accept H_0 .

OR

The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value(s) $\pm t_{1-\alpha/2}$:

If $|t_{\text{obs}}| > t_{1-\alpha/2}$ we reject H_0 , otherwise we accept H_0 .

^aWe are often interested in the test where $\delta_0 = 0$

Example - nutrition study

The hypothesis of no difference is in focus:

$$H_0: \delta = \mu_B - \mu_A = 0$$

versus the non-directional (= two-sided) alternative:

$$H_1: \delta = \mu_B - \mu_A \neq 0$$

First, the computations of t_{obs} and v :

$$t_{\text{obs}} = \frac{10.298 - 8.293}{\sqrt{2.0394/9 + 1.954/9}} = 3.01$$

and

$$v = \frac{\left(\frac{2.0394}{9} + \frac{1.954}{9}\right)^2}{\frac{(2.0394/9)^2}{8} + \frac{(1.954/9)^2}{8}} = 15.99$$

Example - nutrition study

Next, the *p*-value is found:

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
## Nutrition study example: P(T > 3.01)
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Example - nutrition study

Next, the *p*-value is found:

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
## Nutrition study example: P(T > 3.01)
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Evaluate the evidence (Table 3.1):

There is strong evidence against the null hypothesis.

Example - nutrition study

Next, the *p*-value is found:

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
## Nutrition study example: P(T > 3.01)
1 - pt(3.01, df = 15.99)

## [1] 0.004154
```

Evaluate the evidence (Table 3.1):

There is strong evidence against the null hypothesis.

Conclude based on $\alpha = 0.05$:

We reject the null hypothesis. There is a significant difference between the two groups. Nurses in Hospital *B* can be said to have a larger (mean) energy usage than nurses in Hospital *A*.

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 **Confidence interval for the mean difference**
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

Method 3.47: Confidence interval for $\mu_1 - \mu_2$

The confidence interval for the mean difference:

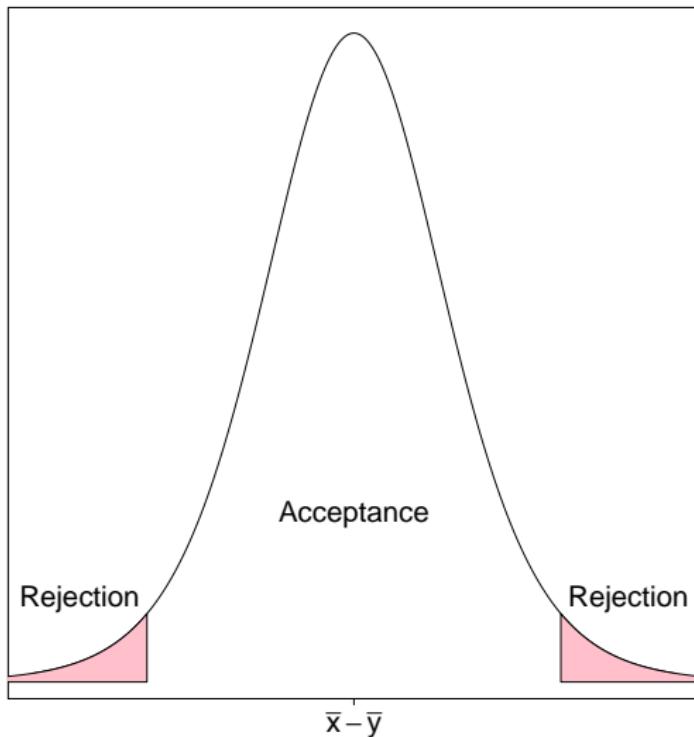
For two samples x_1, \dots, x_n and y_1, \dots, y_n the $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{1-\alpha/2}$ is the $100(1 - \alpha/2)\%$ -quantile from the t -distribution with v degrees of freedom given by Theorem 3.50 (see above).

Confidence interval and hypothesis test (repetition)

The acceptance region consists of the potential values for $\mu_1 - \mu_2$ that are not too far away from the data:



Example - nutrition study:

Let's find the 95% confidence interval for $\mu_B - \mu_A$. Using $v = 15.99$, the relevant t -quantile is

$$t_{0.975} = 2.120$$

and the confidence interval becomes:

$$10.298 - 8.293 \pm 2.120 \cdot \sqrt{\frac{2.0394}{9} + \frac{1.954}{9}}.$$

This gives the confidence interval also shown above:

$$[0.59; 3.42]$$

Example - nutrition study - everything in R:

```
# Read the two samples into R
xA = c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9)
xB = c(9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68, 9.19)

# Perform Welch two-sample t-test
t.test(xB, xA)

##
##  Welch Two Sample t-test
##
## data: xB and xA
## t = 3, df = 16, p-value = 0.008
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5923 3.4166
## sample estimates:
## mean of x mean of y
## 10.298     8.293
```

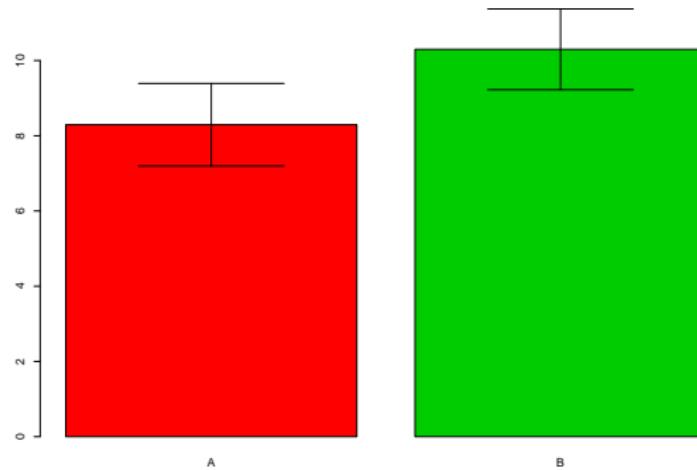
Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

Example - nutrition study - presentation of result

Barplots with *error bars* are often seen:

A grouped barplot with some "error bars" - below, the 95%-confidence intervals for the mean of each group are shown:



Be careful about using "overlapping confidence intervals"

The approach is actually using an incorrect variation for evaluation of the difference:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} \neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

$$\text{Var}(\bar{X}_A - \bar{X}_B) = \text{Var}(\bar{X}_A) + \text{Var}(\bar{X}_B)$$

Assume that the two standard-errors are 3 and 4: The sum is 7, but $\sqrt{3^2 + 4^2} = 5$

Be careful about using "overlapping confidence intervals"

The approach is actually using an incorrect variation for evaluation of the difference:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} \neq \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

$$\text{Var}(\bar{X}_A - \bar{X}_B) = \text{Var}(\bar{X}_A) + \text{Var}(\bar{X}_B)$$

Assume that the two standard-errors are 3 and 4: The sum is 7, but $\sqrt{3^2 + 4^2} = 5$

The correct relation between the standard deviations is:

$$\sigma_{(\bar{X}_A - \bar{X}_B)} < \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

Be careful about using "overlapping confidence intervals"

Remark 3.59. Rule for using "overlapping confidence intervals":

When two CIs do NOT overlap: The two groups are significantly different.

When two CIs DO overlap: We do not know what the conclusion is – but we could, e.g., make a CI for the mean *difference* instead, to investigate.

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 **The paired setup**
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

Motivating example - sleep medicine

Difference between sleep medicines?

In a study, the aim is to compare two kinds of sleep medicine, A and B. 10 test persons tried both kinds of medicine, and the following results were obtained, in terms of prolonged sleep length (in hours) for each medicine type:

Sample, $n = 10$:

Person	A	B	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

The paired setup and analysis = one-sample analysis

```
# Read the two samples into R
x1 = c(.7,-1.6,-.2,-1.2,-1,3.4,3.7,.8,0,2)
x2 = c(1.9,.8,1.1,.1,-.1,4.4,5.5,1.6,4.6,3.4)

# Compute differences to get a paired t-test
dif = x2 - x1

# Perform paired t-test
t.test(dif)

## 
##  One Sample t-test
##
## data:  dif
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of x
##          1.67
```

The paired setup and analysis = one-sample analysis

```
# Another way to perform the paired t-test
t.test(x2, x1, paired = TRUE)

##
##  Paired t-test
##
## data: x2 and x1
## t = 4.7, df = 9, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8613 2.4787
## sample estimates:
## mean of the differences
##                           1.67
```

Paired vs. independent experiment

Completely randomized (independent samples)

20 patients are used and allocated, completely at random, to one of the two treatments (usually making sure to have 10 patients in each group). That is, there are different patients in the different treatment groups.

Paired (dependent samples)

10 patients are used, and each of them tests both of the treatments. Usually this will involve some time in between treatments, to make sure that the experiment is meaningful, and also one would typically make sure that some patients try A before B and others try B before A (order allocated at random). That is, the same patients are included in both treatment groups.

Example - sleep medicine - WRONG analysis

```
# WRONG analysis
t.test(x1, x2)

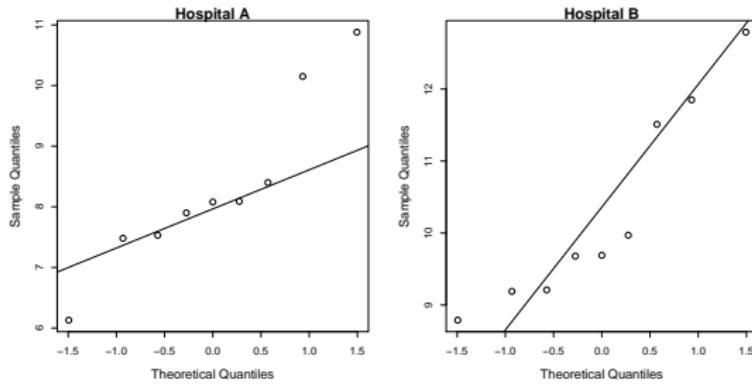
##
##  Welch Two Sample t-test
##
## data: x1 and x2
## t = -1.9, df = 18, p-value = 0.07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.4854 0.1454
## sample estimates:
## mean of x mean of y
## 0.66 2.33
```

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 **Checking the normality assumptions**
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

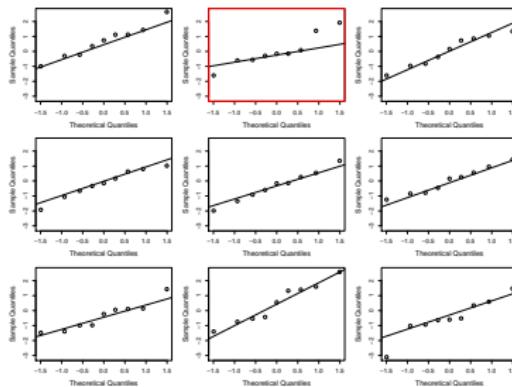
Example - Q-Q plot for EACH sample:

```
# Q-Q plots separately for each sample
qqnorm(xA, main = "Hospital A")
qqline(xA)
qqnorm(xB, main = "Hospital B")
qqline(xB)
```



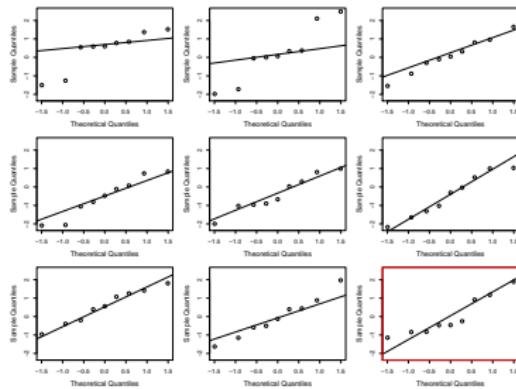
Example - Comparing with simulated data, A

```
# Multiple (simulated) Q-Q plots and sample A
require(MESS)
fitA <- lm(xA ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...) }
wallyplot(fitA, FUN = qqnorm.wally, main = "")
```



Example - Comparing with simulated data, B

```
# Multiple (simulated) Q-Q plots and sample B
fitB <- lm(xB ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...) }
wallyplot(fitB, FUN = qqnorm.wally, main = "")
```



Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

Planning of study with requirements to the precision

The one-sample $100 \cdot (1 - \alpha)\%$ CI: $\bar{x} \pm t_{1-\alpha/2} \cdot s / \sqrt{n}$.

The *margin of error (ME)* is defined as

$$t_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Method 3.63: The one-sample CI sample size formula:

When σ is known, or guessed to be some value, we can calculate the sample size n needed to achieve a given margin of error, ME , with probability $1 - \alpha$, as:

$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME} \right)^2$$

Example, height data again

Sample mean og standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimate the population mean and standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

Example, height data again

Sample mean og standard deviation:

$$\bar{x} = 178$$

$$s = 12.21$$

Estimate the population mean and standard deviation:

$$\hat{\mu} = 178$$

$$\hat{\sigma} = 12.21$$

If we want $ME = 3$ cm with 95% confidence, how large should n be?

$$n = \left(\frac{1.96 \cdot 12.21}{3} \right)^2 = 63.64 \approx 64$$

Planning, power

What is the power of a future study/experiment:

- The probability of detecting an (assumed) effect.
- $P(\text{reject } H_0)$ when H_1 is true.
- Probability of correct rejection of H_0 .
- Challenge: The null hypothesis can be wrong in many ways!
- Practically: Scenario-based approach
 - E.g. "What if $\mu = 86$, how good will my study be to detect this?"
 - E.g. "What if $\mu = 84$, how good will my study be to detect this?"
 - etc.

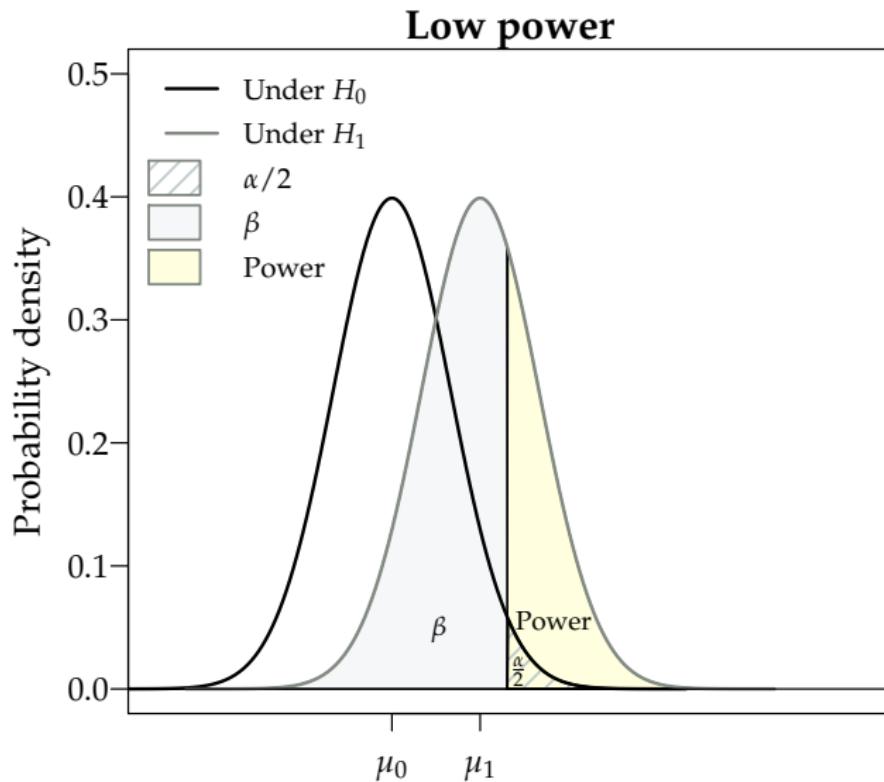
Planning and power

When the null hypothesis has been decided on:

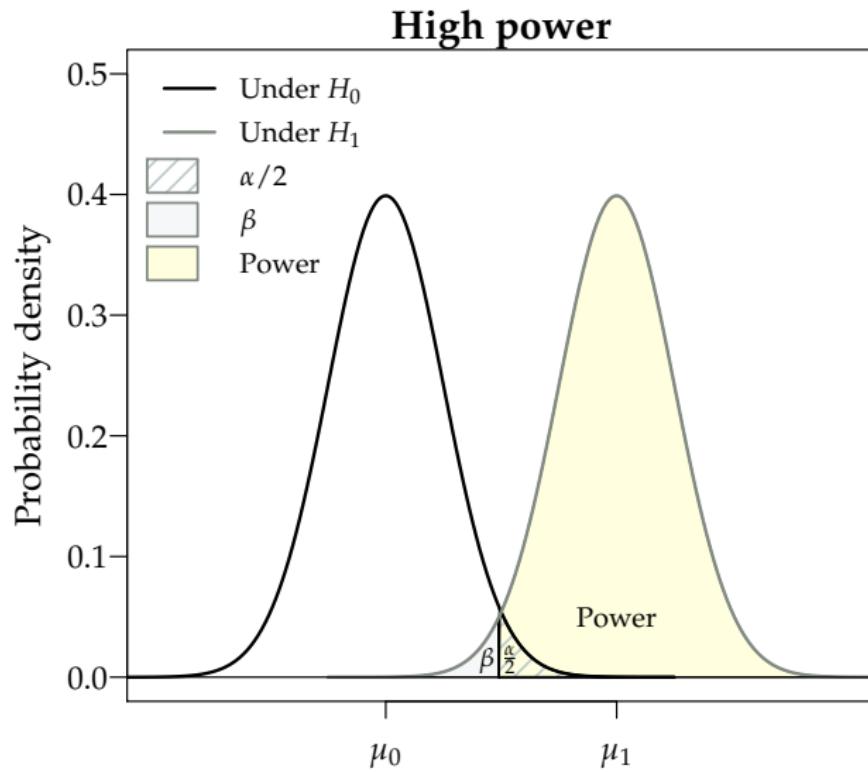
If you know (or set/guess) four out of the following five pieces of information, you can find the fifth:

- The sample size n .
- Significance level α of the test.
- A difference in mean that you would want to detect (effect size) $\mu_0 - \mu_1$.
- The population standard deviation, σ .
- The power $(1 - \beta)$.

Low power example



High power example



Planning, sample size n

The big practical question: What should n be?

The experiment should be large enough to detect a relevant effect with high power $1 - \beta$ (usually at least 80%):

Planning, sample size n

The big practical question: What should n be?

The experiment should be large enough to detect a relevant effect with high power $1 - \beta$ (usually at least 80%):

Metode 3.65: The one-sample sample size formula:

For the one-sample t-test for given α , β and σ :

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{\mu_0 - \mu_1} \right)^2$$

Here, $\mu_0 - \mu_1$ is the difference in mean that we would like to detect, and $z_{1-\beta}$, $z_{1-\alpha/2}$ are quantiles of the standard normal distribution.

Example - The power for $n = 40$

```
# Power calculation (one-sample)
power.t.test(n = 40, delta = 4, sd = 12.21, type = "one.sample")

##
##      One-sample t test power calculation
##
##              n = 40
##              delta = 4
##              sd = 12.21
##              sig.level = 0.05
##              power = 0.5242
##              alternative = two.sided
```

Example - The sample size for power = 0.80

```
# Sample size calculation (one-sample)
power.t.test(power = .80, delta = 4, sd = 12.21, type = "one.sample")

##
##      One-sample t test power calculation
##
##              n = 75.08
##              delta = 4
##              sd = 12.21
##              sig.level = 0.05
##              power = 0.8
##      alternative = two.sided
```

Power and sample size - two-sample

Finding the power of detecting a group difference of 2 with $\sigma = 1$ for $n = 10$:

```
# Power calculation (two-sample)
power.t.test(n = 10, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##              n = 10
##              delta = 2
##              sd = 1
##              sig.level = 0.05
##              power = 0.9882
##              alternative = two.sided
##
## NOTE: n is number in *each* group
```

Power and sample size - two-sample

Finding the sample size for detecting a group mean difference of 2 with $\sigma = 1$ and power = 0.9:

```
# Sample size calculation (two-sample)
power.t.test(power = 0.90, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##              n = 6.387
##              delta = 2
##              sd = 1
##              sig.level = 0.05
##              power = 0.9
##              alternative = two.sided
##
## NOTE: n is number in *each* group
```

Power and sample size - two-sample

Finding the detectable effect size (delta) with $\sigma = 1$, $n = 10$ and power = 0.9:

```
## Detectable effect size (two-sample)
power.t.test(power = 0.90, n = 10, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##              n = 10
##              delta = 1.534
##              sd = 1
##              sig.level = 0.05
##              power = 0.9
##              alternative = two.sided
##
## NOTE: n is number in *each* group
```

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

The pooled two-sample *t*-test statistic

The *pooled* estimate of variance (assuming $\sigma_1^2 = \sigma_2^2$)

Method 3.52

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The pooled t-test statistic, Method 3.53

When considering the null hypothesis about the difference between the means of two *independent* samples:

$$\delta = \mu_2 - \mu_1$$

$$H_0 : \delta = \delta_0$$

the pooled two-sample *t*-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

Theorem 3.54: The distribution of the pooled test-statistic

... is a t -distribution:

The pooled two-sample statistic seen as a random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}$$

follows, under the null hypothesis and under the assumption that $\sigma_1^2 = \sigma_2^2$, a t -distribution with $n_1 + n_2 - 2$ degrees of freedom if the two population distributions are normal.

We always use the "Welch" t-test

Almost (fool)proof to always use the Welch-version:

- If $s_1^2 = s_2^2$, the Welch and the pooled t-test statistics are the same.
- Only when the two variances become really different, the two test-statistics may differ in an important way. Furthermore, if this is the case, we would not tend to favour the pooled version, since the assumption of equal variances appears questionable then.
- Only for cases with a small sample size in at least one of the two groups, the pooled approach may provide slightly higher power (if you believe in the equal variance assumption). For these cases, the Welch approach is then a somewhat cautious approach.

Overview

- 1 Motivating example: Nutrition study
- 2 Repetition: p -values and hypothesis tests
- 3 Two-sample t -test and p -value
- 4 Confidence interval for the mean difference
- 5 Overlapping confidence intervals?
- 6 The paired setup
- 7 Checking the normality assumptions
- 8 Planning for wanted precision or power
 - Precision requirements
 - Power and sample size - one-sample
 - Power and sample size - two-sample
- 9 The pooled t-test - a possible alternative

Course 02402 Introduction to Statistics

Lecture 7: Simulation-based statistics

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- ① Introduction to simulation - what is it really?
 - Example: Area of plates
- ② Propagation of error
- ③ Parametric bootstrap
 - Introduction to bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval assuming any distributions
- ④ Non-parametric bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval

Overview

- ➊ Introduction to simulation - what is it really?
 - Example: Area of plates
- ➋ Propagation of error
- ➌ Parametric bootstrap
 - Introduction to bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval assuming any distributions
- ➍ Non-parametric bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval

Motivation

- Many (most?) relevant statistics (“computed features”) have complicated sampling distributions. One might want to do statistical inference for, e.g.:
 - The median
 - Quantiles in general, or perhaps $IQR = Q_3 - Q_1$
 - The coefficient of variation
 - Any non-linear function of one or more input variables
 - (The standard deviation)
- The distribution of the data itself may be non-normal, complicating the statistical theory for even the simple mean.
- We may hope for the magic of the CLT (Central Limit Theorem).
- But: We never *really* know whether the CLT is good enough in a given situation - simulation can tell us!
- Requires: Use of a computer with software that can do simulations. R is a super tool for this!

What is simulation really?

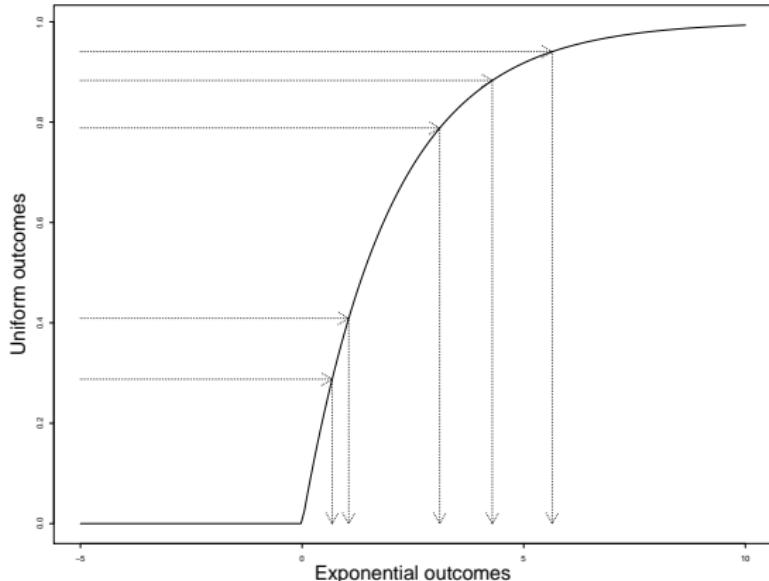
- (Pseudo) random numbers are generated using a computer.
- A random number generator is an algorithm that can generate x_{i+1} from x_i .
- The resulting sequence of numbers appears random.
- Requires a “starting point” called a *seed*.
- Basically, the uniform distribution is simulated in this manner, and then:

Theorem 2.51: All distributions can be extracted from the uniform

If $U \sim \text{Uniform}(0,1)$ and F is a distribution function for any probability distribution, then $F^{-1}(U)$ follows the distribution given by F .

Example: The exponential distribution with $\lambda = 0.5$:

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



In practice, in R

Many distributions are ready for simulation, for instance:

<code>rbinom</code>	The binomial distribution
<code>rpois</code>	The Poisson distribution
<code>rhyper</code>	The hypergeometric distribution
<code>rnorm</code>	The normal distribution
<code>rlnorm</code>	The log-normal distributions
<code>rexp</code>	The exponential distribution
<code>runif</code>	The uniform distribution
<code>rt</code>	The t-distribution
<code>rchisq</code>	The χ^2 -distribution
<code>rf</code>	The F-distribution

Example: Area of plates

A company produces rectangular plates. The length of a plate (in meters), X , is assumed to follow a normal distribution $N(2, 0.01^2)$. The width of a plate (in meters), Y , is assumed to follow a normal distribution $N(3, 0.02^2)$. We are interested in the area of the plates, which is given by $A = XY$.

- What is the mean area?
- What is the standard deviation of the area?
- How often do such plates have an area that differs by more than 0.1 m^2 from the targeted 6 m^2 ?
- (The probability of other events?)
- Generally: What is the probability distribution of the random variable A ?

Example: Area of plates, solution by simulation

```
k = 10000 # Number of simulations
X = rnorm(k, 2, 0.01)
Y = rnorm(k, 3, 0.02)
A = X*Y

mean(A)
```

```
[1] 6
```

```
var(A)
```

```
[1] 0.002458
```

```
mean(abs(A - 6) > 0.1)
```

```
[1] 0.0439
```

Overview

1 Introduction to simulation - what is it really?

- Example: Area of plates

2 Propagation of error

3 Parametric bootstrap

- Introduction to bootstrap
- One-sample confidence interval for any feature
- Two-sample confidence interval assuming any distributions

4 Non-parametric bootstrap

- One-sample confidence interval for any feature
- Two-sample confidence interval

Propagation of error

Must be able to find:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

Propagation of error

Must be able to find:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

We already know:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \sum_{i=1}^n a_i^2 \sigma_i^2 \quad \text{if} \quad f(X_1, \dots, X_n) = \sum_{i=1}^n a_i X_i \text{ (and independence)}$$

Propagation of error

Must be able to find:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \text{Var}(f(X_1, \dots, X_n))$$

We already know:

$$\sigma_{f(X_1, \dots, X_n)}^2 = \sum_{i=1}^n a_i^2 \sigma_i^2 \quad \text{if} \quad f(X_1, \dots, X_n) = \sum_{i=1}^n a_i X_i \text{ (and independence)}$$

Method ??: For non-linear functions, if X_1, \dots, X_n are independent,

$$\sigma_{f(X_1, \dots, X_n)}^2 \approx \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_i^2$$

Example: Area of plates (continued)

We used a simulation method in the first part of the example.

Now, given two specific measurements of X and Y , $x = 2.00$ m and $y = 3.00$ m: What is the variance of $A = XY$, using the error propagation law?

Example: Area of plates (continued)

The variances are:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ and } \sigma_2^2 = \text{Var}(Y) = 0.02^2$$

Example: Area of plates (continued)

The variances are:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ and } \sigma_2^2 = \text{Var}(Y) = 0.02^2$$

The function and its derivatives are:

$$f(x, y) = xy, \frac{\partial f}{\partial x} = y, \frac{\partial f}{\partial y} = x$$

Example: Area of plates (continued)

The variances are:

$$\sigma_1^2 = \text{Var}(X) = 0.01^2 \text{ and } \sigma_2^2 = \text{Var}(Y) = 0.02^2$$

The function and its derivatives are:

$$f(x, y) = xy, \frac{\partial f}{\partial x} = y, \frac{\partial f}{\partial y} = x$$

So the result becomes:

$$\begin{aligned} \text{Var}(A) &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_2^2 \\ &= y^2 \sigma_1^2 + x^2 \sigma_2^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 \\ &= 0.0025 \end{aligned}$$

Propagation of error - by simulation

Method ??: Error propagation by simulation

Assume that we have actual measurements x_1, \dots, x_n with known/assumed error variances $\sigma_1^2, \dots, \sigma_n^2$.

- 1 Simulate k outcomes of all n measurements from assumed error distributions, e.g. $N(x_i, \sigma_i^2)$: $X_i^{(j)}, j = 1 \dots, k$.
- 2 Calculate the standard deviation directly as the observed standard deviation of the k simulated values of f :

$$s_{f(X_1, \dots, X_n)}^{\text{sim}} = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (f_j - \bar{f})^2}$$

where

$$f_j = f(X_1^{(j)}, \dots, X_n^{(j)})$$

Example: Area of plates (continued)

Actually, in this example, one *could* deduce the variance of A theoretically:

$$\begin{aligned}\text{Var}(XY) &= \mathbb{E}[(XY)^2] - [\mathbb{E}(XY)]^2 \\ &= \mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(X)^2\mathbb{E}(Y)^2 \\ &= [\text{Var}(X) + \mathbb{E}(X)^2][\text{Var}(Y) + \mathbb{E}(Y)^2] - \mathbb{E}(X)^2\mathbb{E}(Y)^2 \\ &= \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\mathbb{E}(Y)^2 + \text{Var}(Y)\mathbb{E}(X)^2 \\ &= 0.01^2 \times 0.02^2 + 0.01^2 \times 3^2 + 0.02^2 \times 2^2 \\ &= 0.00000004 + 0.0009 + 0.0016 \\ &= 0.00250004\end{aligned}$$

Example: Area of plates (continued)

Three different approaches:

- ① The simulation based approach.
- ② A theoretical derivation.
- ③ The analytical, but approximate, error propagation method.

Example: Area of plates (continued)

Three different approaches:

- ① The simulation based approach.
- ② A theoretical derivation.
- ③ The analytical, but approximate, error propagation method.

The simulation approach has a number of crucial advantages:

- ① It offers a simple tool to compute many other quantities than just the standard deviation. (The theoretical derivations of these could be much more complicated than what was shown for the variance).
- ② It offers a simple tool to use any other distributions than the normal, if we believe that they reflect reality better.
- ③ It does not rely on linear approximations of the true non-linear relations.

Overview

- ➊ Introduction to simulation - what is it really?
 - Example: Area of plates
- ➋ Propagation of error
- ➌ Parametric bootstrap
 - Introduction to bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval assuming any distributions
- ➍ Non-parametric bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval

Bootstrapping

Bootstrapping exists in two versions:

- ① Parametric bootstrap: Simulate multiple samples from the assumed (and estimated) distribution.
- ② Non-parametric bootstrap: Simulate multiple samples directly from the data.

Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data, we estimate

$$\hat{\mu} = \bar{x} = 26.08 \text{ and hence: } \hat{\lambda} = 1/26.08 = 0.03834356$$

Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data, we estimate

$$\hat{\mu} = \bar{x} = 26.08 \text{ and hence: } \hat{\lambda} = 1/26.08 = 0.03834356$$

Our distributional assumption:

The waiting times come from an exponential distribution.

Example: Confidence interval for an exponential mean

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data, we estimate

$$\hat{\mu} = \bar{x} = 26.08 \text{ and hence: } \hat{\lambda} = 1/26.08 = 0.03834356$$

Our distributional assumption:

The waiting times come from an exponential distribution.

What is the confidence interval for μ ?

Based on previous knowledge in this course: We don't know!

Example: Confidence interval for an exponential mean

```
# Number of simulations
k <- 100000

# Simulate 10 exponentials with the 'right' mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

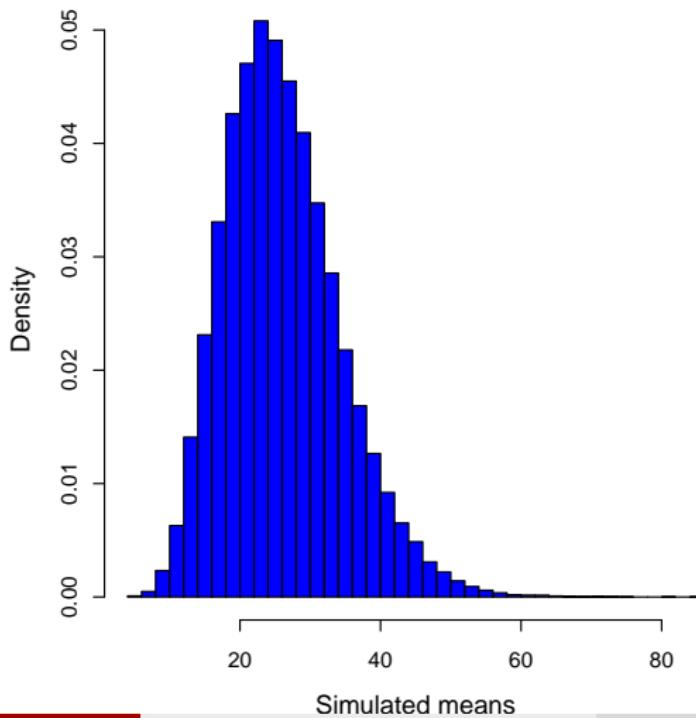
# Compute the mean of the 10 simulated observations k times
sim_means <- apply(sim_samples, 2, mean)

# Find relevant quantiles of the k simulated means
quantile(sim_means, c(0.025, 0.975))

## 2.5% 97.5%
## 12.59 44.63
```

Example: Confidence interval for an exponential mean

```
# Make histogram of simulated means
hist(sim_means, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Simulated means")
```



Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data we estimate

Median = 21.4 and $\hat{\mu} = \bar{x} = 26.08$

Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data we estimate

Median = 21.4 and $\hat{\mu} = \bar{x} = 26.08$

Our distributional assumption:

The waiting times come from an exponential distribution.

Example: Confidence interval for an exponential median

Assume that we observed the following 10 call waiting times (in seconds) in a call center:

32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0

From the data we estimate

$$\text{Median} = 21.4 \text{ and } \hat{\mu} = \bar{x} = 26.08$$

Our distributional assumption:

The waiting times come from an exponential distribution.

What is the confidence interval for the median?

Based on previous knowledge in this course: We don't know!

Example: Confidence interval for an exponential median

```
# Number of simulations
k <- 100000

# Simulate 10 exponentials with the 'right' mean k times
sim_samples <- replicate(k, rexp(10, 1/26.08))

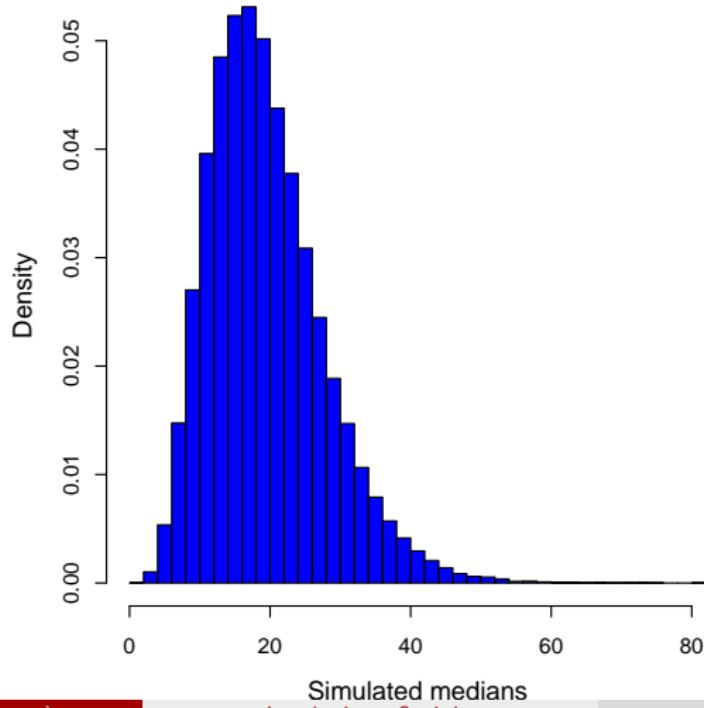
# Compute the median of the 10 simulated observations k times
simmedians <- apply(sim_samples, 2, median)

# Find relevant quantiles of the k simulated medians
quantile(simmedians, c(0.025, 0.975))

##    2.5% 97.5%
## 7.038 38.465
```

Example: Confidence interval for an exponential median

```
# Make histogram of simulated medians
hist(sim_medians, col = "blue", nclass = 30, main = "", prob = TRUE, xlab = "Simulated medians")
```



Confidence interval for any feature (including μ)

Method 4.7: Confidence interval for any feature θ by parametric bootstrap

Assume we have actual observations x_1, \dots, x_n , and that they come from some probability distribution with density f .

- ① Simulate k samples of n observations from the assumed distribution f where the mean^a is set to \bar{x} .
- ② Calculate the statistic $\hat{\theta}$ in each of the k samples to obtain $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$.
- ③ Find the $100(\alpha/2)\%$ and $100(1 - \alpha/2)\%$ quantiles of $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$, $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$, to obtain the $100(1 - \alpha)\%$ confidence interval:
$$[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$$

^aAnd otherwise chosen to match the data as well as possible: Some distributions have more than one mean related parameter, e.g. the normal or the log-normal. For these one should use a distribution with a variance that matches the sample variance of the data. Even more generally, the approach would be to match the chosen distribution to the data using the so-called *maximum likelihood* approach.

Example: 99% CI for Q_3 assuming a normal distribution

```
# Heights data
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
n <- length(x)

# Define a Q3-function
Q3 <- function(x){ quantile(x, 0.75) }

# Set number of simulations
k <- 100000

# Simulate k samples of n = 10 normals with the 'right' mean and variance
sim_samples <- replicate(k, rnorm(n, mean(x), sd(x)))

# Compute the Q3 of the n = 10 simulated observations k times
simQ3s <- apply(sim_samples, 2, Q3)

# Find the two relevant quantiles of the k simulated Q3s
quantile(simQ3s, c(0.005, 0.995))

## 0.5% 99.5%
## 172.8 198.0
```

Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$)

Method 4.10: Two-sample confidence interval for any feature comparison $\theta_1 - \theta_2$ by parametric bootstrap

Assume we have actual observations x_1, \dots, x_n and y_1, \dots, y_n , and that they stem from probability distributions with densities f_1 and f_2 .

- ① Simulate k sets of 2 samples of n_1 and n_2 observations from the assumed distributions, setting the means^a to $\hat{\mu}_1 = \bar{x}$ and $\hat{\mu}_2 = \bar{y}$, respectively.
- ② Calculate the difference between the features in each of the k samples: $\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$.
- ③ Find the $100(\alpha/2)\%$ and $100(1 - \alpha/2)\%$ quantiles for these, $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$, to obtain the $100(1 - \alpha)\%$ confidence interval:
$$[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$$

Example: Confidence interval for the difference of exponential means

```
# Day 1 data
x <- c(32.6, 1.6, 42.1, 29.2, 53.4, 79.3, 2.3, 4.7, 13.6, 2.0)
n1 <- length(x)

# Day 2 data
y <- c(9.6, 22.2, 52.5, 12.6, 33.0, 15.2, 76.6, 36.3, 110.2,
      18.0, 62.4, 10.3)
n2 <- length(y)
```

Example: Confidence interval for the difference of exponential means

```
# Set number of simulations:  
k <- 100000  
  
# Simulate k samples of each n1 = 10 and n2 = 12 exponentials  
# with the 'right' means  
  
simX_samples <- replicate(k, rexp(n1, 1/mean(x)))  
simY_samples <- replicate(k, rexp(n2, 1/mean(y)))  
  
# Compute the difference between the simulated means k times  
sim_dif_means <- apply(simX_samples, 2, mean) -  
  apply(simY_samples, 2, mean)  
  
# Find the relevant quantiles of the k simulated differences of means:  
quantile(sim_dif_means, c(0.025, 0.975))  
  
##    2.5% 97.5%  
## -40.74 14.12
```

Parametric bootstrap - an overview

We assume *some* distribution!

Two confidence interval method boxes were given:

	One-sample	Two-sample
For any feature	Method 4.7	Method 4.10

Overview

- ➊ Introduction to simulation - what is it really?
 - Example: Area of plates
- ➋ Propagation of error
- ➌ Parametric bootstrap
 - Introduction to bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval assuming any distributions
- ➍ Non-parametric bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval

Non-parametric bootstrap - an overview

We do *not* assume *any* distribution!

Two confidence interval method boxes will be given:

	One-sample	Two-sample
For any feature	Method 4.15	Method 4.17

Example: Womens' cigarette consumption

In a study, womens' cigarette consumption before and after giving birth is explored. The following observations of the number of smoked cigarettes per day were obtained:

before	after	before	after
8	5	13	15
24	11	15	19
7	0	11	12
20	15	22	0
6	0	15	6
20	20		

Compare the before and after means! (Are they different?)

Example: Womens' cigarette consumption

A paired t -test setting, *but* with clearly non-normal data!

```
# Data
x1 <- c(8, 24, 7, 20, 6, 20, 13, 15, 11, 22, 15)
x2 <- c(5, 11, 0, 15, 0, 20, 15, 19, 12, 0, 6)

# Compute differences
dif <- x1-x2
dif

## [1]  3 13  7  5  6  0 -2 -4 -1 22  9

# Compute average difference
mean(dif)

## [1] 5.273
```

Example: Women's cigarette consumption - bootstrapping

```
t(replicate(5, sample(dif, replace = TRUE)))  
  
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  
## [1,]     3    6    0    9    3    9   -4    0    0   -1     6  
## [2,]    -1    9    5    5    6    9    3   13    3   22    22  
## [3,]    -4   -2    3   -1    3   -1    7    3    9    6     0  
## [4,]     6    3   -4    9    3   22    3   -1   -1   -4     7  
## [5,]    13    0    5   22    0    9    9    5    0   22   -1
```

Example: Womens' cigarette consumption - the non-parametric results

Let us find the 95% confidence interval for the *mean* change in cigarette consumption.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_means = apply(sim_samples, 2, mean)
quantile(sim_means, c(0.025,0.975))

##  2.5% 97.5%
## 1.364 9.818
```

One-sample confidence interval for any feature θ (including μ)

Method 4.15: Confidence interval for any feature θ by non-parametric bootstrap

Assume we have actual observations x_1, \dots, x_n .

- ① Simulate k samples of size n by randomly sampling from the available data (with replacement).
- ② Calculate the statistic $\hat{\theta}$ for each of the k samples: $\hat{\theta}_1^*, \dots, \hat{\theta}_k^*$.
- ③ Find the $100(\alpha/2)\%$ and $100(1 - \alpha/2)\%$ quantiles for these, $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$, as the $100(1 - \alpha)\%$ confidence interval: $[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$

Example: Womens' cigarette consumption

Let us find the 95% confidence interval for the *median* change in cigarette consumption in the example from above.

```
k = 100000
sim_samples = replicate(k, sample(dif, replace = TRUE))
sim_mediants = apply(sim_samples, 2, median)
quantile(sim_mediants, c(0.025,0.975))

##  2.5% 97.5%
##      -1      9
```

Example: Tooth health and infant bottle use

In a study, it was explored whether children who had received milk from a bottle had worse or better tooth health than those who had *not* received milk from a bottle. For 19 randomly selected children, it was recorded when they had had their first incident of caries:

bottle	age	bottle	age	bottle	age
no	9	no	10	yes	16
yes	14	no	8	yes	14
yes	15	no	6	yes	9
no	10	yes	12	no	12
no	12	yes	13	yes	12
no	6	no	20		
yes	19	yes	13		

Example: Tooth health and infant bottle use - a 95% confidence interval for $\mu_1 - \mu_2$

```
# Reading in data
x <- c(9, 10, 12, 6, 10, 8, 6, 20, 12)
y <- c(14, 15, 19, 12, 13, 13, 16, 14, 9, 12)

# 95% CI for mean difference by non-parametric bootstrap
k <- 100000

simx_samples <- replicate(k, sample(x, replace = TRUE))
simy_samples <- replicate(k, sample(y, replace = TRUE))
sim_mean_difs <- apply(simx_samples, 2, mean)-
                        apply(simy_samples, 2, mean)
quantile(sim_mean_difs, c(0.025,0.975))

##      2.5%    97.5%
## -6.2333 -0.1444
```

Two-sample confidence interval for $\theta_1 - \theta_2$ (including $\mu_1 - \mu_2$) by non-parametric bootstrap

Method 4.17: Two-sample confidence interval for $\theta_1 - \theta_2$ by non-parametric bootstrap

Assume we have actual observations x_1, \dots, x_n and y_1, \dots, y_n .

- ① Randomly draw k sets of 2 samples of n_1 and n_2 observations from the respective groups of data (with replacement).
 $\hat{\theta}_{x1}^*, \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^*, \hat{\theta}_{yk}^*$
- ② Calculate the difference between the features in each of the k samples:
 $\hat{\theta}_{x1}^* - \hat{\theta}_{y1}^*, \dots, \hat{\theta}_{xk}^* - \hat{\theta}_{yk}^*$.
- ③ Find the $100(\alpha/2)\%$ and $100(1 - \alpha/2)\%$ quantiles for these, $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$, to obtain the $100(1 - \alpha)\%$ confidence interval:
$$[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$$

Example: Tooth health and infant bottle use - a 99% confidence interval for the difference of medians

```
k <- 100000

simx_samples <- replicate(k, sample(x, replace = TRUE))
simy_samples <- replicate(k, sample(y, replace = TRUE))
sim_median_difs <- apply(simx_samples, 2, median)-
                      apply(simy_samples, 2, median)
quantile(sim_median_difs, c(0.005,0.995))

## 0.5% 99.5%
##      -8      0
```

Bootstrapping - an overview

We were given 4 similar method boxes

- ① With distribution assumptions or not (parametric or non-parametric).
- ② For one- or two-sample analysis.

Bootstrapping - an overview

We were given 4 similar method boxes

- ① With distribution assumptions or not (parametric or non-parametric).
- ② For one- or two-sample analysis.

Note:

Means also included in *other features*. Or: These methods may be used *not only* for means!

Bootstrapping - an overview

We were given 4 similar method boxes

- ① With distribution assumptions or not (parametric or non-parametric).
- ② For one- or two-sample analysis.

Note:

Means also included in *other features*. Or: These methods may be used *not only* for means!

Hypothesis testing also possible

We can do hypothesis testing by looking at the confidence intervals!

Overview

- ➊ Introduction to simulation - what is it really?
 - Example: Area of plates
- ➋ Propagation of error
- ➌ Parametric bootstrap
 - Introduction to bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval assuming any distributions
- ➍ Non-parametric bootstrap
 - One-sample confidence interval for any feature
 - Two-sample confidence interval

Course 02402 Introduction to Statistics

Lecture 8: Simple linear regression

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

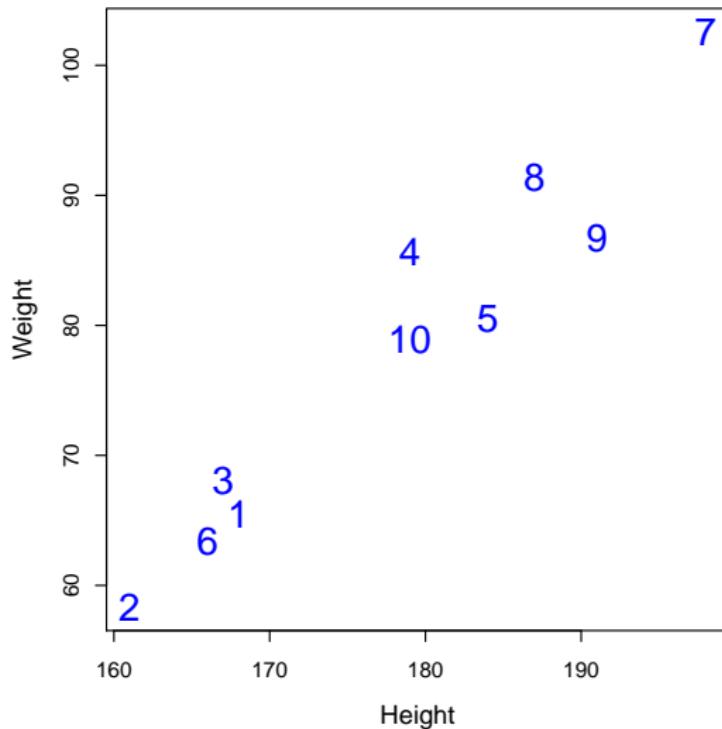
Overview

1 Example: Height-Weight

- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

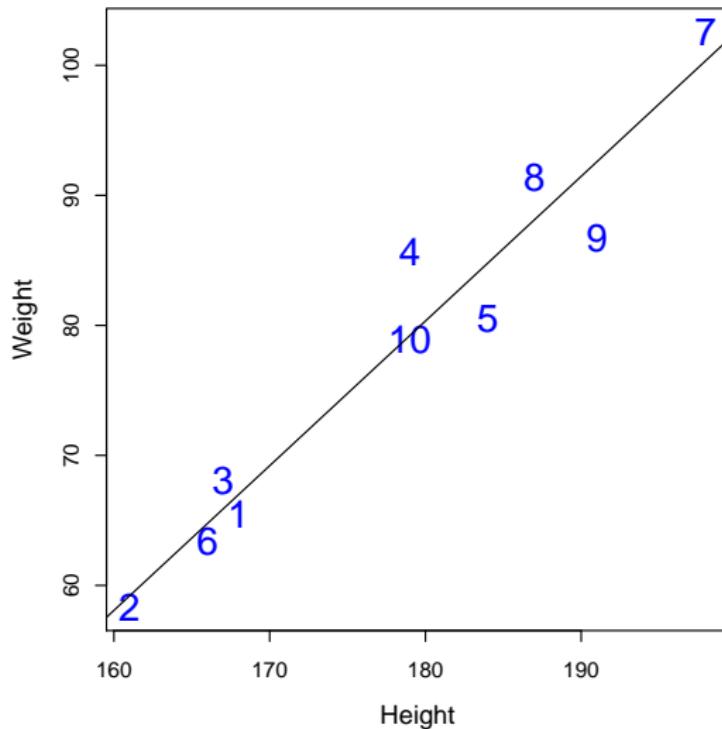
Example: Height-Weight

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Example: Height-Weight

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

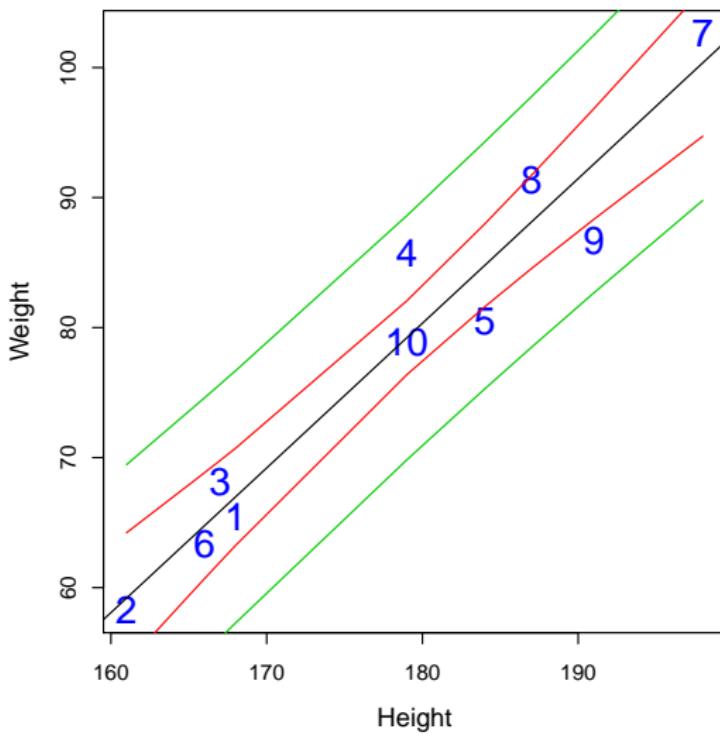


Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
summary(lm(y ~ x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.876 -1.451 -0.608  2.234  6.477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -119.958     18.897   -6.35  0.00022 ***
## x             1.113      0.106   10.50  5.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.88 on 8 degrees of freedom
## Multiple R-squared:  0.932, Adjusted R-squared:  0.924
## F-statistic: 110 on 1 and 8 DF,  p-value: 5.87e-06
```

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

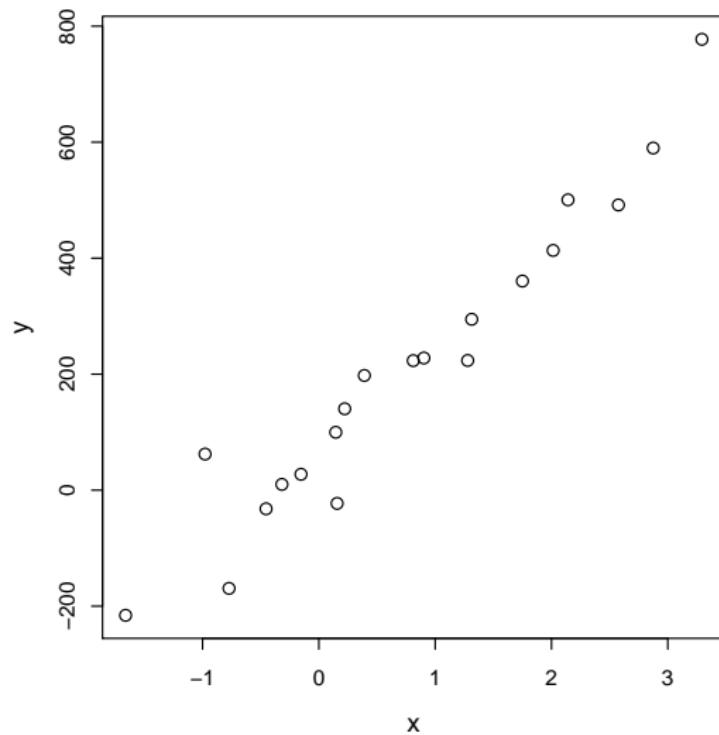


Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

A scatter plot of some data

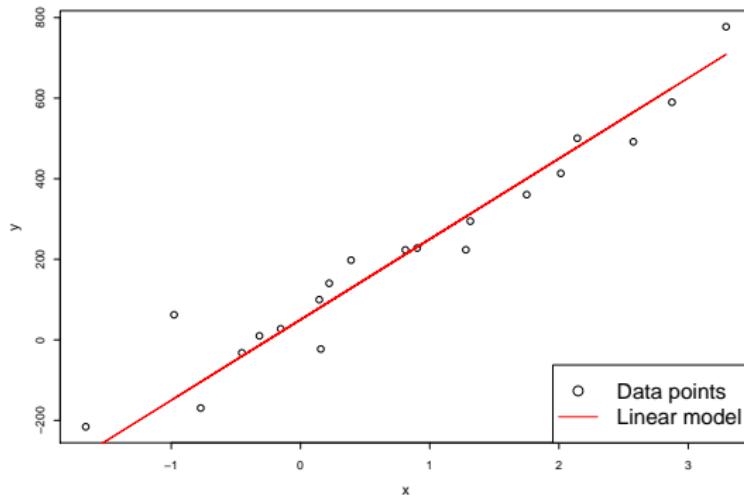
- We have n pairs of data points (x_i, y_i) .



Express a linear model

- Express a linear model:

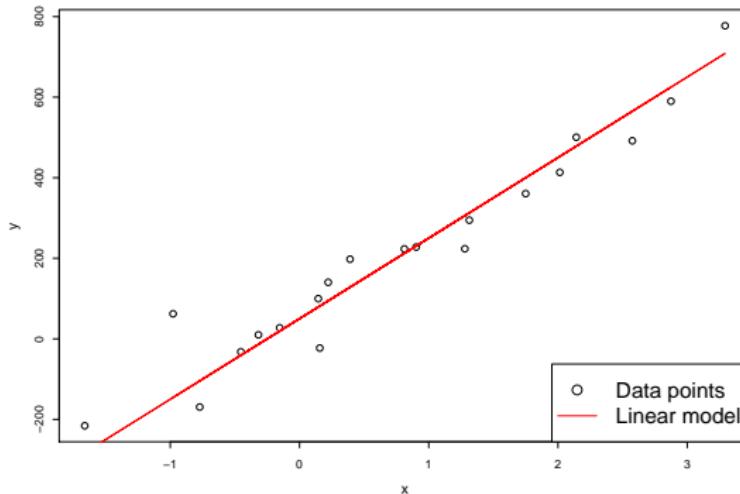
$$y_i = \beta_0 + \beta_1 x_i + ?$$



Express a linear model

- Express a linear model:

$$y_i = \beta_0 + \beta_1 x_i + ?$$



- Something is missing: Description of the *random variation*.

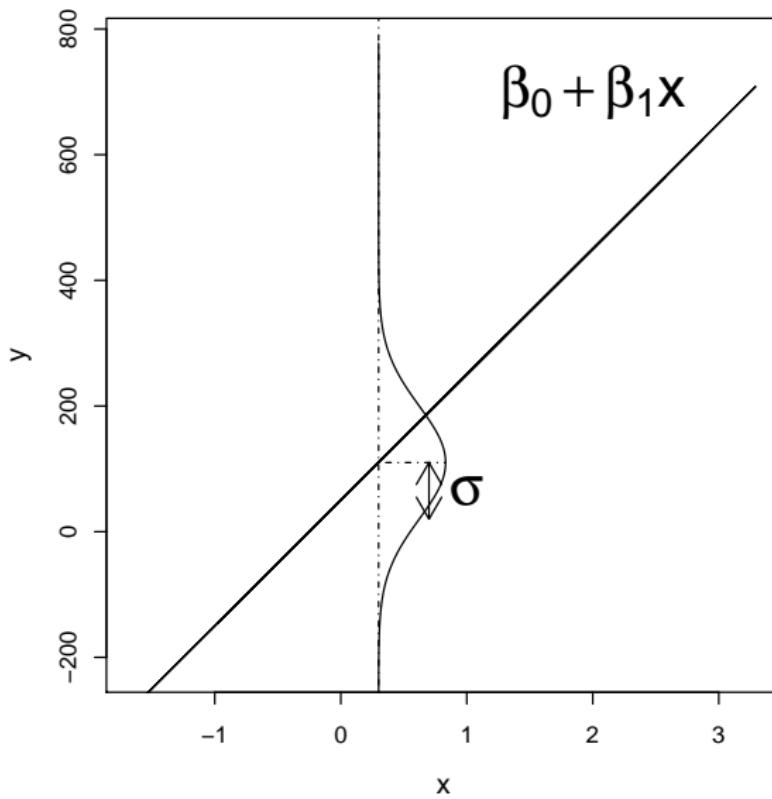
Express a linear regression model

- Express the *linear regression model*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

- Y_i is the *dependent/outcome variable*. A random variable.
- x_i is an *independent/explanatory variable*. Deterministic numbers.
- ε_i is the deviation/error. A random variable.
- We assume that the ε_i , $i = 1, \dots, n$, are *independent and identically distributed (i.i.d.)*, with $\varepsilon_i \sim N(0, \sigma^2)$.

Illustration of statistical model



Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

Least squares method

- How can we estimate the parameters β_0 and β_1 ?

Least squares method

- How can we estimate the parameters β_0 and β_1 ?
- Good idea: Minimize the variance σ^2 of the residuals.

Least squares method

- How can we estimate the parameters β_0 and β_1 ?
- Good idea: Minimize the variance σ^2 of the residuals.
- But how?

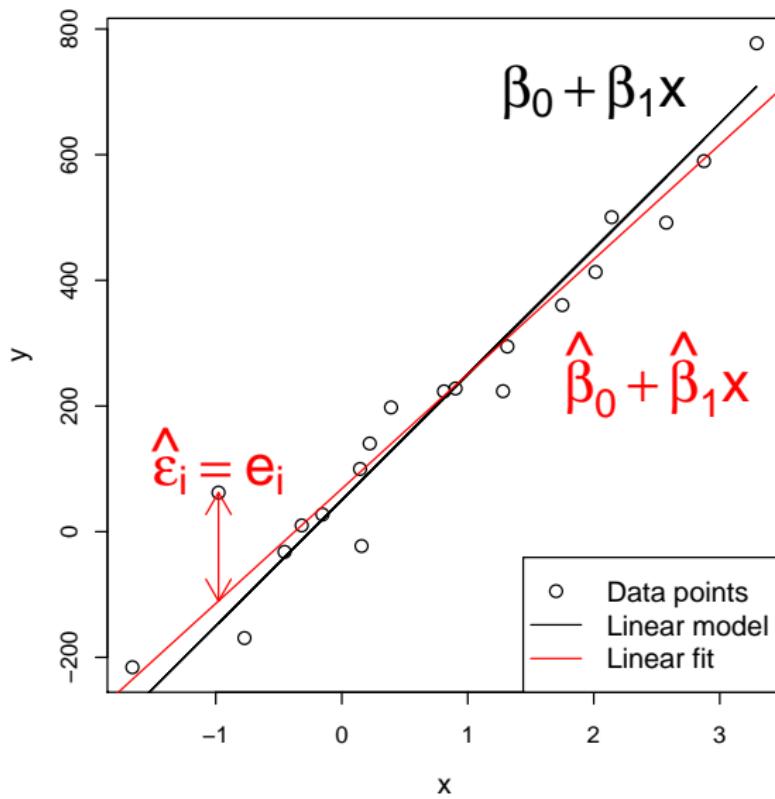
Least squares method

- How can we estimate the parameters β_0 and β_1 ?
- Good idea: Minimize the variance σ^2 of the residuals.
- But how?
- Minimize the Residual Sum of Squares (RSS),

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the RSS.

Illustration of model, data and fit



Least squares estimator

Theorem 5.4 (here as estimators, as in the book)

The least squares estimators of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.

Least squares estimates

Theorem 5.4 (here as *estimates*)

The least squares estimates of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.

R example

```
set.seed(100)

# Generate x
x <- runif(n = 20, min = -2, max = 4)

# Simulate y
beta0 <- 50; beta1 <- 200; sigma <- 90
y <- beta0 + beta1 * x + rnorm(n = length(x), mean = 0, sd = sigma)

# From here: like for the analysis of 'real data', we have data in x and y:
# Scatter plot of y against x
plot(x, y)

# Find the least squares estimates, use Theorem 5.4
(beta1hat <- sum( (y - mean(y))*(x-mean(x)) ) / sum( (x-mean(x))^2 ))
(beta0hat <- mean(y) - beta1hat*mean(x))

# Use lm() to find the estimates
lm(y ~ x)

# Plot the fitted line
abline(lm(y ~ x), col="red")
```

Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

The parameter estimates are random variables

What if we took a new sample?

Would the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ be the same?

The parameter estimates are random variables

What if we took a new sample?

Would the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ be the same?

No, they are random variables!

If we took a new sample, we would get another realisation.

The parameter estimates are random variables

What if we took a new sample?

Would the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ be the same?

No, they are random variables!

If we took a new sample, we would get another realisation.

What are the (sampling) distributions of the parameter estimates ...

... in a linear regression model w. normal distributed errors?

The parameter estimates are random variables

What if we took a new sample?

Would the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ be the same?

No, they are random variables!

If we took a new sample, we would get another realisation.

What are the (sampling) distributions of the parameter estimates ...

... in a linear regression model w. normal distributed errors?

This may be investigated using simulation ...

Let's go to R!

The distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are normal distributed and their variance can be estimated:

Theorem 5.8 (first part)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$$

$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x} \sigma^2}{S_{xx}}$$

- We won't use the covariance $Cov[\hat{\beta}_0, \hat{\beta}_1]$ for now.

Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$

Theorem 5.8 (second part)

σ^2 is usually replaced by its estimate, $\hat{\sigma}^2$, the *central estimator of σ^2* :

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of σ^2 is used, the variances also become estimates. We'll refer to them as $\hat{\sigma}_{\beta_0}^2$ and $\hat{\sigma}_{\beta_1}^2$.

Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$

Theorem 5.8 (second part)

σ^2 is usually replaced by its estimate, $\hat{\sigma}^2$, the *central estimator of σ^2* :

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of σ^2 is used, the variances also become estimates. We'll refer to them as $\hat{\sigma}_{\beta_0}^2$ and $\hat{\sigma}_{\beta_1}^2$.

Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ (equations 5-43 and 5-44):

$$\hat{\sigma}_{\beta_0} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}; \quad \hat{\sigma}_{\beta_1} = \hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

Hypothesis tests for β_0 and β_1

We can carry out hypothesis tests for the parameters in a linear regression model:

$$\begin{aligned} H_{0,i} : \quad \beta_i &= \beta_{0,i} \\ H_{1,i} : \quad \beta_i &\neq \beta_{1,i} \end{aligned}$$

Hypothesis tests for β_0 and β_1

We can carry out hypothesis tests for the parameters in a linear regression model:

$$\begin{aligned} H_{0,i} : \quad \beta_i &= \beta_{0,i} \\ H_{1,i} : \quad \beta_i &\neq \beta_{1,i} \end{aligned}$$

Theorem 5.12

Under the null-hypotheses ($\beta_0 = \beta_{0,0}$ and $\beta_1 = \beta_{0,1}$) the statistics

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}; \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

are t -distributed with $n - 2$ degrees of freedom, and inference should be based on this distribution.

Hypothesis tests for β_0 and β_1

- See Example 5.13 for an example of a hypothesis test.
- Test if the parameters are significantly different from 0:

$$H_{0,i} : \beta_i = 0, \quad H_{1,i} : \beta_i \neq 0$$

```
# Read data into R
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Fit model to data
fit <- lm(y ~ x)

# Look at model summary to find Tobs-values and p-values
summary(fit)
```

Confidence intervals for β_0 and β_1

Method 5.15

$(1 - \alpha)$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_0}$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2} \hat{\sigma}_{\beta_1}$$

where $t_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of a t -distribution with $n - 2$ degrees of freedom.

- Remember that $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ may be found using equations 5-43 and 5-44.
- In R, we can find $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ under "Std. Error" from `summary(fit)`.

Illustration of CIs by simulation

```
# Number of repetitions (here: CIs)
nRepeat <- 1000

# Empty logical vector of length nRepeat
TrueValInCI <- logical(nRepeat)

# Repeat the simulation and estimation nRepeat times:
for(i in 1:nRepeat){
  # Generate x
  x <- runif(n = 20, min = -2, max = 4)
  # Simulate y
  beta0 = 50; beta1 = 200; sigma = 90
  y <- beta0 + beta1 * x + rnorm(n = length(x), mean = 0, sd = sigma)
  # Use lm() to fit model
  fit <- lm(y ~ x)
  # Use confint() to compute 95% CI for intercept
  ci <- confint(fit, "(Intercept)", level=0.95)
  # Was the 'true' intercept included in the interval? (covered)
  (TrueValInCI[i] <- ci[1] < beta0 & beta0 < ci[2])
}

# How often was the true intercept included in the CI?
sum(TrueValInCI) / nRepeat
```

Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line**
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

Method 5.18 Confidence interval for $\beta_0 + \beta_1 x_0$

- The confidence interval for $\beta_0 + \beta_1 x_0$ corresponds to a confidence interval for the line at the point x_0 .
- The $100(1 - \alpha)\%$ CI is computed by

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

Method 5.18 Prediction interval for $\beta_0 + \beta_1 x_0 + \varepsilon_0$

- The prediction interval for Y_0 is found using a value x_0 .
- This is done *before* Y_0 is observed, using

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

- In $100(1 - \alpha)\%$ of cases, the prediction interval will contain the observed y_0 .
- For a given α , a prediction interval is wider than a confidence interval.

Example of confidence intervals for the line

```
# Generate x
x <- runif(n = 20, min = -2, max = 4)

# Simulate y
beta0 = 50; beta1 = 200; sigma = 90
y <- beta0 + beta1 * x + rnorm(n = length(x), sd = sigma)

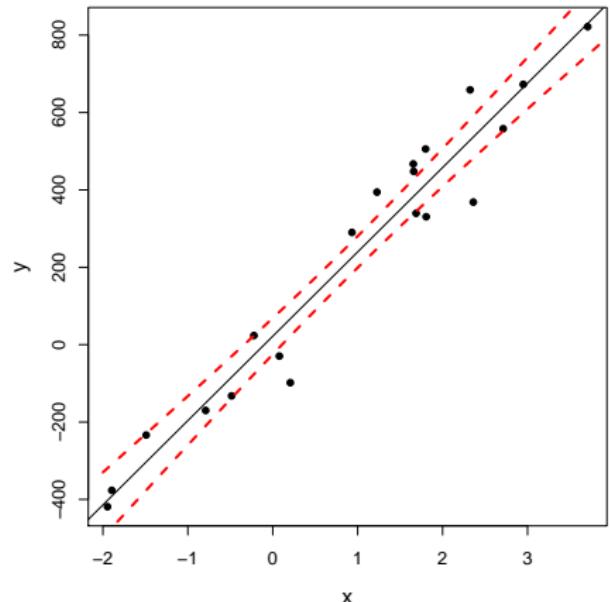
# Use lm() to fit model
fit <- lm(y ~ x)

# Make a sequence of 100 x-values
xval <- seq(from = -2, to = 6, length.out = 100)

# Use the predict function
CI <- predict(fit, newdata = data.frame(x = xval),
  interval = "confidence",
  level = 0.95)

# Check what we got
head(CI)

# Plot the data, model fit and intervals
plot(x, y, pch = 20)
abline(fit)
lines(xval, CI[, "lwr"], lty=2, col = "red", lwd = 2)
lines(xval, CI[, "upr"], lty=2, col = "red", lwd = 2)
```



Example of prediction intervals for the line

```

# Generate x
x <- runif(n = 20, min = -2, max = 4)

# Simulate y
beta0 = 50; beta1 = 200; sigma = 90
y <- beta0 + beta1 * x + rnorm(n = length(x), sd = sigma)

# Use lm() to fit model
fit <- lm(y ~ x)

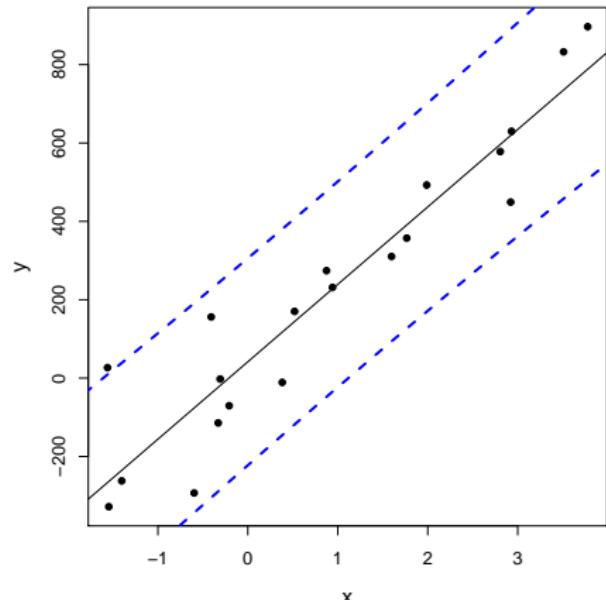
# Make a sequence of 100 x-values
xval <- seq(from = -2, to = 6, length.out = 100)

# Use the predict function
PI <- predict(fit, newdata = data.frame(x = xval),
              interval = "prediction",
              level = 0.95)

# Check what we got
head(PI)

# Plot the data, model fit and intervals
plot(x, y, pch = 20)
abline(fit)
lines(xval, PI[, "lwr"], lty = 2, col = "blue", lwd = 2)
lines(xval, PI[, "upr"], lty = 2, col = "blue", lwd = 2)

```



Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

What more do we get from `summary()`?

```
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -216.86  -66.09   -7.16   58.48  293.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  41.8       30.9     1.35    0.19    
## x            197.6      16.4    12.05  4.7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122 on 18 degrees of freedom
## Multiple R-squared:  0.89, Adjusted R-squared:  0.884 
## F-statistic: 145 on 1 and 18 DF,  p-value: 4.73e-10
```

summary(lm(y~x))

• Residuals: Min 1Q Median 3Q Max

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

The coefficients':

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	$p\text{-value}$
-----------------	--------------------------	------------------	------------------

- The test is $H_{0,i} : \beta_i = 0$ vs. $H_{1,i} : \beta_i \neq 0$
- The stars indicate which size category the p -value belongs to.

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

The coefficients':

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	$p\text{-value}$
-----------------	--------------------------	------------------	------------------

- The test is $H_{0,i} : \beta_i = 0$ vs. $H_{1,i} : \beta_i \neq 0$
- The stars indicate which size category the p -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

The coefficients':

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	$p\text{-value}$
-----------------	--------------------------	------------------	------------------

- The test is $H_{0,i} : \beta_i = 0$ vs. $H_{1,i} : \beta_i \neq 0$
- The stars indicate which size category the p -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom
 $\varepsilon_i \sim N(0, \sigma^2)$, the output shows $\hat{\sigma}$ and v degrees of freedom (used for hypothesis tests, CIs, PIs etc.)

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

The coefficients':

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	$p\text{-value}$
-----------------	--------------------------	------------------	------------------

- The test is $H_{0,i} : \beta_i = 0$ vs. $H_{1,i} : \beta_i \neq 0$
- The stars indicate which size category the p -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom
 $\varepsilon_i \sim N(0, \sigma^2)$, the output shows $\hat{\sigma}$ and v degrees of freedom (used for hypothesis tests, CIs, PIs etc.)
- Multiple R-squared: XXX

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

The coefficients':

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	$p\text{-value}$
-----------------	--------------------------	------------------	------------------

- The test is $H_{0,i} : \beta_i = 0$ vs. $H_{1,i} : \beta_i \neq 0$
- The stars indicate which size category the p -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom
 $\varepsilon_i \sim N(0, \sigma^2)$, the output shows $\hat{\sigma}$ and v degrees of freedom (used for hypothesis tests, CIs, PIs etc.)
- Multiple R-squared: XXX
Explained variation r^2 .

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max

The residuals': minimum, 1st quartile, median, 3rd quartile, maximum

- Coefficients:

Estimate	Std. Error	t value	Pr(> t)	"stars"
----------	------------	---------	----------	---------

The coefficients':

$\hat{\beta}_i$	$\hat{\sigma}_{\beta_i}$	t_{obs}	$p\text{-value}$
-----------------	--------------------------	------------------	------------------

- The test is $H_{0,i} : \beta_i = 0$ vs. $H_{1,i} : \beta_i \neq 0$
- The stars indicate which size category the p -value belongs to.
- Residual standard error: XXX on XXX degrees of freedom
 $\varepsilon_i \sim N(0, \sigma^2)$, the output shows $\hat{\sigma}$ and v degrees of freedom (used for hypothesis tests, CIs, PIs etc.)
- Multiple R-squared: XXX
Explained variation r^2 .
- The rest we don't use in this course.

Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

Explained variation and correlation

- Explained variation in a model is r^2 , in summary "Multiple R-squared".
- Found as

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

- The proportion of the total variability explained by the model.

Explained variation and correlation

- The correlationen ρ is a measure of *linear relation* between two random variables.
- Estimated (i.e. empirical) correlation satisfies that

$$\hat{\rho} = r = \sqrt{r^2} \operatorname{sgn}(\hat{\beta}_1)$$

where $\operatorname{sgn}(\hat{\beta}_1)$ is: -1 for $\hat{\beta}_1 \leq 0$ and 1 for $\hat{\beta}_1 > 0$

Explained variation and correlation

- The correlationen ρ is a measure of *linear relation* between two random variables.
- Estimated (i.e. empirical) correlation satisfies that

$$\hat{\rho} = r = \sqrt{r^2} \operatorname{sgn}(\hat{\beta}_1)$$

where $\operatorname{sgn}(\hat{\beta}_1)$ is: -1 for $\hat{\beta}_1 \leq 0$ and 1 for $\hat{\beta}_1 > 0$

- Hence:
 - Positive correlation when positive slope.
 - Negative correlation when negative slope.

Test for significance of correlation

- Test for significance of correlation (linear relation) between two variables

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

is equivalent to

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

where $\hat{\beta}_1$ is the estimated slope in a simple linear regression model

Example: Correlation and R^2 for height-weight data

```
# Read data into R

x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Fit model to data
fit <- lm(y ~ x)

# Scatter plot of data with fitted line
plot(x,y, xlab = "Height", ylab = "Weight")
abline(fit, col="red")

# See summary
summary(fit)

# Correlation between x and y
cor(x,y)

# Squared correlation is the "Multiple R-squared" from summary(fit)
cor(x,y)^2
```

Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

Residual Analysis

Method 5.28

- Check normality assumptions with a qq-plot.
- Check (non-)systematic behavior by plotting the residuals, e_i , as a function of the fitted values \hat{y}_i .

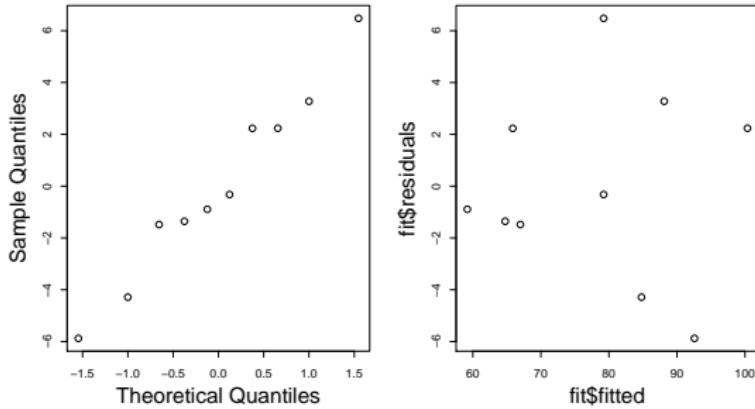
(Method 5.29)

- Is the independence assumption reasonable?

Residual analysis in R

```
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)
fit <- lm(y ~ x)

par(mfrow = c(1, 2))
qqnorm(fit$residuals, main = "", cex.lab = 1.5)
plot(fit$fitted, fit$residuals, cex.lab = 1.5)
```



Overview

- 1 Example: Height-Weight
- 2 Linear regression model
- 3 Least squares method
- 4 Statistics and linear regression?
- 5 Hypothesis tests and confidence intervals for β_0 and β_1
- 6 Confidence and prediction intervals for the line
- 7 Summary of 'summary(lm(y~x))'
- 8 Correlation
- 9 Residual Analysis: Model validation

Course 02402 Introduction to Statistics

Lecture 9: Multiple linear regression

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Agenda

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Example: Ozon concentration

We have a set of observations of: logarithm to ozone concentration ($\log(\text{ppm})$), temperature, radiation and wind speed:

ozone	radiation	wind	temperature	month	day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
:	:	:	:	:	:
18	131	8.0	76	9	29
20	223	11.5	68	9	30

Example: Ozone concentration

```
## See info about data
?airquality
## Copy the data
Air <- airquality
## Remove rows with at least one NA value
Air <- na.omit(Air)

## Remove one outlier
Air <- Air[-which(Air$Ozone == 1), ]

## Check the empirical density
hist(Air$Ozone, probability=TRUE, xlab="Ozon", main="")

## Concentrations are positive and very skewed, let's
## log-transform right away:
## (although really one could wait and check residuals from models)
Air$logOzone <- log(Air$Ozone)
## Bedre epdf?
hist(Air$logOzone, probability=TRUE, xlab="log Ozone", main="")

## Make a time variable (R timeclass, see ?POSIXct)
Air$t <- ISOdate(1973, Air$Month, Air$Day)
## Keep only some of the columns
Air <- Air[,c(7,4,3,2,8)]
## New names of the columns
names(Air) <- c("logOzone", "temperature", "wind", "radiation", "t")

## What's in Air?
str(Air)
Air
head(Air)
tail(Air)

## Typically one would begin with a pairs plot
pairs(Air, panel = panel.smooth, main = "airquality data")
```

Example: Ozone concentration

- Let us first analyse the relation between ozone and temperature
- Apply a *simple linear regressions model*

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$

where

- Y_i is the (logarithm of) ozone concentration of observation i
- x_i is the temperature at observation i

Fit the model in R

```
#####
## See the relation between ozone and temperature
plot(Air$temperature, Air$logOzone, xlab="Temperature", ylab="Ozon")

## Correlation
cor(Air$logOzone, Air$temperature)

## Fit a simple linear regression model
summary(lm(logOzone ~ temperature, data=Air))

## Add a vector with random values, is there a significant linear relation?
## ONLY for ILLUSTRATION purposes
Air$noise <- rnorm(nrow(Air))
plot(Air$logOzone, Air$noise, xlab="Noise", ylab="Ozon")
cor(Air$logOzone, Air$noise)
summary(lm(logOzone ~ noise, data=Air))
```

Simple linear regression model for the other two

We can also make a simple linear regression model with each of the other two independent variables

```
#####
## With each of the other two independent variables

## Simple linear regression model with the wind speed
plot(Air$logOzone, Air$wind, xlab="logOzone", ylab="Wind speed")
cor(Air$logOzone, Air$wind)
summary(lm(logOzone ~ wind, data=Air))

## Simple linear regression model with the radiation
plot(Air$logOzone, Air$radiation, xlab="logOzone", ylab="Radiation")
cor(Air$logOzone, Air$radiation)
summary(lm(logOzone ~ radiation, data=Air))
```

Overview

- 1 Warm up with some simple linear regression
- 2 **Multiple linear regression**
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Multiple linear regression

- Y is the *dependent variable*
- We are interested in modelling the Y 's dependency of the *independent* or *explanatory* variables x_1, x_2, \dots, x_p
- We are modelling a *linear relation* between Y and x_1, x_2, \dots, x_p , described with the regression model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

- Y_i og ε_i are random variables and $x_{j,i}$ are variables

Least squares estimates

- The coefficient estimates are found by minimizing:

$$RSS(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p})]^2$$

- The "predicted" (= "fitted") are found as

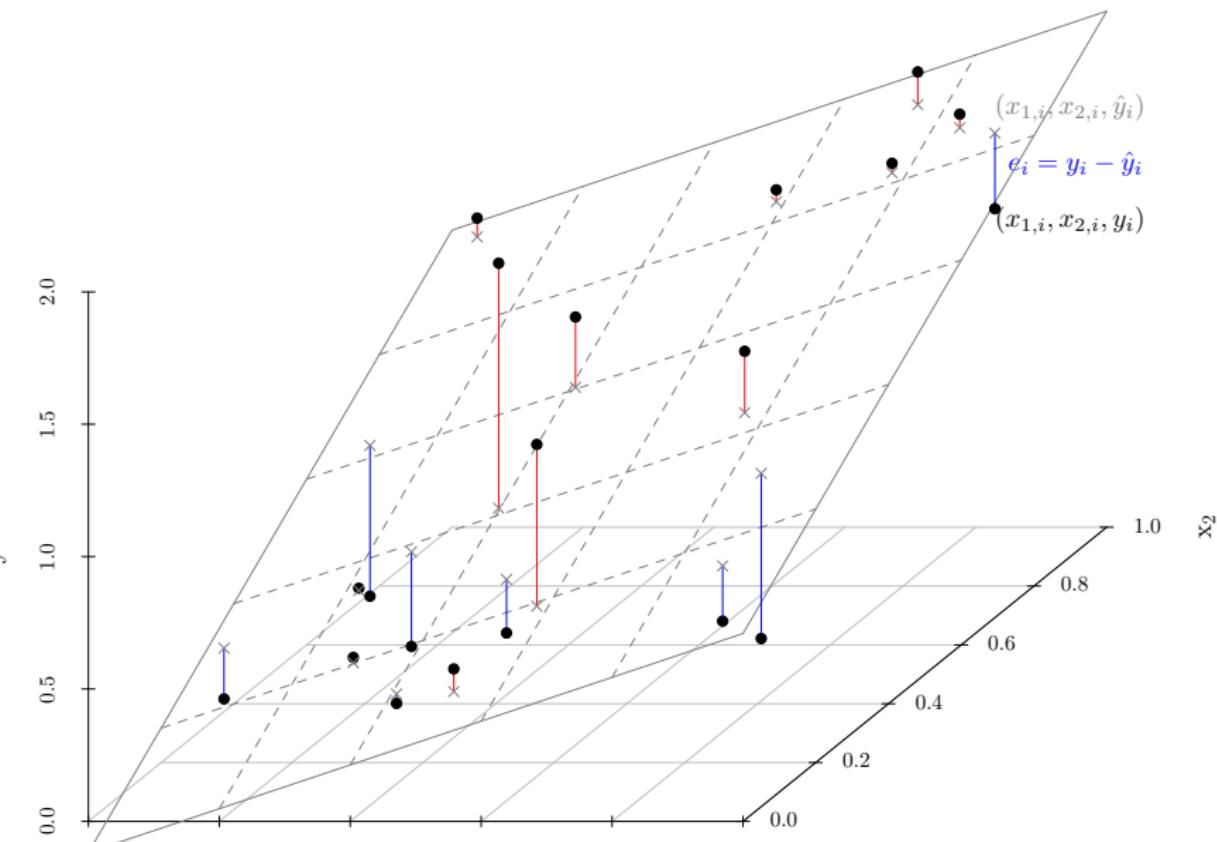
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_p x_{i,p}$$

- And then the residuals are found as

$$e_i = y_i - \hat{y}_i$$

residual = observation – prediction

Least squares estimates - The concept!



Computations for MLR - no explicit formulas given!

- Remark 6.6: Extract $\hat{\beta}_i$ and $\hat{\sigma}_{\beta_i}$ from R-output (`summary(myfit)`)

Computations for MLR - no explicit formulas given!

- Remark 6.6: Extract $\hat{\beta}_i$ and $\hat{\sigma}_{\beta_i}$ from R-output (`summary(myfit)`)
- Theorem 6.2: The t-distribution can be used for inference for parameters

Computations for MLR - no explicit formulas given!

- Remark 6.6: Extract $\hat{\beta}_i$ and $\hat{\sigma}_{\beta_i}$ from R-output (`summary(myfit)`)
- Theorem 6.2: The t-distribution can be used for inference for parameters
- Methods 6.4 and 6.5: Hypothesis tests and Confidence intervals for parameters based on R-output.

Computations for MLR - no explicit formulas given!

- Remark 6.6: Extract $\hat{\beta}_i$ and $\hat{\sigma}_{\beta_i}$ from R-output (`summary(myfit)`)
- Theorem 6.2: The t-distribution can be used for inference for parameters
- Methods 6.4 and 6.5: Hypothesis tests and Confidence intervals for parameters based on R-output.
- Everything: **THE SAME as for SIMPLE linear regression!**

Computations for MLR - no explicit formulas given!

- Remark 6.6: Extract $\hat{\beta}_i$ and $\hat{\sigma}_{\beta_i}$ from R-output (`summary(myfit)`)
- Theorem 6.2: The t-distribution can be used for inference for parameters
- Methods 6.4 and 6.5: Hypothesis tests and Confidence intervals for parameters based on R-output.
- Everything: **THE SAME as for SIMPLE linear regression!**
- (In Section 6.6: Mathematical matrix based expressions including explicit formulas. Not syllabus in course 02402)

Parameter interpretation in MLR (Remark 6.14)

What dose $\hat{\beta}_i$ express?

- The expected y -change with 1 unit x_i -change

Parameter interpretation in MLR (Remark 6.14)

What does $\hat{\beta}_i$ express?

- The expected y -change with 1 unit x_i -change
- The effect of x_i given the other variables
- The effect of x_i corrected for the other variables
- The effect of x_i "other variables being equal"
- The unique effect of x_i

Parameter interpretation in MLR (Remark 6.14)

What dose $\hat{\beta}_i$ express?

- The expected y -change with 1 unit x_i -change
- The effect of x_i given the other variables
- The effect of x_i corrected for the other variables
- The effect of x_i "other variables being equal"
- The unique effect of x_i
- Depends on what else is in the model!!

Parameter interpretation in MLR (Remark 6.14)

What does $\hat{\beta}_i$ express?

- The expected y -change with 1 unit x_i -change
- The effect of x_i given the other variables
- The effect of x_i corrected for the other variables
- The effect of x_i "other variables being equal"
- The unique effect of x_i
- Depends on what else is in the model!!
- Generally: NOT a causal/intervention effect!!

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 **Model selection**
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Extend the model (forward selection)

- *Not included in the eNote*
- Start with the *linear regression model* with the most significant independent variable
- Extend the model with the remaining independent variables (inputs) one at a time
- Stop when there is not any significant extensions possible

```
#####
## Extend the model

## Forward selection:
## Add wind to the model
summary(lm(logOzone ~ temperature + wind, data=Air))
## Add radiation to the model
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))
```

Reduce the model (model reduction or backward selection)

- *Described in the eNote, section 6.5*
- Start with the full model
- Remove the most insignificant independent variable
- Stop when all prm. estimates are significant

```
#####
## Backward selection

## Fit the full model
summary(lm(logOzone ~ temperature + wind + radiation + noise, data=Air))
## Remove the most non-significant input, are all now significant?
summary(lm(logOzone ~ temperature + wind + radiation, data=Air))
```

Model selection

- There is no fully certain method for finding the best model!
- It will require subjective decisions to select a model
- Different procedures: either forward or backward selection (or both), depends on the circumstances
- Statistical measures and tests to compare model fits
- In this course only backward selection is described

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Residual analysis (model validation)

- Model validation: Analyze the residuals to check that the assumptions is met
- $e_i \sim N(0, \sigma^2)$ is independent and identically distributed (i.i.d.)
- Same as for the simple linear regression model

Assumption of normal distributed residuals

- Make a qq-normalplot (normal score plot) to see if they seem normal distributed

```
#####
## Assumption of normal distributed residuals

## Save the selected fit
fitSel <- lm(logOzone ~ temperature + wind + radiation, data=Air)

## qq-normalplot
qqnorm(fitSel$residuals)
qqline(fitSel$residuals)
```

Assumption of identical distribution of residuals

- Plot the residuals (e_i) versus the predicted (fitted) values (\hat{y}_i)

```
#####
## Plot the residuals vs. predicted values

plot(fitSel$fitted.values, fitSel$residuals, xlab="Predicted values",
      ylab="Residuals")
```

- Seems like the model kan be improved!
- Plot the residuals vs. the independent variables

```
#####
## Plot the residuals vs. the independent variables

par(mfrow=c(1,3))
plot(Air$temperature, fitSel$residuals, xlab="Temperature")
plot(Air$wind, fitSel$residuals, xlab="Wind speed")
plot(Air$radiation, fitSel$residuals, xlab="Radiation")
```

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Curvilinear model

If we want to estimate a model of the type

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

we can use a multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$$

where

- $x_{i,1} = x_i$
- $x_{i,2} = x_i^2$

and apply the same methods as for multiple linear regression.

Extend the ozone model with appropriate curvilinear regression

```
#####
## Extend the ozone model with appropriate curvilinear regression

## Make the squared wind speed
Air$windSq <- Air$wind^2
## Add it to the model
fitWindSq <- lm(logOzone ~ temperature + wind + windSq + radiation, data=Air)
summary(fitWindSq)

## Equivalently for the temperature
Air$temperature2 <- Air$temperature^2
## Add it
fitTemperatureSq <- lm(logOzone ~ temperature + temperature2 + wind + radiation, data=Air)
summary(fitTemperatureSq)

## Equivalently for the radiation
Air$radiation2 <- Air$radiation^2
## Add it
fitRadiationSq <- lm(logOzone ~ temperature + wind + radiation + radiation2, data=Air)
summary(fitRadiationSq)

## Which one was best?
## One could try to extend the model further
fitWindSqTemperatureSq <- lm(logOzone ~ temperature + temperature2 + wind + windSq + radiation, data=Air)
summary(fitWindSqTemperatureSq)

## Model validation
qqnorm(fitWindSq$residuals)
qqline(fitWindSq$residuals)
plot(fitWindSq$residuals, fitWindSq$fitted.values, pch=19)
```

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 **Confidence and prediction intervals**
- 7 Colinearity
- 8 The overall regression method

Confidence and prediction intervals for the plane, Method 6.9:

Extract Confidence and prediction intervals for the plane by R-function `predict`. Options for `confidence` og `prediction` exist.

```
#####
## Confidence and prediction intervals for the curvilinear model

## Generate a new data.frame with constant temperature and radiation, but with varying wind speed
wind<-seq(1,20.3,by=0.1)
AirForPred <- data.frame(temperature=mean(Air$temperature), wind=wind,
                           windSq=wind^2, radiation=mean(Air$radiation))

## Calculate confidence and prediction intervals (actually bands)
CI <- predict(fitWindSq, newdata=AirForPred, interval="confidence", level=0.95)
PI <- predict(fitWindSq, newdata=AirForPred, interval="prediction", level=0.95)

## Plot them
plot(wind, CI[, "fit"], ylim=range(CI,PI), type="l",
      main=paste("At temperature =", format(mean(Air$temperature), digits=3),
                 "and radiation =", format(mean(Air$radiation), digits=3)))
lines(wind, CI[, "lwr"], lty=2, col=2)
lines(wind, CI[, "upr"], lty=2, col=2)
lines(wind, PI[, "lwr"], lty=2, col=3)
lines(wind, PI[, "upr"], lty=2, col=3)
## legend
legend("topright", c("Prediction", "95% confidence band", "95% prediction band"), lty=c(1,2,2), col=1:3)
```

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 **Colinearity**
- 8 The overall regression method

Colinearity

- MLR breaks down if X-data has "exact linear redundancy"
 - Example: Both height in cm and height in m is in the data.

Colinearity

- MLR breaks down if X-data has "exact linear redundancy"
 - Example: Both height in cm and height in m is in the data.
- Interpretation and model stability is challenged if X-data has "near redundancy" patterns
 - Example: Both weight and BMI are in the X-data (highly correlated)

Colinearity

- MLR breaks down if X-data has "exact linear redundancy"
 - Example: Both height in cm and height in m is in the data.
- Interpretation and model stability is challenged if X-data has "near redundancy" patterns
 - Example: Both weight and BMI are in the X-data (highly correlated)
- With e.g. two highly correlated x -variables:
 - Together in the model for y none of them may have a unique effect
 - Separately they may have a strong effect each of them

Colinearity - an illustration in R

```
#####
## See problems with highly correlated inputs

## Generate values for MLR
n <- 100
## First variable
x1 <- sin(0:(n-1)/(n-1)*2*2*pi) + rnorm(n, 0, 0.1)
plot(x1, type="b")
## The second variable is the first plus a little noise
x2 <- x1 + rnorm(n, 0, 0.1)
## x1 and x2 are highly correlated
plot(x1,x2)
cor(x1,x2)
## Simulate an MLR
beta0=20; beta1=1; beta2=1; sigma=1
y <- beta0 + beta1 * x1 + beta2 * x2 + rnorm(n,0,sigma)
## See scatter plots for y vs. x1, and y vs. x2
par(mfrow=c(1,2))
plot(x1,y)
plot(x2,y)
## Fit an MLR
summary(lm(y ~ x1 + x2))

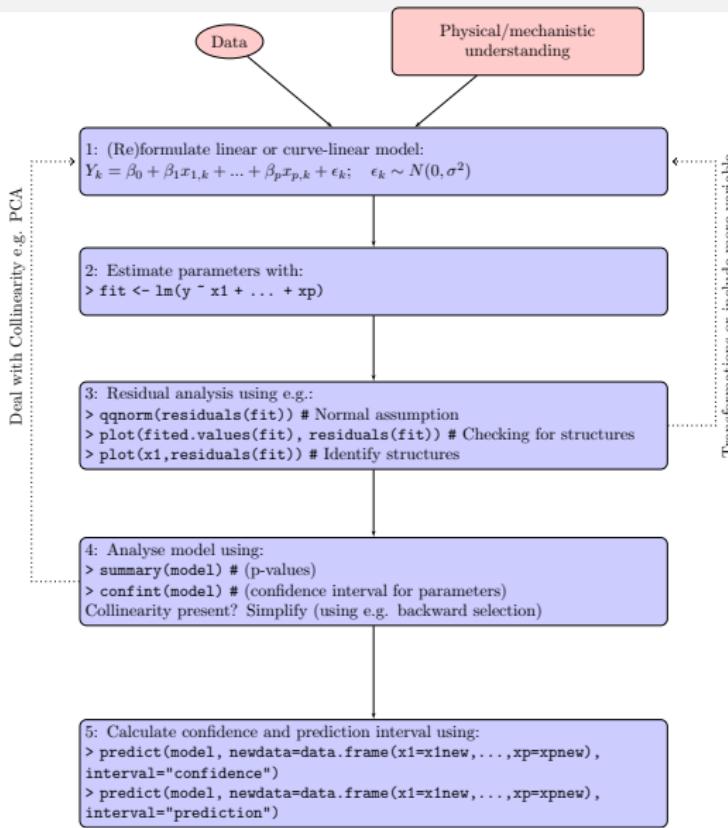
## If it was an experiment and the effects could be separated in the design
x1[1:(n/2)] <- 0
x2[(n/2):n] <- 0
## Plot them
plot(x1, type="b")
lines(x2, type="b", col="red")
## Now very low correlation
cor(x1,x2)
## Simulate MLR again
y <- beta0 + beta1 * x1 + beta2 * x2 + rnorm(n,0,sigma)
## and fit MLR
summary(lm(y ~ x1 + x2))
```

It is important how experiments
are designed!

Overview

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

The overall regression method box 6.16



Agenda

- 1 Warm up with some simple linear regression
- 2 Multiple linear regression
- 3 Model selection
- 4 Residual analysis (model validation)
- 5 Curvilinearity
- 6 Confidence and prediction intervals
- 7 Colinearity
- 8 The overall regression method

Course 02402 Introduction to Statistics

Lecture 10: Inference for proportions

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 Hypothesis test for several proportions
- 6 Analysis of contingency tables

Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 Hypothesis test for several proportions
- 6 Analysis of contingency tables

Different analysis/data situations in 02402

Mean of quantitative data

- Hypothesis test/CI for one mean (one sample)
- Hypothesis test/CI for two means (two samples)
- Hypothesis test/CI for several means (K samples)

Different analysis/data situations in 02402

Mean of quantitative data

- Hypothesis test/CI for one mean (one sample)
- Hypothesis test/CI for two means (two samples)
- Hypothesis test/CI for several means (K samples)

Today: Proportions

- Hypothesis test/CI for one proportion
- Hypothesis test/CI for two proportions
- Hypothesis test for several proportions
- Hypothesis test for several “multi-categorical” proportions

Estimation of proportions

- Estimation of a proportion/probability, by observing how many times x an event has occurred in n (independent) trials:

$$\hat{p} = \frac{x}{n}$$

- Note that $\hat{p} \in [0; 1]$.
- Example: A dice is thrown $n = 100$ times. In $x = 20$ cases the outcome was . Then, \hat{p} is the estimated probability of throwing a .

Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 Hypothesis test for several proportions
- 6 Analysis of contingency tables

Confidence interval for one proportion

Method 7.3

If we have a **large sample**, then a $(1 - \alpha)100\%$ confidence interval for p is:

$$\frac{x}{n} - z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}}$$

Confidence interval for one proportion

Method 7.3

If we have a **large sample**, then a $(1 - \alpha)100\%$ confidence interval for p is:

$$\frac{x}{n} - z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}}$$

How?

Follows from **approximating** the binomial distribution by the normal distribution.

Confidence interval for one proportion

Method 7.3

If we have a **large sample**, then a $(1 - \alpha)100\%$ confidence interval for p is:

$$\frac{x}{n} - z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}} < p < \frac{x}{n} + z_{1-\alpha/2} \sqrt{\frac{\frac{x}{n}(1 - \frac{x}{n})}{n}}$$

How?

Follows from **approximating** the binomial distribution by the normal distribution.

A rule of thumb

Suppose that $X \sim \text{binom}(n, p)$. The normal distribution is a good approximation of the binomial distribution if np and $n(1 - p)$ (expected no. of successes and failures, respectively) are both greater than 15.

Confidence interval for one proportion

Mean and variance of binomial distribution, Chapter 2.21

$$E(X) = np$$

$$Var(X) = np(1-p)$$

This means that

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X) = \frac{p(1-p)}{n}$$

Example 1

Left-handedness:

p = Proportion of left-handed people in Denmark

and/or:

Female engineering students:

p = Proportion of female engineering students

Example 1

Left-handedness:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 = 0.10 \pm 0.059 = [0.041, 0.159]$$

Example 1

Left-handedness:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 = 0.10 \pm 0.059 = [0.041, 0.159]$$

Better “small sample” method - the “plus 2-approach” (Remark 7.7)

Use the same formula on $\tilde{x} = 10 + 2 = 12$ and $\tilde{n} = 100 + 2 + 2 = 104$:

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.031$$

$$0.115 \pm 1.96 \cdot 0.031 = 0.115 \pm 0.061 = [0.054, 0.177]$$

The Margin of Error (ME)

The Margin of Error

with $(1 - \alpha)100\%$ confidence becomes:

$$ME = z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

where p may be estimated using $\hat{p} = \frac{x}{n}$.

The Margin of Error:

- Corresponds to half the width of the $(1 - \alpha)100\%$ confidence interval.
- Describes the “minimum desired precision” of the estimate \hat{p} .

Sample size determination

Design of experiments:

How large should the sample size n be in order to obtain the desired precision?

Method 7.13

If you want a Margin of Error, ME, with $(1 - \alpha)100\%$ confidence, then you need the following sample size:

$$n = p(1 - p) \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$$

Sample size determination

Method 7.13

If you want a Margin of Error, ME , with $(1 - \alpha)100\%$ confidence, and you do *not* have a reasonable guess of p , then you need the following sample size:

$$n = \frac{1}{4} \left(\frac{z_{1-\alpha/2}}{ME} \right)^2$$

since the worst case approach is given by: $p = \frac{1}{2}$

Example 1 - continued

Left-handedness:

Suppose that we want $ME = 0.01$ (with $\alpha = 0.05$) – what should n be?

Example 1 - continued

Left-handedness:

Suppose that we want $ME = 0.01$ (with $\alpha = 0.05$) – what should n be?

Assume $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01} \right)^2 = 3467.4 \approx 3468$$

Example 1 - continued

Left-handedness:

Suppose that we want $ME = 0.01$ (with $\alpha = 0.05$) – what should n be?

Assume $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01} \right)^2 = 3467.4 \approx 3468$$

Without any assumption on the size of p :

$$n = \frac{1}{4} \left(\frac{1.96}{0.01} \right)^2 = 9604$$

Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 Hypothesis test for several proportions
- 6 Analysis of contingency tables

Steps of a hypothesis test - an overview (repetition)

- ① Formulate the hypothesis and choose the level of significance α (i.e. the “risk-level”).
- ② Use the data to calculate the value of the test statistic.
- ③ Calculate the p-value using the test statistic and the relevant distribution. Compare the p -value to the significance level α and draw a conclusion.
- ④ (Alternatively, draw a conclusion based on the relevant critical value(s)).

Hypothesis test for one proportion

The null and alternative hypothesis for one proportion p :

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

We either accept H_0 or reject H_0 .

Testing the hypothesis: The test statistic

Theorem 7.10 and Method 7.11

If the sample size is sufficiently large (if $np_0 > 15$ and $n(1 - p_0) > 15$), we use the following test statistic:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

Under the null hypothesis, the random variable Z (approximately) follows a standard normal distribution, $Z \sim N(0, 1^2)$.

Testing the hypothesis: p -value and conclusion (Method 7.11)

Find the p -value (evidence against the null hypothesis):

- $2P(Z > |z_{\text{obs}}|)$

Test using the critical value:

Reject null hypothesis if $z_{\text{obs}} < -z_{1-\alpha/2}$ or $z_{\text{obs}} > z_{1-\alpha/2}$.

Example 1 - continued

Is half of the Danish population left handed?

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Example 1 - continued

Is half of the Danish population left handed?

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Test statistic:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1 - 0.5)}} = -8$$

Example 1 - continued

Is half of the Danish population left handed?

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5$$

Test statistic:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1 - 0.5)}} = -8$$

p-value:

$$2 \cdot P(Z > 8) = 1.2 \cdot 10^{-15}$$

There is very strong evidence against the null hypothesis - we reject it (with $\alpha = 0.05$).

Example 1 - continued

Testing the hypothesis in R

```
prop.test(10, 100, p = 0.5, correct = FALSE)

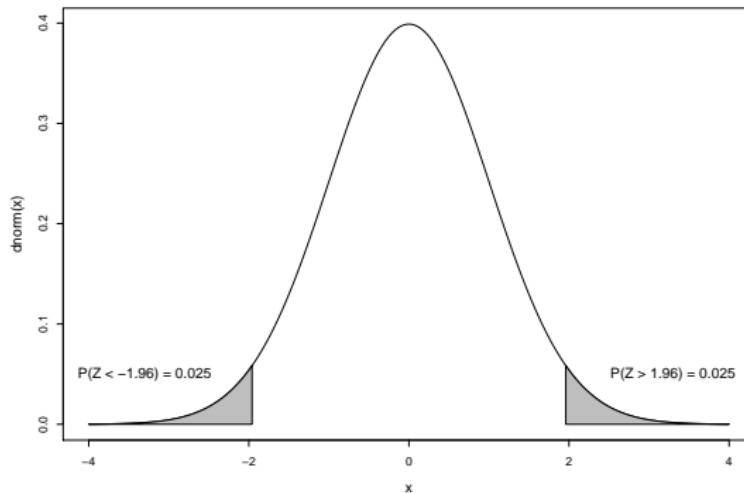
##
## 1-sample proportions test without continuity correction
##
## data: 10 out of 100, null probability 0.5
## X-squared = 64, df = 1, p-value = 1e-15
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
## 0.05523 0.17437
## sample estimates:
## p
## 0.1
```

Example 1 - continued

Using the critical value instead:

$$z_{0.975} = 1.96$$

As $z_{\text{obs}} = -8$ is (much) less than -1.96 we reject the hypothesis.



Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 Hypothesis test for several proportions
- 6 Analysis of contingency tables

Confidence interval for (the difference between) two proportions

Method 7.15

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$$

where

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Rule of thumb:

Both $n_i p_i \geq 10$ and $n_i(1 - p_i) \geq 10$ for $i = 1, 2$.

Hypothesis test for two proportions, Method 7.18

Two sample proportions hypothesis test

Comparing two proportions (shown here for a two-sided alternative)

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

The test statistic:

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

And for large samples:

Use the standard normal distribution again.

Example 2

Is there a relation between the use of birth control pills and the risk of a blood clot in the heart?

In a study (USA, 1975) the connection between birth control pills and the risk of a blood clot in the heart was investigated.

	Blood clot	No blood clot
B. C. pill	23	34
No B. C. pill	35	132

Carry out a test to check if there is any connection between the use of birth control pills and the risk of a blood clot in the heart. Use a significance level of $\alpha = 0.05$.

Example 2

In a study (USA, 1975) the connection between birth control pills and the risk of blood clot in the heart was investigated.

	Blood clot	No blood clot
B. C. pill	23	34
No B. C. pill	35	132

Estimates within each sample

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \quad \hat{p}_2 = \frac{35}{167} = 0.2096$$

Common estimate:

$$\hat{p} = \frac{23 + 35}{57 + 167} = \frac{58}{224} = 0.2589$$

Example 2 - continued

prop.test for equality of two proportions in R

```
# Read data table into R
pill.study <- matrix(c(23, 34, 35, 132),
                      ncol = 2, byrow = TRUE)
colnames(pill.study) <- c("Blood Clot", "No Clot")
rownames(pill.study) <- c("Pill", "No pill")
pill.study

# Test whether probabilities are equal for the two groups
prop.test(pill.study, correct = FALSE)
```

Example 2 - continued

prop.test for equality of two proportions in R

```
##          Blood Clot No Clot
## Pill          23      34
## No pill      35     132
```

```
##
## 2-sample test for equality of proportions without continuity
## correction
##
## data: pill.study
## X-squared = 8.3, df = 1, p-value = 0.004
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## 0.05239 0.33546
## sample estimates:
## prop 1 prop 2
## 0.4035 0.2096
```

Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 **Hypothesis test for several proportions**
- 6 Analysis of contingency tables

Hypothesis test for several proportions

The comparison of c proportions

In some cases, we might be interested in determining whether two or more binomial distributions have the same parameter p . That is, we are interested in testing the null hypothesis:

$$H_0 : p_1 = p_2 = \dots = p_c = p$$

vs. the alternative that at least two proportions are different.

Hypothesis test for several proportions

Table of observed counts for c samples:

	Sample 1	Sample 2	...	Sample c	Total
Success	x_1	x_2	...	x_c	x
Failure	$n_1 - x_1$	$n_2 - x_2$...	$n_c - x_c$	$n - x$
Total	n_1	n_2	...	n_c	n

Common (average) estimate:

Under the null hypothesis, the estimate of p is:

$$\hat{p} = \frac{x}{n}$$

Hypothesis test for several proportions

Common (average) estimate:

Under the null hypothesis the estimate of p is:

$$\hat{p} = \frac{x}{n}$$

"Use" this common estimate in each group:

If the null hypothesis is true, we expect that the j 'th group/sample has e_{1j} successes and e_{2j} failures, where

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

$$e_{2j} = n_j(1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}$$

Hypothesis test for several proportions

Make a table with the *expected* counts for the c samples:

e_{ij}	Sample 1	Sample 2	...	Sample c	Total
Success	e_{11}	e_{12}	...	e_{1c}	x
Failure	e_{21}	e_{22}	...	e_{2c}	$n - x$
Total	n_1	n_2	...	n_c	n

General way to find the expected counts in frequency tables:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Computation of the test statistic - Method 7.20

The test statistic becomes

$$\chi^2_{\text{obs}} = \sum_{i=1}^2 \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency in cell (i,j) and e_{ij} is the expected frequency in cell (i,j) .

Find the p -value or use the critical value - Method 7.20

Sampling distribution for test statistic under H_0 :
 χ^2 -distribution with $(c - 1)$ degrees of freedom (approx.)

Critical value method:

If $\chi^2_{\text{obs}} > \chi^2_{\alpha}(c - 1)$ the null hypothesis is rejected.

Rule of thumb for validity of the test:

All expected values $e_{ij} \geq 5$.

Example 2 - continued

The *observed* values o_{ij}

Observed	Blood clot	No Blood clot
B. C. pill	23	34
No B. C. pill	35	132

Example 2 - continued

Compute the *expected* values e_{ij}

Expected	Blood clot	No Blood clot	Total
B. C. pill			57
No B. C. pill			167
Total	58	166	224

Example 2 - continued

Use “the rule” for expected values four times, e.g.:

$$e_{22} = \frac{167 \cdot 166}{224} = 123.76$$

The *expected* values e_{ij} :

Expected	Blood clot	No Blood clot	Total
B. C. pill	14.76	42.24	57
No B. C. pill	43.24	123.76	167
Total	58	166	224

Example 2 - continued

The test statistic (remember to include all cells):

$$\chi^2_{\text{obs}} = \frac{(23 - 14.76)^2}{14.76} + \frac{(34 - 42.24)^2}{42.24} + \frac{(35 - 43.24)^2}{43.24} + \frac{(132 - 123.76)^2}{123.76}$$
$$= 8.33$$

Critical value:

```
qchisq(0.95, 1)
```

```
[1] 3.841
```

Conclusion:

We reject the null hypothesis - there *is* a significantly higher risk of blood clots in the birth control pill group.

Example 2 - continued

chisq.test for equality of two proportions in R

```
# Test whether probabilities are equal for the two groups
chisq.test(pill.study, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data: pill.study
## X-squared = 8.3, df = 1, p-value = 0.004

# Expected values
chisq.test(pill.study, correct = FALSE)$expected

##          Blood Clot No Clot
## Pill        14.76   42.24
## No pill    43.24  123.76
```

Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 Hypothesis test for several proportions
- 6 Analysis of contingency tables

Example 3: Analysis of contingency tables

A 3×3 table - 3 samples, 3-category outcomes

	4 weeks bef	2 weeks bef	1 week bef
Candidate I	79	91	93
Candidate II	84	66	60
Undecided	37	43	47
	$n_1 = 200$	$n_2 = 200$	$n_3 = 200$

Are the votes equally distributed?

$$H_0 : p_{i1} = p_{i2} = p_{i3}, \quad i = 1, 2, 3.$$

Analysis of contingency tables

A 3×3 table - 1 sample, two 3-category variables:

	bad	average	good
bad	23	60	29
average	28	79	60
good	9	49	63

Is there independence between the row and column variables?

$$H_0 : p_{ij} = p_{i \cdot} p_{\cdot j}$$

Computation of the test statistic – no matter the type of table 7.22

In a contingency table with r rows and c columns, the test statistic is:

$$\chi^2_{\text{obs}} = \sum_{i=1}^r \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed value in cell (i,j) and e_{ij} is the expected value in cell (i,j) .

General way to find the expected counts in frequency tables:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{\text{total}}$$

Find p -value or use critical value - Method 7.22

Sampling distribution for test-statistic under H_0 :

χ^2 -distribution with $(r - 1)(c - 1)$ degrees of freedom (approx.).

Critical value method:

If $\chi^2_{\text{obs}} > \chi^2_{\alpha}$ with $(r - 1)(c - 1)$ degrees of freedom, the null hypothesis is rejected.

Rule of thumb for validity of the test:

All expected values $e_{ij} \geq 5$.

Example 3 - continued

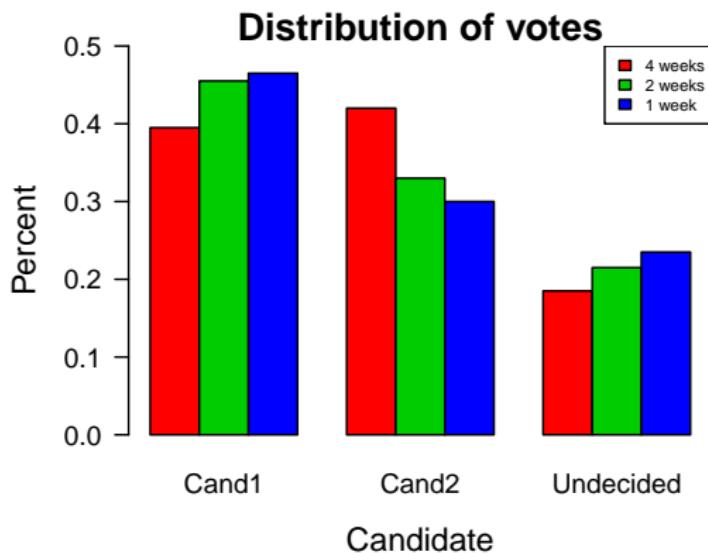
chisq.test for contingency tables

```
# Read data table into R
poll <- matrix(c(79, 91, 93, 84, 66, 60, 37, 43, 47),
                 ncol = 3, byrow = TRUE)
colnames(poll) <- c("4 weeks", "2 weeks", "1 week")
rownames(poll) <- c("Cand1", "Cand2", "Undecided")

# Show column percentages
prop.table(poll, 2)

##          4 weeks 2 weeks 1 week
## Cand1      0.395   0.455   0.465
## Cand2      0.420   0.330   0.300
## Undecided  0.185   0.215   0.235
```

Example 3 - continued



Example 3 - continued

```
# Testing for same distribution in the three populations
chisq.test(poll, correct = FALSE)

##
##  Pearson's Chi-squared test
##
## data: poll
## X-squared = 7, df = 4, p-value = 0.1

# Expected values
chisq.test(poll, correct = FALSE)$expected

##          4 weeks 2 weeks 1 week
## Cand1      87.67   87.67   87.67
## Cand2      70.00   70.00   70.00
## Undecided  42.33   42.33   42.33
```

Overview

- 1 Introduction
- 2 Confidence interval for one proportion
 - Sample size determination (planning)
- 3 Hypothesis test for one proportion
- 4 Confidence interval and hypothesis test for two proportions
- 5 Hypothesis test for several proportions
- 6 Analysis of contingency tables

Course 02402 Introduction to Statistics

Lecture 11: One-way Analysis of Variance, ANOVA

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

One-way ANOVA - simple example

Group A	Group B	Group C
2.8	5.5	5.8
3.6	6.3	8.3
3.4	6.1	6.9
2.3	5.7	6.1

Is there a difference (in means) between the groups A, B and C?

Analysis of variance (ANOVA) can be used for the analysis, if the observations in each group can be assumed to be normally distributed.

TV set development at Bang & Olufsen

Sound and image quality measured by the human perceptual instrument.



Bang & Olufsen data in R

```
# Get the B&O data from the lmerTest-package
library(lmerTest)
data(TVbo)
head(TVbo) # First rows of the data

# Define factor identifying the 12 TV set and picture combinations
TVbo$TVPic <- factor(TVbo$TVset:TVbo$Picture)

# Each of 8 assessors scored each of the 12 combinations twice.
# Average the two replicates for each assessor and combination of
# TV set and picture
library(doBy)
TVbonoise <- summaryBy(Noise ~ Assessor + TVPic, data = TVbo,
                        keep.names = T)

# One-way ANOVA of the noise (not the correct analysis!)
anova(lm(Noise ~ TVPic, data = TVbonoise))

# Two-way ANOVA of the noise (better analysis, week 12)
anova(lm(Noise ~ Assessor + TVPic, data = TVbonoise))
```

One-way ANOVA - simple example in R

```
# Input data
y <- c(2.8, 3.6, 3.4, 2.3,
      5.5, 6.3, 6.1, 5.7,
      5.8, 8.3, 6.9, 6.1)

## Define treatment groups
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

## Plot data by treatment groups
par(mfrow = c(1,2))
plot(y ~ as.numeric(treatm), xlab = "Treatment", ylab = "y")
boxplot(y ~ treatm, xlab = "Treatment", ylab = "y")
```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 **Model and hypothesis**
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

One-way ANOVA, model

- The model may be formulated as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where the ε_{ij} are assumed to be independent and identically distributed (i.i.d.) with

$$\varepsilon_{ij} \sim N(0, \sigma^2).$$

- μ : overall mean.
- α_i : effect of group (treatment) i .
- Y_{ij} : j th measurement in group i (j runs from 1 to n).

One-way ANOVA, hypothesis

- We want to compare the (more than 2) means $\mu + \alpha_i$ in the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

- The hypothesis may be formulated as

$$H_0: \quad \alpha_i = 0 \quad \text{for all } i$$

$$H_1: \quad \alpha_i \neq 0 \quad \text{for at least one } i$$

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

One-way ANOVA, decomposition and the ANOVA table

- With the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

the total variation in the data can be decomposed:

$$SST = SS(Tr) + SSE.$$

- 'One-way' refers to the fact that there is only one factor in the experiment on k levels.
- The method is called analysis of variance, because the testing is carried out by comparing certain variances.

Formulas for sums of squares

- Total sum of squares ("the total variance")

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

Formulas for sums of squares

- Total sum of squares ("the total variance")

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- The sum of squares for the residuals ("residual variance after model fit")

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Formulas for sums of squares

- Total sum of squares ("the total variance")

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

- The sum of squares for the residuals ("residual variance after model fit")

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- Sum of squares of treatment ("variance explained by the model")

$$SS(Tr) = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$$

The ANOVA table

<i>Source of variation</i>	Deg. of freedom	Sums of squares	Mean sum of squares
<i>Treatment</i>	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$
<i>Residual</i>	$n - k$	SSE	$MSE = \frac{SSE}{n-k}$
<i>Total</i>	$n - 1$	SST	

```
# One-way ANOVA using anova() and lm()
anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## treatm     2   30.8   15.40    26.7 0.00017 ***
## Residuals  9    5.2    0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 **Hypothesis test (F-test)**
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

One-way ANOVA, F-test

- We have: (Theorem 8.2)

$$SST = SS(Tr) + SSE$$

- and we can find the test statistic

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)} = \frac{MS(Tr)}{MSE}$$

where

- k is the number of levels of the factor,
- n is the total number of observations.
- Choose the significance level α , and compute the test statistic F .
- Compare the test statistic to the relevant quantile of the F -distribution:

$$F \sim F_{\alpha}(k-1, n-k) \text{ (Theorem 8.6)}$$

The F -distribution and the F -test

```
# Remember, this is "under  $H_0$ " (i.e. we compute as if  $H_0$  is true)

# Number of groups
k <- 3

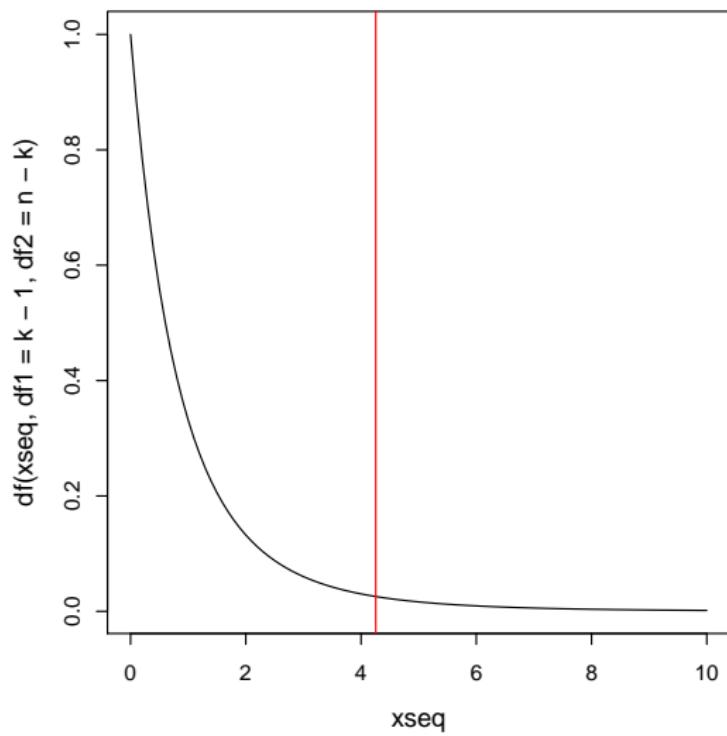
# Total number of observations
n <- 12

# Sequence for plot
xseq <- seq(0, 10, by = 0.1)

# Plot density of the  $F$ -distribution
plot(xseq, df(xseq, df1 = k-1, df2 = n-k), type = "l")

# Plot critical value for significance level 5%
cr <- qf(0.95, df1 = k-1, df2 = n-k)
abline(v = cr, col = "red")
```

An F-distribution with a critical value



The ANOVA table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p -value
<i>treatment</i>	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
<i>Residual</i>	$n - k$	SSE	$MSE = \frac{SSE}{n-k}$		
<i>Total</i>	$n - 1$	SST			

```
anova(lm(y ~ treatm))

## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## treatm      2  30.8   15.40   26.7 0.00017 ***
## Residuals   9   5.2    0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-way ANOVA F-test “by hand”

```
k <- 3; n <- 12 # Number of groups k, total number of observations n

# Total variation, SST
(SST <- sum( (y - mean(y))^2 ))

# Residual variance after model fit, SSE
y1 <- y[1:4]; y2 <- y[5:8]; y3 <- y[9:12]

(SSE <- sum( (y1 - mean(y1))^2 ) +
  sum( (y2 - mean(y2))^2 ) +
  sum( (y3 - mean(y3))^2 ))

# Variance explained by the model, SS(Tr)
(SSTr <- SST - SSE)

# Test statistic
(Fobs <- (SSTr/(k-1)) / (SSE/(n-k)))

# P-value
(1 - pf(Fobs, df1 = k-1, df2 = n-k))
```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

Within-group variability and relation to the 2-sample t-test (Theorem 8.4)

The residual sum of squares, SSE , divided by $n - k$, also called residual mean square, $MSE = SSE/(n - k)$, is the average within-group variability:

$$MSE = \frac{SSE}{n - k} = \frac{(n_1 - 1)s_1^2 + \cdots + (n_k - 1)s_k^2}{n - k} \quad (1)$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

ONLY when $k = 2$: (cf. Method 3.52)

$$MSE = s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2}$$

$$F_{\text{obs}} = t_{\text{obs}}^2$$

where t_{obs} is the pooled t-test statistic from Methods 3.52 and 3.53.

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis**
- 7 Model control / model validation
- 8 A complete example - from the book

Post hoc confidence interval - Method 8.9

- A single pre-planned confidence interval for the difference between treatment i and j is found as:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{n-k} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (2)$$

where $t_{1-\alpha/2}$ is based on the t-distribution with $n - k$ degrees of freedom.

- Note the fewer degrees of freedom as more unknowns are estimated in the computation of $MSE = SSE/(n - k) = s_p^2$ (i.e. pooled variance estimate)
- If all $M = k(k - 1)/2$ combinations of pairwise confidence intervals are found use the formula M times, but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$.

Post hoc pairwise hypothesis test- Method 8.10

- A single pre-planned level α hypothesis test:

$$H_0 : \mu_i = \mu_j, \quad H_1 : \mu_i \neq \mu_j$$

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad (3)$$

and

$$p\text{-value} = 2P(t > |t_{\text{obs}}|)$$

where the t -distribution with $n - k$ degrees of freedom is used.

- If all $M = k(k - 1)/2$ combinations of pairwise hypothesis tests are carried out use the approach M times, but each time with significance level $\alpha_{\text{Bonferroni}} = \alpha/M$.

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

Variance homogeneity

Look at a box plot to check whether the variability seems different across the groups.

```
# Check assumption of homogeneous variance using, e.g.,  
# a box plot.  
plot(treatm, y)
```

Normal assumption

Look at a normal QQ-plot of the residuals

```
# Check normality of residuals using a normal QQ-plot
fit1 <- lm(y ~ treatm)
qqnorm(fit1$residuals)
qqline(fit1$residuals)
```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

A complete example - from the book

Introduction to Statistics

Agendas eNotes Course Material Podcast Forum Quiz Admin

Dokumentegenkskaper...

8.2.5 A complete worked through example: plastic types for lamps

||| Example 8.17 Plastic types for lamps

On a lamp two plastic screens are to be mounted. It is essential that these plastic screens have a good impact strength. Therefore an experiment is carried out for 5 different types of plastic. 6 samples in each plastic type are tested. The strengths of these items are determined. The following measurement data was found (strength in kJ/m^2):

Type of plastic				
I	II	III	IV	V
44.6	52.8	53.1	51.5	48.2
50.5	58.3	50.0	53.7	40.8
46.3	55.4	54.4	50.5	44.5
48.5	57.4	55.3	54.4	43.9
45.2	58.1	50.6	47.5	45.9
52.3	54.6	53.4	47.8	42.5

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model and hypothesis
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Within-group variability and relation to the 2-sample t-test
- 6 Post hoc analysis
- 7 Model control / model validation
- 8 A complete example - from the book

Course 02402 Introduction to Statistics

Lecture 12: Two-way Analysis of Variance, ANOVA

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book

TV set development at Bang & Olufsen

Sound and image quality is measured by the human perceptual instrument:



Bang & Olufsen data in R

```
# Get the B&O data from the lmerTest-package
library(lmerTest)
data(TVbo)

# Each of 8 assessors scored each of 12 combinations 2 times.
# Take a look at the sharpness scores for one single picture
# and one of the two repetitions
TVbo_sub <- subset(TVbo, Picture == 1 & Repeat == 1)[, c(1, 2, 9)]
sharp <- matrix(TVbo_sub$Sharpness, nrow = 8, byrow = T)
colnames(sharp) <- c("TV3", "TV2", "TV1")
rownames(sharp) <- c("Person 1", "Person 2", "Person 3",
                      "Person 4", "Person 5", "Person 6",
                      "Person 7", "Person 8")
library(xtable)
xtable(sharp)
```

Bang & Olufsen data in R

	TV3	TV2	TV1
Person 1	9.30	4.70	6.60
Person 2	10.20	7.00	8.80
Person 3	11.50	9.50	8.00
Person 4	11.90	6.60	8.20
Person 5	10.70	4.20	5.40
Person 6	10.90	9.10	7.10
Person 7	8.50	5.00	6.30
Person 8	12.60	8.90	10.70

Two-way ANOVA - example

- Same data as for one-way, but now we know that the experiment was split into blocks:

	Group A	Group B	Group C
Block 1	2.8	5.5	5.8
Block 2	3.6	6.3	8.3
Block 3	3.4	6.1	6.9
Block 4	2.3	5.7	6.1

- Hence three *groups* on four *blocks*,
- or three *treatments* on four *persons*,
- or three *varieties* on four *fields* (hence blocks),
- or something similar.

Two-way ANOVA - example

- Same data as for one-way, but now we know that the experiment was split into blocks:

	Group A	Group B	Group C
Block 1	2.8	5.5	5.8
Block 2	3.6	6.3	8.3
Block 3	3.4	6.1	6.9
Block 4	2.3	5.7	6.1

- Hence three *groups* on four *blocks*,
- or three *treatments* on four *persons*,
- or three *varieties* on four *fields* (hence blocks),
- or something similar.
- *One-way* vs. *two-way* ANOVA
- *Completely randomized design* vs. *Randomized block design*

Two-way ANOVA - example

- Same data as for one-way, but now we know that the experiment was split into blocks:

	Group A	Group B	Group C
Block 1	2.8	5.5	5.8
Block 2	3.6	6.3	8.3
Block 3	3.4	6.1	6.9
Block 4	2.3	5.7	6.1

Two-way ANOVA - example

- Same data as for one-way, but now we know that the experiment was split into blocks:

	Group A	Group B	Group C
Block 1	2.8	5.5	5.8
Block 2	3.6	6.3	8.3
Block 3	3.4	6.1	6.9
Block 4	2.3	5.7	6.1

- Question: Is there a significant difference (in means) between the groups A, B and C?
- ANOVA can be used if the observations in each group are (approximately) normal distributed or if the n_i s are large enough (CLT).

The toy data in R

```
# Observations
y <- c(2.8, 3.6, 3.4, 2.3,
      5.5, 6.3, 6.1, 5.7,
      5.8, 8.3, 6.9, 6.1)

# Treatments (groups, varieties)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

# Blocks (persons, fields)
block <- factor(c(1, 2, 3, 4,
                  1, 2, 3, 4,
                  1, 2, 3, 4))

# No. of treatments and no. of blocks (for later formulas)
(k <- length(unique(treatm)))
(l <- length(unique(block)))

# Box plots by treatment
plot(treatm, y, xlab = "Treatment", ylab = "y")

# Box plots by block
plot(block, y, xlab = "Block", ylab="y")
```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 **Model**
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book

Two-way ANOVA, model

- The model may be formulated as

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

where the errors are i.i.d. with

$$\varepsilon_{ij} \sim N(0, \sigma^2).$$

- μ is the overall mean
- α_i is the effect of treatment i
- β_j is the level for block j
- There are k treatments and l blocks

Estimates of parameters in the model

- We can compute the estimates of the parameters ($\hat{\mu}$, $\hat{\alpha}_i$, and $\hat{\beta}_j$)

$$\hat{\mu} = \bar{y} = \frac{1}{k \cdot l} \sum_{i=1}^k \sum_{j=1}^l y_{ij}$$

$$\hat{\alpha}_i = \left(\frac{1}{k} \sum_{j=1}^k y_{ij} \right) - \hat{\mu}$$

$$\hat{\beta}_j = \left(\frac{1}{l} \sum_{i=1}^l y_{ij} \right) - \hat{\mu}$$

```

# Sample mean
(mu_hat <- mean(y))

# Sample mean deviation for each treatment
(alpha_hat <- tapply(y, treatm, mean) - mu_hat)

# Sample mean deviation for each block
(beta_hat <- tapply(y, block, mean) - mu_hat)

```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book

Two-way ANOVA, decomposition and the ANOVA table, Theorem 8.20

- With the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

the total variation in the data can be decomposed:

$$SST = SS(Tr) + SS(Bl) + SSE$$

- 'Two-way' refers to the fact that there are two factors (grouping variables) in the experiment.
- The method is called analysis of variance, because hypothesis testing is carried out by comparing certain variances.

Formulas for sums of squares

- Total sum of squares (or “the total variance”, same as for one-way)

$$SST = \sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\mu})^2$$

Formulas for sums of squares

- Total sum of squares (or “the total variance”, same as for one-way)

$$SST = \sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\mu})^2$$

- Treatment sum of squares (or “variance explained by the treatment part of the model”)

$$SS(Tr) = l \cdot \sum_{i=1}^k \hat{\alpha}_i^2$$

Formulas for sums of squares

- Sum of squares for blocks/persons ("variance explained by the block part of the model")

$$SS(Bl) = k \cdot \sum_{j=1}^l \hat{\beta}_j^2$$

- Sum of squares for the residuals ("residual variance after model fit")

$$SSE = \sum_{i=1}^k \sum_{j=1}^l (y_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu})^2$$

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book

Two-way ANOVA: Hypothesis of no effect of treatment, Theorem 8.22

- We want to compare (more than 2) means $\mu + \alpha_i$ in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- The hypothesis of no difference between treatment means may be formulated as

$$H_{0,Tr} : \alpha_i = 0 \quad \text{for all } i$$

$$H_{1,Tr} : \alpha_i \neq 0 \quad \text{for at least one } i$$

Two-way ANOVA: Hypothesis of no effect of treatment, Theorem 8.22

- We want to compare (more than 2) means $\mu + \alpha_i$ in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- The hypothesis of no difference between treatment means may be formulated as

$$H_{0,Tr} : \alpha_i = 0 \quad \text{for all } i$$

$$H_{1,Tr} : \alpha_i \neq 0 \quad \text{for at least one } i$$

- Under $H_{0,Tr}$ the following is true:

$$F_{Tr} = \frac{SS(Tr)/(k-1)}{SSE/((k-1)(l-1))}$$

is F -distributed with $k-1$ and $(k-1)(l-1)$ degrees of freedom.

Two-way ANOVA: hypothesis of no effect of blocks/persons, Theorem 8.22

- We want to compare (more than 2) means $\mu + \beta_j$ in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- The hypothesis of no difference between block means may be formulated as

$$H_{0,Bl} : \beta_j = 0 \quad \text{for all } j$$

$$H_{1,Bl} : \beta_j \neq 0 \quad \text{for at least one } j$$

Two-way ANOVA: hypothesis of no effect of blocks/persons, Theorem 8.22

- We want to compare (more than 2) means $\mu + \beta_j$ in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

- The hypothesis of no difference between block means may be formulated as

$$H_{0,Bl} : \beta_j = 0 \quad \text{for all } j$$

$$H_{1,Bl} : \beta_j \neq 0 \quad \text{for at least one } j$$

- Under $H_{0,Bl}$ the following is true:

$$F_{Bl} = \frac{SS(Bl)/(l-1)}{SSE/((k-1)(l-1))}$$

follows an F -distribution with $l-1$ and $(k-1)(l-1)$ degrees of freedom.

F-distribution and treatment hypothesis

```
# Plot density of relevant F-distribution. Remember that this is "under H0"
# (computed as if H0 were true)
xseq <- seq(0, 10, by = 0.1)
plot(xseq, df(xseq, df1 = k-1, df2 = (k-1)*(l-1)), type = "l")

# Show critical value (5% signif. level) for test of treatment hypothesis
critical_value <- qf(0.95, df1 = k-1, df2 = (k-1)*(l-1))
abline(v = critical_value, col = "red")

# Compute value of the test statistic
(FTr <- (SSTr/(k-1)) / (SSE/((k-1)*(l-1)))))

# Compute p-value for the test
1 - pf(FTr, df1 = k-1, df2 = (k-1)*(l-1))
```

F-distribution and block hypothesis

```
# Plot density of relevant F-distribution. Remember that this is "under H0"
# (computed as if H0 were true)
xseq <- seq(0, 10, by = 0.1)
plot(xseq, df(xseq, df1 = l-1, df2 = (k-1)*(l-1)), type = "l")

# Show critical value (5% signif. level) for test of treatment hypothesis
critical_value <- qf(0.95, df1 = l-1, df2 = (k-1)*(l-1))
abline(v = critical_value, col = "red")

# Compute value of the test statistic
(FB1 <- (SSB1/(l-1)) / (SSE/((k-1)*(l-1)))))

# Compute p-value for the test
1 - pf(FB1, df1 = l-1, df2 = (k-1)*(l-1))
```

The two-way ANOVA table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p -value
Treatment	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{Tr} = \frac{MS(Tr)}{MSE}$	$P(F > F_{Tr})$
Block	$l - 1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{Bl} = \frac{MS(Bl)}{MSE}$	$P(F > F_{Bl})$
Residual	$(k - 1)(l - 1)$	SSE	$MSE = \frac{SSE}{(k-1)(l-1)}$		
Total	$n - 1$	SST			

```
anova(lm(y ~ treatm + block))

## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## treatm      2  30.79   15.40   74.40 5.8e-05 ***
## block       3   3.95    1.32    6.37   0.027 *
## Residuals   6   1.24    0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book

Post hoc confidence interval

- Like for one-way ANOVA (use methods 8.9 and 8.10) but substitute $n - k$ degrees of freedom with $(k - 1)(l - 1)$ (and use MSE from the two-way ANOVA).
- Can be done with either treatments or blocks.

Post hoc confidence interval

- Like for one-way ANOVA (use methods 8.9 and 8.10) but substitute $n - k$ degrees of freedom with $(k - 1)(l - 1)$ (and use MSE from the two-way ANOVA).
- Can be done with either treatments or blocks.
- A single pre-planned CI for the difference between treatment i and j :

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{(k-1)(l-1)} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (1)$$

where $t_{1-\alpha/2}$ is based on the t-distribution with $(k - 1)(l - 1)$ degrees of freedom.

Post hoc confidence interval

- Like for one-way ANOVA (use methods 8.9 and 8.10) but substitute $n - k$ degrees of freedom with $(k - 1)(l - 1)$ (and use MSE from the two-way ANOVA).
- Can be done with either treatments or blocks.
- A single pre-planned CI for the difference between treatment i and j :

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{(k-1)(l-1)} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \quad (1)$$

where $t_{1-\alpha/2}$ is based on the t-distribution with $(k - 1)(l - 1)$ degrees of freedom.

- If all $M = k(k - 1)/2$ combinations of pairwise confidence intervals are found use the formula M times but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$.

Post hoc pairwise hypothesis test

- A single pre-planned level α hypothesis tests:

$$H_0: \mu_i = \mu_j, \quad H_1: \mu_i \neq \mu_j$$

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad (2)$$

and:

$$p\text{-value} = 2P(t > |t_{\text{obs}}|)$$

where the t -distribution with $(k-1)(l-1)$ degrees of freedom is used.

Post hoc pairwise hypothesis test

- A single pre-planned level α hypothesis tests:

$$H_0: \mu_i = \mu_j, \quad H_1: \mu_i \neq \mu_j$$

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \quad (2)$$

and:

$$p\text{-value} = 2P(t > |t_{\text{obs}}|)$$

where the t -distribution with $(k-1)(l-1)$ degrees of freedom is used.

- If all $M = k(k-1)/2$ combinations of pairwise confidence intervals are found use the formula M times but each time with $\alpha_{\text{Bonferroni}} = \alpha/M$.

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 **Model control / model validation**
- 7 A complete example - from the book

Model validation: Variance homogeneity

Make box plots of the *residuals* to check whether the variability seems different across the groups.

```
# Save the fitted model
fit <- lm(y ~ treatm + block)

# Make box plots of residuals
par(mfrow = c(1,2))
plot(treatm, fit$residuals, xlab = "Treatment")
plot(block, fit$residuals, xlab = "Block")
```

Model validation: Normality

Make a normal QQ-plot to check whether the distribution of the residuals seems normal.

```
# Normal QQ-plot of the residuals
qqnorm(fit$residuals)
qqline(fit$residuals)
```

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book

A complete example - from the book

The screenshot shows a navigation bar with the DTU logo, course title, and links for Agendas, eNotes, Course Material, Podcast, Forum, Quiz, Admin, perbb, and Logout. Below the bar, a breadcrumb trail shows '0.3.3 A complete worked through example: Car tires'. The main content area is titled 'Example 8.26 Car tires'.

Example 8.26 Car tires

In a study of 3 different types of tires ("treatment") effect on the fuel economy, drives of 1000 km in 4 different cars ("blocks") were carried out. The results are listed in the following table in km/l.

	Car 1	Car 2	Car 3	Car 4	Mean
Tire 1	22.5	24.3	24.9	22.4	22.525
Tire 2	21.5	21.3	23.9	18.4	21.275
Tire 3	22.2	21.9	21.7	17.9	20.925
Mean	21.400	22.167	23.167	19.567	21.575

Let us analyse these data with a two-way ANOVA model, but first some explorative plotting:

http://onlinestatbook.com/statistics/analysis_of_variance.html

Overview

- 1 Intro: Small example and TV-data from B&O
- 2 Model
- 3 Computation - decomposition and the ANOVA table
- 4 Hypothesis test (F-test)
- 5 Post hoc analysis
- 6 Model control / model validation
- 7 A complete example - from the book