Course 02402 Introduction to Statistics Lecture 1:

Introduction to Statistics

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

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, quantiles
   - Covariance and correlation
5. Software: R

Oversigt

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, quantiles
   - Covariance and correlation
5. Software: R

Practical course information

- Teaching module: Tuesdays 13-17
- Generic weekly agenda:
  - BEFORE teaching module: Read announced stuff
  - 2 hours long lectures (curriculum of the week)
  - 2 hours of exercises (Mix of: Enote and online quiz-questions)
  - AFTER teaching module: Test yourself by online exam quiz.
- Exam: 4 hour multiple choice, Sunday 28/05
- MANDATORY projects: 2 must be approved to be able to go to the exam.
  - Each project will have 4 optional versions!

Practical Information

- Homepage: 02402.compute.dtu.dk
  - Online book (website or via lix)
  - Syllabus, Lecture plan
  - Exercises & solutions
  - Slides
  - Podcasts of lectures (In English AND Danish)
  - Quizzes
- Campusnet: www.campusnet.dtu.dk
  - Messages and (certain) file sharings
  - Links to interesting stories
  - Projects - description AND submission
- The lix eBook portal
  - eBook portal with marking and annotation features.

Oversigt

Introduction to Statistics - a primer

New England Journal of medicine:

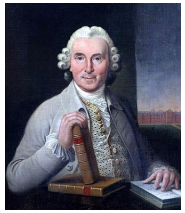
EDITORIAL: Looking Back on the Millennium in Medicine, *N Engl J Med*, 342:42-49, January 6, 2000.
`http://www.nejm.org/doi/full/10.1056/NEJM200001063420108`

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)`).

Google - *Big Data*

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

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.

(And Politiken, 12/2 2014 - see links in CampusNet)

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.

Oversigt

Statistics and Engineers

- How to treat (or analyse) data?
- What is random variation?
- Statistics is an important tool in problem solving
- Data analysis
- Quality improvement
- Design of experiments
- Predictions of future values
- .. and much more!

Statistics at DTU (mostly Compute)

- Energy Systems:
  - Prognoses of sun and wind power
  - Optimization of storage of energy e.g. in buildings
  - Modelling of human behaviour waste water treatment plant
- Control:
  - Robot navigation
  - Mechanical systems (e.g. cars, ships, wind turbines, etc)
- Medicine, Food and Pharma (Compute):
  - Statistics of clinical trials
  - Artificial pancreas
  - Human perceptual data in industrial product development
  - Pharmakokinetic and dynamic modelling
- Image Analysis
  - Image data is used more and more
  - X-rays, scannings, satelite photos, etc – videos

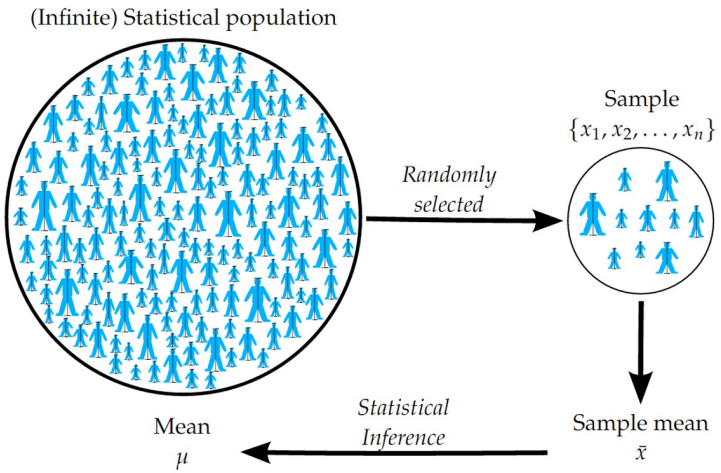
Statistics at DTU (mostly Compute)

- Signal processing:
  - Electrical systemes (filters, amplifiers, ...)
- Computer science:
  - Internet data (trafic, Google, Facebook, etc.)
  - Text recognition and mining
  - Security: Server attacks etc.
  - Software testing
- Civil Engineering
  - Tests of material properties and constructions
  - Production methods, e.g. casting of concrete
  - Energy systemes and indoor climate testing
- Management:
  - Financal Engineering, questionnaire surveys, ...
- Chemistry, Physics, Environment, Food, Vet, Aqua, etc

Statistics

- Statistics is often about analyzing a *sample*, that is taken from a *population*
- Based on the sample, we try to generalize (or comment on) the population
- Therefore it is important that the sample is *representative* of the population

Statistics



Oversigt

Summary statistics

We use a number of summary statistics to summarize and describe data (stochastic variables)

- Mean $\bar{x}$ and Median
- Variance s^2 and Standard deviation s
- Percentiles/quantiles
- Covariance and Correlation

Mean, Definition 1.4

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

- The mean:

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

We say that $\bar{x}$ is an *estimate* of the mean value

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

- Median:

  The observation in the middle (in sorted order)

Example: Student heights:

- Sample:

```
x <- c(185, 184, 194, 180, 182)
```

  n=5

- **mean**:

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

- **Median**, first order data: 180 182 184 185 194

  And then chosse the middle (since n is uneven)(3'th) number: 184

- What if a person on 235cm is added to the data:

  Median = 184

  Mean = 193

Variance and standard deviation, Definition 1.10

The variance (or the standard deviation) indicates the 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:

- **Data** n=5:
  185 184 194 180 182
- **Varians**, $s^2 =$

$$\frac{1}{4}((185-185)^2 + (184-185)^2 + (194-185)^2 + (180-185)^2 + (182-185)^2) = 29$$

- **Standard deviation**, $s = \sqrt{s^2} =$

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

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, the coefficient of variation:

$$V = \frac{s}{\bar{x}} \cdot 100$$

Percentiles, quantiles

The median it the point that divides the data into two halves. It is of course possible to find other points that divide the data in other parts, they are called percentiles.

Often calculated percentiles are

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

Note: the 50% percentile is the median

Quantiles, Definition 1.7

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

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

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

4. If pn is a non-integer, take the "next one" in the ordered list:

$$\text{The } p\text{'th quantile } = x_{(\lceil np \rceil)} \tag{2}$$

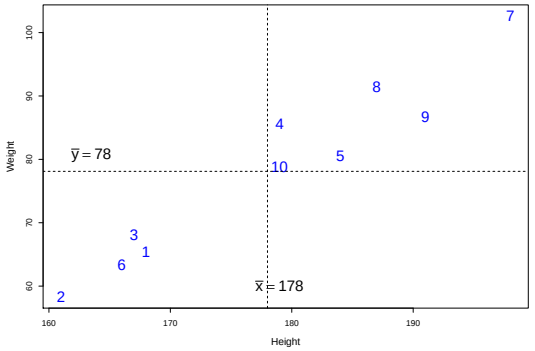
where $\lceil np \rceil$ is the *ceiling* of np, that is, the smallest integer larger than np.

Example: Student heights:

- **Data,** n=5:
  185 184 194 180 182
- **Lower quartile**, Q_1, first order the data: 180 182 184 185 194
  Then choose the right based on $np = 1.25$:
  $Q_1 = 182$
- **Upper quartile**, Q_3, first order the data: 180 182 184 185 194
  Then choose the right based on $np = 3.75$:
  $Q_3 = 185$

Covariance and correlation - measuring relation

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



Covariance and correlation, Definitions 1.18 and 1.19

The sample covariance is given by

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

The sample correlation coefficient is given by

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{s_{xy}}{s_x \cdot s_y} \tag{4}$$

where s_x and s_y is the sample standard deviation for x and y respectively.

Covariance and correlation - measuring relation

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

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

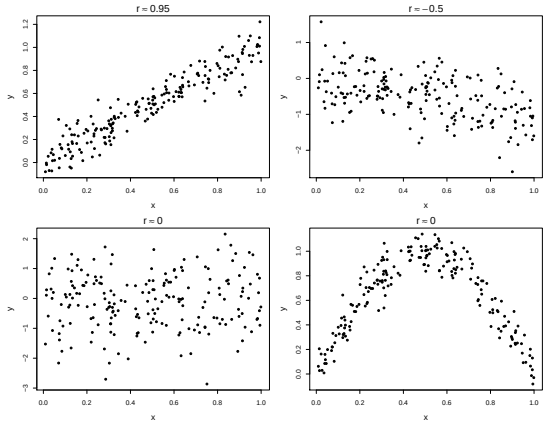
$$s_x = 12.21, \quad \text{and} \quad s_y = 14.07$$

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

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

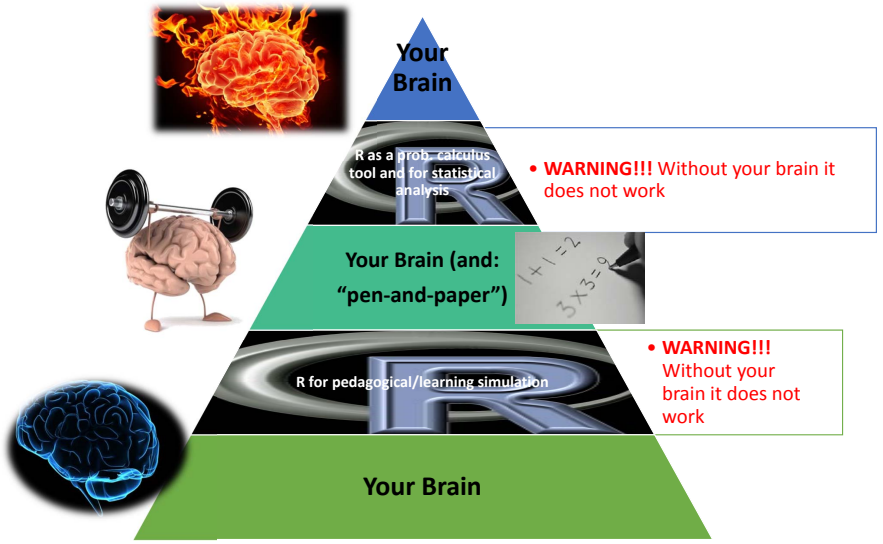
- Quantitative data:
  - Scatter plot (xy plot)
  - Histogram
  - Cumulative distribution
  - Boxplots
- Count data:
  - Bar charts
  - Pie charts

Oversigt

Software: R

- Install R and Rstudio
- Intro to basic computing
- Introduced in the eNote
- We use in an integrated way throughout the course and material
- Globalt rapidly growing open source computing environment
- WAARRRNIING: R CANNOT substitute our brains!!!! (Note Section 1.5)

Use your Brain!!!



Software: R

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

```
[1] 5
```

```
> y <- 3
```

```
> x <- c(1, 4, 6, 2)
> x
```

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

```
> x <- 1:10
> x
```

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

Software: R

```
## Sample Mean and Median (data from eNote)
x <- c(168,161,167,179,184,166,198,187,191,179)
mean(x)
```

[1] 178

```
median(x)
```

[1] 179

```
## Sample variance and standard deviation
var(x)
```

[1] 149

```
sd(x)
```

[1] 12

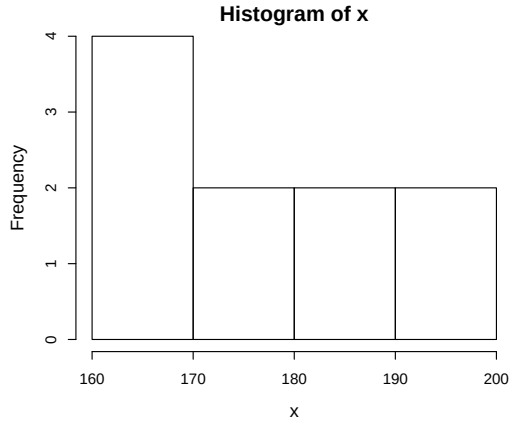
Software: R

```
## 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
```

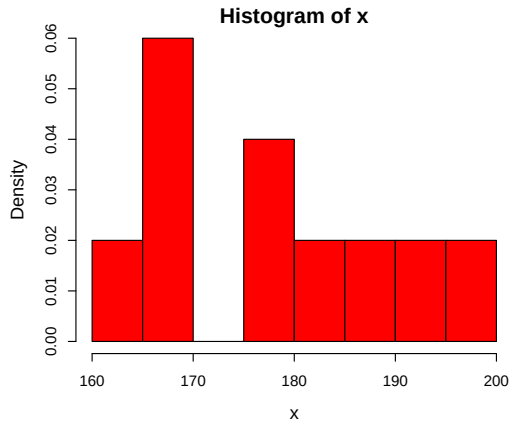

Software: R

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



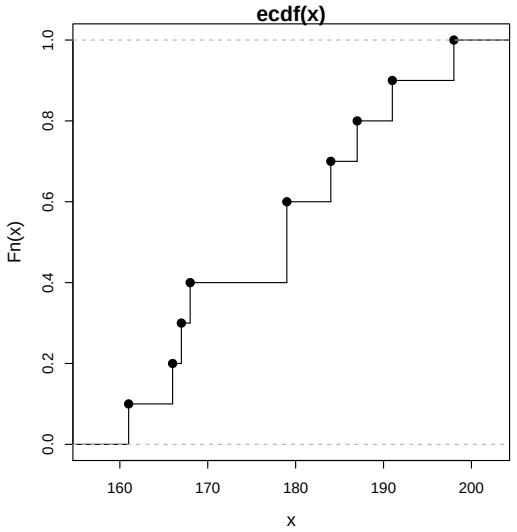
Software: R

```
## A density histogram of the heights:
hist(x,freq=FALSE,col="red",nclass=8)
```



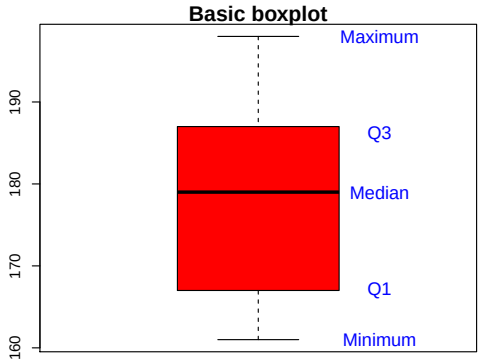
Software: R

```
plot(ecdf(x),verticals=TRUE)
```



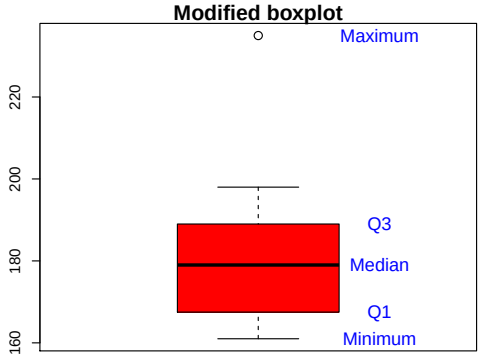
Software: R

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

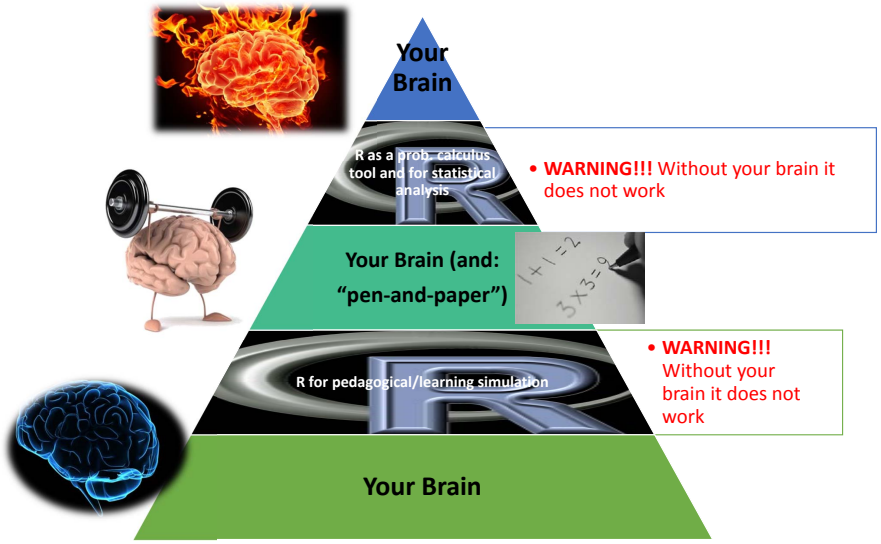


Software: R

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



Use your Brain!!!



Next week:

- Probability, part 1 - eNote 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, quantiles
   - Covariance and correlation
5. Software: R

Course 02402 Introduction to Statistics Lecture 2:

Random variables and discrete distributions

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

Oversigt

Random Variables

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

- A dice throw
- The number six'es in 10 dice throws
- km/l for for a car
- Measurement of glucose level in blood sample
- ...

Discrete or continuous

- We distinguish between discrete and continuous
- Discrete are countable:
  - How many use glassse in this room
  - The number of plains departing the next hour
- Kontinuert:
  - Wind speed measurement
  - Transport time to DTU

Today: discrete. Next week: Continuous

Random variable

Before the experiment is carried out, random variable:

$$X \text{ (or } X_1, \ldots, X_n)$$

indicated with capital letters
Then the experiment is carried out, and then we have a *realization* or *observation*

$$x \text{ (or } x_1, \ldots, x_n)$$

indicated with small letters

Simulate rolling a dice

Make a random draw from (1,2,3,4,5,6) with equal probability for each outcome

Discrete distributions

- The concept is to described the experiment before it is carried out
- What to do when we do not know the outcome?
- Solution: us the density function

Density function

A random variable has a *density function* (probability density function (pdf))

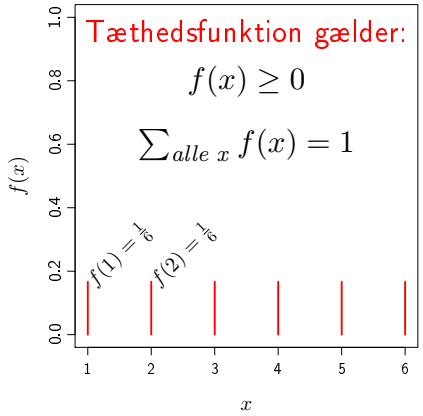
Definition

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

The probability that the X becomes x when the experiment is carried out

Density function

A fair dice density function



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*

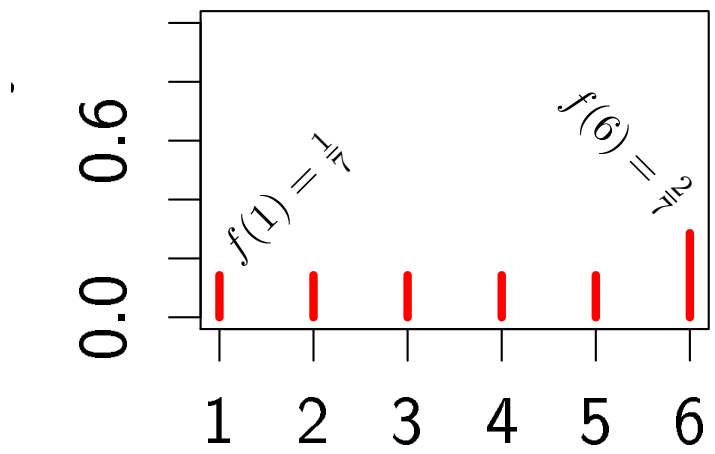
$$\{x_1, x_2, ..., x_n\}$$

and we can begin to "see" the distribution.

Simulate n rolls with a fair dice

```
n <- 30

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

## Print the values
xFair

## Count the 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")

## Add the 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","pdf"), lty=1, col=c(1,2),
       lwd=c(5,2), cex=0.8)
```

An unfair dice density function



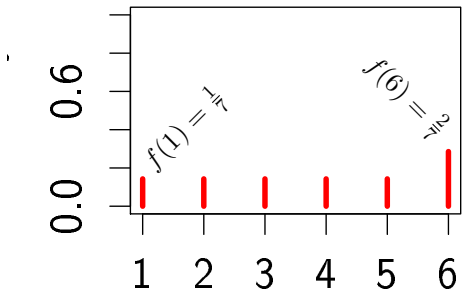
Simulate n rolls with an unfair dice

```r
## Number of simulated realizations
n <- 30

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

## Plot the empirical density function
plot(table(xUnfair)/n, lwd=10, ylim=c(0,1), xlab="x", ylab="Density")

## Add the pdf to the plot
lines(c(rep(1/7,5),2/7), lwd=4, type="h", col=2)

## Add a legend
legend("topright", c("Empirical pdf","pdf"), lty=1, col=c(1,2), lwd=c(5,2))
```


Some questions

Find some probabilities for X^{unFair}:

- The probability to get a 4?
- The probability to get a 5 or a 6?
- The probability to get less than 3?



Oversigt

Distribution function or cumulative density function (cdf))

Definition

The distribution function(cdf) is the cumulated density function:

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

Fair dice example

Let X represent one throw with a fair dice
Find the probability to get below 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 to above or equal to 3:

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

Oversigt

Specific discrete distributions

- A number of statistical distributions exists that can be used to describe and analyse different kind of problems
- Today we consider discrete distributions:
  - The binomial distribution
  - The hypergeometric distribution
  - The Poisson distribution

The Binomial distribution

- An experiment with two outcomes (succes or failure) is repeated
- X is the number of successes after n repeats
- So X follows a binomial distribution

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

- n number of repeats
- p the probability of success in each repeat

The density function for 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

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

Binomial distribution simulation

```
## Probability  of success
p <- 0.1
## Number of repeats
nRepeat <- 30
## Simulate Bernoulli experiment nRepeat times
tmp <- sample(c(0,1), size=nRepeat, prob=c(1-p,p), replace=TRUE)
## x is now
sum(tmp)

## Make similar with binomial distribution simulation function
rbinom(1, size=30, prob=p)

################
## Fair dice example

## Number of simulated realizations
n <- 30
## Sample independent from the set (1,2,3,4,5,6) with same probabilities
xFair <- sample(1:6, size=n, replace=TRUE)
## Count the number of 6'es
sum(xFair == 6)

## Make similar with rbinom()
rbinom(n=1, size=30, prob=1/6)
```

Example 1

In a call center in a phone company the costumer satisfaction is an issue. It is especially important that when errors/faults occur, then they are corrected within the same day.
Assume that the probability of an error being corrected within the same is $p = 0.7$.
Assume that the probability of an error being corrected within the same is $p = 0.7$.

- Step 1) What is the random variable: X is number of corrected errors
- Step 2) What dstribution: X follows The binomial distribution
- Step 3) What probability: $P(X = x) = f(x;n,p)$ $P(X = 6) = f(6;n,p)$
- Step 4)
  - What is the number of repeats? $n = 6$
  - What is the probability of success? $p = 0.7$

Example 1

What is the probability that 2 or less of the errors is corrected within the same day?

- Step 1) What is the random variable: X is number of corrected errors
- Step 2) What dstribution: X follows The binomial distribution
- Step 3) What probability: $P(X\ ?\ ?)$ $P(X \leq 2) = F(2;n,p)$
- Step 4)
  - What is the number of repeats? $n = 6$
  - What is the probability of success? $p = 0.7$

Oversigt

The hypergeometric distribution

- X is again 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 (repeats)
- a the number of successes in the population
- N is the number of elements in the (entire) population

The hypergeometric distribution

- The probability to get 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 (repeats)
- a the number of successes in the population
- N is the number of elements in the (entire) population

Example 2

In a shipment of 10 hard disks 2 of them have small scrathes.

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

- Step 1) What is the random variable: X is number with scratches
- Step 2) What distribution: X follows the hypergeometric distribution
- Step 3) What probability:
  $P(X\ ?\ ?)\ P(X \geq 1) = 1 - P(X=0) = 1 - f(0;n,a,N)$
- Step 4)
  - What is number of draws? $n=3$
  - How many successes is there? $a=2$
  - How many disks all together? $N=10$

Binomial vs. hypergeometric

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

Oversigt

The Poisson distribution

- The Poisson distribution is often use as distribution (model) for counts which do not have a natural upper bound
- The Poisson distribution is often characterized as intensity, that is on the form number/unit
- The parameter λ gives the gives the intensity in the Poisson distribution

The Poisson distribution

$$X \sim P(\lambda)$$

The density function:

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

The distribution function:

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

Example 3.1

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

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

- Step 1) What is the random variable: X is the number of patients on a day
- Step 2) What distribution: X follows the Poisson distribution
- Step 3) What probability: $P(X \, ? \, ?) P(X \leq 2)$
- Step 4) What is the intensity: $\lambda = 0.3$ patients per day

Example 3.2

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

What is the probability that exactly two patients are put in hospital in Copenhagen due to air pollution on a given day?

- Step 3) What probability: $P(X \, ? \, ?) P(X = 2)$

Example 3.3

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

What is the probability that at least 2 patients are put in hospital in Copenhagen due to air pollution on a given day?

- Step 3) What probability:
  $P(X\ ?\ ?)P(X \geq 2) = 1 - P(X \leq 1)$

Example 3.4

What is the probability that exactly 1 patient is put in hospital in Copenhagen due to air pollution within 3 days?

- Step 1) What is the random variable:
  - From X number per day
  - To $X^{3\text{days}}$ which is *patients per 3 days*
- Step 2) What distribution has $X^{3\text{days}}$:
  The Poisson distribution
- Step 3) What probability: $P(X^{3\text{days}} = 1)$
- Step 4) Scale the intensity
  - From $\lambda = 0.3$ patientes/day to $\lambda_{3days} = 0.9$ patients/3days

Oversigt

R	Name
`binom`	Binomial
`hyper`	Hypergeometric
`pois`	Poisson

- `d` $f(x)$ (probability density function).
- `p` $F(x)$ (cumulative distribution function).
- `r` Random numbers from the distribution
- `q` quantiles (the inverse of $F(x)$)

Remember that function help etc. is achieved by putting '?' in front of the name.

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

```
pbinom(q=5, size=10, prob=0.6)
## Get the hep with
?pbinom
```

Oversigt

Mean (Expected value))

Mean of discrete random variable:

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

- The "correct mean"
- Expresses the "center" of X

Example: Mean of a dice throw

$$\begin{aligned}\mu = E(X) &= \\ &= 1\cdot\frac{1}{6}+2\cdot\frac{1}{6}+3\cdot\frac{1}{6}+4\cdot\frac{1}{6}+5\cdot\frac{1}{6}+6\cdot\frac{1}{6} \\ &= 3.5\end{aligned}$$

Link to sample mean - simulation learning

```
## NUmber of simulated realizations
n <- 30
## Sample independently from the set (1,2,3,4,5,6) with
## equal probability
xFair <- sample(1:6, size=n, replace=TRUE)

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

The more observations, the close you get to the right mean (expected value)

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

- Try it in R

Variance

Definition

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

- Measures average spread
- The "correct standard deviation" of X (as opposed to sample variance))

Variance, example

Variance of dice throw

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

Link to sample variance - simulation learning

```
  ## NUmber of simulated realizations
  n <- 30
  ## Sample independently from the set (1,2,3,4,5,6) with
  ## equal probability
  xFair <- sample(1:6, size=n, replace=TRUE)

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

Mean and variances for specific discrete distributions

The binomial distribution:

- Mean:
  $\mu = n \cdot p$
- Variance:
  $\sigma^2 = n \cdot p \cdot (1-p)$

Mean and variances for specific discrete distributions

The hypergeometric distribution:

- Mean:
  $\mu = n \cdot \frac{a}{N}$
- Variance:
  $\sigma^2 = \frac{na \cdot (N-a) \cdot (N-n)}{N^2 \cdot (N-1)}$

Mean and variances for specific discrete distributions

The poisson distribution:

- Mean:
  $\mu = \lambda$
- Variance:
  $\sigma^2 = \lambda$

Agenda

Course 02402 Introduction to Statistics Lecture 3:

Continuous Distributions

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

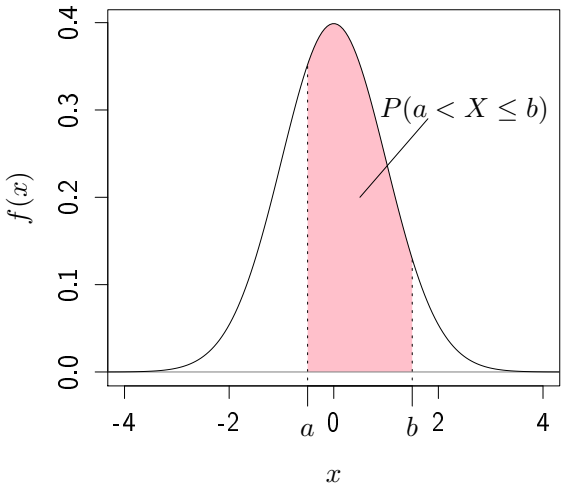
Agenda

Oversigt

The Density Function (pdf)

- The density function for a stochastic variable is denoted by $f(x)$
- $f(x)$ says something about the frequency of the outcome x for the stochastic variable X
- The density function for continuous variables does not correspond to the probability, that is $f(x) \neq P(X = x)$
- A nice plot of $f(x)$ is a histogram

The Density Function for Continuous Variables



The Density Function for Continuous Variables

The density function for a continuous variable is written as:

$$f(x)$$

The following is valid:

$$f(x) \geq 0 \quad \text{for all } x$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

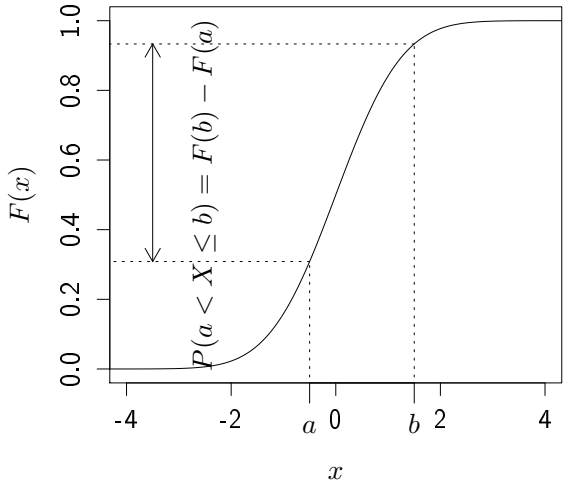
Distribution function or cumulative density function (cdf))

- The distribution function for a continuous stochastic variable is denoted by $F(x)$.
- The distribution function corresponds to the cumulative density function:

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

- $$F(x) = \int_{t=-\infty}^{x} f(t)dt$$
- A nice plot of $F(x)$ is the cumulative distribution plot
- $$f(x) = F'(x)$$

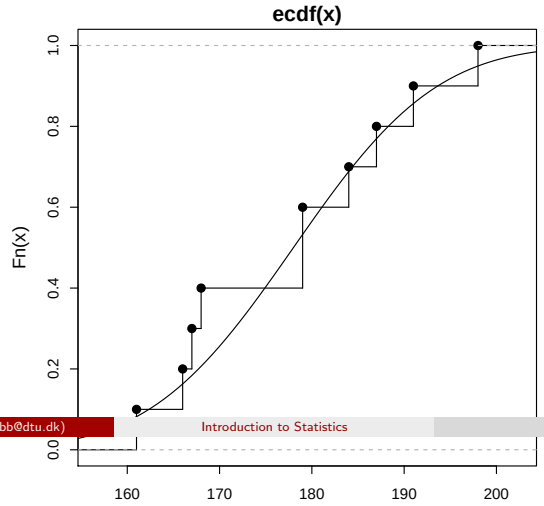
The distribution function(cdf))



The empirical cumulative distribution function - ecdf

Student height example from Chapter 1:

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



The Mean of a Continuous Random Variable

The Mean of a Continuous Random Variable

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

Compare with the discrete definition:

$$\mu = \sum_{i=1}^{\infty} x_i f(x_i)$$

The Variance of a Continuous Random Variable

The Variance of a Continuous Random Variable:

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

Compare with the discrete definition:

$$\sigma^2 = \sum_{i=1}^{\infty} (x_i - \mu)^2 f(x_i)$$

The Covariance of two random variables

The Covariance of two random variables:

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

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

Oversigt

Specific Statistical Distributions

- A number of statistical distributions exist that can be used to describe and analyze different kind of problems

Now we consider continuous distributions

- The uniform distribution
- The normal distribution
- The log-normal distribution
- The Exponential distribution

The Uniform Distribution

Syntax:

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

Density function:

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

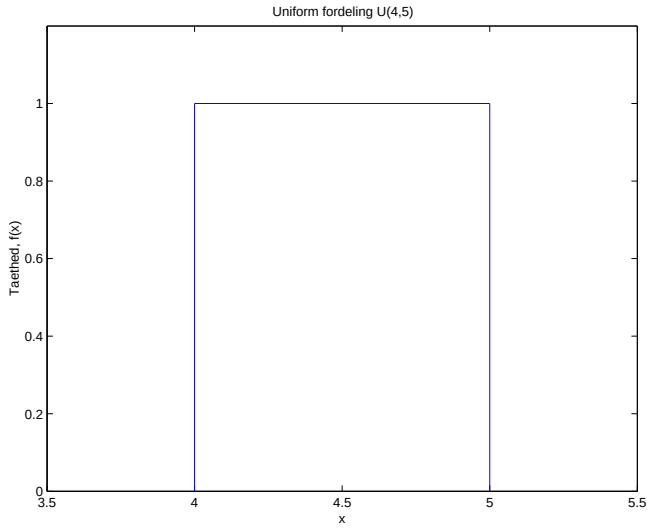
Mean:

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

Variance:

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

The Uniform distribution



Example 1

Students in a course arrive to a lecture between 8.00 and 8.30. It is assumed that the arival times can be described by a uniform distribution.

Question:

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

Answer:

10/30=1/3

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

[1] 0.33

Example 1 - cont.

Question:

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

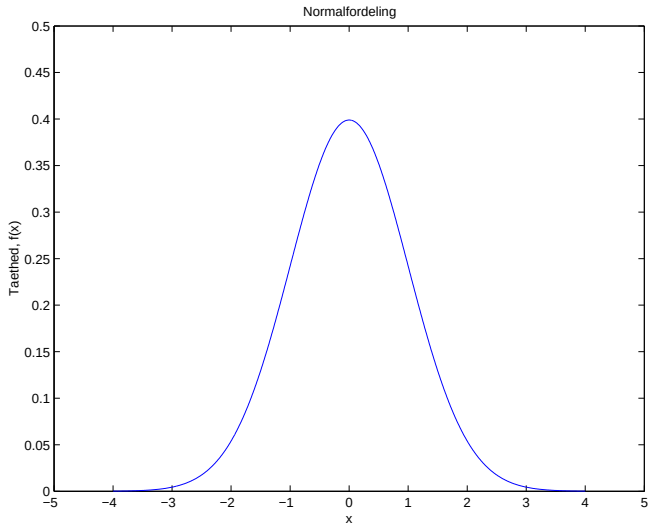
Answer:

0

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

[1] 0

The Normal Distribution



The Normal Distribution

Syntax:

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

Dnsity function:

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

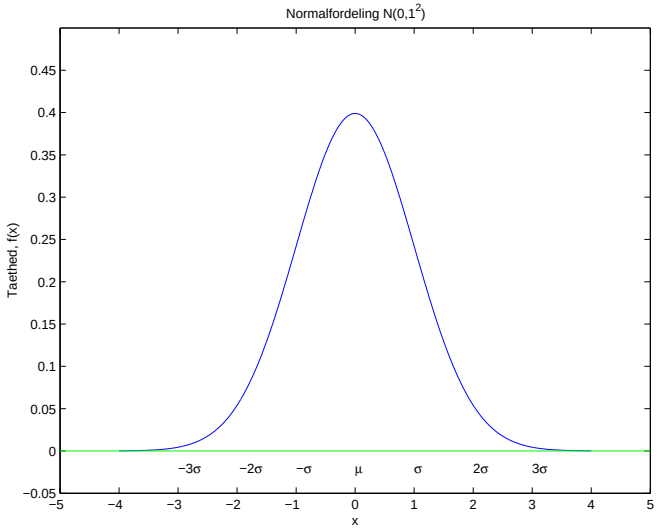
Mean:

$$\mu = \mu$$

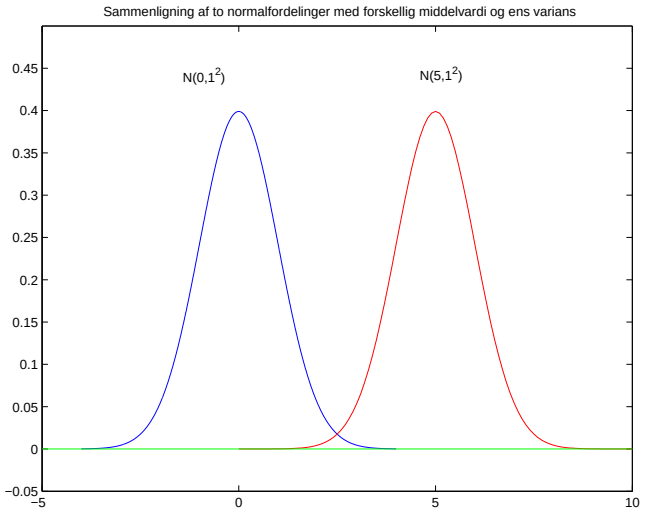
Variance:

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

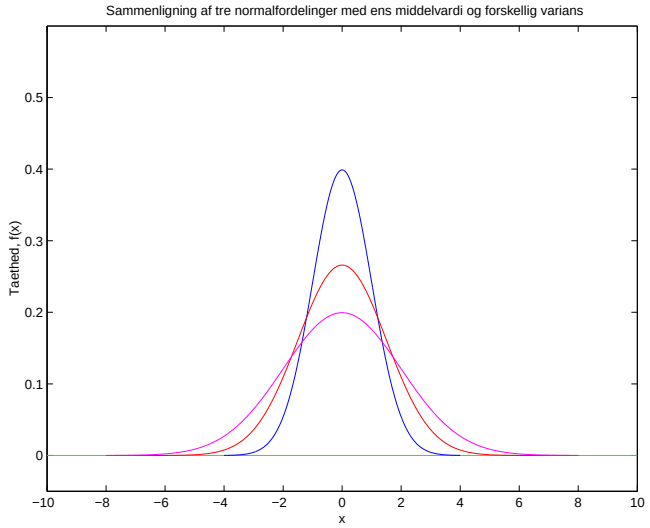
The Normal Distribution



The Normal Distribution



The Normal Distribution



The Normal Distribution

A standard normal distribution:

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

A normal distribution with mean 0 and variance 1.

Standardization:

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

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

Example 2

Measurement error:

A weight has a measurement error, Z, that can be described by a standard normal disrtribution, i.e.

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

that is, mean $\mu = 0$ and standard deviation $\sigma = 1$ gram.

We now measure the weight of a single piece

Question a):

What is the probability that the weight measures at least 2 grams too little?

Answer:

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

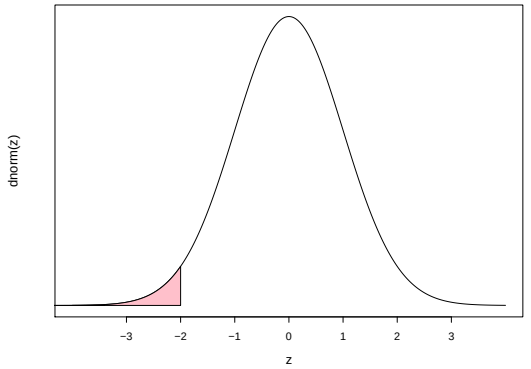
```
pnorm(-2)
```


Example 2

Answer:

```
pnorm(-2)
```

[1] 0.023



Example 2

Question b):

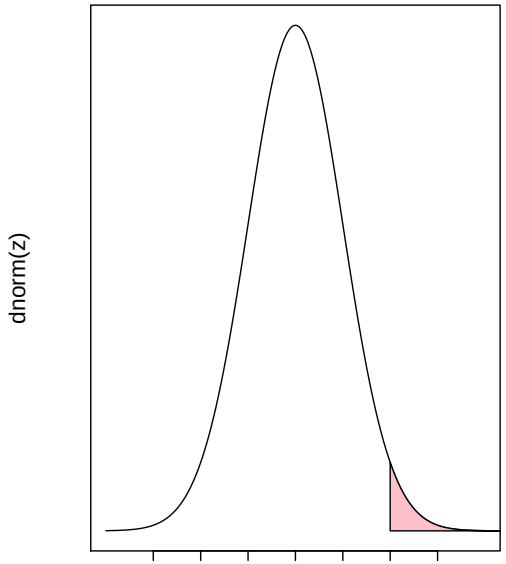
What is the proba ... *much?*

Answer:

$P(Z \geq 2) = 0.022$

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

[1] 0.023



Example 2

Question c):

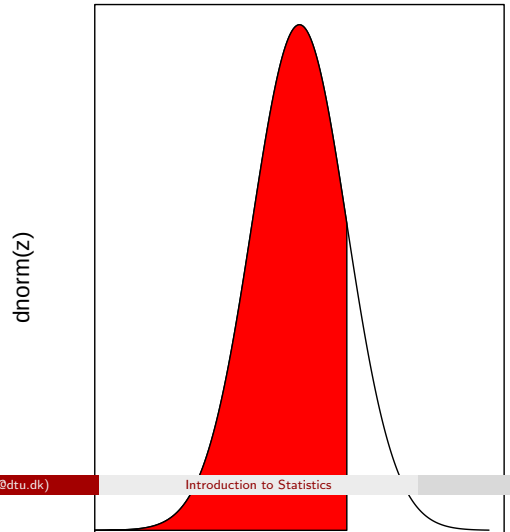
What is the probability that the weight measures at most ± 1 gram wrong?

Answer:

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

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

[1] 0.68



Example 2

Question c):

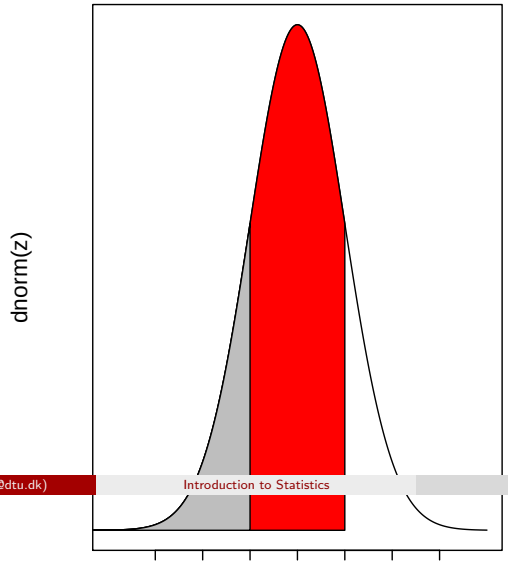
What is the probability that the weight measures at most ± 1 gram wrong?

Answer:

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

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

[1] 0.68



Example 3

Indkomstfordeling:

It is assumed that among a group af elementary school teachers, the salary distribution can be described as a normal distribution with mean $\mu = 280.000$ and standard deviation $\sigma = 10.000$..

Question a):

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

Answer:

$P(X > 300) = P(Z > \frac{300-280}{10}) = P(Z > 2) = 0.023$

$$X \sim N(300, 10^2) \Rightarrow Z = \frac{X - 280}{10} \sim N(0, 1^2)$$

Example 3

Question a):

What is the proba *an 300.000?*

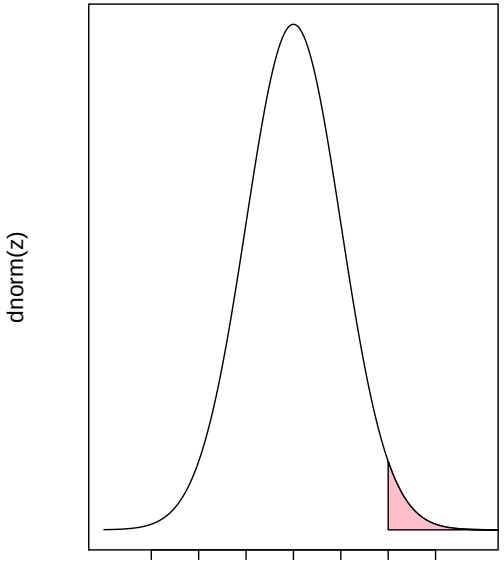
Answer:

```
1-pnorm(300, m
```

[1] 0.023

```
1-pnorm((300-28
```

[1] 0.023



Example 4

A more narrow distribution:

It is assumed that among a group af elementary school teachers, the salary distribution can be described as a normal distribution with mean $\mu = 290.000$ and standard deviation $\sigma = 4.000$.

Question a):

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

Example 4

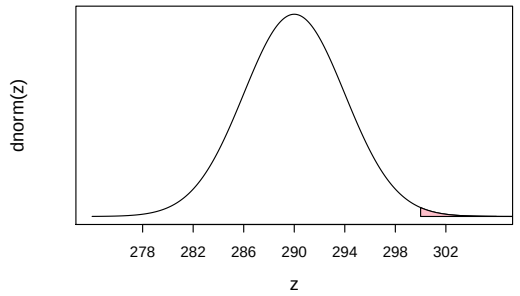
Question a):

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

Answer:

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

[1] 0.0062



Example 5

Same income distribution

It is assumed that among a group af elementary school teachers, the salary distribution can be described as a normal distribution with mean $\mu = 290.000$ and standard deviation $\sigma = 4.000$.

"Opposite question"

Give the salary interval that covers 95% of all teachers' salary

Answer:

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

[1] 282 298

The Log-Normal distribution

Syntax:

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

Density function:

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

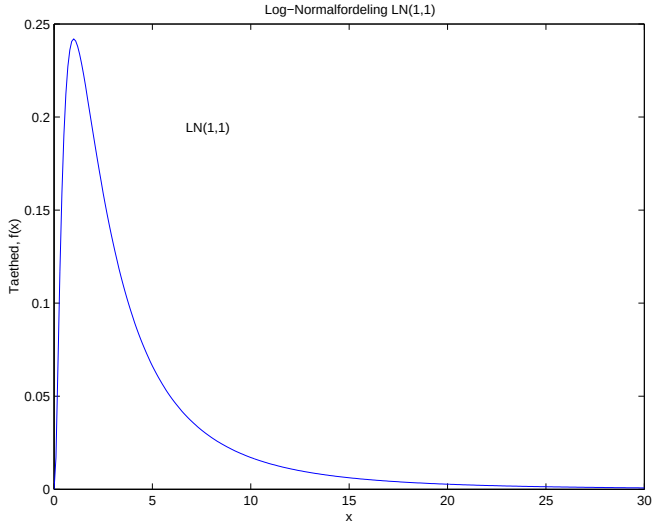
Mean:

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

Variance:

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

The Log-Normal distribution



The Log-Normal distribution

Log-normal and Normal distributions:

A log-normally distributed variable $Y \sim LN(\alpha, \beta)$ can be transformed into a standard normally distributed variable X by:

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

dvs.

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

Continuous distributions in R

R	Distribution
`norm`	The normal distribution
`unif`	The uniform distribution
`lnorm`	The log-normal distribution
`exp`	The exponential distribution

- `d` $(f(x))$ probability density function.
- `p` $(F(x))$cumulative distribution function.
- `q` Quantile in distribution.
- `r` Random numbers from distribution

Oversigt

The Exponential Distribution

Density function

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

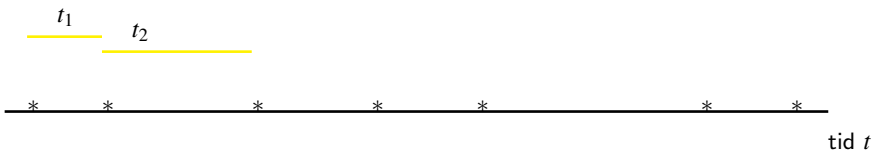
The Exponential Distribution

- 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
- Mean $\mu = \beta$
- Variance $\sigma^2 = \beta^2$

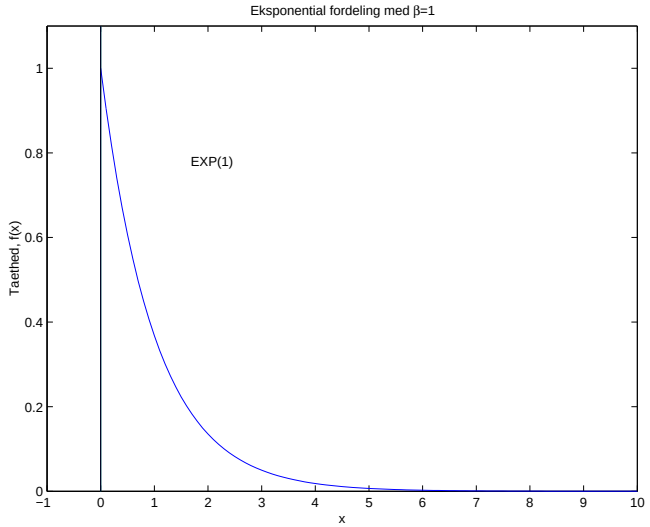
Connection between the Exponetial- and Poisson Distribution

Poisson: Discrete events pr./ unit

Exponential: Continuous distance between events



The Exponential Distribution



Example 6

Queing model - poisson proces

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

Question:

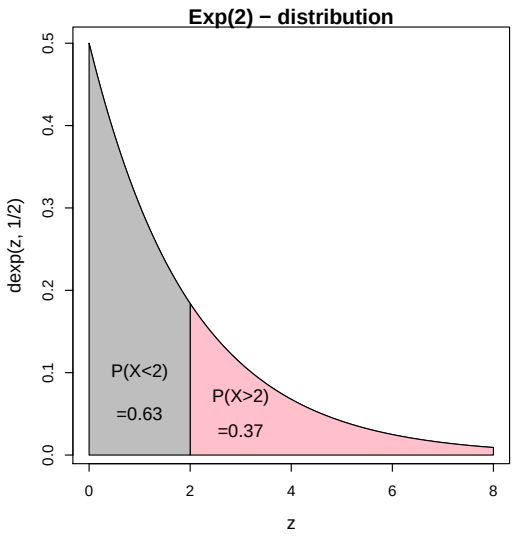
One customer is just arrived. What is the probability that no other costumers will arrive in the next period of 2 minutes?

Answer:

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

[1] 0.37

Example 6



Example 6

```
z=seq(0,8,by=0.01)

plot(z,dexp(z, 1/2),type = "l", main = "Exp(2) - distribution")

polygon(c(2, seq(2, 8, by = 0.01), 8, 2),
        c(0, dexp(seq(2, 8, by = 0.01), 1/2), 0, 0),
        col = "pink")

text(3,0.07,"P(X>2)")
text(3,0.03,"=0.37")

polygon(c(2, seq(2, 0, by = -0.01), 0, 2),
        c(0, dexp(seq(2, 0, by =- 0.01), 1/2), 0, 0),
        col = "grey")

text(1,0.1,"P(X<2)")
text(1,0.05,"=0.63")
```


Example 7

Question:

One customer is just arrived. Using the Poisson distribution, calculate the probability that no other costumers will arrive in the next period of 2

Answer:

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

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

[1] 0.37

```
exp(-1)
```

[1] 0.37

Example 8

Other time periods:

The time between customer arrivals at a post office is exponentially distributed with mean $\mu = 2$ minutes. Now consider a period of 10 minutes

Question:

Using the Poisson distribution, calculate the probability that no other costumers will arrive in this period

Answer:

$\lambda_{10min} = 5$, $P(X = 0) = \frac{e^{-5}}{1!} 5^0 = e^{-5}$

```
dpois(0,5)
```

[1] 0.0067

Oversigt

Calculation Rules for Random Variables

(Holds for AS WELL continuous as discrete variables)

X is a random variable

. We assume that a and b are constants.Then we have:

Mean rule:

$$E(aX+b) = aE(X)+b$$

Variance rule:

$$Var(aX+b) = a^2 Var(X)$$

Example 9

X is a random variable

. A random variable X has mean 4 and variance 6.

Question:

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

Answer:

$$\mathrm{E}(Y) = -3\mathrm{E}(X)+2 = -3\cdot 4+2 = -10$$
$$\mathrm{Var}(Y) = (-3)^2\mathrm{Var}(X) = 9\cdot 6 = 54$$

Calculation Rules for Random Variables

$X_1, \ldots, X_n$ are random variables

Then (when independent)

Mean rule:

$$E(a_1X_1 + a_2X_2 + .. + a_nX_n)$$
$$= a_1E(X_1) + a_2E(X_2) + .. + a_nE(X_n)$$

Variance rule::

$$Var(a_1X_1 + a_2X_2 + .. + a_nX_n)$$
$$= a_1^2Var(X_1) + .. + a_n^2Var(X_n)$$

Example 10

Airline Planning

The weight of the passengers on a flight is assumed Normal distributed $X \sim N(70, 10^2)$.

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

Question:

Calculate the probability that the plain is overloaded

What is Y=Total passenger weight?

What is Y?

Definitely NOT: $Y = 55 \cdot X$!!!!!!

Example 10

What is Y=Total passenger weight?

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

Mean and variance of Y:

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

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

We use a normal distribution for Y:

```
1-pnorm(4000, m = 3850, s = sqrt(5500))
```

[1] 0.022

Example 10 - WRONG ANALYSIS

What is Y?

Definitely NOT: $Y = 55 \cdot X$!!!!!!

Mean and variance of Y:

$$E(Y) = 55 \cdot 70 = 3850$$

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

We use a normal distribution for Y:

```
1-pnorm(4000, m = 3850, s = 550)
```

[1] 0.39

Consequence of wrong calculation:

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

Agenda

Overview

Oversigt

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**:

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

NEW:**Confidence interval, σ**:

$$[8.4;\ 22.3]$$

Oversigt

Let's simulate the key challenge of statistics!

```
## Mean
mu <- 178
## Standard deviation
sigma <- 12
## Sample size
n <- 10
## Simulate  normally distributed  X_i
x <- rnorm(n=n, mean=mu, sd=sigma)
## See the realizations
x
## Empirical density
hist(x, prob=TRUE, col='blue')
## Find the sample mean
mean(x)
##  Find the sample variance
var(x)
## Repeat the simulated sampling many times
mat <- replicate(100, rnorm(n=n, mean=mu, sd=sigma))
## Find the sample mean for each of them
xbar <- apply(mat, 2, mean)
## Now we have many realizations of the sample mean
xbar
## See their distribution
hist(xbar, prob=TRUE, col='blue')
## There mean
mean(xbar)
## and sample variance
var(xbar)
```


Theorem 3.2: The distribution of the mean of normal random variables

(Sample-) Distribution/ The (sampling) distribution for $\bar{X}$

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

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

Mean and variance follow from 'rules':

The Mean of $\bar{X}$

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

The variance of $\bar{X}$

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

We now know the distribution of the error we make:

(When using $\bar{x}$ as an estimate of μ)

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 same thing, Corollary 3.3:

Distribution for the standardized error we make:

Assume that $X_1, \ldots, X_n$ are independent and identically normally distributed random variables, $X_i \sim N\left(\mu, \sigma^2\right)$ where $i = 1, \ldots, n$, then:

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

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

Practical problem in all this, so far:

How to transform this into a specific interval for μ?

When the populations standard deviation σ is in all the formulas?

Obvious solution:

Use the estimtate s in stead of σ in formulas!

BUT BUT:

The given theory then breaks down!!

Luckily:

We have en extended theory to handle it for us!!

Theorem 3.4: More applicable extension of the same stuff: (copy of Theorem 2.49)

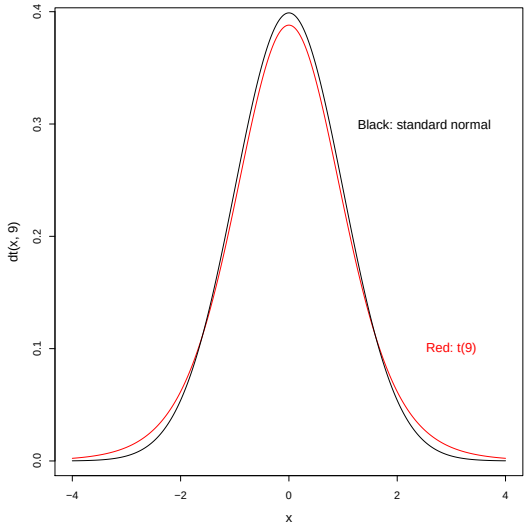
The t-Distribution takes the uncertainty of s into account:

Assume that $X_1, \ldots, X_n$ are independent and identically normally distributed random variables, where $X_i \sim N\left(\mu, \sigma^2\right)$ and $i = 1, \ldots, n$, then:

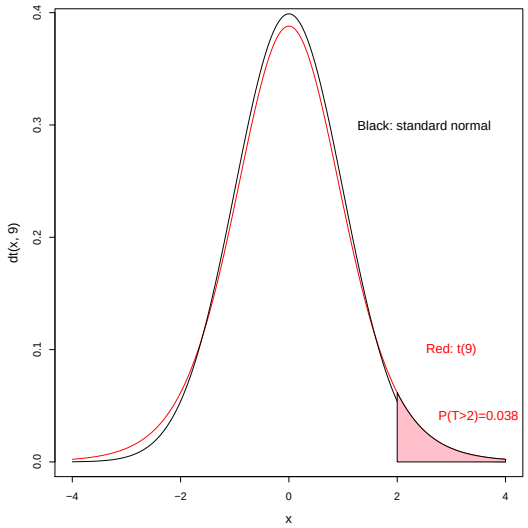
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t$$

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

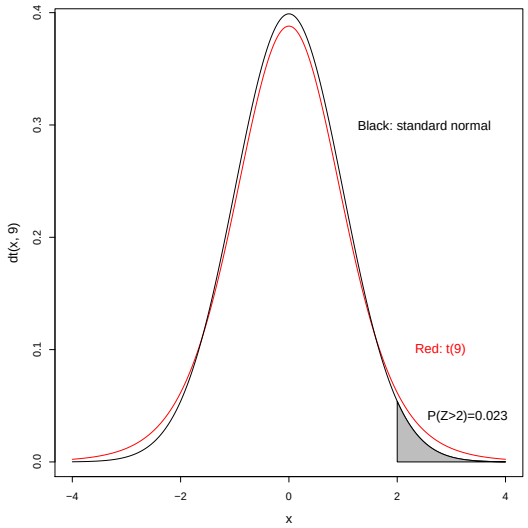
t-Distribution with 9 degrees of freedom ($n = 10$):



t-Distribution with 9 degrees of freedom and standard normal distribution:



t-Distribution with 9 degrees of freedom and standard normal distribution:



Oversigt

Method box 3.8: One-sample Confidence interval for μ

Use the right t-distribution to make the confidence interval:

For a sample $x_1, \ldots, x_n$ 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 $100(1-\alpha)\%$ 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:

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

Student height Example

```
## The t-quantiles for n=10:
 qt(0.975,9)
```

[1] 2.3
and we can recognize the already given result:

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

which is:

$$178 \pm 8.74 = [169.3;\ 186.7]$$

Student height example, 99% Confidence interval (CI)

```
qt(0.995,9)
```

[1] 3.2

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

giving

$$178 \pm 12.55 = [165.4;\ 190.6]$$

There is an R-function, that can do it all (and more than that):

```
x <- c(168,161,167,179,184,166,198,187,191,179)
t.test(x,conf.level=0.99)

##
##  One Sample t-test
##
## data:  x
## t = 50, df = 9, p-value = 5e-12
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  165 191
## sample estimates:
## mean of x
##       178
```


Oversigt

The formal framework for *statistical inference*

From eNote, Chapter 1:

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

Language and concepts:

- μ and σ are parameters describing th populationen
- $\bar{x}$ 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, heights example

We measure the heights of 10 randomly selected persons in Demark

The sample:

The 10 specific numbers: $x_1, \ldots, x_{10}$

The population:

The heights for all people in Dnemark

Observational unit:

A person

Statistical inference = Learning from data

Learning from data:

Is learning about parameters of distributions that describe populations.

Important for this:

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

How to ensure this:

F.ex. by making sure that the sample is taken completely at random

Random Sampling

Definition 3.11:

- A random sample from an (infinite) population: A set of observations $X_1, X_2, ..., X_n$ 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. if we sampled entire families)

Oversigt

Theorem 3.13: The Central Limit Theorem

No matter what, the distribution of the mean becomes 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

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

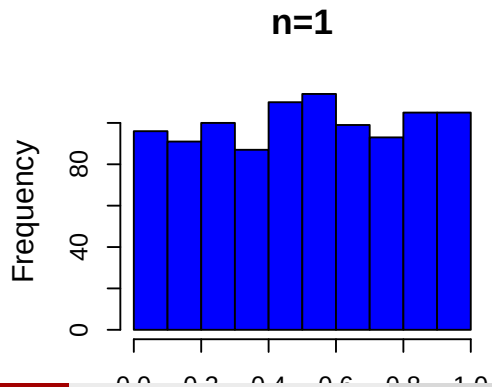
is a random variable whose distribution function approaches that of the standard normal distribution, $N(0, 1^2)$, as $n \to \infty$

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

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

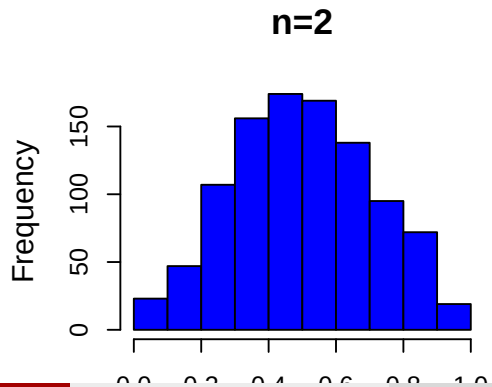
CLT in action - mean of uniformly distributed observations

```
n=1
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=1",xlab="Means")
```



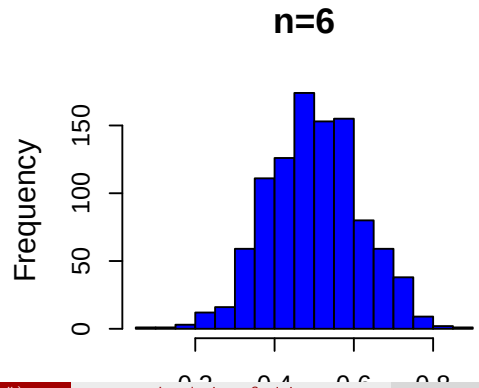
CLT in action - mean of uniformly distributed observations

```
n=2
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=2",xlab="Means")
```



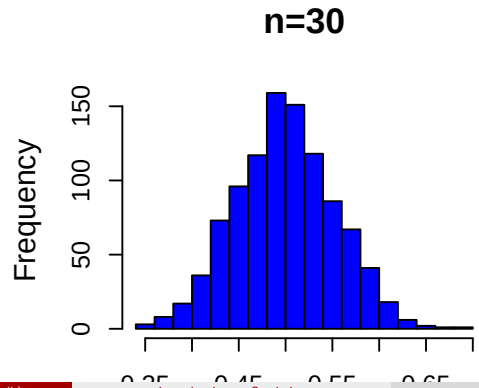
CLT in action - mean of uniformly distributed observations

```
n=6
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=6",xlab="Means")
```



CLT in action - mean of uniformly distributed observations

```
n=30
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=30",xlab="Means", nclass=15)
```



Consequence of 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.

What is "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.

Oversigt

'Repeated sampling' interpretation

In the long run we catch the true value in 95% of cases:

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

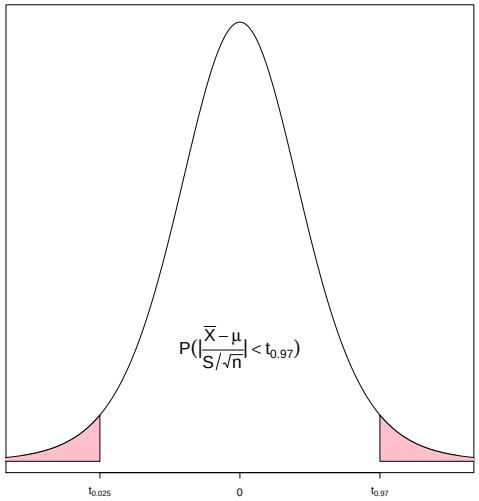
More formally expressed (Theorem 3.4 and 2.49):

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

Which is equivalent to:

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

'Repeated sampling' interpretation



Oversigt

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 2.53)

Variance estimators behaves like a χ^2-distribution:

Let

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

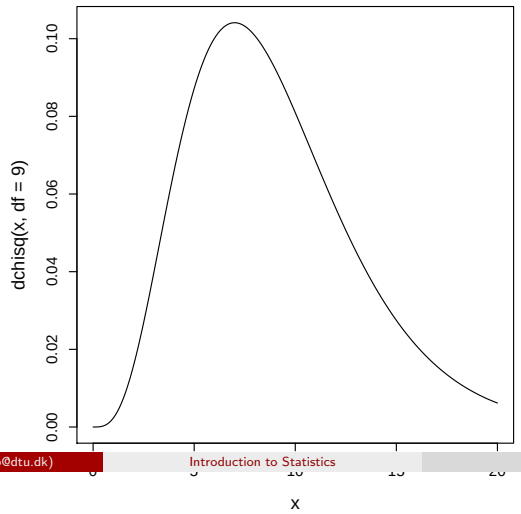
then:

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

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 3.18: Confidence interval for sample variance and standard deviation

The variance:

A $100(1-\alpha)\%$ confidence interval for a sample variance $\hat{\sigma}^2$ 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 sample standard deviation $\hat{\sigma}$ 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:

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

95%-Confidence interval for the variance - we need the χ^2-quantiles:

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

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

```
[1] 8.9 32.9
```

Example

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

$$\left[\frac{19 \cdot 0.7^2}{32.85};\ \frac{19 \cdot 0.7^2}{8.907}\right] = [0.002834;\ 0.01045]$$

and the confidence interval for the standard deviation σ becomes:

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

Heights example

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

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

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

```
[1] 2.7 19.0
```

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- recap:

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**:

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

NEW:**Confidence interval, σ**:

$$[8.4;\ 22.3]$$

Overview

Course 02402 Introduction to Statistics Lecture 5:

One-sample hypothesis test and model control

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

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

Oversigt

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

Motivating example - sleeping medicine

Difference of sleeping medicines?

In a study the aim is to compare two kinds of sleeping medicine A and B. 10 test persons tried both kinds of medicine and the following 10 DIFFERENCES between the two medicine types were measured:
(For person 1, sleep medicine B was 1.2 sleep hour better than medicine A, etc.):

Sample, $n = 10$:

person	$x =$ Beffect - Aeffect
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 - sleeping medicine

The hypothesis of no difference:

$$H_0: \ \mu = 0$$

Sample mean and standard deviation:

$$\bar{x} = 1.670 = \hat{\mu}$$
$$s = 1.13 = \hat{\sigma}$$

Is data in acoordance with the null hypothesis H_0?

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

NEW:p-**value:**

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

(Computed under the scenario, that H_0 is true)

NEW:**Conclusion**:

As the data is unlike far away from H_0, we **reject** H_0 - we have found a **significant effect** of

A.

Oversigt

Method 3.22: 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 the *evidence* against the null hypothesis – Table **??**:

$p < 0.001$	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 3.21 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 - sleeping medicine

The hypothesis of no difference:

$$H_0: \ \mu = 0$$

Compute the test-statistic:

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

Compute the p-value:

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

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

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

There is strong evidence agains the null hypothesis.

Example - sleeping 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)
## Compute the tobs - the observed test statistic:
tobs <- (mean(x) - 0) / (sd(x) / sqrt(n))
## Compute the p-value as a tail-probability
## in the t-distribution:
pvalue <- 2 * (1-pt(abs(tobs), df=n-1))
pvalue

## [1] 0.0012
```


Example - sleeping medicine - in R - with inbuilt function

```
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 5, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.86 2.48
## sample estimates:
## mean of x
##       1.7
```


The definition of hypothesis test and significance (generally)

Definition 3.23. 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-specifed (so-called) *significance level*. And if not, then the null hypothesis is said to be *accepted*.

Definition 3.28. 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 - sleeping medicine

With $\alpha = 0.05$ we can conclude:

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

And hence:

We have found a **significant effect** af medicine B as compared to A. (And hence that B works better than A)

Oversigt

Critical value

Definition 3.30 - 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:

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

Metode 3.31: One-sample t-test by critical value:

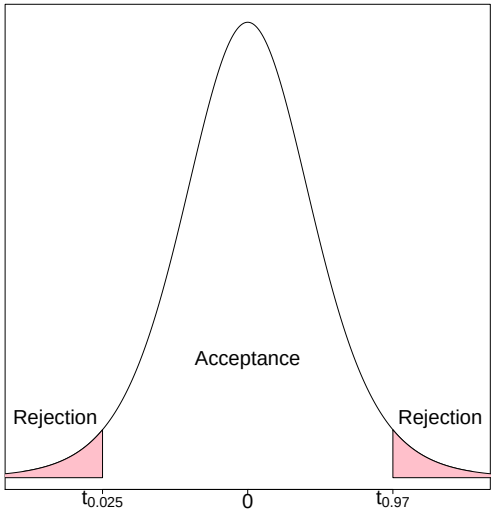
A null hypothesis is *rejected* if the observed test-statistic is more extreme than the critical values:

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

otherwise *accept*.

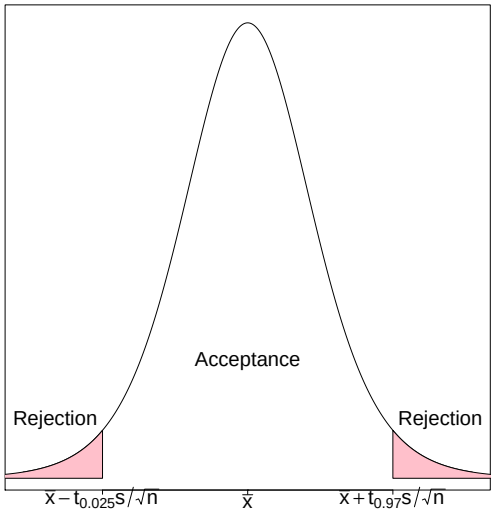
Critical value and hypothesis test

The acceptance region are the values for μ not too far away from the data - here on the standardized scale:



Critical value and hypothesis test

The acceptance region are the values for μ not too far away from the data - now on the original scale:



Critical value, confidence interval and hypothesis test

Theorem ??: 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

$$H_0: \quad \mu = \mu_0$$

(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 3.33

A μ_0 inside the confidence interval will fullfill 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.

Oversigt

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 : $H_1: \ \mu < \mu_0$

But we stick to the "non-directional"in this course

Steps by hypothesis tests - an overview

Generelly a hypothesis test consists of the foloowing steps:

1. Formulate the hypotheses 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 sampling distribution, and compare the p-value and 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 3.35 The level α test is:

1. Compute t_{obs} as before
2. Compute the 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–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 or errors:

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

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 the alternative hypotheses are:

H_0 : The man is not guilty

H_1 : The man is guilty

That you cannot be proved guilty is not the same as being proved innocent

Or differently put:
Accepting a null hypothesis is NOT a statistical proof of the null hypothesis being true!

Errors in hypothesis testing

Theorem 3.38: Significance level = The risk of a Type I error

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

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

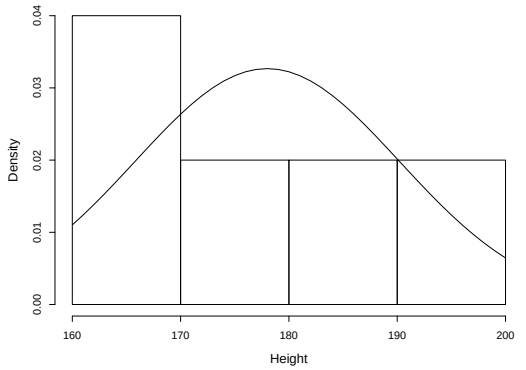
Two possible truths vs. two possible conclusions:

	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)

Oversigt

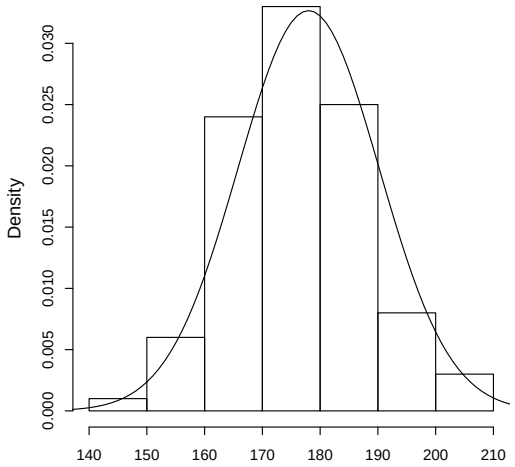
Example - student heights - are they normally distributed?

```
x <- c(168,161,167,179,184,166,198,187,191,179)
hist(x, xlab="Height", main="", freq = FALSE)
lines(seq(160, 200, 1), dnorm(seq(160, 200, 1), mean(x), sd(x)))
```



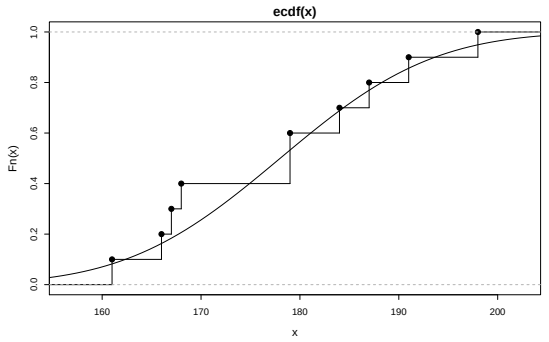
Example - 100 observations from a normal distribution:

```
xr <- rnorm(100, mean(x), sd(x))
hist(xr, xlab="Height", main="", freq = FALSE)
lines(seq(130, 230, 1), dnorm(seq(130, 230, 1), mean(x), sd(x)))
```



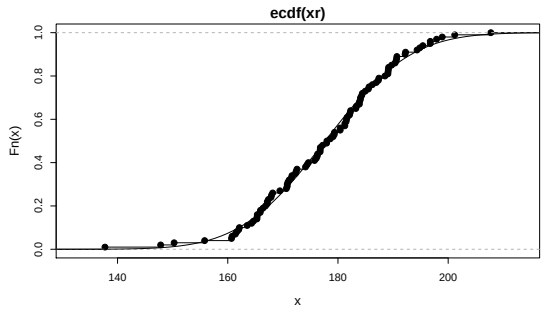
Example - student heights - ecdf

```
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)))
```



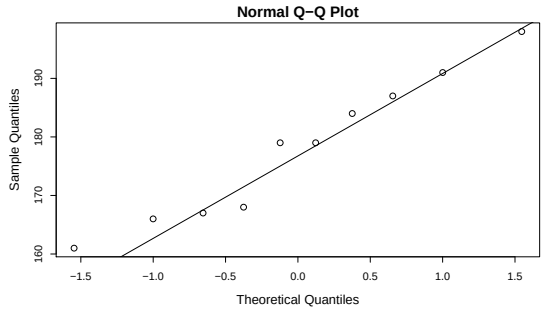
Example - 100 observations from a normal distribution, ecdf:

```
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)))
```

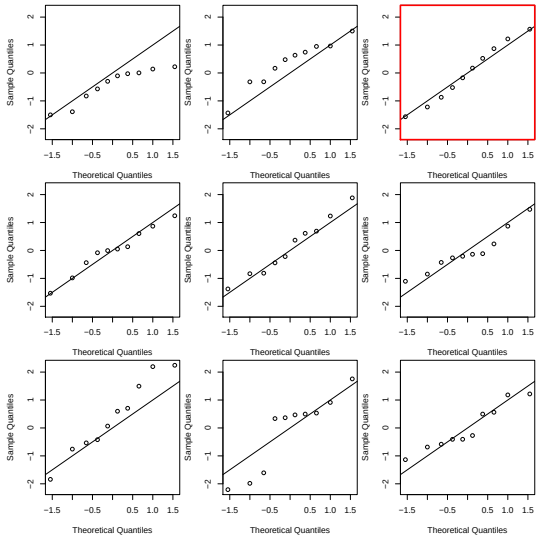


Example - student heights - Normal Q-Q plot

```
qqnorm(x)
qqline(x)
```



Example - student heights - Normal Q-Q plot - compare with other simulated normally distributed data



Normal Q-Q plot

Metode 3.41- The formal definition

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

- In R, when $n > 10$:

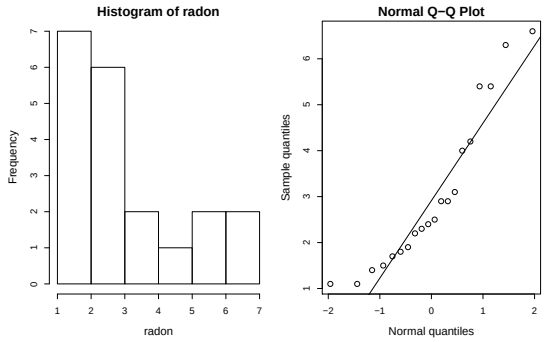
$$p_i = \frac{i-0.5}{n+1},\ i = 1,\ldots,n$$

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

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

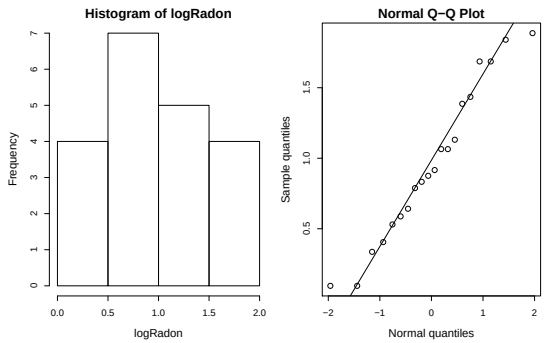
Example - Radon data

```
## 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)
##A HISTOGRAM AND A QQ-PLOT
par(mfrow=c(1,2))
hist(radon)
qqnorm(radon,ylab = 'Sample quantiles',xlab = "Normal quantiles")
qqline(radon)
```



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

```
##TRANSFORM USING NATURAL LOGARITHM
logRadon<-log(radon)

hist(logRadon)
qqnorm(logRadon,ylab = 'Sample quantiles',xlab = "Normal quantiles")
qqline(logRadon)
```



Agenda

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

Course 02402 Introduction to Statistics Lecture 6:

Two-sample comparisons and power/sample size

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

Oversigt

Motivating example - nutrition study

Nutrition study

In a nutrition study the aim is to investigate if 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):

Sample from each hospital, $n_1 = n_2 = 9$:

Hospital A	Hospital B
7.53	9.21
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 is in focus:

$$H_0: \ \mu_1 = \mu_2$$

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

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

p-value $= 0.0083$

(Found in the scenario that H_0 is true)

NYT:**Confidence interval for difference**:

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

Oversigt

The definition of hypothesis test and significance (Repetition)

Definition 3.23. 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-specifed (so-called) *significance level*. And if not, then the null hypothesis is said to be *accepted*.

Definition 3.28. Statistical significance:

An *effect* is said to be *(statistically) significant* if the p-value is less than the significance level α.
(OFTEN we use $\alpha = 0.05$)

Steps by hypothesis tests - an overview (Repetition)

Generelly a hypothesis test consists of the foloowing steps:

1. Formulate the hypotheses 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 sampling distribution, and compare the p-value and the significance level α and make a conclusion

   OR:

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

The definition and interpretation of the p-value (Repetition)

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

$p < 0.001$	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 3.21 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 **??**: 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

$$H_0: \quad \mu = \mu_0$$

(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.

Oversigt

Method 3.48: 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 3.49: The distribution of the (Welch) t-test statistic

Welch t-test statistic is t-distributed

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, under the null hypothesis, follows a t-distribution with ν degrees of freedom, where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

if the two population distributions are normal or if the two sample sizes are large enough.

Method 3.50: Two-sample t-test

The level α test is

1. Compute t_{obs} and ν 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 ν degrees of freedom is used.
3. If p–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.

[a] We 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:

$$H_0 : \ \delta = \mu_B - \mu_A \neq 0$$

First the computations of t_{obs} and ν:

$$t_{\text{obs}} = \frac{10.298 - 8.293}{\sqrt{2.0394/9 + 1.954/9}} = 3.01$$

and

$$\nu = \frac{\left(\frac{2.0394}{9} + \frac{1.954}{9}\right)^2}{\frac{(2.0394/9)^2}{8} + \frac{(1.954/9)^2}{8}} = 15.99$$

Example - nutrition study

Next the p-value is found:

$$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|) = 2P(T > 3.01) = 2 \cdot 0.00415 = 0.0083$$

```
## p-value for nutrition study example
1 - pt(3.01, df = 15.99)

## [1] 0.0042
```

Evaluate the evidence (Table ??):

There is strong evidence AGAINST the null hypothesis.

Conclude based on $\alpha = 0.05$:

We reject the null hypothesis, as there is a significant difference of the two

usage than nurses on Hospital A.

Oversigt

Method 3.46: Confidence interval for $\mu_1 - \mu_2$

The Confidence interval for the mean difference:

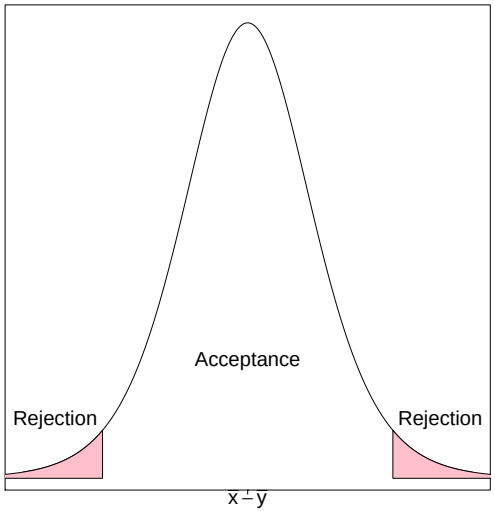
For two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$-quantile from the t-distribution with ν degrees of freedom given from Theorem 3.49 (as above)

The Confidence interval and hypothesis test (Repetition)

The acceptance region is the potential values for $\mu_1 - \mu_2$ that are not too far away from the data:



Example - nutrition study - everything in R:

Let us find the 95% confidence interval for $\mu_B - \mu_A$. Since the relevant t-quantile is, using $\nu = 15.99$,

$$t_{0.975} = 2.120$$

the confidence interval becomes:

$$10.298 - 8.293 \pm 2.120 \cdot \sqrt{\frac{2.0394}{9} + \frac{1.954}{9}}$$

which then gives the result as also seen above:

$$[0.59;\ 3.42]$$

Example - nutrition study - everything in R:

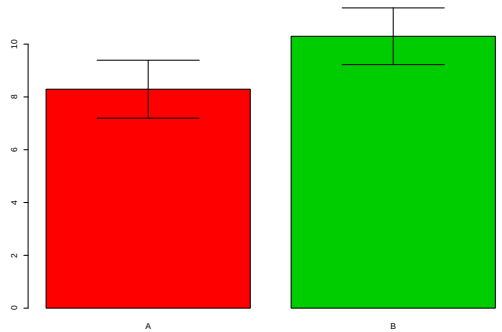
```
## Read the two-sample in R
xA=c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4, 10.88, 6.13, 7.9)
xB=c(9.21, 11.51, 12.79, 11.85, 9.97, 8.79, 9.69, 9.68, 9.19)
## A two sample Welch t-test
t.test(xB, xA)

##
##  Welch Two Sample t-test
##
## data:  xB and xA
## t = 3, df = 20, p-value = 0.008
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.59 3.42
## sample estimates:
## mean of x mean of y
##      10.3       8.3
```


Oversigt

Example - nutrition study - presentation of result

Barplot with *error bars* are often seen

A grouped barplot with some "error bars" - below the 95%-confidence intervals for each group is shown:



Be careful about using "overlapping confidence intervals"

The approach actually is 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 two hence is:

$$\sigma_{(\bar{X}_A-\bar{X}_B)} < \sigma_{\bar{X}_A} + \sigma_{\bar{X}_B}$$

Be careful about using "overlapping confidence intervals"

Remark 3.58. 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

Oversigt

Motivating example - sleeping medicine

Difference of sleeping medicines?

In a study the aim is to compare two kinds of sleeping medicine A and B. 10 test persons tried both kinds of medicine and the following results are obtained, given in 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
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)
## Take the difference to get a paired t-test
dif=x2-x1
## Calculate the test and results
t.test(dif)

##
##  One Sample t-test
##
## data:  dif
## t = 5, df = 9, p-value = 0.001
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.86 2.48
## sample estimates:
## mean of x
##       1.7
```

The paired setup and analysis = one-sample analysis

```
## Another way to calculate the paired setup
t.test(x2, x1, paired=TRUE)

##
##  Paired t-test
##
## data:  x2 and x1
## t = 5, df = 9, p-value = 0.001
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.86 2.48
## sample estimates:
## mean of the differences
##                     1.7
```


Paired versus independent experiment

Completely Randomized (independent samples)

20 patients are used and completely at random allocated to one of the two treatments (but usually making sure to have 10 patients in each group). So: different persons in the different 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 it becomes meaningful, and also one would typically make sure that some patients do A before B and others B before A. (and doing this allocation at random). So: the same persons in the different groups.

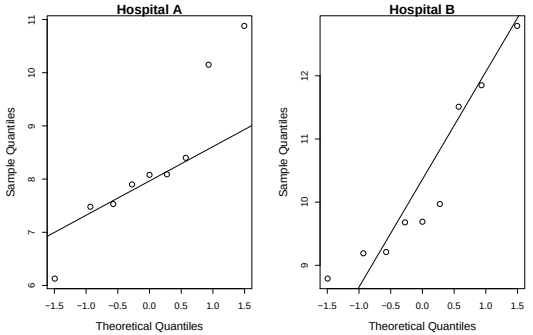
Example - Sleeping medicine - WRONG analysis

```
t.test(x1,x2)

##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -2, df = 20, p-value = 0.07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.49  0.15
## sample estimates:
## mean of x mean of y
##      0.66      2.33
```


Oversigt

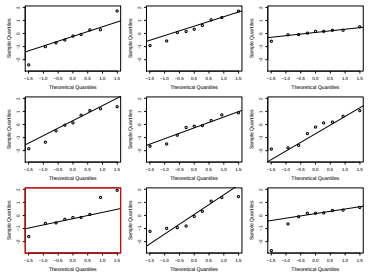
Example - Q-Q plot for EACH sample:

```
## Q-Q plot for each sample
par(mfrow=c(1,2))
qqnorm(xA, main="Hospital A")
qqline(xA)
qqnorm(xB, main="Hospital B")
qqline(xB)
```



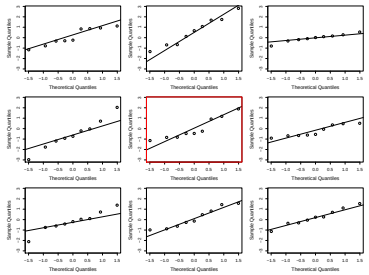
Example - Comparing with simulated, A

```
require(MESS)
fit1 <- lm(xA ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...)}
wallyplot(fit1, FUN=qqnorm.wally, main="")
```



Example - Comparing with simulated, B

```
## Multiple (simulated) Q-Q plots for each sample
fit1 <- lm(xB ~ 1)
qqnorm.wally <- function(x, y, ...) { qqnorm(y, ...); qqline(y, ...)}
wallyplot(fit1, FUN=qqnorm.wally, main="")
```



Oversigt

Planning of study with requirements to the precision

Method 3.62: The one-sample CI sample size formula:

When σ is known or guessed at 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:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

If we want that $ME = 3$cm with 95% confidence, how large should n then be?

$$n = \left(\frac{1.96 \cdot 12.21}{3} \right)^2 = 63.64$$

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

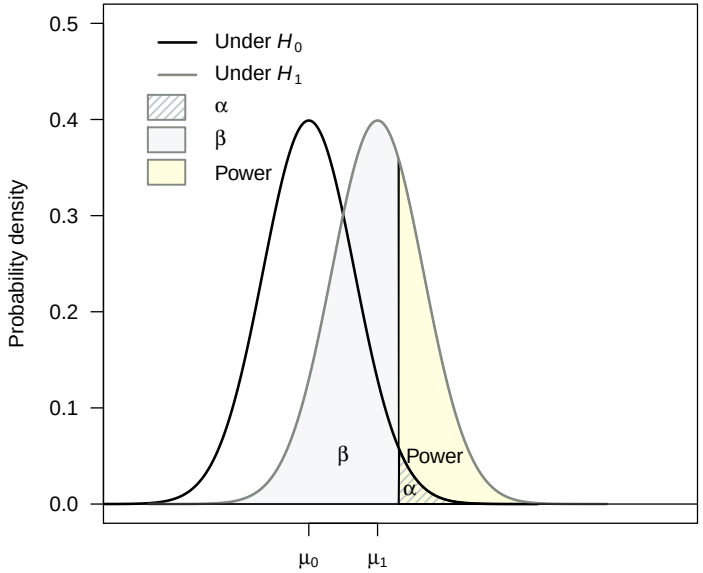
Planning and power

When the test to use has been set:

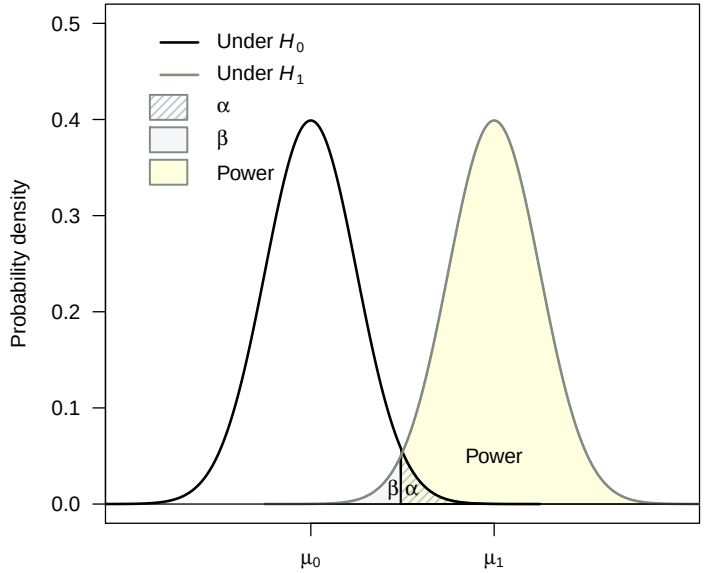
If you know (or set/guess) four out of the following five pieces of information, you can find the fifth:

- The sample size n
- Significance level α of the test.
- A change 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 (usually at least 80%):

Metode 3.64: The one-sample sample size formula:

For the one-sample t-test for given α, β and σ:

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)} \right)^2$$

Where $\mu_0 - \mu_1$ is the change in means that we would want to detect and $z_{1-\beta}$, $z_{1-\alpha/2}$ are quantiles of the standard normal distribution.

Example - The power for $n = 40$

```
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
##       sig.level = 0.05
##           power = 0.52
##     alternative = two.sided
```

Example - The sample size for power$= 0.80$

```
power.t.test(power = .80, delta = 4, sd = 12.21,
             type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 75
##           delta = 4
##              sd = 12
##       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 for 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 = 0.99
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```


Power and sample size - two-sample

Finding the sample size for detecting a group difference of 2 with $\sigma = 1$ and power$= 0.9$:

```
## Sample size calculation for two-sample
power.t.test(power = 0.90, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 6.4
##           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 calculation for 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 = 1.5
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Oversigt

The pooled two-sample t-test statistic

The *pooled* estimate of variance (assuming $\sigma_1^2 = \sigma_2^2$)

Method 3.51

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

The pooled test statistic, Method 3.52

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 3.53: 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' version

Almost (fool)proof to use the Welch-version always:

- if $s_1^2 = s_2^2$ the Welch and the Pooled test statistics are the same.
- Only when the two variances become really different the two test-statistics may differ in any important way, and 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 sizes in at least one of the two groups the pooled approach may provide slightly higher power if you believe in the equal variance assumption. And for these cases the Welch approach is then a somewhat cautious approach.

Agenda

1. Motivating example - nutrition study
2. p-values and hypothesis tests - repetition
3. Two-sample t-test and p-value
4. The confidence interval for the 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:

Statistics - Simulation based

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

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 intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

Oversigt

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 intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

Motivation

- Many (most?) relevant statistics("computed features") have complicated sampling distributions:
  - A trimmed mean
  - The median
  - Quantiles in general, i.e. IQR$= Q_3 - Q_1$
  - The coefficient of variation
  - ANY non-linear function of one or more input variables
  - (The standard deviation)
- The data distribution itself may be non-normal, complicating the statistical theory for even the simple mean
- We may HOPE for the magic of CLT (Central Limit Theorem)
- BUT but: We NEVER really no whether CLT is good enough - simulation can tell us!!
- Require : Use of computer - R is a super tool for this!

What is simulation really?

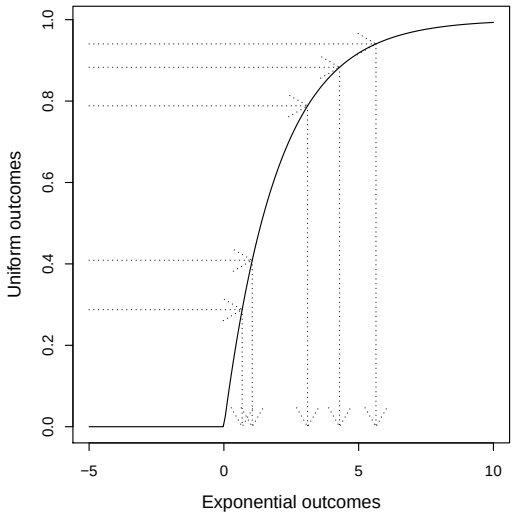
- (Pseudo) random numbers generated from a computer
- A random number generator is an algorithm that can generate x_{i+1} from x_i
- A sequence of numbers appears random
- Require a "start" called a "seed" (Using e.g. the computer clock)
- Basically the uniform distribution is simulated in this way, 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 $F^{-1}(U)$ follow the distribution given by F

Example: the exponential distribution, $\lambda = 0.5$:

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



In practice in R

Most distributions are ready for simulation, for instance:

`rbinom`	Binomial distribution
`rpois`	Poisson distribution
`rhyper`	The hypergeometric distribution
`rnorm`	normal distribution
`rlnorm`	log-normal distributions
`rexp`	exponential
`runif`	The uniform distribution
`rt`	t-distribution
`rchisq`	χ^2-distribution
`rf`	F distribution

Example: Area of plates

A company produces rectangular plates. The length of plates (in meters), X is assumed to follow a normal distribution $N(2, 0.01^2)$ and the width of the plates (in meters), Y are assumed to follow a normal distribution $N(3, 0.02^2)$. We are interested in the area of the plates which of course is given by $A = XY$.

- What is the mean area?
- What is the standard deviation in the areas from plate to plate?
- how often such plates have an area that differ by more than $0.1m^2$ from the targeted $6m^2$?
- The probability of other events?
- Generally: what is the probability distribution of the random variable A

Example: Area of plates, Solution by simulation

```
set.seed(345)
k = 10000 # Number of simulations
X = rnorm(k, 2, 0.01)
Y = rnorm(k, 3, 0.02)
A = X*Y
```

```
mean(A)

## [1] 6

sd(A)

## [1] 0.05

mean(abs(A-6)>0.1)

## [1] 0.044
```


Oversigt

Propagation of error

Must be able to find:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \text{Var}(f(X_1,\ldots,X_n))$$

We already know:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \sum_{i=1}^{n} a_i^2\sigma_i^2, \text{ if } f(X_1,\ldots,X_n) = \sum_{i=1}^{n} a_iX_i$$

Method 4.3: for non-linear functions:

$$\sigma^2_{f(X_1,\ldots,X_n)} \approx \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2$$

Example, cont.

We already used the simulation method in the first part of the example. Given two specific measurements of X and Y, $X = 2.00m$ and $y = 3.00m$. What is the variance of $A = 2.00 \times 3.00 = 6.00$ using the error propagation law?

Example, cont.

The varianses are:

$$\sigma_1^2 = Var(X) = 0.01^2 \text{ og } \sigma_2^2 = Var(Y) = 0.02^2$$

The function andn the derivarives are:

$$f(x,y) = xy, \ \frac{\partial f}{\partial x} = y, \ \frac{\partial f}{\partial y} = x$$

So the result becomes:

$$\begin{aligned} Var(A) &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_2^2 \\ &= y^2\sigma_1^2 + x^2\sigma_2^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 \\ &= 0.0025 \end{aligned}$$

Propagation of error - by simulation

Method 4.4: Error propagation by simulation

Assume we have actual measurements $x_1, \ldots, x_n$ with known/assumed error variances $\sigma_1^2, \ldots, \sigma_n^2$.

1. Simulate k outcomes of all n measurements from assumed error distributions, e.g. $N(x_i, \sigma_i^2)$: $X_i^{(j)}, j = 1 \ldots, k$
2. Calculate the standard deviation directly as the observed standard deviation of the k simulated values of f:

$$s^{\text{sim}}_{f(X_1,\ldots,X_n)} = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(f_j - \bar{f})^2}$$

where

$$f_j = f(X_1^{(j)}, \ldots, X_n^{(j)})$$

Example, Area, cont.

Actually one can deduce the variance of A theoretically,

$$\begin{aligned} \text{Var}(XY) &= \text{E}\left[(XY)^2\right] - [\text{E}(XY)]^2 \\ &= \text{E}(X^2)\text{E}(Y^2) - \text{E}(X)^2\text{E}(Y)^2 \\ &= \left[\text{Var}(X) + \text{E}(X)^2\right]\left[\text{Var}(Y) + \text{E}(Y)^2\right] - \text{E}(X)^2\text{E}(Y)^2 \\ &= \text{Var}(X)\text{Var}(Y) + \text{Var}(X)\text{E}(Y)^2 + \text{Var}(Y)\text{E}(X)^2 \\ &= 0.01^2 \times 0.02^2 + 0.01^2 \times 3^2 + 0.02^2 \times 2^2 \\ &= 0.00000004 + 0.0009 + 0.0016 \\ &= 0.00250004 \end{aligned}$$

Example, Area, cont. - in summary

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 such other quantities could be much more complicated than what was shown for the variance here)
2. It offers a simple tool to use any other distribution than the normal, if we believe such better reflect reality.
3. It does not rely on any linear approximations of the true non-linear

Oversigt

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 the exponential rate or 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?

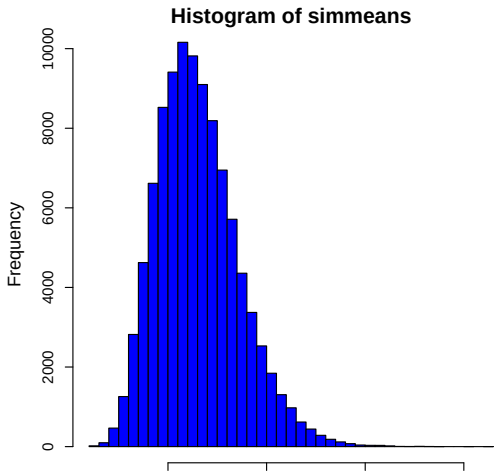
Based on previous knowledge in this course: We don't know!

Example: Confidence interval for the exponential rate or mean

```
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876.543)
simsamples <- replicate(k, rexp(10, 1/26.08))
## 2. Compute the mean of the 10 simulated observations k times:
simmeans <- apply(simsamples, 2, mean)
## 3. Find the two relevant quantiles of the k simulated means:
quantile(simmeans, c(0.025, 0.975))

## 2.5%  98%
##   13   45
```

Example: Confidence interval for the exponential rate or mean

```
hist(simmeans, col="blue", nclass=30)
```



Example: Confidence interval for the median of an exponential distribution

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 a an exponential distribution

What is the confidence interval for the median?

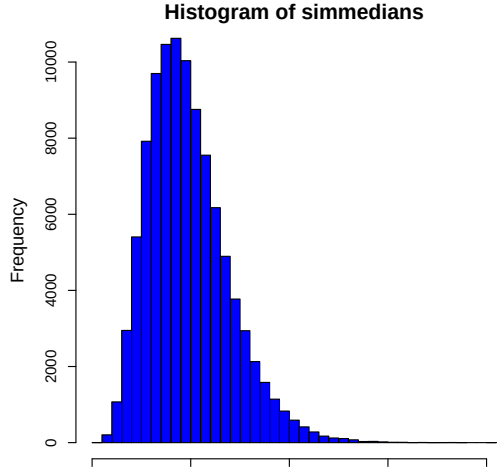
Based on previous knowledge in this course: We don't know!

Example: Confidence interval for the median of an exponential

```
## Set the number of simulations:
k <- 100000
## 1. Simulate 10 exponentials with the right mean k times:
set.seed(9876.543)
simsamples <- replicate(k, rexp(10, 1/26.08))
## 2. Compute the median of the n=1010 simulated observations k times:
simmedians <- apply(simsamples, 2, median)
## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simmedians, c(0.025, 0.975))

## 2.5%  98%
##    7   38
```


Example: Confidence interval for the median of an exponential

```
hist(simmedians, col="blue", nclass=30)
```



Confidence interval for any feature (including μ)

Method 4.7: Confidence interval for any feature θ by parametric bootstrap

Assume we have actual observations $x_1, \ldots, x_n$ and assume that they stem 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 $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_k$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$

[a] And otherwise chosen to match the data as good as possible: Some distributions have more than just a single 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 by the so-called *maximum likelihood* approach

Another example: 99% confidence interval for Q_3 assuming a normal distribution

```
## Read in the 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 the number of simulations:
k <- 100000
## 1. Simulate k samples of n=10 normals with the right mean and variance:
set.seed(9876.543)
simsamples <- replicate(k, rnorm(n, mean(x), sd(x)))
## 2. Compute the Q3 of the n=10 simulated observations k times:
simQ3s <- apply(simsamples, 2, Q3)
## 3. Find the two relevant quantiles of the k simulated medians:
quantile(simQ3s, c(0.005, 0.995))

## 0.5% 100%
##  173  198
```


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 $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ and assume that they stem from some probability distributions with density 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 $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$

[a] As before

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)
## 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)
n1 <- length(x)
n2 <- length(y)
```

Example: Confidence interval for the difference of exponential means

```
## Set the number of simulations:
k <- 100000
## 1. Simulate k samples of each n1=10 and n2=12
## exponentials with the right means:
set.seed(9876.543)
simXsamples <- replicate(k, rexp(n1, 1/mean(x)))
simYsamples <- replicate(k, rexp(n2, 1/mean(y)))
## 2. Compute the difference between the simulated
## means k times:
simDifmeans <- apply(simXsamples, 2, mean) -
                    apply(simYsamples, 2, mean)
## 3. Find the two relevant quantiles of the
## k simulated differences of means:
quantile(simDifmeans, c(0.025, 0.975))

## 2.5%  98%
##  -41   14
```


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

Oversigt

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: Women's cigarette consumption

In a study women's cigarette consumption before and after giving birth is explored. The following observations of the number of smoked cigarettes per day were the results:

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: Women's cigarette consumption

A paired t-test setting, BUT with clearly non-normal 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)
dif <- x1-x2
dif

##  [1]  3 13  7  5  6  0 -2 -4 -1 22  9

mean(dif)

## [1] 5.3
```


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: Women's cigarette consumption - the non-parametric results:

```
k = 100000

simsamples = replicate(k, sample(dif, replace = TRUE))
simmeans = apply(simsamples, 2, mean)
quantile(simmeans, c(0.025,0.975))

## 2.5%  98%
##  1.4  9.8
```

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 $x_1, \ldots, x_n$.

1. Simulate k samples of size n by randomly sampling among the available data (with replacement)
2. Calculate the statistic $\hat{\theta}$ in each of the k samples $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%},\ q^*_{100(1-\alpha/2)\%}\right]$

Example: Women's cigarette consumption

Let us find the 95% confidence interval for the median cigarette consumption change in the example from above:

```
k = 100000
simsamples = replicate(k, sample(dif, replace = TRUE))
simmedians = apply(simsamples, 2, median)
quantile(simmedians, c(0.025,0.975))

## 2.5%  98%
##   -1    9
```


Example: Tooth health and infant bottle use

In a study it was explored whether children who received milk from bottle as a child had worse or better teeth health conditions than those who had not received milk from the bottle. For 19 randomly selected children is was recorded when they 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 no group:
 x <- c(9, 10, 12, 6, 10, 8, 6, 20, 12)
## Reading in yes group:
 y <- c(14,15,19,12,13,13,16,14,9,12)

k <- 100000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
simmeandifs <- apply(simxsamples, 2, mean)-
                          apply(simysamples, 2, mean)
quantile(simmeandifs, c(0.025,0.975))

##  2.5%   98%
## -6.23 -0.14
```

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 $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.

1. Simulate k sets of 2 samples of n_1 and n_2 observations from the respective groups (with replacement)
2. Calculate the difference between the features in each of the k samples $\hat{\theta}^*_{x1} - \hat{\theta}^*_{y1}, \ldots, \hat{\theta}^*_{xk} - \hat{\theta}^*_{yk}$.
3. Find the $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles for these, $q^*_{100(\alpha/2)\%}$ and $q^*_{100(1-\alpha/2)\%}$ as the $100(1-\alpha)\%$ confidence interval: $\left[q^*_{100(\alpha/2)\%}, q^*_{100(1-\alpha/2)\%}\right]$

Example: Tooth health and infant bottle use - a 99% confidence interval for the difference of medians

```
k <- 100000
simxsamples <- replicate(k, sample(x, replace = TRUE))
simysamples <- replicate(k, sample(y, replace = TRUE))
simmediandifs <- apply(simxsamples, 2, median)-
                         apply(simysamples, 2, median)
quantile(simmediandifs, c(0.005,0.995))

## 0.5% 100%
##   -8    0
```


Bootstrapping - an overview

We were given 4 similar method boxes

1. With distribution or not (parametric or non-parametric)
2. For one- or two-sample analysis

Note:

Means also included in *other features*. Or: These methods can be used not only for means!!

Hypothesis testing also possible

We can do hypothesis testing by looking at the confidence intervals!

Agenda

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 intervals assuming any distributions
4. Non-parametric bootstrap
   - One-sample confidence interval for any feature
   - Two-sample confidence intervals

Course 02402 Introduction to Statistics Lecture 8:

Simple linear regression

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

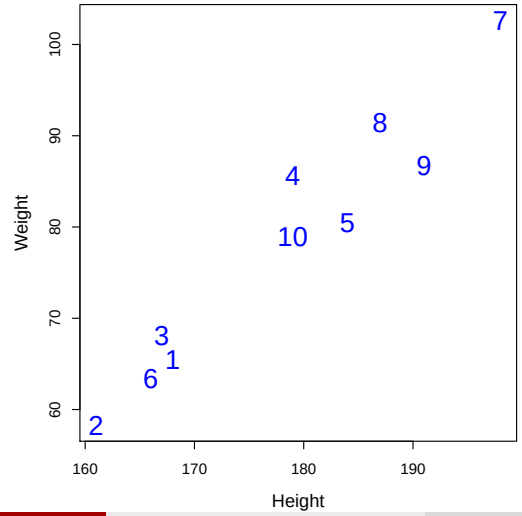
Agenda

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 interval for the line
7. Summary of summary(lm(y∼x))
8. Correlation
9. Residual Analysis: Model control

Oversigt

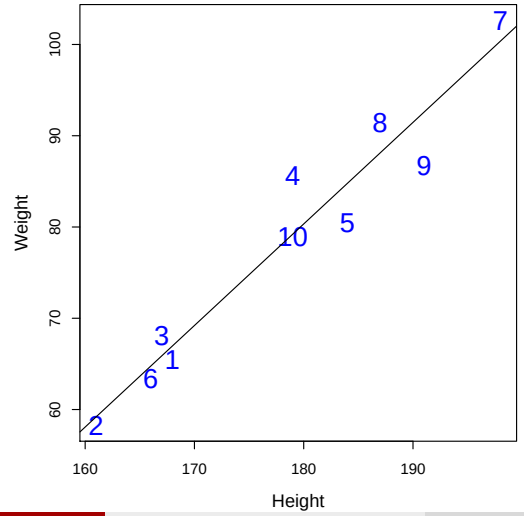
Example: Height-Weight

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

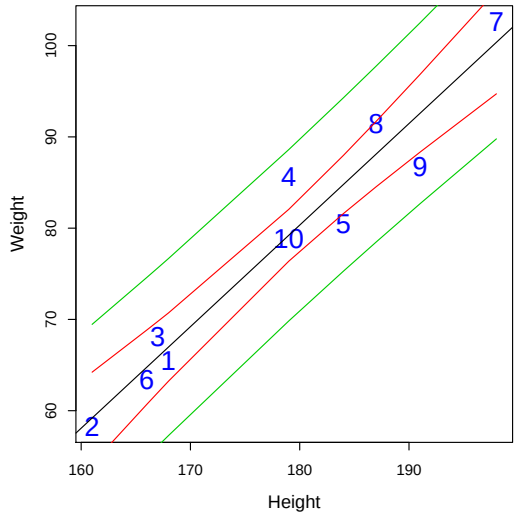
Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



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: 3.9 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
```

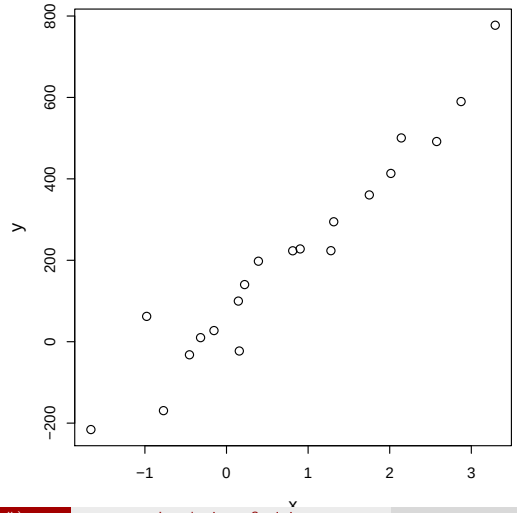

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Oversigt

A scatter plot of some data

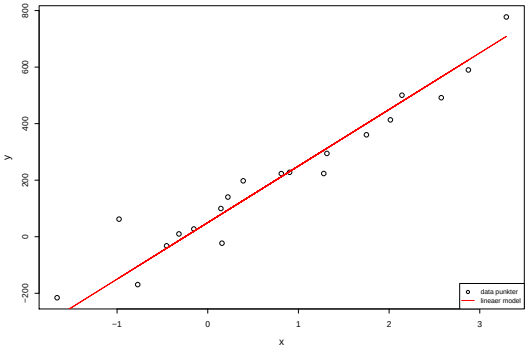
- We have n pairs of data points (x_i, y_i)



Express a linear model

- Express a linear model

$$y_i = \beta_0 + \beta_1 x_i$$



but something is missing in the desrcription of the *random variation*!

Express a linear regression model

- Express the *linear regression model*

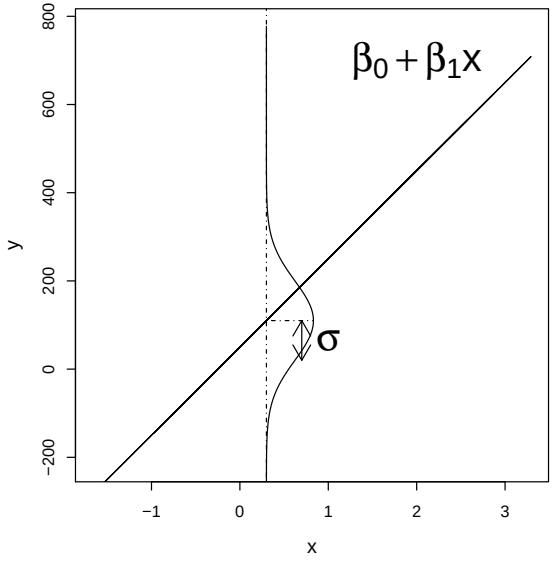
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Y_i is the *dependent variabel*. A random variable.
- x_i er en *explanatory variable*. Given numbers.
- ε_i is the deviation (error). A random variable.

and we assume

ε_i is independent and identically distributed (i.i.d.) and $N(0, \sigma^2)$

Model illustration



Oversigt

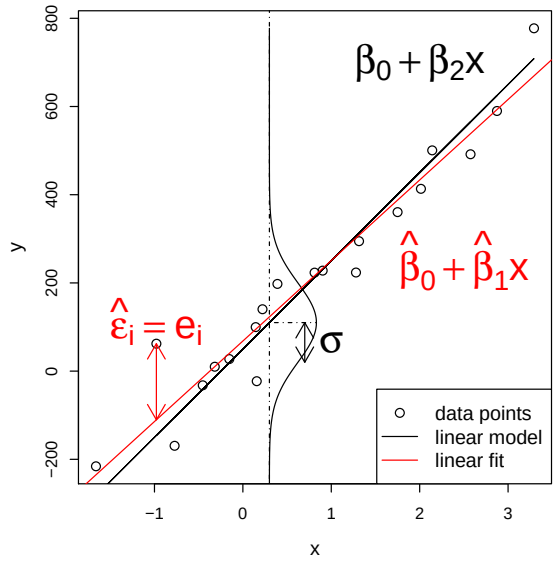
Least Squares Method

- How can we estimate the parameters β_0 and β_1?
- Good idea: Minimize the variance σ^2 of the residuals. It is in almost any way the best choice in this setup.
- But how!?
- Minimize the sum of the Residual Sum of Squares (RSS))

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ minimizes 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 estimatates 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 $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Don't think too much about this for now!

R example

```
## Simulate a  linear model with  normally distrubuted
## errors and estimate the parameters

## FiRST MAKE DATA:
## Generates 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: as  real data analysis, we have th data in x and y:
## A scatter plot of x and y
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")
```


Oversigt

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 is the (sampling) distribution of the parameter estimators?

in a linear regression model (given normal distributed errors)?

Try to simulate to have a look at this...

Let's go to R!!

- What is the (sampling) distribution of the parameter estimates in a linear regression model (given normal distributed errors)?
- Answer: They are normally distributed (for $n < 30$ use the t-distribution) and their variance can be estimated:

Theorem 5.7 (first part)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}}$$
$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$$
$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}$$

- The Covariance $Cov[\hat{\beta}_0, \hat{\beta}_1]$ we do not use for anything for now..

Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$

Theorem 5.7 (second part)

Where σ^2 is usually replaced by its estimate ($\hat{\sigma}^2$). The central estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of σ^2 is used the variances also become estimates and we'll refer to them as $\hat{\sigma}^2_{\hat{\beta}_0}$ and $\hat{\sigma}^2_{\hat{\beta}_1}$.

Estimates of standard deviations of $\hat{\beta}_0$ and $\hat{\beta}_1$ (equations 5-41 and 5-42)

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}; \quad \hat{\sigma}_{\beta_1} = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Oversigt

Hypothesis tests for the parameters

- We can carry out hypothesis tests for the parameters in a linear regression model:

$$H_{0,i}: \quad \beta_i = \beta_{0,i}$$
$$H_{1,i}: \quad \beta_i \neq \beta_{1,i}$$

- We use the t-distributed statistics:

Theorem 5.11

Under the null-hypothesis ($\beta_0 = \beta_{0,0}$ 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.

- See Example 5.12 for example of hypothesis test.
- Test if the parameters are signifikantly different from 0

$$H_{0,i}: \quad \beta_i = 0$$
$$H_{1,i}: \quad \beta_i \neq 0$$

- See the resultats in R

```
## Hypothesis tests om signifikante parametre

## 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 find the estimates
fit <- lm(y ~ x)

## See summary - what we need
summary(fit)
```


Confidence intervals for the parameters

Method 5.14

$(1-\alpha)$ 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 $\hat{\sigma}_{\beta_1}$ are found from equations 5-41 and 5-42
- in R we can read off $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ udner "Std. Error" from "summary(fit)"

Simulation illustration of CIs

```
## Make confidence intervals for the parameters

## number of repeats
nRepeat <- 100

## Did we catch the correct parameter
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 find the estimates
  fit <- lm(y ~ x)

  ## Luckily  R can compute the  confidence interval (level=1-alpha)
  (ci <- confint(fit, "(Intercept)", level=0.95))

  ## Was the correct parameter value "caught" by the interval? (covered)
  (TrueValInCI[i] <-  ci[1] < beta0  &  beta0 < ci[2])
}

## How often did this happen?
sum(TrueValInCI) / nRepeat
```


Oversigt

Method 5.17 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 in the point x_0
- Is computed by

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- The confidence interval will in $100(1-\alpha)\%$ of the times contain the correct line, that is $\beta_0 + \beta_1 x_0$

Method 5.17 Prediction interval for $\beta_0 + \beta_1 x_0 + \varepsilon_0$

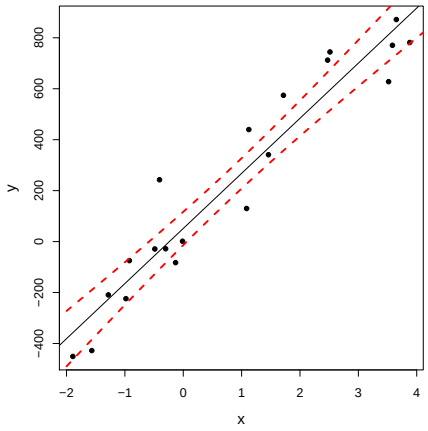
- The prediction interval for Y_0 is found using a value x_0
- This is done *before* Y_0 is observered with

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- The prediction intervallet will in $100(1-\alpha)\%$ of the times contain the observered y_0
- A prediction intervalis wider than a confidence interval for a given α

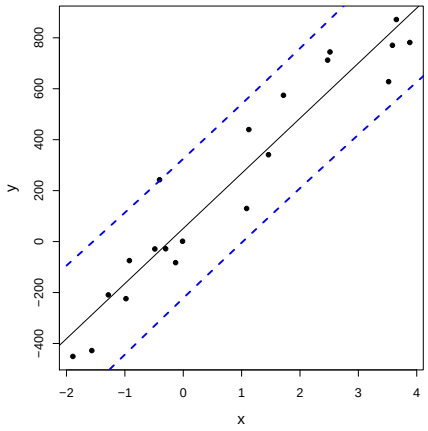
Example of confidence interval for the line

```
## Example of confidence interval for the line

## Make a sequence of 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=.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 interval

```
## Example with prediction interval

## Make a sequence of 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=.95)

## Check what we got
head(CI)

## 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)
```



Oversigt

What more do we get from summary?

```
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -184.7  -96.4  -20.3   86.6  279.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     51.5       31.1    1.66     0.12
## x              216.3       15.2   14.22  3.1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126 on 18 degrees of freedom
## Multiple R-squared:  0.918,Adjusted R-squared:  0.914
## F-statistic:  202 on 1 and 18 DF,  p-value: 3.14e-11
```


summary(lm(y∼x)) wrap up

- `Residuals: Min 1Q Median 3Q Max:`
  The residuals': Minimum, 1. quartile, Median, 3. quartile, Maximum
- `Coefficients:`
  `Estimate Std. Error t value Pr(>|t|) "stars"`
  The coefficients':
  Estimate $\hat{\sigma}_{\beta_i}$ t_{obs} p-value
  - The test is $H_{0,i}: \beta_i = 0$ vs. $H_{1,i}: \beta_i \neq 0$
  - The stars is showing the size categories of the p-value
- `Residual standard error: XXX on XXX degrees of freedom`
  $\varepsilon_i \sim N(0, \sigma^2)$ printed is $\hat{\sigma}$ and ν degrees of freedom (used for hypothesis test)
- `Multiple R-squared: XXX`
  Explained variation r^2
- The rest we do not use in this course

Oversigt

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

$$\hat{\rho} = r = \sqrt{r^2}\, sgn(\hat{\beta}_1)$$

where $sgn(\hat{\beta}_1)$ er: -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

R Illustration

```
## Generates 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)

## Scatter plot
plot(x,y)

## Use lm() to find the estimates
fit <- lm(y ~ x)

## The "true" line
abline(beta0, beta1)
## Plot of fit
abline(fit, col="red")

## See summary
summary(fit)

## Correlation between x and y
cor(x,y)

## Squared becomes the "Multiple R-squared" from summary(fit)
cor(x,y)^2
```

Oversigt

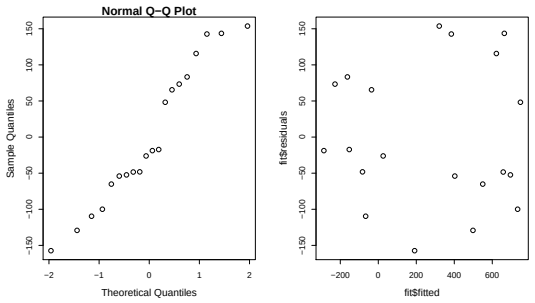
Residual Analysis

Method 5.26

- Check normality assumption with qq-plot.
- Check (non)systematic behavior by plotting the residuals e_i as a function of fitted values $\hat{y}_i$

Residual Analysis in R

```
fit <- lm(y ~ x)
par(mfrow = c(1, 2))
qqnorm(fit$residuals)
plot(fit$fitted, fit$residuals)
```



OR: Wally plot again!

Agenda

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 interval for the line
7. Summary of summary(lm(y∼x))
8. Correlation
9. Residual Analysis: Model control

Course 02402 Introduction to Statistics Lecture 9:

Multiple linear regression

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

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

Oversigt

Example: Ozon concentration

We have a set of observations of: logarithm to ozone concentration (log(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

```r
## Se 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, se ?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

```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

```r
################################
## 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))
```

Oversigt

Multiple linear regression

- Y is the *dependent variable*
- We are interested in modelling the Y's dependency of the *independent* or *explanatory* variables $x_1, x_2, ..., x_p$
- We are modelling a *linear relation* between Y and $x_1, x_2, ..., x_p$, described with the regression model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ and i.i.d.}$$

- Y_i og ε_i are random variables and $x_{j,i}$ are variables

Least squares estimates

- The coefficient estimates are found by minimizing:

$$RSS(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})]^2$$

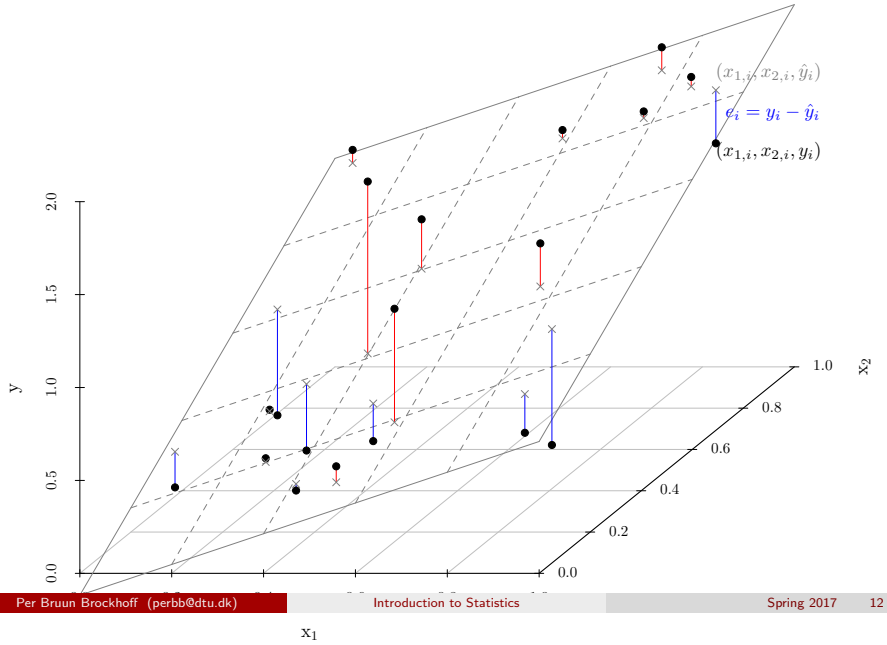
- The "predicted" (= "fitted") are found as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}$$

- And then the residuals are found as

$$e_i = y_i - \hat{y}_i$$
$$\text{residual} = \text{observation} - \text{prediction}$$

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)`)
- 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
- 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!!

Oversigt

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

Oversigt

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")
```


Oversigt

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
fitWindSqTemperaturSq <- lm(logOzone ~ temperature + temperature2 + wind + windSq + radiation, data=Air)
summary(fitWindSqTemperaturSq)

## Model validation
qqnorm(fitWindSq$residuals)
qqline(fitWindSq$residuals)
plot(fitWindSq$residuals, fitWindSq$fitted.values, pch=19)
```


Oversigt

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)
```


Oversigt

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

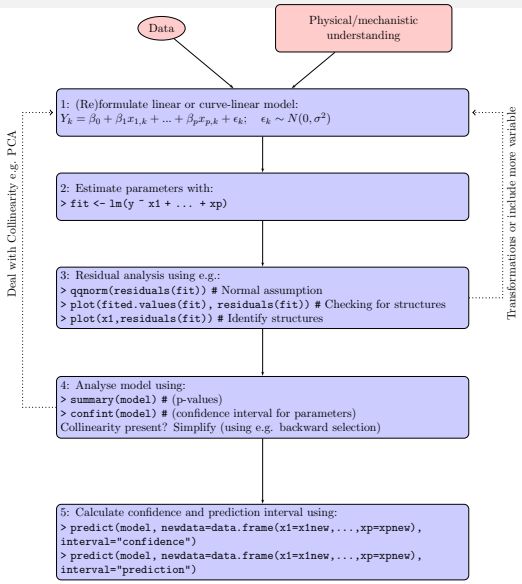
Colinearity - an illustration in R

```
##############################
## 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!

Oversigt

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:

Oneway Analysis of Variance, ANOVA

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

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 the relation to 2-Group t-test
6. Post hoc analysis
7. Model control
8. A complete example - from the book

Oversigt

Oneway ANOVA - 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 eacg group can bve assumed to be normally distributed.

TV set development at Bang & Olufsen

Sound and image quality is measured by th human perceptual instrument:

We developed a tool that is used by B&O to ANOVA (among other things) PanelCheck (*Show Panelcheck programme with TV data*)

Bang & Olufsen data in R:

```
 # Getting the Bang and Olufsen data from the lmerTest-package:
library(lmerTest) # (Developed by us)
data(TVbo)
head(TVbo)
# Defining the factor identifying the 12 TVset and Picture combs:
TVbo$TVPic <- factor(TVbo$TVset:TVbo$Picture)
# Each of 8 assessors scored each of 12 combinations 2 times
# Averaging the two replicates for each Assessor and TVpic:
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: (Much better analysis - next week)
anova(lm(Noise ~ Assessor + TVPic, data = TVbonoise))
```


Oneway ANOVA - example

```
####################################
## Input data and plot

## 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)

## Groups (treatments)
treatm <- factor(c(1, 1, 1, 1,
                   2, 2, 2, 2,
                   3, 3, 3, 3))

## Plot
par(mfrow=c(1,2))
plot(as.numeric(treatm), y, xlab="Treatment", ylab="y")
##
plot(treatm, y, xlab="Treatment", ylab="y")
```


Oversigt

Oneway ANOVA, model

- Express the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

  where it is assumed that

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

- μ is the overall mean
- α_i is the effect of Group (treatment) i
- j indicates the measurements in the groups, from 1 to n_i in each Group

Oneway ANOVA, hypothesis

- We want to compare (more than 2) means $\mu + \alpha_i$ in the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- So we can express the hypothesis:

$$\begin{aligned} H_0: &\quad \alpha_i = 0 \quad \text{for all } i \\ H_1: &\quad \alpha_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

Oversigt

Oneway ANOVA, decomposition and the ANOVA table

- With the model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- the total variation in the data can be decomposed:

$$SST = SS(Tr) + SSE$$

- 'Oneway' 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")

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

- 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

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	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
```


Example

```
## Number of Groups
k <- 3
## Number in each  Group
ni <- 10
## Simulate data from model with 3 means
yModel1 <- rep( c(4, 5, -3), each=ni) + rnorm(ni*k, sd=1)
## Simulate data from model with 3 other means
yModel2 <- rep( c(1, 3, 1), each=ni) + rnorm(ni*k, sd=1)
## 3 Groups
group <- rep(1:k, each=ni)
## Plot them
par(mfrow=c(1,2))
plot(group, yModel1, ylim=range(yModel1,yModel2))
plot(group, yModel2, ylim=range(yModel1,yModel2))

## Compute SST: total variance, which is highest?
(SST1 <- sum( (yModel1 - mean(yModel1))^2 ))
(SST2 <- sum( (yModel2 - mean(yModel2))^2 ))

## Compute SSE: total residual variation, which is highest?
(SSE1 <- sum(tapply(yModel1, group, function(x){ sum((x - mean(x))^2) })))
(SSE2 <- sum(tapply(yModel2, group, function(x){ sum((x - mean(x))^2) })))
```


Oversigt

Oneway ANOVA, F-test

- We have: (Theorem 8.2)

$$SST = SS(Tr) + SSE$$

- and can find the test statistic:

$$F = \frac{SS(Tr)/(k-1)}{SSE/(n-k)}$$

where

- k is the number of levels of the factor
- n is the total number of observations

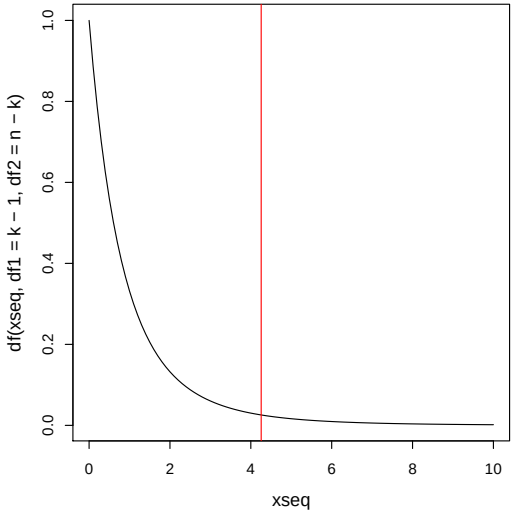
- The significance level α is chosen and the test statistic F is computed
- The test statistic is compared with a quantile in the F distribution

$$F \sim F_\alpha(k-1, n-k) \text{ (Theorem 8.6)}$$

The F-distribution

```
################################
## Plot the F distribution  and see the critical value

## Remember, this is "under H0" (that is we compute as if H0 is true):
## Number of Groups
k <- 3
## number of observations
n <- 12
## Sequence for plot
xseq <- seq(0, 10, by=0.1)
## Plot the density of the  F distribution
plot(xseq, df(xseq, df1=k-1, df2=n-k), type="l")
##The critical value for significance level 5 %
cr <- qf(0.95, df1=k-1, df2=n-k)
## Mark it in the plot
abline(v=cr, col="red")
## The value of the  test statistic
(F <- (SSTr/(k-1)) / (SSE/(n-k)))
## The p-value hence is:
(1 - pf(F, df1=k-1, df2=n-k))
```


The F-distribution



The ANOVA table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{obs}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{obs}})$
Residual	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	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
```

Oversigt

Within-Group variability and the relation to 2-Group 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:

$$MSE = \frac{SSE}{n-k} = \frac{(n_1-1)s_1^2 + \cdots + (n_k-1)s_k^2}{n-k} \quad (1)$$

$$s_i^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

IF $k = 2$:(cf. Method 3.51)

$$\text{For } k = 2: \; MSE = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n-2}$$

$$\text{For } k = 2: \; F_{\text{obs}} = t_{\text{obs}}^2$$

where t_{obs} is the pooled version coming from Methods 3.51 and 3.52.

Oversigt

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 computationen 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 tests:

$$H_0: \mu_i = \mu_j, \; H_1: \mu_i \neq \mu_j$$

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{3}$$

and:

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

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

- HIf all $M = k(k-1)/2$ combinations of pairwise hypothesis tests are carried out use the approach M times but each time with test level $\alpha_{\text{Bonferroni}} = \alpha/M$.

Oversigt

Variance homogeneity

Look at box-plot to check whether the variability seems different for the groups

```
#################################
## Check assumption of  homogeneous variance

## Box plot
plot(treatm,y)
```


Normal assumption

Look at qq-normal plot

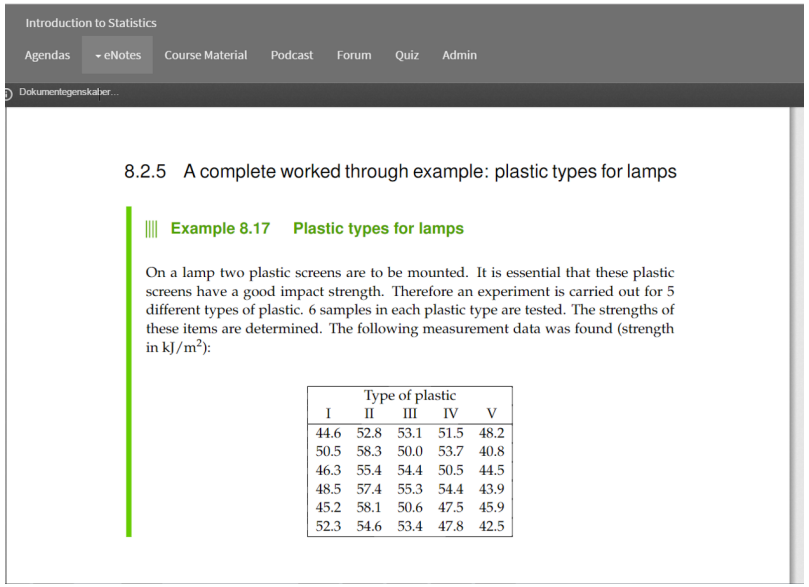
```
#################################
## Check the assumption  of  normality of residuals

## qq-normal plot of residuals
fit1 <- lm(y ~ treatm)
qqnorm(fit1$residuals)
qqline(fit1$residuals)

## Or with a  Wally plot
require(MESS)
qqwrap <- function(x, y, ...) {qqnorm(y, main="",...);
qqline(y)}
## Can we see a deviating qq-norm plot?
wallyplot(fit1$residuals, FUN = qqwrap)
```

Next week: Two-way ANOVA

```
 # Getting the Bang and Olufsen data from the lmerTest-package:
library(lmerTest) # (Developed by us)
data(TVbo)
head(TVbo)
# Defining the factor identifying the 12 TVset and Picture combs:

TVbo$TVPic <- factor(TVbo$TVset:TVbo$Picture)
# Each of 8 assessors scored each of 12 combinations 2 times
# Averaging the two replicates for each Assessor and TVpic:
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: (Much better analysis - next week)
anova(lm(Noise ~ Assessor + TVPic, data = TVbonoise))
```


Oversigt

A complete example - from the book



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

Agenda

Course 02402 Introduction to Statistics Lecture 11:

Twoway Analysis of Variance, ANOVA

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

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
7. A complete example - from the book

Oversigt

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
7. A complete example - from the book

TV set development at Bang & Olufsen

Sound and image quality is measured by th human perceptual instrument:

We developed a tool that is used by B&O to ANOVA (among other things) PanelCheck (*Show Panelcheck programme with TV data*)

Bang & Olufsen data in R:

```
##  # Getting the Bang and Olufsen data from the lmerTest-package:
library(lmerTest) # (Developed by us)
data(TVbo)

# Each of 8 assessors scored each of 12 combinations 2 times
# Let's look at only a single picture and one of the two reps:
# And let us look at the sharpness
TVbosubset <- subset(TVbo,Picture==1 & Repeat==1)[,c(1, 2, 9)]

sharp <- matrix(TVbosubset$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

Twoway ANOVA - example

- Same data as for oneway, but now we know that the experiment was split in 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 similarly
- *oneway* vs. *twoway* ANOVA
- *Completely randomized design* vs. *Randomized block design*

Twoway ANOVA - example

- Same data as for oneway, but now we know that the experiment was split in 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) normally distributed - OR if n_is are large enough (CLT)

The toy data in R

```
###################################
## Input data and plot

## 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))

## for later formulas
(k <- length(unique(treatm)))
(l <- length(unique(block)))

## Plots
par(mfrow=c(1,2))

## Plot histogramms by treatments
plot(treatm, y, xlab="Treatments", ylab="y")
## Plot histogrammer by  blocks
plot(block, y, xlab="Blocks", ylab="y")
```


Oversigt

Twoway ANOVA, model

- Express a model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

where the deviations

$$\varepsilon_{ij} \sim N(0, \sigma^2) \text{ and i.i.d.}$$

  - μ is the overall mean
  - α_i is the effect of treatment i
  - β_j is the level for Block i
  - there are k treatments and l blocks
  - j indicates the observations in the groups, from 1 to n_i for treatment i

Estimates of parameters in the model

- We can compute the estimates of the parameters ($\hat{\mu}$ and $\hat{\alpha}_i$, and $\hat{\beta}_j$)

$$\hat{\mu} = \bar{y} = \frac{1}{k \cdot l} \sum_{i=1}^{k} \sum_{j=1}^{l} y_{ij}$$

$$\hat{\alpha}_i = \left(\frac{1}{l} \sum_{j=1}^{l} y_{ij} \right) - \hat{\mu}$$

$$\hat{\beta}_j = \left(\frac{1}{k} \sum_{i=1}^{k} y_{ij} \right) - \hat{\mu}$$

```
###################################
## Compute estimates of parameters in the model

## Sample mean
(muHat <- mean(y))
## Sample mean for each treatment
(alphaHat <- tapply(y, treatm, mean) - muHat)
## Sample mean for each  Block
(betaHat <- tapply(y, block, mean) - muHat)
```

Oversigt

Twoway ANOVA, decomposition and the ANOVA table, Theorem 8.20

- With the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- the total variation in the data can be decomposed:

$$SST = SS(Tr) + SS(Bl) + SSE$$

- 'twoway' refers to the fact that there are two factors in the experiment(Two "ways"of the data table)
- 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") (same as for oneway)

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij} - \hat{\mu})^2$$

- treatment sum of squares ("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")

$$SS(Bl) = k \cdot \sum_{j=1}^{l}\hat{\beta}_j^2$$

- The sum of squares for the residuals ("residual variance after model fit")

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{l}(y_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu})^2$$

Oversigt

Twoway ANOVA: hypothesis of no effect of treatment, Theorem 8.22

- We want to compare (more than 2) means $\mu + \alpha_i$ in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- So we can express the hypothesis:

$$\begin{aligned} H_{0,Tr}: &\quad \alpha_i = 0 \quad \text{for all } i \\ H_{1,Tr}: &\quad \alpha_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

- 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

Twoway ANOVA: hypothesis of no effect of persons (blocks), Theorem 8.22

- We want to compare (more than 2) means $\mu + \beta_i$ in the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- So we can express the hypothesis

$$\begin{aligned} H_{0,Bl}: &\quad \beta_i = 0 \quad \text{for all } i \\ H_{1,Bl}: &\quad \beta_i \neq 0 \quad \text{for at least one } i \end{aligned}$$

- 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 treatments hypothesis

```
################################
## Plot the F distribution and see the critical value for treatments

## Remember, this is "under H0" (that is we compute as if H0 is true):
## Sequence for plot
xseq <- seq(0, 10, by=0.1)
## Plot the density of the  F distribution
plot(xseq, df(xseq, df1=k-1, df2=(k-1)*(l-1)), type="l")
##The critical value for significance level 5 %
cr <- qf(0.95, df1=k-1, df2=(k-1)*(l-1))
## Mark it in the plot
abline(v=cr, col="red")
## The value of the  test statistic
  (Ftr <- (SSTr/(k-1)) / (SSE/((k-1)*(l-1))))
## The p-value hence is:
(1 - pf(Ftr, df1=k-1, df2=(k-1)*(l-1)))
```

F-distribution and blocks hypothesis

```
################################
## Plot the F distribution  and see the critical value

## Remember, this is "under H0" (that is we compute as if H0 is true):
## Sequence for plot
xseq <- seq(0, 10, by=0.1)
## Plot the density of the  F distribution
plot(xseq, df(xseq, df1=l-1, df2=(k-1)*(l-1)), type="l")
##The critical value for significance level 5 %
cr <- qf(0.95, df1=l-1, df2=(k-1)*(l-1))
## Mark it in the plot
abline(v=cr, col="red")
## The value of the  test statistic
(Fbl <- (SSBl/(l-1)) / (SSE/((k-1)*(l-1))))
## The p-value hence is:
(1 - pf(Fbl, df1=l-1, df2=(k-1)*(l-1)))
```


The twoway ANOVA table

Source of variation	Deg. of freedom	Sums of squares	Mean sum of squares	Test-statistic F	p-value
Treatment	$k-1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k-1}$	$F_{\text{Tr}} = \frac{MS(Tr)}{MSE}$	$P(F > F_{\text{Tr}})$
Block	$l-1$	$SS(Bl)$	$MS(Bl) = \frac{SS(Bl)}{l-1}$	$F_{\text{Bl}} = \frac{MS(Bl)}{MSE}$	$P(F > F_{\text{Bl}})$
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
```


Oversigt

Post hoc confidence interval

- As for oneway (Use methods 8.9 and 8.10) substitute $(n-k)$ degrees of freedom with $(k-1)(l-1)$ (and use MSE from twoway).
- Can be done with either treatments or blocks
- A single pre-planned confidence interval for the difference between treatment i and j is found as:

$$\bar{y}_i - \bar{y}_j \pm t_{1-\alpha/2} \sqrt{\frac{SSE}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \quad (1)$$

  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:

$$H_0 : \mu_i = \mu_j, \; H_1 : \mu_i \neq \mu_j$$

is carried out as:

$$t_{\text{obs}} = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{2}$$

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 $\alpha_{\text{Bonferroni}} = \alpha/M$.

Oversigt

Variance homogeneity

Look at box-plot to check whether the variability seems different for the groups

```
################################
## Check assumption of  homogeneous variance


## Save the  fit
fit <- lm(y ~ treatm + block)
## Box plot
par(mfrow=c(1,2))
plot(treatm, fit$residuals, y, xlab="Treatment")
## Box plot
plot(block, fit$residuals, xlab="Block")
```

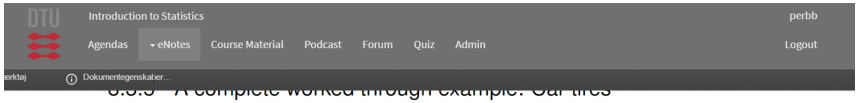

Normal assumption

Look at qq-normal plot

```
################################
## Check the assumption  of  normality of residuals

## qq-normal plot of residuals
qqnorm(fit$residuals)
qqline(fit$residuals)

## Or with a  Wally plot
require(MESS)
qqwrap <- function(x, y, ...) {qqnorm(y, main="",...);
  qqline(y)}
## Can we see a deviating qq-norm plot?
wallyplot(fit$residuals, FUN = qqwrap)
```

Oversigt

A complete example - from the book



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:

Agenda

Course 02402 Introduction to Statistics Lecture 12:

Inference for proportions

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

1. Intro
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
7. R

Oversigt

1. Intro
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
7. R

Different analysis/data-situations in course 02402

Mean for 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 proportions:

$$\hat{p} = \frac{x}{n}$$

$$\hat{p} \in [0;1]$$

Oversigt

Confidence interval for one proportion

Method 7.3

If we have a large sample , then an $(1-\alpha)\%$ 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.

As a rule of thumb

the normal distribution gives a good approximation of the binomial distrinution if np and $n(1-p)$ are both greater than 15

Confidence interval for one proportion

Mean and variance in binomial distribution, Chapter 2.21

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1-p) \end{aligned}$$

This means that

$$\begin{aligned} E(\hat{p}) = E(\frac{X}{n}) &= \frac{np}{n} = p \\ Var(\hat{p}) = Var(\frac{X}{n}) &= \frac{1}{n^2}Var(X) = \frac{p(1-p)}{n} \end{aligned}$$

Example 1

Left handed:

$p =$ proportion of left handed in Denmark

and/or:

Female engineering students:

$p =$ Proportion of female engineering students

Example 1

Left handed:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 \Leftrightarrow 0.10 \pm 0.059 \Leftrightarrow [0.041, 0.159]$$

Better "small sample" method - "plus 2-approach":(Remark 7.7)

Use the same formula on $\tilde{x} = 10+2 = 12$ and $\tilde{n} = 104$:

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.031328$$

$$0.1154 \pm 1.96 \cdot 0.03132 \Leftrightarrow 0.1154 \pm 0.0614 \Leftrightarrow [0.054, 0.177]$$

"Margin of Error" on estimate

Margin of Error

with $(1-\alpha)\%$ confidence becomes:

$$ME = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where an estimate of p comes from $p = \frac{x}{n}$

Sample size determination

Method 7.13

If you want a Margin of Error ME with $(1-\alpha)\%$ confidence, then you need the following sample size:

$$n = p(1-p)[\frac{z_{1-\alpha/2}}{ME}]^2$$

Sample size determination

Method 7.13

If you want a Margin of Error ME with $(1-\alpha)\%$ confidence, and you have NO reasonable guess of p, then you need the following sample size:

$$n = \frac{1}{4}[\frac{z_{1-\alpha/2}}{ME}]^2$$

since the worst case approach is given by: $p = \frac{1}{2}$

Example 1 - continued

Left handed:

Assume 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:

$$n = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604$$

Oversigt

Steps by hypothesis testing - an overview (Repetition)

1. Formulate the hypotheses 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 sampling distribution, and compare the p-value and the significance level α and make a conclusion
4. (Alternatively, make a conclusion based on the relevant critical value(s))

Hypothesis test for one proportion

The null and alternative hypothesis for one proportion p:

$$H_0: \quad p = p_0$$
$$H_1: \quad p \neq p_0$$

We either accept H_0 or reject H_0

Calculation of test statistic

Theorem 7.10 and Method 7.11

If the sample size is sufficiently large, we use the test statistic: (If $np_0 > 15$ and $n(1-p_0) > 15$)

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

Under the null hypothesis the random variable Z follows a standard normal distribution, $Z \sim N(0, 1^2)$

Finishing the test (Method 7.11)

Find the p-value (evidence against the null hypothesis):

- $2P(Z > |z_{\text{obs}}|)$

Test using the critical value

Alternative hypothesis	reject null hypothesis if
$p \neq p_0$	$z_{\text{obs}} < -z_{1-\alpha/2}$ or $z_{\text{obs}} > z_{1-\alpha/2}$

Example 1 - continued

Is half of all people in Denmark 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:

$$2 \cdot P(Z > 8) = 1.2 \cdot 10^{-15}$$

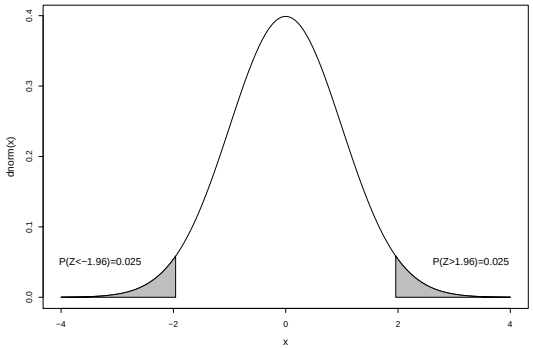
There is very strong evidence against the null hypothesis - we reject this (with $\alpha = 0.05$).

Example 1 - continued

Using the critical value in stead:

$$z_{0.975} = 1.96$$

As $z_{\text{obs}} = -8$ is (much) less than -1.96 we reject the hypothesis.



Oversigt

Confidence interval for 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 (here shown for a two-sided alternative)

$$H_0: \quad p_1 = p_2$$
$$H_1: \quad p_1 \neq p_2$$

The test statistic:

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad \textit{where} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

And for large samples:

Use the standard normal distribution again.

Example 2

Is there a relation between ths use of birth control pills and the risk of blood clot in the heart

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

Is there a relation between ths use of birth control pills and the risk of blood clot in the heart

Carry out a test to check if there is any connection between the use of birth control pills and the risk of blood clot in the heart. Use a significance level of $\alpha = 5\%$

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 in each sample

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \ \hat{p}_2 = \frac{35}{167} = 0.2096$$

Common estimate:

$$\hat{p} = \frac{23+35}{57+167} = \frac{58}{224} = 0.2589$$

Oversigt

Hypothesis test for several proportions

The comparison of c proportions

In some cases we might be interested in determining if two or more binomial distributions have the same parameter p, that is we are interested in testing the null hypothesis:

$$H_0: \quad p_1 = p_2 = \ldots = p_c = p$$

vs. the alternative that the proportions are not equal.

Hypothesis test for several proportions

Table of observed counts for k samples:

	sample 1	sample 2	...	sample c	Total
Succes	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 has e_{1j} successes and e_{2j} failure, 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

We fill compute table of EXPECTED counts for k samples:

e_{ij}	sample 1	sample 2	...	sample c	Total
Succes	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 freqency 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:

χ^2-distribution with $(c-1)$ degrees of freedom

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

Find 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 vaues 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			57
No B. C. pill			167
Total	58	166	224

Example 2 - continued

The test statistic:

$$\chi^2_{\text{obs}} = \frac{(23-14.76)^2}{14.76} + \frac{(34-42.24)^2}{42.24} + \frac{(35-43.24)^2}{43.24} + \frac{(132-123.76)^2}{123.76}$$

$$= 8.33$$

Critical value:

```
qchisq(0.95, 1)
```

[1] 3.8

Conclusion:

We reject the null hypothesis - there IS a significant higher risk in the BC pill group.

Oversigt

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

Are the votes equally distributed?

$$H_0: \quad p_{i1} = p_{i2} = p_{i3}, i = 1,2,3.$$

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 a dependency between the rows and columns?

$$H_0: \; p_{ij} = p_{i\cdot} p_{\cdot j}$$

Computation of the test statistic – no matter 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 freqency 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:

χ^2-distribution with $(r-1)(c-1)$ degrees of freedom

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 $e_{ij} \geq 5$.

Oversigt

R: `prop.test` - one proportion

```
# TESTING THE PROBABILITY = 0.5 WITH A TWO-SIDED ALTERNATIVE
# WE HAVE OBSERVED 518 OUT OF 1154
# WITHOUT CONTINUITY CORRECTIONS

prop.test(518, 1154, p = 0.5, correct = FALSE)
```

R: prop.test - two proportions

```
#READING THE 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")

# TESTING THAT THE PROBABILITIES FOR THE TWO GROUPS ARE EQUAL
prop.test(pill.study, correct = FALSE)
```


R: chisq.test - two proportions

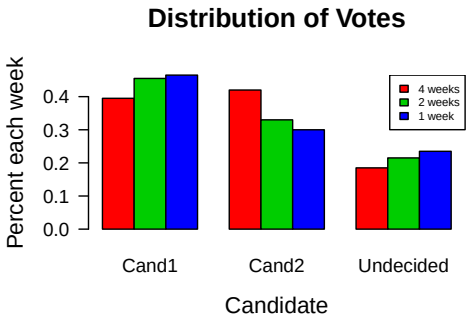
```
# CHI2 TEST FOR TESTING THE PROBABILITIES FOR THE TWO GROUPS ARE EQ
chisq.test(pill.study, correct = FALSE)
#IF WE WANT THE EXPECTED NUMBERS SAVE THE TEST IN AN OBJECT
chi <- chisq.test(pill.study, correct = FALSE)
#THE EXPECTED VALUES
chi$expected
```


R: chisq.test - contingency tables

```
#READING THE 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")

#COLUMN PERCENTAGES
colpercent<-prop.table(poll, 2)
colpercent
```


R: chisq.test - contingency tables

```
# Plotting percentages
par(mar=c(5,4,4.1,2)+0.1)
barplot(t(colpercent), beside = TRUE, col = 2:4, las = 1,
        ylab = "Percent each week", xlab = "Candidate",
        main = "Distribution of Votes")
legend( legend = colnames(poll), fill = 2:4,"topright", cex = 0.5)
par(mar=c(5,4,4,2)+0.1)
```



R: chisq.test - contingency tables

```
#TESTING SAME DISTRIBUTION IN THE THREE POPULATIONS
chi <- chisq.test(poll, correct = FALSE)
chi

#EXPECTED VALUES
chi$expected
```


Agenda

Course 02402 Introduction to Statistics Lecture 13:

A course summary

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda - the 12 lectures

1. Chapter 1: Simple Graphics and Summary Statistics
2. Chapter 2: Discrete Distributions
3. Chapter 2: Continuous Distributions
4. Chapter 3: One sample confidence intervals
5. Chapter 3: One sample hypothesis testing
6. Chapter 3: Two Sample statistics
7. Chapter 4: Statistics by simulation
8. Chapter 5: Simple linear Regression Analysis
9. Chapter 6: Multiple linear Regression Analysis
10. Chapter 8: One-way Analysis of Variance
11. Chapter 8: Two-way Analysis of Variance
12. Chapter 7: Inferences for Proportions

Oversigt

Chapter 1: Simple Graphics and Summary Statistics

- Look at data as it is! (descriptive statistics)
- Summary Statistics
  - Mean $\bar{x}$
  - Standard deviation s, variance s^2
  - Median, upper- and lower quartiles
- Simple graphics
  - Scatter plot (xy plot)
  - Histogram, cumulative distribution
  - Boxplots, Bar charts, Pie charts

Oversigt

Chapter 2: Discrete Distributions

- General concepts:
  - Definition of a stochastic variable
  - Density function
  - Distribution function
  - Mean and variance
- Specific distributions:
  - The binomial distribution
  - The hypergeometric distribution
  - The Poisson distribution

Oversigt

Chapter 2: Continuous Distributions

- General concepts:
  - Density function, distribution function
  - Mean, variance
  - Calculation rules for stochastic variables
- Specific distributions:
  - Normal
  - Log-Normal, Uniform, Exponential

Oversigt

Chapter 3: One sample confidence intervals

- General concepts
  - Estimation, confidence intervals
  - Population and a random sample
  - Sampling distributions (t and χ^2)
  - Central Limit Theorem
- Specific methods, one sample:
  - Confidence intervals for the mean
  - Confidence intervals for the variance (and standard deviation)

Oversigt

Chapter 3: One sample hypothesis testing

- General concepts:
  - Hypotheses, p-value, Significance level
  - Type I and Type II error, Power
- Specific methods, One sample:
  - t-test for mean difference
  - Sample size for wanted power
  - Normal qq-plot

Oversigt

Chapter 3: Two Samples

- Specific methods, two samples:
  - Test and confidence interval for the mean difference (t-test)
- Specific methods, two PAIRED samples:
  - "Take difference" $\Rightarrow$ "One sample"
- Planning for precision and/or power
  - One-sample Confidence interval: sample size for wanted precision
  - One-sample hypothesis test: sample size for wanted power (or other combinations)
  - Two-sample hypothesis test: sample size for wanted power (or other combinations)

Oversigt

Chapter 4, Statistics by simulation

- Introduction to simulation
- Error propagation rules
- Bootstrapping
  - Parametric
  - Non-parametric
  - Confidence intervals (and hence also hypothesis testing)
- Specific situations: (4 versions of confidence intervals)
  - One-sample and Two-sample data
  - Parametric and Non-parametric

Oversigt

Chapter 5: Simple linear Regression Analysis

- Two quantitative variables, x and y.
- Calculating least squares line
- Inferences for a simple linear regression model
  - Statistical model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  - Interval estimation and test for β_0 and β_1.
  - Confidence interval for the expected line.
  - Prediction interval.
- r and r^2
  - r describes the strength of a linear relation.
  - r^2 expresses the proportion of the y variability explained by the linear relation.

Oversigt

Chapter 6: Multiple linear Regression Analysis

- Many quantitative variables, x_1, x_2 and y.
- Calculating least squares fit
- Inferences for a the multiple linear regression model
  - Statistical model: $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} \varepsilon_i$
  - Interval estimation and test for β_0 and β_i.
  - Confidence interval for the expected fit.
  - Prediction interval.
- r^2 expresses the proportion of the y variability explained by the linear relation.

Oversigt

Chapter 8: One-way Analysis of Variance

- Specific methods, k INDEPENDENT samples
- One-way analysis of variance
  - Compares the means of the groups
  - ANOVA-table: $SST = SS(Tr) + SSE$
  - F-test.
  - Post hoc test: pairwise t-test with/without Bonferroni correction

Oversigt

Chapter 8: Two-way Analysis of Variance

- Block design - two-way analysis of variance
- ANOVA-tabel: $SST = SS(Tr) + SS(Bl) + SSE$
  - SST, $SS(Tr)$ and $SS(Bl)$ calculated as one-way ANOVA
  - $SSE = SST - SS(Tr) - SS(Bl)$
- F-test.
- Post hoc test: pairwise t-test with/without Bonferroni correction

Oversigt

Chapter 7: Inferences for Proportions

- Specific methods, one, two and $k > 2$ samples
  - Binary/categorical response
- Estimation and confidence interval of proportions
  - Large sample vs. small sample methods
- Hypotheses for one proportion
- Hypotheses for two proportions
- Analysis of contingency tables (χ^2-test) (All expected > 5)

Agenda