Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

JADE HOCHSCHULE
Wilhelmshaven Oldenburg Elsfleth

(Advanced) Statistics A

Sommersemester 2022

Prof. Dr. Bernhard Köster

- Due to the pandemic the organizational form of this lecture is not fixed right now!
- Start is Thursday, 2 March 2022 online via zoom (see moodle)
- Classroom lectures may be possible from 21 March 2022 on
- Further information will be given in the lecture. Times are volatile nowadays!
- During the pandemic, I always had ideas, who to deal with the uncertain situation, but always I hade to bow force majeure. Therefore I do not make any annoucements in advance.
- But I can assure you, that we will find a good teaching concept!
- Kind regards Bernhard Köster

Prof. Dr. Bernhard Köster

Room:	S 113
Street:	Friedrich-Paffrath-Straße 101
location:	26389 Wilhelmshaven
Tel.	+49 4421 985-2766
Email:	bernhard.koester@jade-hs.de
Consultation hour:	by arrangement or just have a look into my office! or Webex/Zoom …

Literature

- **Statistics for Business and Economics**, Anderson, Sweeney, Williams, Camm, Cochran
- **Mathematical Statistics for Economics and Business**, Mittelhammer
- **Statistics for Business and Economics, Global Edition**, Newbold, Thorne, Carlson
- **A Guide to Modern Econometrics**, Verbeek

Some open resources

[OpenIntro Statistics]

[Statistical Thinking for the 21st Century]

[MIT OPENCOURSEWARE]

-> there is much more!!!

Excel

Statistics

X	Y
1	2
4	7
8	7
6	5

via calculator

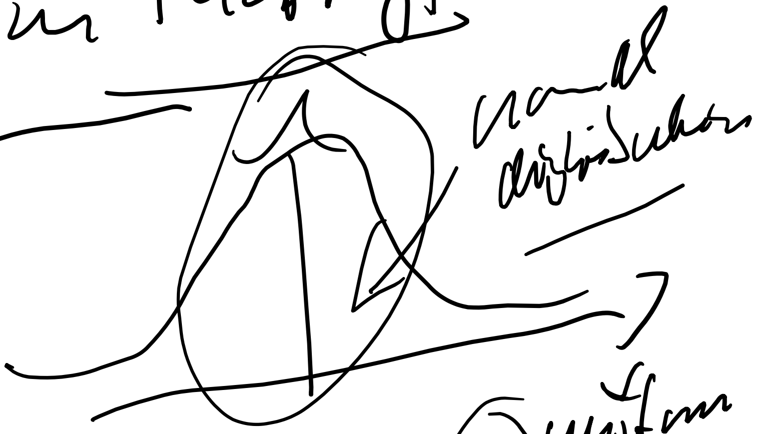
$Y = a + bx$

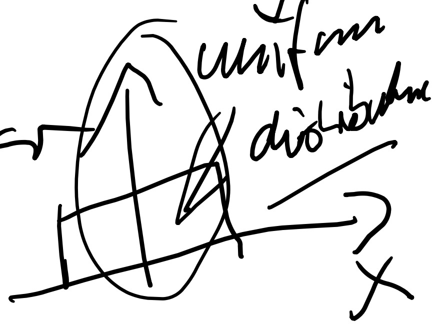
- Best unbiased estimator
- Efficiency
- Consistency
- Sufficiency
- Confidence Interval
- One and two sided tests
- alpha/beta error
- Significance
- Hypothesis testing
  t, F, χ^2
- ML-estimator

Decision Theory

1. Kind 2 Kind error

Probability distribution





Introduction and Revision

- **Means and Median**
- **Dispersion: MAD, Variance, Standard Deviation, Coefficiant of Variation**
- **Contingency table and independence**
- **Bayes` Theorem**
- **Random Variable**
- **Probability Distributions, density function and cumulative distribution function**
- **Central limit theorem**

Example

$\bar{x}_a = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$ $\bar{x}_h = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\dots+\frac{1}{x_n}}$

1) Generate in Excel a sample of the length of 15 male students around 175cm and a maximum deviation of 25cm.

WHV 100 km/h, 60 km/h OL → What is the mean velocity

2) Calculate the arithmetic, geometric, harmonic mean and the median of the distribution.

3) Calculate the MAD, variance, standard deviation, and the coefficient of variation of the distribution.

$MAD = \frac{1}{n}\sum_{i=1}^{15} |\bar{x}_a - x_i|$ $\sigma^2 = \frac{1}{n}\left(\sum (\bar{x}_a - x_i)^2\right)$ $f(x) = \frac{1}{\sqrt{2\pi}} e^{\dots}$ $\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

4) Generate in Excel a sample of the length of 11 female students around 165cm and a maximum deviation of 20cm.

5) Calculate the contingency table of the attributes gender and lenght ((0-155),[155,170),[170,185),[185,200] and the marignal distributions

1) Are the attributes gender and length statistically independent?

Statistics A

Wilhelmshaven

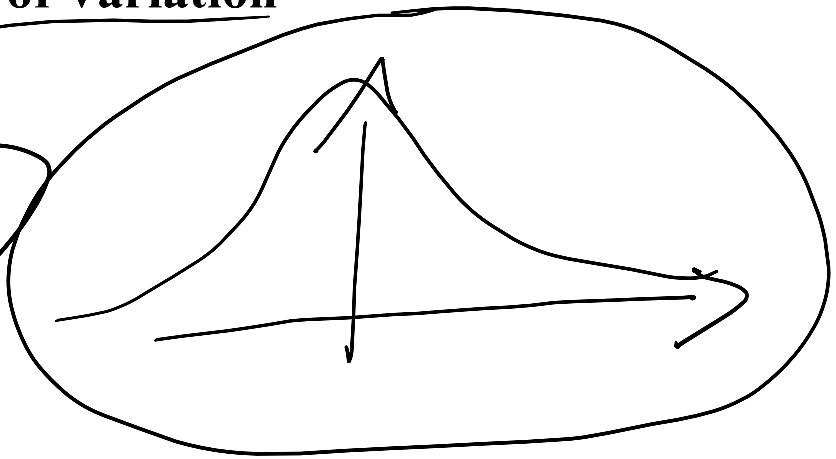
<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

arithmetic, geometric, harmonic

Introduction and Revision

- **Means and Median**
- **Dispersion: MAD, Variance, Standard Deviation, Coefficiant of Variation**
- **Contingency table and independence**
- **Bayes` Theorem**
- **Random Variable**
- **Probability Distributions, density function and cumulative distribution function**
- **Central limit theorem**



Example

1) Generate in Excel a sample of the length of 15 male students around 175cm and a maximum deviation of 25cm.

2) Calculate the arithmetic, geometric, harmonic mean and the median of the distribution.

3) Calculate the MAD, variance, standard deviation, and the coefficient of variation of the distribution.

4) Generate in Excel a sample of the length of 11 female students around 165cm and a maximum deviation of 20cm.

5) Calculate the contingency table of the attributes gender and lenght ((0-155),[155,170),[170,185),[185,200] and the marignal distributions

1) Are the attributes gender and length statistically independent?

Conditional probability

The conditional probability of an event A is the probablity of the occurence the event A given event B has happened (or happens simultaneously with A)

→ conditional probability of A given: P(A | B).

Definition:

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$$

Bayes Theorem:

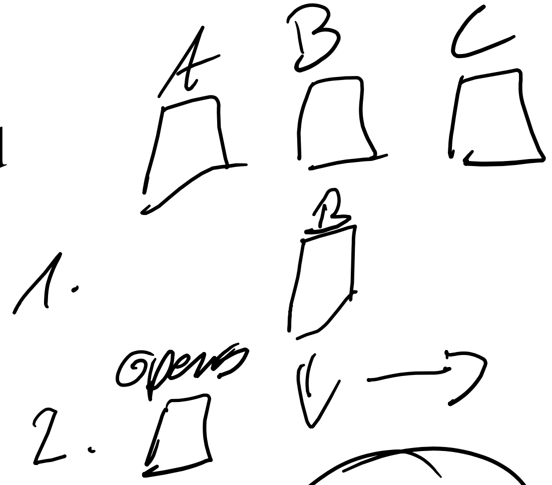
$$\boxed{P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)}}$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Symmetrically

Goats and Cars

Suppose in a game show you are sitting infront of 3 doors. You know behind on door you`ll win a car and behind the two other doors you get only a goat



Round 1: You choose a door

After that, the showmaster, knowing the door with the car, opens a door with a goat

Marty Hall Problem

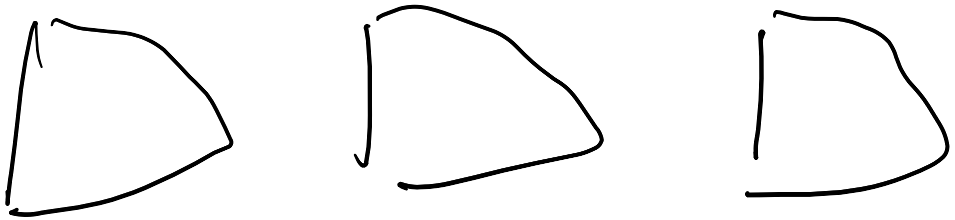
Round 2: You`re asked if you want to change doors

What is your decision?

If you don`t believe it, try to simulate this in Excel!

$* \Rightarrow$ winning the by switching the door $= 1 \cdot \frac{\frac{1}{3}}{\frac{1}{3}}$

A B C



choose B

open A

$1 \cdot$

2

$= \frac{2}{3}$

⇓

Should I switch ?

shouldn't I switch ?

It doesn't matter?

First Decision had a winning probability of 1/3

$P(\text{car}) = \frac{1}{3}$

A B

why

P(behind not opened door) $= \frac{1}{2}$

$P(\text{Car} \mid \text{not opened door in Round 2}) = \frac{P(B|A) \cdot P(A)}{P(B)}$

what we need

P(door ist not opened | a car is behind the door) $= 1$

B A

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Random Variable

Definition:
A **Random Variable X** is the mapping of the sample space Ω into the real numbers R. That means every elementary event A_i maps onto a real number x_i and the probability $p_i = P(A_i) = P(X=x_i) = p(x_i)$ is known.

→ p_i ist the probality that random variable X equals the outcome x_i.

Random Variable – Explainations

- Every possible event of a random experiment can be expressed via a random variable.

- Random variables are expressed in capital letters.
  Example: dice with sample space X is element of {1,2,3,4,5,6}

- The realization of a random variable is expressed in lowercase letters:
  Example rolling dice: X realizes the value x = 5

- Probality in the example of rolling dice:
  $P(\underbrace{x < 5}) = ??$

$P(X=4) = P(X=3) = P(X=2) = P(X=1) = \frac{1}{6}$

$= \frac{4}{6} = \frac{2}{3} \approx 66\%$

Discrete and continous random variables (RV)

Discrete RV:

RV, which have only a finite or countably infinite realizations.

Continous RV:

RV which can realize within an intervall every possible real number.

Realization in the Interval
12:15 and 12:45 → 30 minutes
~~Ma~~ A Bus is stopping within this
time intervall

Discret

Probability mass function:

$$\boxed{P(X = x_i) = p_i}$$

Cumulative distribution function:

$$\boxed{P(X \leq x) = \sum_{x \leq x_i} p(x_i)}$$

Expectated value:

$$\boxed{E(X) = \mu = \sum_{i=1}^{n} x_i p(x_i)}$$

Variance:

$$\boxed{Var(X) = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i)}$$

Continous

Probability density function:

$$\boxed{f(x)}$$

Bus stops at same specific time i.e. 12^{34}

$P(X = 12^{34})$ = ? = 0

Cumulative distribution function:

$$\boxed{F(x) = \int_{x'=-\infty}^{x} f(x')dx'}$$

~ Probability can be interpreted as the Area below the density function

Expectated value :

$$\boxed{E(X) = \mu = \int_{x=-\infty}^{x=\infty} x f(x)dx}$$

Variance:

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

Calculating with the Expectated Value and the Variance

Constant:

$$E(a) = a$$

$$Var(a) = 0$$

Constant factor:

$$E(a \cdot X) = a \cdot E(X)$$

$$Var(aX) = a^2 \cdot Var(X)$$

Linear Transformation:

$$E(a \cdot X + b) = a \cdot E(X) + b$$

$$Var(a \cdot X + b) = a^2 \cdot Var(X)$$

$E(X) = \sum x_i p_i \Rightarrow E(aX) = \sum a x_i p_i = a \sum x_i p_i = a\mu$

$Var(aX) = \sum (a x_i - a\mu)^2 p_i$ $E(X)$

$= \sum a^2 (x_i - \mu)^2 p_i$

$= a^2 \sum (x_i - \mu)^2 p_i$

$\Rightarrow$ Expected Value Operator is a linear operator

Probability Axioms

Axiom 1: $\boldsymbol{P(A) \geq 0}$ **for all** $\boldsymbol{A \in \Omega}$
Every event has a non negative probability

Axiom 2: $\boldsymbol{P(\Omega) = 1}$
The certain event has probability 1

Axiom 3: $\boldsymbol{P(A \cup B) = \boxed{P(A) + P(B)}}$**, if** $\boldsymbol{A \cap B = \varnothing}$
Addition rule for disjoint events

Rules

& because in general this could be different from the emptyset

Rule 1: $\boldsymbol{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$

Addtion rule for any events

Rule 2: $\mathbf{P(A \backslash} \boldsymbol{B}\mathbf{)} = \boldsymbol{P(B) - P(A \cap B)}$

A without B

Rule 3: $\mathbf{P(}\bar{\boldsymbol{A}}\mathbf{)} = \mathbf{1} - \boldsymbol{P(A)}$

Probality of complementary events

Conditional probability

The conditional probability of an event A is the probablity of the occurence the event A given event B has happened (or happens simultaneously with A)

→ conditional probability of A given B: P(A | B).

Definition:

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

Bayes Theorem:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Example

Suppose the following probabilities are known:

$P(A) = 0{,}5$; $P(B) = 0{,}3$; $P(A \cap B) = 0{,}2$

Calculate

conditional Probability

A without B

a) $P(A \cup B)$ b) $P(A|B)$ c) $P(\overline{A \cap B})$ d) $P(A \setminus B)$

a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 50\% + 30\% - 20\% = 60\%$

b) $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0{,}2}{0{,}3} = \dfrac{2}{3} = 66\%$

c) $P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - 0{,}2 = 80\%$

d) $P(A \setminus B) = P(B) - P(A \cap B) = 0{,}3 - 0{,}2$
$= 10\%$

Statistically independent events

P(A|B) = P(A)   or   P(B|A) = P(B)

→

$$\boxed{P(A \cap B) = P(A) \cdot P(B)}$$

Statistically independent events: Example

$n=2$ First student 365 second 365 $\boxed{365 \cdot 365}$

Sylvester 1988 in Casino in Konstanz at the roulette table, the event $A = \{0,3\}$ occured 9 times in a row. Calculate the probability of this event

$P = \frac{2}{37}$ $P(X \in \{0,3\} \text{ 9 times in a row}) = \underbrace{\frac{2}{37} \cdot \frac{2}{37} \cdot \ldots \cdot \frac{2}{37}}_{\text{9 times}}$

1–36 + 0 → 37 numbers

$\left(\frac{2}{37}\right)^9 = 4 \cdot 10^{-12}$

Suppose in class two students have birthday at the same day. What is the the minimum number of students in class, that the probability for this event is 50%?

n: Number of students

Just guess 182

365^n possible Distributions of Birthday

First has at some specific day his Birthday

then the second has only 364 possibilities left

3rd Student 363 ....

$365 \cdot 364 \cdot \ldots \cdot 365-(n-1)$ possibilities that no students have at the same day their birthday

$\boxed{1 - \frac{365 \cdot 364 \cdot \ldots \cdot 365-(n-1)}{365^n}} = 50\%$ ⇒ we need $\boxed{\text{23 Students}}$

Total Probability

If the sample space Ω consists of in k disjoint elemtary events A_i, then the probability of event B is:

→ $$\boxed{P(B) = \sum_{i=1}^{k} P(B \mid A_i) \cdot P(A_i)}$$

Total Probability: Example

Every day a small village is visited by a postman. If he is in good mood he is in time with probability 90%. If he is in bad mood he is late with probability 40%.

What is the probability that he is late at any day, if he is on average in good at 7 out of 10 days?

$$P(\text{Postman is late}) = \overbrace{P(\text{Postman is late} \mid \text{good})}^{10\%} \cdot \overbrace{P(\text{good})}^{70\%}$$

$$+ \overbrace{P(\text{Postman is late} \mid \text{bad})}^{40\%} \cdot \overbrace{P(\text{bad})}^{30\%}$$

$$= 0{,}07 + 0{,}12 = 19\%$$

Total Probability and Bayes Theorem

Suppose you are testing for a rare illness and you have the following probabilities:

A = {patient is ill}
B = {test is positive}
Since we have a rare illness: P(A) = 0,1%.

Since a test does not have 100% accurancy, from surveys we have:

P(B|A) = 0,98
$P(B|\bar{A}) = 0{,}03$

Suppose your test result is positive. What is the probability, that you are really ill?

$P(\bar{A}) = ? = 0{,}999$

$P(A) = 0{,}001$

just have a guess around 90%

$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$ Total probability

what I want to know is

$$\underset{\text{I'm ill}}{P(A}|\underset{\text{Test positive}}{B)} = \frac{P(B|A)\cdot P(A)}{P(B)} = \frac{0{,}98\cdot 0{,}001}{0{,}98\cdot 0{,}001 + 0{,}03\cdot 0{,}999}$$

$\approx 3{,}2\%$

quite surprisingly a very low probability

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Binomial distribution

- A random experiment has two possible outcomes (success and failure).
- The random experiment will be repeated n times.
- What is the probability to be k times successful within $n \geq k$ repetions
- Every repetition is independent from each other. Within every repetition succes has probability p und failure probability 1-p.

Binomial Distribution

Suppose p of „success“ is known, define the random variable X =„number of succes k“ within n ≥ k repetitions.

$n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot \ldots \cdot n$

→ X is Binomial distributed $X \sim B(n;p)$ with

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

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$E(X) = np$$

$$Var(X) = np(1-p)$$

Example

$$\binom{7}{3} = \frac{7!}{3!(7-3)!} = \frac{\cancel{1 \cdot 2 \cdot 3 \cdot 4} \cdot \ldots \cdot 7}{\cancel{1 \cdot 2 \cdot 3}(\cancel{1 \cdot 2 \cdot 3 \cdot 4})} = \frac{5 \cdot 6 \cdot 7}{1 \cdot 2 \cdot 3} = 35$$

1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

Binomial coefficients

Examples Binomial Distribution

$$\boxed{\mathbf{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}}$$

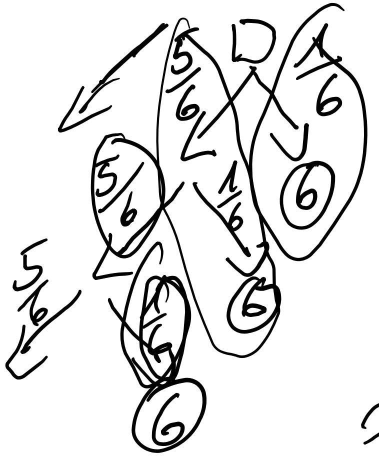
1. Probability within 10 throws of a dice to obtain 3 times a 6?

$n=10 \quad k=3 \quad p=\frac{1}{6}$

$$\binom{10}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^7=\frac{1\cdot 5^7}{6^{10}}\binom{10}{3}$$

$\approx 15{,}5\%$

2. Probability to move the first time in the game „Mensch ärgere Dich nicht“ (Hombre, no te enfades)?



$$\frac{1}{6}+\frac{5}{6}\cdot\frac{1}{6}+\left(\frac{5}{6}\right)^2\cdot\frac{1}{6}$$

$\approx 42\%$

3. You play four different statistically independent Lotteries. The probability of success is always 20 %.

   a) What is the expectated value and the variance and standard deviation?
   b) What is the expected payoff, if every successs counts 10 Euros?
   c) Calculate the probability to win exactly two times
   d) Calculate the probability that you loose at least three times.

a) $n=4 \quad p=\frac{2}{10}$

$E(x)=n\cdot p=\frac{8}{10}=0{,}8$

$Var(x)=np(1-p)=\frac{4\cdot 2}{10}\cdot\frac{8}{10}=0{,}64$

$Stdev=\sqrt{0{,}64}=0{,}08$

b) $E(10x)=10E(x)=8$

c) $k=2 \quad \binom{4}{2}(0{,}2)^2(0{,}8)^2=\frac{1\cdot2\cdot3\cdot4}{1\cdot2\cdot1\cdot2}\,\frac{4}{100}\,\frac{64}{100}\approx 15\%$

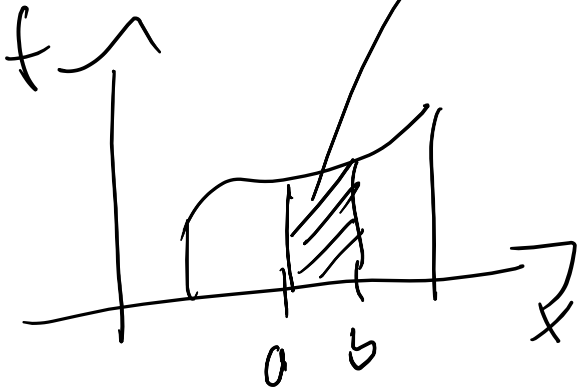
d) switch to $p=0{,}8 \quad \binom{4}{3}(0{,}8)^3\,0{,}2+\binom{4}{4}0{,}8^4\cdot 0{,}2^0\approx 81\%$

Continous Random Variables

probability that $X \in [a,b]$

The probability of the realization of X within some specific interval [a,b] is the area below the density function f(x):

$$P(a \le X \le b) = \int_a^b f(x)\,dx$$



Properties of the density function

- Values of f(x) cannot be interpreted!
- If the length of the intervall reaches zero (a = b), also the area reaches zero:

$$\rightarrow \quad \boxed{P(X = x) = 0}$$

- The whole area below the density function equals 1 (certain event):

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

Example Density Function

Student Mike is always late in the statistics lecture. His delay is a continous random variable with the following density function

$$f(x) = \begin{cases} a - \frac{1}{8}x & \text{for } 0 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

$$\int_0^4 \left(a - \frac{1}{8}x\right) dx = \int_0^4 a\,dx - \int_0^4 \frac{1}{8}x\,dx$$

$$= ax\Big|_0^4 - \frac{1}{16}x^2\Big|_0^4 = 1$$

$$4a - \frac{1}{16}\cdot 4^2 = 1 \quad 4a - 1 = 1 \quad \Rightarrow a = \frac{1}{2}$$

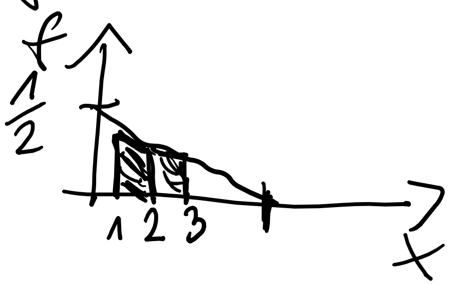
1. What is a?
2. Sketch the density graphically
3. Calculate the probability that Mike is between 1 und 2 minutes late at some specific day.
4. Sketch the cumulative distribution function graphically
5. Calculate the Expected value and Variance

Example Density Function

a [1,2] density, cumulative distribution, E(x), Var(x)

$$f(x) = \begin{cases} a - \frac{1}{8}x & \text{for } 0 \le x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow f(x) = \frac{1}{2} - \frac{1}{8}x \Rightarrow \frac{1}{2} = \frac{1}{8}x \Rightarrow x = 4$




```
solve( int(a-(1/8)x dx; 0..4)=1,a)
plot(1/2-x/8,0..4)
int(1/2-x/8),0..y
plot(-1/16 (-8 + y) y,0..4)
int(x(1/2-x/8)),0..4
int((x-4/3)^2(1/2-x/8)),0..4
```

3.) $\int_1^2 \left(\frac{1}{2} - \frac{1}{8}x\right) dx = \frac{1}{2}x\Big|_1^2 - \frac{x^2}{16}\Big|_1^2 = \left(1 - \frac{1}{2}\right) - \left[\frac{4}{16} - \frac{1}{16}\right]$

$= \frac{1}{2} - \frac{3}{16} = \frac{8-3}{16} = \frac{5}{16}$

4.) $F(x) = \int_0^x f(y)\,dy$

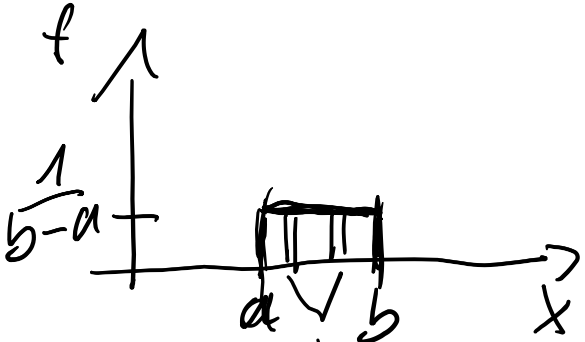
https://www.wolframalpha.com/

5.) $E(x) = \int_0^4 x f(x)\,dx \qquad Var(x) = \int (x - E(x))^2 f(x)\,dx$

Uniform distribution

The density is constant over a given Interval [a,b]. This means, that every subinterval with same length have same probability

$$f(x) = \begin{cases} \frac{1}{b-a}x & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

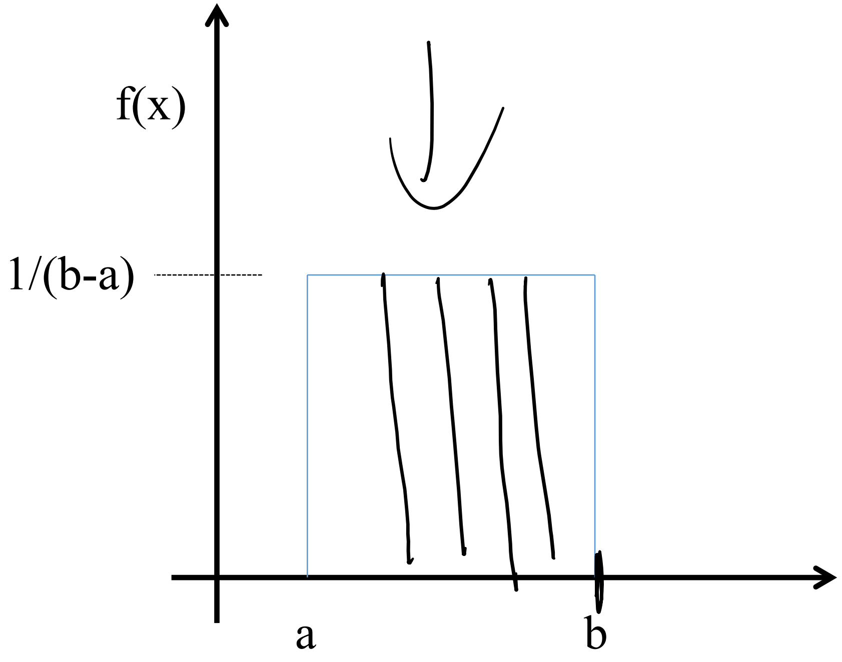


$E(x) = \int x f(x)$

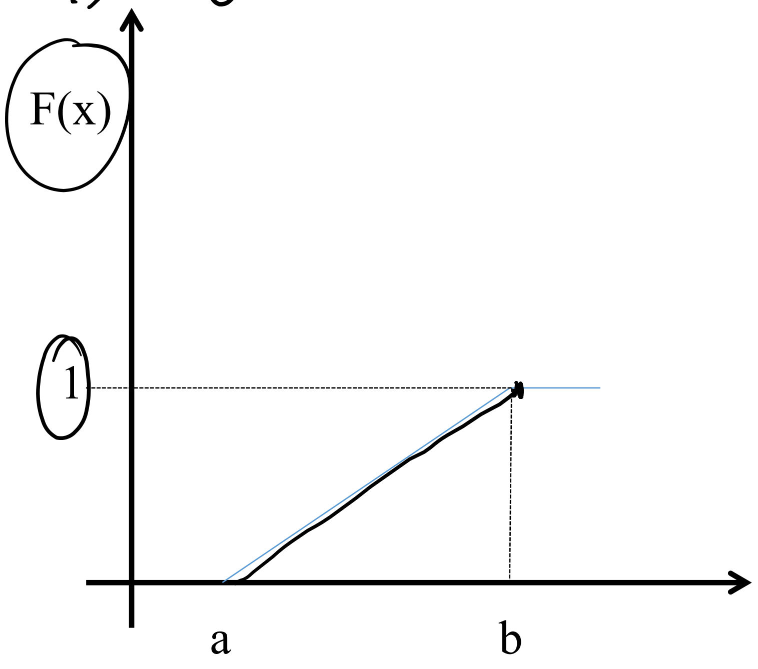
$$\boxed{E(X) = \frac{a+b}{2}}$$

$\int (x - E(x))^2 f(x) dx =$ $$\boxed{Var(X) = \frac{1}{12}(b-a)^2}$$

Uniform distribution



$= \int_0^x f(y)dx$



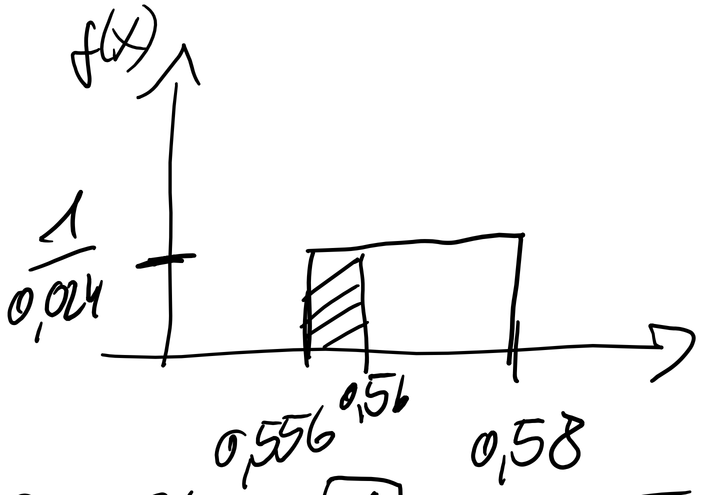
Example

Suppose you are the COP (Chief of Production) within a brewery. Furthermore the maschine shows a filling quanatity of the bottles is 0,568L. But you suppose, that real filling quantity is uniformly distributed around 0,556L and 0,58L.

What is the probability that the filling quantity is less than 0,56L?

$$f(x) = \frac{1}{0{,}58 - 0{,}556} \quad x \qquad 0{,}556 \leq x \leq 0{,}58 \qquad 0 \text{ otherwise}$$

$$= \frac{1}{0{,}024} \quad x$$



$$\Rightarrow P(x \leq 0{,}56L) = \frac{0{,}004}{0{,}024} = \boxed{\frac{1}{6}} \approx 16{,}\bar{6}\,\%$$

Normal Distribution

density function

The normal distribution or gaussian distribution is the most important distribution.

We write:

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

The density function is given by:

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

The cumulative distribution function by:

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^x e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sqrt{2\pi\sigma^2}}\right)\right)$$

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-y^2}\, dy$$

(Error function)

$$\boxed{E(X) = \mu} \qquad \boxed{Var(X) = \sigma^2}$$

The importance of the normal distribution

- Many empirical distributions follow at least approximately the normal distribution
- Many discrete distribtions can be approximated by the normal distribution. I.e. Binomial distribution
- The distribution of the sample indepent with respect to the underlying true distritibution is approximated by the normal distribution for large sample size N (central limit theorem)
- The normal distribution is the basis theoretical models
  (i.e. white noise)

$f(x) + \epsilon$

often assumed to be normally distributed

Normal distribution

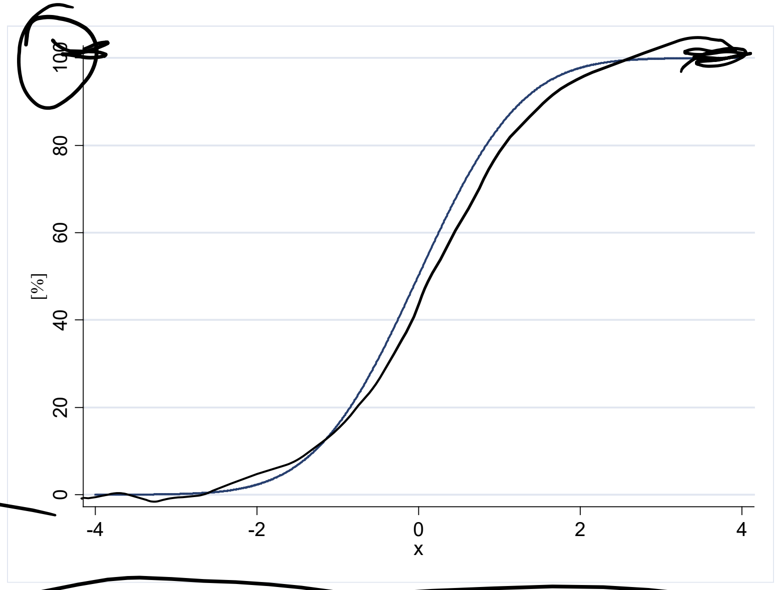
$(x-\mu)^2 \Leftarrow$ symmetric function

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

$$F^{(\mu,\sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

Density function $\mu = 0$ und $\sigma = 1$

$\mu-\sigma$ $\mu+\sigma$



Cumulative distribution function $\mu = 0$ und $\sigma = 1$

Properties of the normal distribution

- The normal distribution is a 2-parameter-distribution in μ and σ
- $N(\mu, \sigma^2)$ is symmetrically around $x = \mu$
- The density function has turning points at $x = \mu + \sigma$ and $x = \mu - \sigma$
- The density function flattens if the variance goes up.
- The density function reaches the x-axes asymptotically at $+\infty$ and $-\infty$.

The standard normal distribution

- Normal distribution with $\boxed{\mu = 0}$ $\boxed{\sigma = 1}$

$$Z \sim \boxed{N(0;1)}$$

- Maximum at z = 0
- Turning points at z = $\boxed{-1 \text{ and } z = 1}$
- In order to calculate probabilities the cumulative distribution function is tabulated and implemented in every standard spreadsheet programm like excel (MS office) or calc (libreoffice)

$$\boxed{\phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{1}{2}x^2}\, dx \cdot}$$

- Every normally distributed random variable can be transformed in a standard normally distributed random variable (Standardisierung)

Normal distribution and standard normal distribution

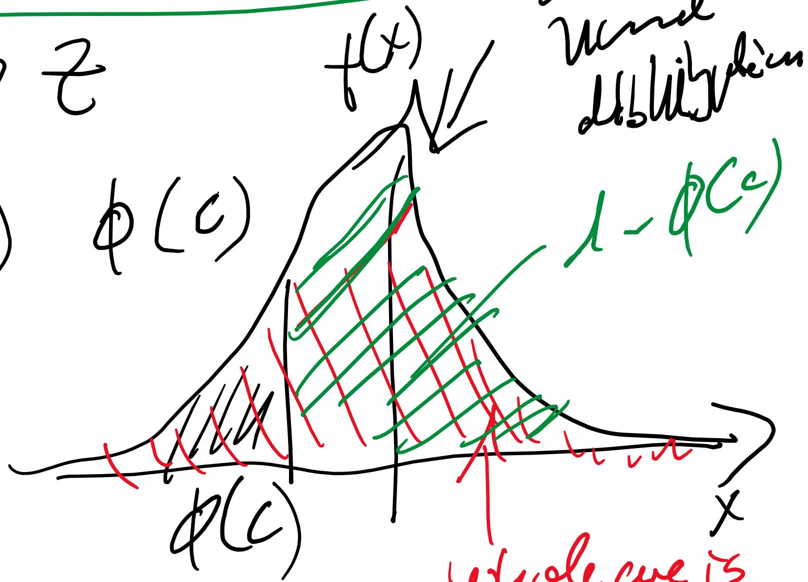
standard normal distribution

- Suppose $X \sim N(\mu;\sigma^2)$ then

$X \rightsquigarrow Z$

$$Z = \frac{x-\mu}{\sigma} \sim N(0;1)$$

$\phi(z)$ $\phi(c)$



- We have $\Phi(-c) = 1 - \Phi(c)$

$x = \sigma z + \mu$

- Generally, we obtain:

$$P(X \leq x) = P(\sigma Z + \mu \leq x) = P(Z \leq \frac{x-\mu}{\sigma})$$

$\sigma z \leq x - \mu$

$z \leq \frac{x-\mu}{\sigma}$

or

$$F^{(\mu,\sigma^2)}(x) = \Phi(\frac{x-\mu}{\sigma})$$

Statistics A

Wilhelmshaven

<u>**This lecture will be recorded and Subsequently uploaded in the world-wide-web**</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Normal distribution and standard normal distribution

- Suppose $X \sim N(\mu;\sigma^2)$ then

$$Z = \frac{x-\mu}{\sigma} \sim N(0;1)$$

- We have $\Phi(-c) = 1 - \Phi(c)$

- Generally, we obtain:

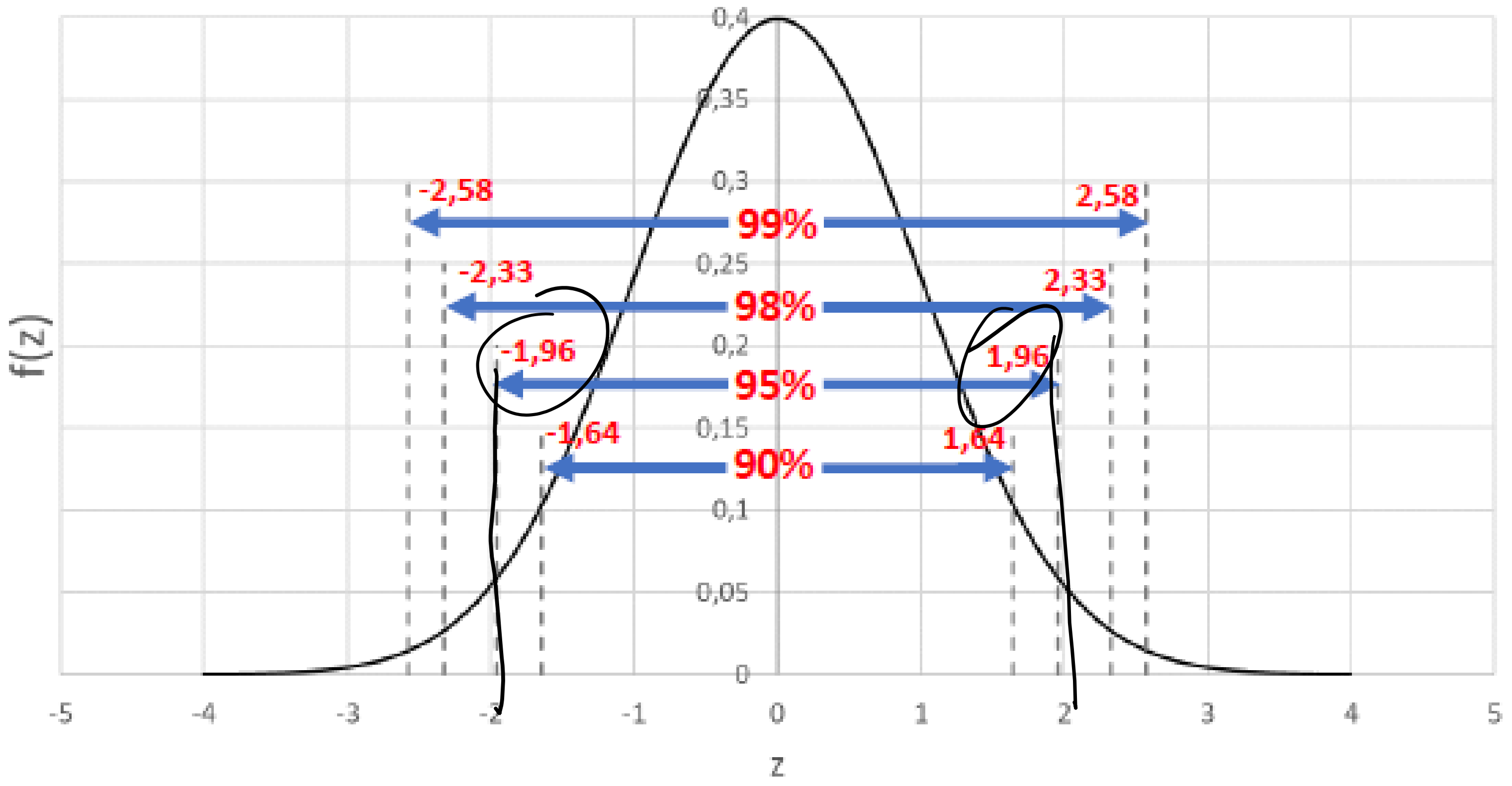
$$P(X \le x) = P(\sigma Z + \mu \le x) = P(Z \le \frac{x-\mu}{\sigma})$$

or

$$F^{(\mu,\sigma^2)}(x) = \Phi(\frac{x-\mu}{\sigma})$$

Relevant Values of the standard normal distribution

z	F(z)
-2.58	0.5 %
-2.33	1 %
-1.96	2.5 %
-1.64	5 %
-1.23	10 %



Point estimator for the mean μ of a parent distribution

General random variable

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

specific sample with concrete values

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

X_i: Capital letters $\rightarrow$ random variables.

x_i: Lowercase letters $\rightarrow$ concrete values of the random variables within a specific sample

$\Longrightarrow$ The arithmetic mean of a sample is the estimator of the true mean of the parent population.

Point estimator for the proportion π of a parent distribution

In analogy to the estimation of the unknown mean via the arithmetic mean, the estimator for the unknown proportion π of a parent population is given by:

$\hat{\pi} = \frac{k}{n}$ k: number of successes out of sample size n

$\Longrightarrow$ $\hat{\pi}$ is the estimator for the true proportion π of the parent population.

Point estimator for the varinace σ^2 of a parent distribution

The estimator $\hat{\sigma}^2$ of the unknown variance σ^2 of a parent population is given by:

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

$\Longrightarrow$ The sample variance $\hat{\sigma}^2$ is the (unbiased) estimator of the true variance σ^2 of the parent population.

Classical criterions of optimal estimators

Unbiased estimator:
An estimator $\hat{\theta}$ of a true parameter θ is called unbiased, if the expected value of the estimator equals the true value of the estimated parameter θ:

$$E(\hat{\theta}) = \theta \quad \forall \theta \in \Theta \quad (\Theta : \text{parameter space})$$

Classical criterions of optimal estimators

Asymptotically unbiased estimator:
An estimator $\hat{\theta}$ of a true parameter θ is called asymptotically unbiased, if a sequence of the expected value of the estimator $E(\hat{\theta}_n)$ reaches in the limit the true value of the estimated parameter θ:

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta \quad \forall\theta \in \Theta \quad (\Theta : \text{parameter space})$$

Classical criterions of optimal estimators

Efficiency:
Efficiency describes the accuracy of an unbiased estimator. The accuracy is generally measured via the variance. The estimator with the minimal variance is called efficient.

$\Longrightarrow$ An estimator $\hat{\theta}^*$ is called efficient if

$$\begin{aligned} E(\hat{\theta}^*) &= \theta \\ Var(\hat{\theta}^*) &\leq Var(\hat{\theta}) \end{aligned}$$

for all unbiased estimators $\hat{\theta}$.

(desired) properties of estimators

Consistency:
An estimator, which approaches the true parameter, if the sample size is raised is called consistent. A sequence $(\hat{\theta}_n)$ of estimators is called consistent if for all $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

this means $\hat{\theta}_n$ is converging stochastically to θ and is located for a given sample size n within an epsilon-neighbourhood $U_\epsilon(\theta)$ of the true value.

(desired) properties of estimators

- **consistency in mean square (MSE: Mean squared error):**
  A sequence $(\hat{\theta}_n)$ of estimators is called consistent in mean square if:
  $$\lim_{n\to\infty} E([\hat{\theta}_n - \theta]^2) = 0$$
  with
  $$MSE := E([\hat{\theta}_n - \theta]^2) = Var(\hat{\theta}_n) - (E(\hat{\theta}_n) - \theta)^2$$
  $\rightarrow$ with raising sample size n the difference between variance and the displacement vanishes.

(desired) properties of estimators

Sufficiency:
An estimator is called sufficient, if all information about the true parameter is used.

Idea:
Suppose you have a data set $\mathbf{S}$ of n independently distributed random variables. Then we look for some mapping $T(\mathbf{S})$ whose values contain all information of original data set. Thus, efficiency is a concept in oder to reduce high-dimensional data-vectors to a usable size.

Exercise

1. $T(X_1, \ldots X_n) = \sum_{i=1}^{n} a_i X_i$ with $\sum_{i=1}^{n} a_i = 1$ $(a_i > 0)$ is an unbiased estimator for μ.
2. $\frac{K}{\bar{X}}$ is an unbiased estimator for p of the binomial distributio $B(n, p)$.
3. $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator for the variance, if μ is unknown.
4. $\hat{s}^2(\mu) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is an unbiased estimator for the variance, if μ is known.
5. $\hat{s}^2(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an asymtotically unbiased estimator for the variance, if μ is unknown.
6. Within the linear unbiased estimators for μ, the arithmetic mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an efficient estimator.

Exercise

7. Suppose $X \sim B(n, p)$: There is no unbiased estimator of the standard deviation of the binomial distribution.
8. Show $MSE := E([\hat{\theta}_n - \theta]^2) = Var(\hat{\theta}_n) - (E(\hat{\theta}_n) - \theta)^2$
9. Suppose $X_1, ..., X_n$ are independently Bernoulli-distributed, then $T(X) = X_1 + ... + X_n$ is a sufficient estimator for p (T is the total number of successes).

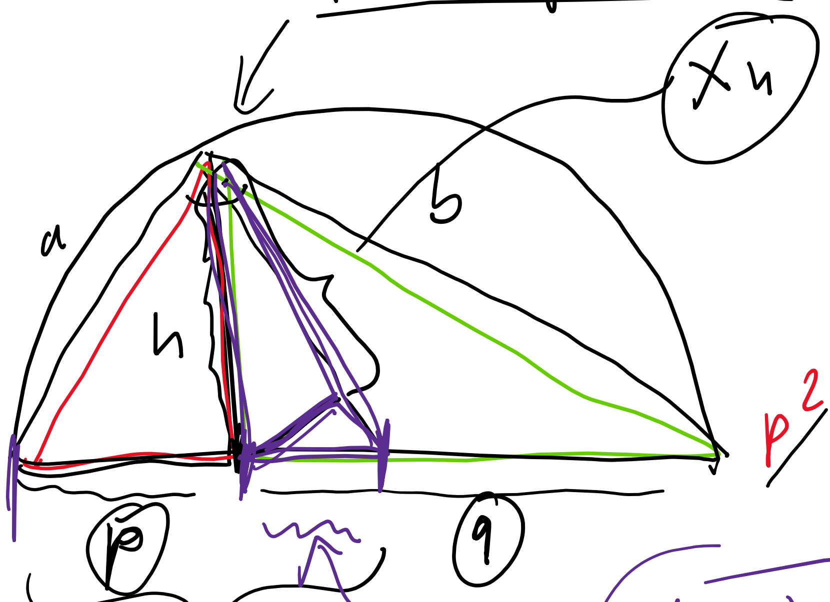
* Suppose $X_1, ..., X_n$ are independently normally distributed and σ^2 is known. Then $\bar{X}$ is a sufficent estimator for μ.

$\Rightarrow x_g = \sqrt{x_h \cdot x_a} \;\Big|^2 \qquad x_g^2 = x_h \cdot x_a \Rightarrow p\cdot q = \frac{2}{\frac{1}{p}+\frac{1}{q}} \cdot \frac{1}{2}(p+q) = \frac{1}{\frac{p+q}{p\cdot q}} \cdot (p+q)$

two numbers p q

$= \frac{1}{\frac{1}{pq}} = p \cdot q = x_g^2$ q.e.d.

theorem of Thales



$h = \sqrt{p \cdot q}$

$a^2 + b^2 = c^2$

$p^2 + h^2 + q^2 + h^2 = (p+q)^2 = p^2 + q^2 + 2pq$

$\Rightarrow 2h^2 = 2pq \Rightarrow h = \sqrt{pq} = x_g$

$x_a = \frac{1}{2}(p+q)$

$x_h = \frac{2}{\frac{1}{p} + \frac{1}{q}}$

$x_g = \sqrt{p \cdot q}$

$x_a = \frac{1}{2}(p+q)$

$\frac{1}{2}(p+q) - p$

$c = p+q$

geometric mean theorem

3) Within a manufacturing process on average 75% of the tools are correct.

a) Calculate the probability, that within a sample of n = 12 you have exactly 3 correct tools.

b) Calculate the probability, that within a sample of n = 9 you have at least 8 correct tools.

c) Calculate the expected value, variance and standard variation of a sample of n=25.

a) $\binom{k}{n} p^k (1-p)^{n-k} = P(k=3)$

$= \binom{3}{12}\left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right)^{12-3} = \frac{12!}{3!(12-3)!}\cdot\frac{3^3\cdot 1}{4^{12}}$

$= \frac{1\cdot 2\cdot 3\cdot 4 \cdots 12}{1\cdot 2\cdot 3\cdot 1\cdot 2\cdot 3 \cdots 9} = \frac{10\cdot 11\cdot 12\cdot 27}{2\cdot 3\cdot 4^{12}} = \frac{5\cdot 11\cdot 3}{4^{11}} = \frac{15\cdot 11}{4^{11}} = \frac{165}{4^{11}}$

b) $P(k \geq 8) = \sum_{k=8}^{12} \binom{k}{12} p^k (1-p)^{12-k} \rightarrow$ do in excel $= 0{,}004\%$

$= \binom{8}{12}(0{,}75)^8(0{,}25)^4 + \binom{9}{12}(0{,}75)^9(0{,}25)^3 + \dots$ $k=12$

c) $E(x) = np = 12\cdot\frac{3}{4} = 9$

$Var(x) = np(1-p) = 9\cdot\frac{1}{4} = \frac{9}{4}$

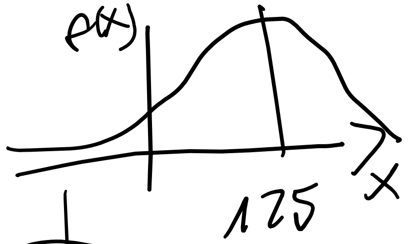
$stdev = \sqrt{Var} = \sqrt{np(1-p)} = \frac{3}{2}$

4) A producer of cocoa knows from experience, that the weight of the 125g-packs is normally distributed with $\mu = 125$ g and variance of $\sigma^2 = 25$ g.

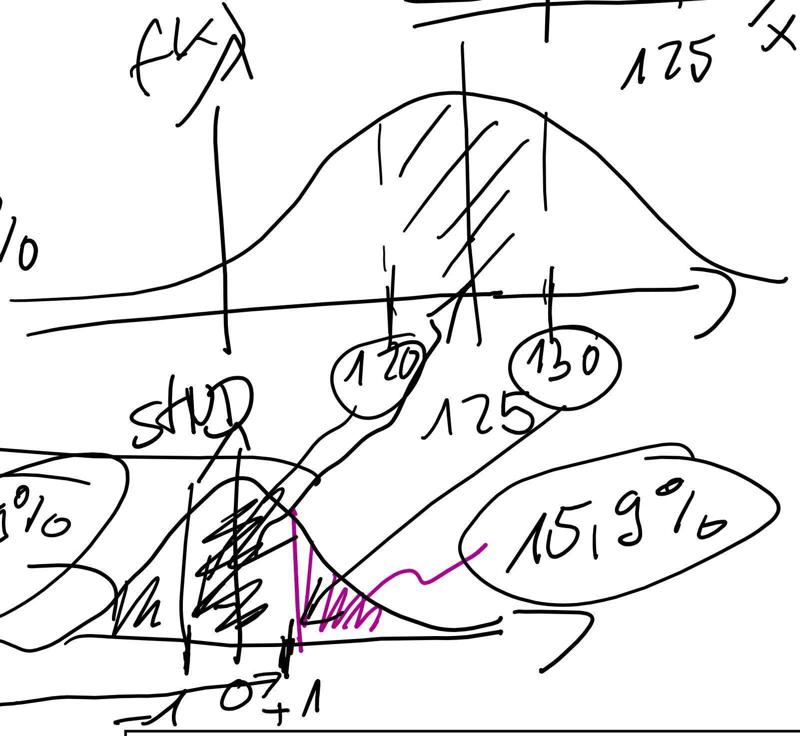
a) What is the probability that the weight of a pack is exactly 125 g (argue)?
b) What is the probability, that the weight of a pack is within 120 g and 130 g?
c) What is the probability, that the weight of a pack is less than 110 g?
d) What is the probability, that the weight of a pack is more than 140 g?
e) Calculate the symmetric interval around the expected value, such that with a probability of 95% the true weight of a pack is within this interval.
f) Sketch your results graphically with the given distribution and the standard normal distribution.

$P(x=125)=0$

$P(120 \le x \le 130)$



$\Phi(1) = 84{,}1\%$



$$\boxed{z = \frac{x-\mu}{\sigma}}$$

$$z_1 = \frac{120-125}{5} = 1$$

$$P(120 \le x \le 130) = 100\% - 2 \cdot 15{,}9\% = 68{,}2\%$$

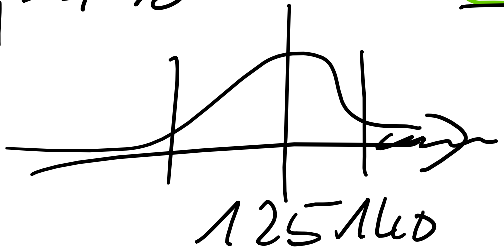
$$= P(-1 \le z_1 \le 1)$$

c) $P(x \le 110) = P(z \le -3) = 1 - \Phi(3)$

$= 1 - 99{,}86\% = 0{,}14\%$

d) $P(X \ge 140)$

$= \Phi(3) = 99{,}86\%$



e) 95%

e) $95\% = P(x_d \le X \le x_u)$

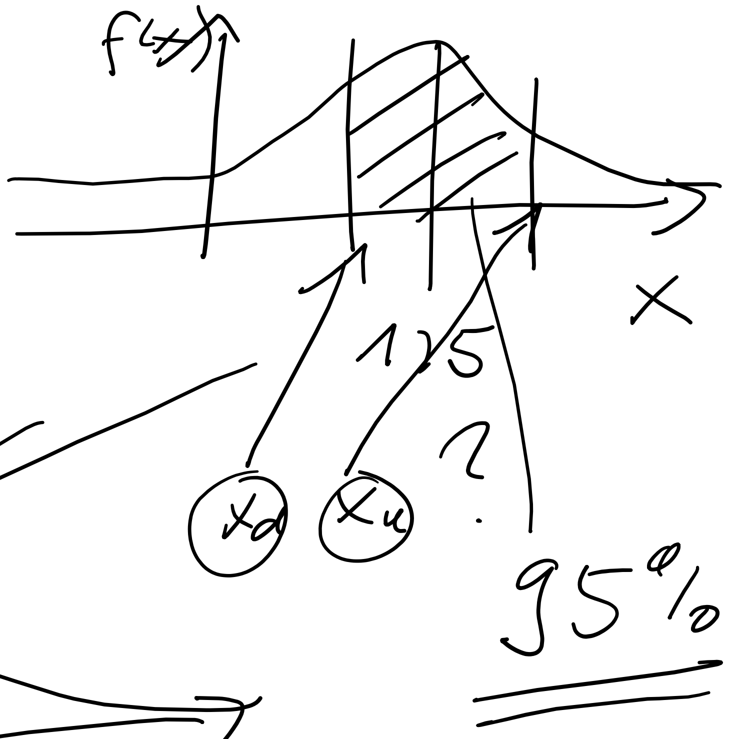
$$-1{,}96 = \frac{x_d - 125}{5}$$

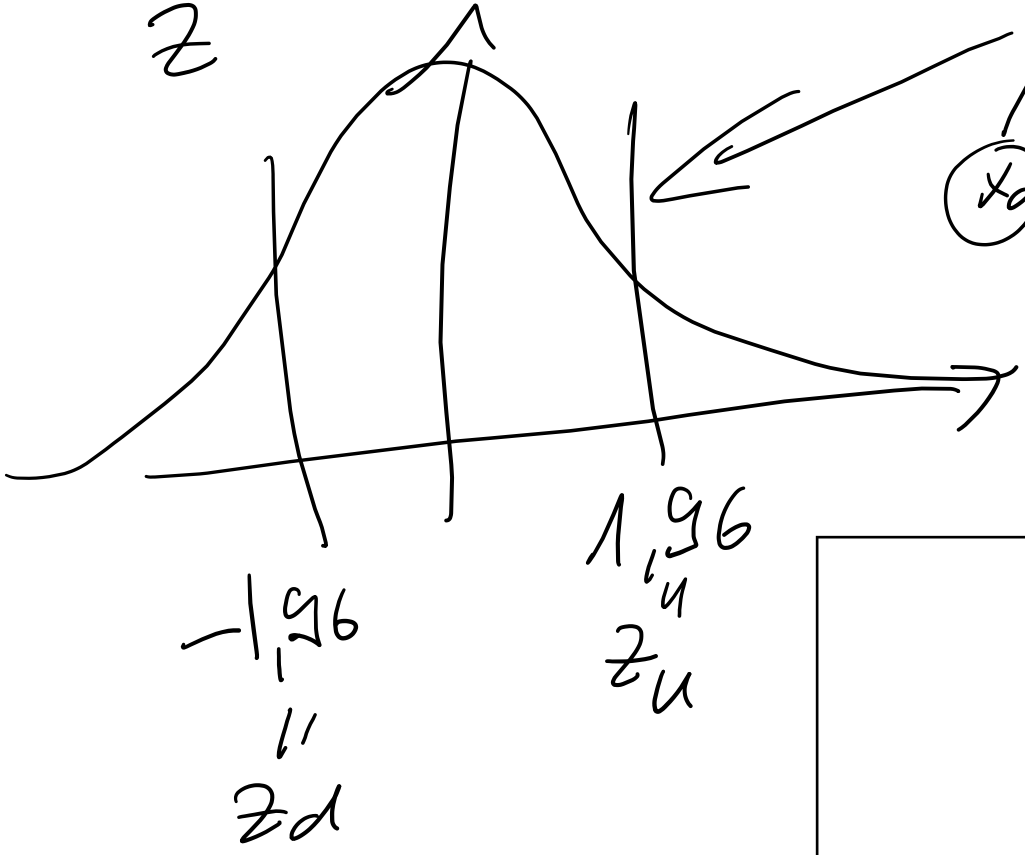
$$\approx 5\cdot(-2) + 125 = x_d$$

$$x_d \approx 115{,}2$$

$$x_u \approx 134{,}8$$

$$z = \frac{x - \mu}{\sigma}$$

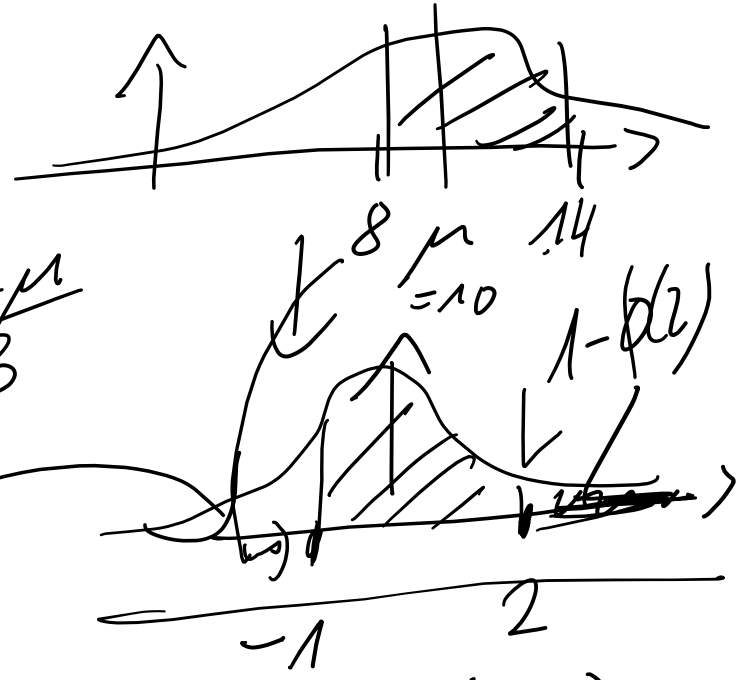




The annual yield [%] of stock investment can be approximated with a normally distributed random variable with μ=10 and σ=2.

a) What is the probability that the yield is within 8% und 14% liegt?
b) Assume that the yields oft wo different years are statistically independent.
   i. What is the probability that the yields in two following years is at least 8%?
   ii. What is the probability that the yields in the next 10 years will be exactly three times less than 11%?
c) Which yield can be maximally expected with a probability of 99%?

a)



$z = \frac{x-\mu}{\sigma}$

$\Rightarrow P(8 \leq x \leq 12)$

$= 1 - [1 - \phi(1) + 1 - \phi(2)]$

$= -(1 - \phi(1) - \phi(2))$

$= \phi(1) + \phi(2) - 1 \approx \underline{81{,}8\%}$

$1 - \phi(1)$

$\phi(1) = 84{,}1\%$

$\phi(2) = 97{,}7\%$

Binomial distribution

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Normal distribution and standard normal distribution

- Suppose $X \sim N(\mu;\sigma^2)$ then

$$Z = \frac{x-\mu}{\sigma} \sim N(0;1)$$

- We have $\Phi(-c) = 1 - \Phi(c)$

- Generally, we obtain:

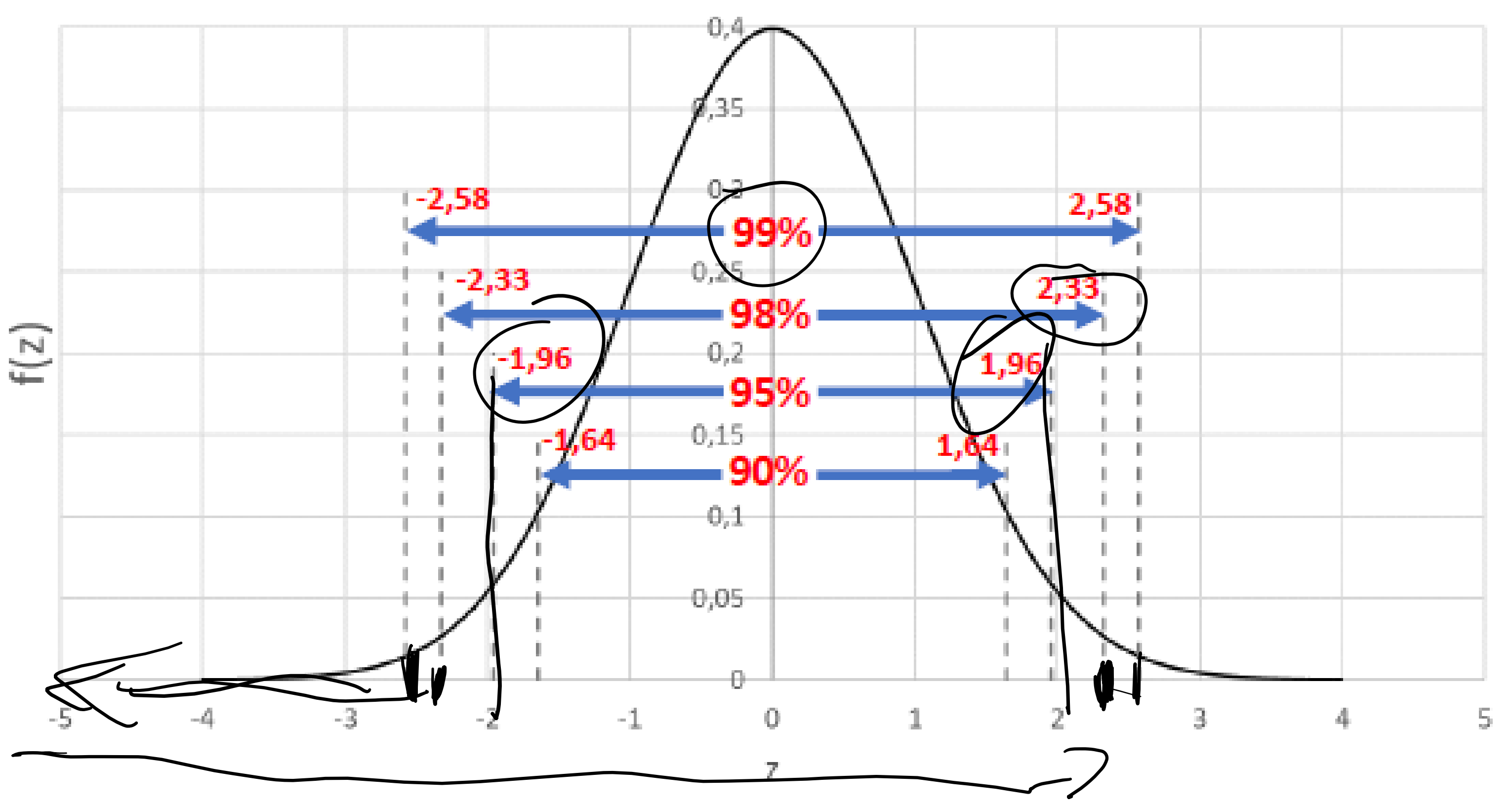
$$P(X \le x) = P(\sigma Z + \mu \le x) = P(Z \le \frac{x-\mu}{\sigma})$$

or

$$F^{(\mu,\sigma^2)}(x) = \Phi(\frac{x-\mu}{\sigma})$$

Relevant Values of the standard normal distribution

z	F(z)
-2.58	0.5 %
-2.33	1 %
-1.96	2.5 %
-1.64	5 %
-1.23	10 %



Point estimator for the mean μ of a parent distribution

General random variable

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

specific sample with concrete values

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

X_i: Capital letters $\rightarrow$ random variables.

x_i: Lowercase letters $\rightarrow$ concrete values of the random variables within a specific sample

$\Longrightarrow$ The arithmetic mean of a sample is the estimator of the true mean of the parent population.

Point estimator for the proportion π of a parent distribution

In analogy to the estimation of the unknown mean via the arithmetic mean, the estimator for the unknown proportion π of a parent population is given by:

$\hat{\pi} = \frac{k}{n}$ k: number of successes out of sample size n

$\Longrightarrow$ $\hat{\pi}$ is the estimator for the true proportion π of the parent population.

Point estimator for the varinace σ^2 of a parent distribution

The estimator $\hat{\sigma}^2$ of the unknown variance σ^2 of a parent population is given by:

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

$\Longrightarrow$ The sample variance $\hat{\sigma}^2$ is the (unbiased) estimator of the true variance σ^2 of the parent population.

Classical criterions of optimal estimators

Unbiased estimator:
An estimator $\hat{\theta}$ of a true parameter θ is called unbiased, if the expected value of the estimator equals the true value of the estimated parameter θ:

$$\boxed{E(\hat{\theta}) = \theta} \quad \forall \theta \in \Theta \quad (\Theta : \text{parameter space})$$

Classical criterions of optimal estimators

Asymptotically unbiased estimator:
An estimator $\hat{\theta}$ of a true parameter θ is called asymptotically unbiased, if a sequence of the expected value of the estimator $E(\hat{\theta}_n)$ reaches in the limit the true value of the estimated parameter θ:

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta \quad \forall\theta \in \Theta \quad (\Theta : \text{parameter space})$$

Classical criterions of optimal estimators

Efficiency:

Efficiency describes the accuracy of an unbiased estimator. The accuracy is generally measured via the variance. The estimator with the minimal variance is called efficient.

$\Longrightarrow$ An estimator $\hat{\theta}^*$ is called efficient if

$$E(\hat{\theta}^*) = \theta$$

$$Var(\hat{\theta}^*) \leq Var(\hat{\theta})$$

for all unbiased estimators $\hat{\theta}$.

(desired) properties of estimators

Consistency:
An estimator, which approaches the true parameter, if the sample size is raised is called consistent. A sequence $(\hat{\theta}_n)$ of estimators is called consistent if for all $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1$$

this means $\hat{\theta}_n$ is converging stochastically to θ and is located for a given sample size n within an epsilon-neighbourhood $U_\epsilon(\theta)$ of the true value.

(desired) properties of estimators

- **consistency in mean square (MSE: Mean squared error):**
  A sequence $(\hat{\theta}_n)$ of estimators is called consistent in mean square if:
  $$\lim_{n\to\infty} E([\hat{\theta}_n - \theta]^2) = 0$$
  with
  $$MSE := E([\hat{\theta}_n - \theta]^2) = Var(\hat{\theta}_n) - (E(\hat{\theta}_n) - \theta)^2$$
  $\rightarrow$ with raising sample size n the difference between variance and the displacement vanishes.

(desired) properties of estimators

Sufficiency:
An estimator is called sufficient, if all information about the true parameter is used.

Idea:
Suppose you have a data set $\mathbf{S}$ of n independently distributed random variables. Then we look for some mapping $T(\mathbf{S})$ whose values contain all information of original data set. Thus, efficiency is a concept in oder to reduce high-dimensional data-vectors to a usable size.

Exercise

1. $T(X_1, ...X_n) = \sum_{i=1}^{n} a_i X_i$ with $\sum_{i=1}^{n} a_i = 1$ $(a_i > 0)$ is an unbiased estimator for μ.
2. $\frac{K}{\bar{X}}$ is an unbiased estimator for p of the binomial distributio $B(n, p)$.
3. $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator for the variance, if μ is unknown.
4. $\hat{s}^2(\mu) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is an unbiased estimator for the variance, if μ is known.
5. $\hat{s}^2(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an asymtotically unbiased estimator for the variance, if μ is unknown.
6. Within the linear unbiased estimators for μ, the arithmetic mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an efficient estimator.

Exercise

7. Suppose $X \sim B(n, p)$: There is no unbiased estimator of the standard deviation of the binomial distribution.
8. Show $MSE := E([\hat{\theta}_n - \theta]^2) = Var(\hat{\theta}_n) - (E(\hat{\theta}_n) - \theta)^2$
9. Suppose $X_1, ..., X_n$ are independently Bernoulli-distributed, then $T(X) = X_1 + ... + X_n$ is a sufficient estimator for p (T is the total number of successes).

* Suppose $X_1, ..., X_n$ are independently normally distributed and σ^2 is known. Then $\bar{X}$ is a sufficent estimator for μ.

$\Rightarrow x_g = \sqrt{x_h \cdot x_a} \;\Big|^2 \qquad x_g^2 = x_h \cdot x_a \Rightarrow p \cdot q = \frac{2}{\frac{1}{p}+\frac{1}{q}} \cdot \frac{1}{2}(p+q) = \frac{1}{\frac{p+q}{p \cdot q}} \cdot (p+q)$ two numbers

$= \frac{1}{\frac{1}{pq}} = p \cdot q = x_g^2$ q.e.d. $\qquad p \quad q$

$x_a = \frac{1}{2}(p+q)$

$x_h = \frac{2}{\frac{1}{p}+\frac{1}{q}}$

$x_g = \sqrt{p \cdot q}$

theorem of Thales

$h = \sqrt{p \cdot q}$

$a^2 + b^2 = c^2$

$p^2 + h^2 + q^2 + h^2 = (p+q)^2 = p^2 + q^2 + 2pq$

$\Rightarrow 2h^2 = 2pq \Rightarrow h = \sqrt{pq} = x_g$

$x_a = \frac{1}{2}(p+q)$

$\frac{1}{2}(p+q) - p$

$c = p+q$

geometric mean theorem

3) Within a manufacturing process on average 75% of the tools are correct.

a) Calculate the probability, that within a sample of n = 12 you have exactly 3 correct tools.

b) Calculate the probability, that within a sample of n = 9 you have at least 8 correct tools.

c) Calculate the expected value, variance and standard variation of a sample of n=25.

a) $\binom{k}{n} p^k (1-p)^{n-k} = P(k=3)$

$= \binom{3}{12} \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right)^{12-3} = \frac{12!}{3!(12-3)!} \cdot \frac{3^3 \cdot 1}{4^{12}}$

$= \frac{\cancel{1} \cdot \cancel{2} \cdot 3 \cdot 4 \cdots \cancel{9} \cdot 10 \cdot 11 \cdot 12}{\cancel{1} \cdot \cancel{2} \cdot \cancel{3} \cdot 1 \cdot 2 \cdot 3 \cdot \cancel{4 \cdots 9}} = \frac{10 \cdot 11 \cdot \cancel{12} \cdot \cancel{27}}{\cancel{2} \cdot \cancel{3} \cdot 4^{\cancel{12}}} = \frac{5 \cdot 11 \cdot 3}{4^{11}} = \frac{15 \cdot 11}{4^{11}} = \frac{165}{4^{11}}$

b) $P(k \geq 8) = \sum_{k=8}^{12} \binom{k}{12} p^k (1-p)^{12-k} \rightarrow$ do in excel $= 0{,}004\%$

$= \binom{8}{12}(0{,}75)^8 (0{,}25)^4 + \binom{9}{12}(0{,}75)^9 (0{,}25)^3 + \ldots k=12$

c) $E(x) = np = \cancel{12} \cdot \frac{3}{\cancel{4}} = 9$

$Var(x) = np(1-p) = 9 \cdot \frac{1}{4} = \frac{9}{4}$

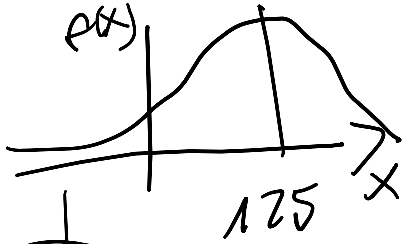
$Stdev = \sqrt{Var} = \sqrt{np(1-p)} = \frac{3}{2}$

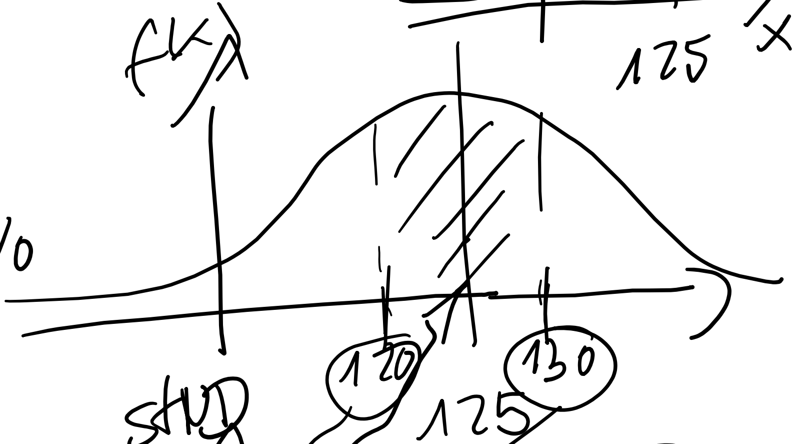
4) A producer of cocoa knows from experience, that the weight of the 125g-packs is normally distributed with $\mu = 125$ g and variance of $\sigma^2 = 25$ g.

a) What is the probability that the weight of a pack is exactly 125 g (argue)?
b) What is the probability, that the weight of a pack is within 120 g and 130 g?
c) What is the probability, that the weight of a pack is less than 110 g?
d) What is the probability, that the weight of a pack is more than 140 g?
e) Calculate the symmetric interval around the expected value, such that with a probability of 95% the true weight of a pack is within this interval.
f) Sketch your results graphically with the given distribution and the standard normal distribution.

$P(x=125)=0$

$P(120 \le x \le 130)$





$\phi(1) = 84{,}1\%$

$\boxed{z = \frac{x-\mu}{\sigma}}$ $\quad z_l = \frac{120-125}{5} = 1$

$P(120 \le x \le 130) = 100\% - 2 \cdot 15{,}9\% = 68{,}2\%$

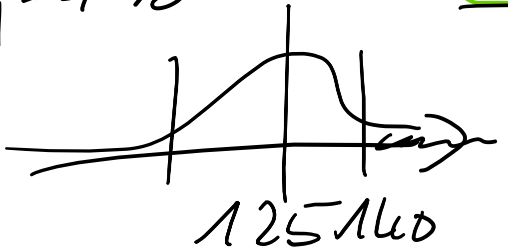
$= P(-1 \le z_i \le 1)$

c) $P(x \le 110) = P(z \le -3) = 1 - \phi(3)$

$= 1 - 99{,}86\% = 0{,}14\%$

d) $P(X \ge 140)$

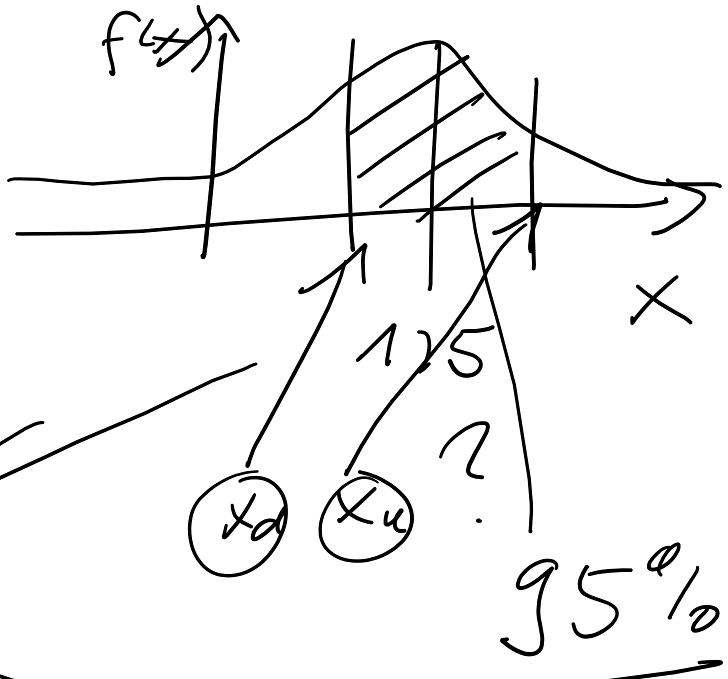
$= \phi(3) = 99{,}86\%$



e) 95%

e) $95\% = P(x_d \le X \le x_u)$

$z = \frac{x - \mu}{\sigma}$

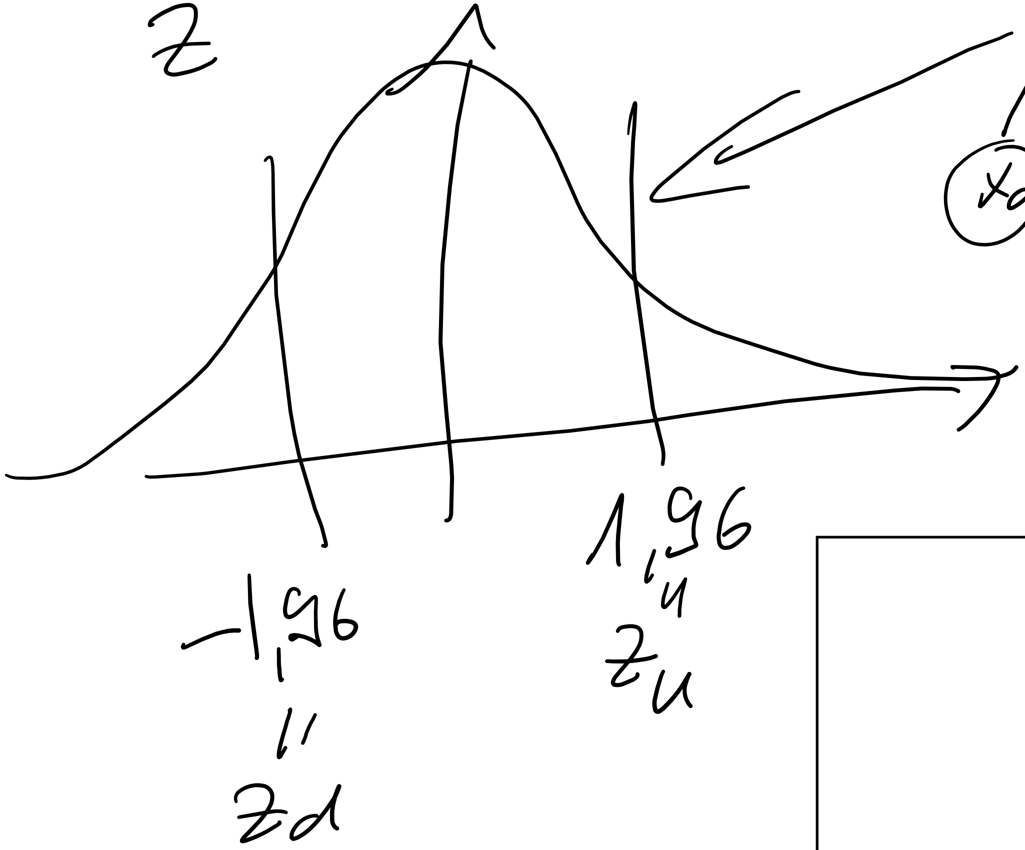


$-1{,}96 = \frac{x_d - 125}{5}$

$\approx 5 \cdot (-2) + 125 = x_d$

$x_d \approx 115{,}2$

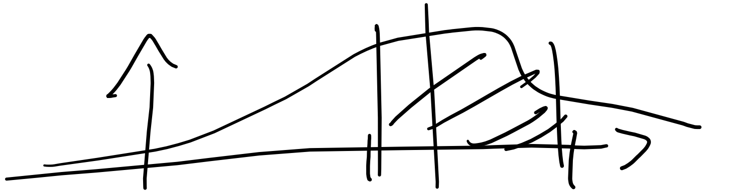
$x_u \approx 134{,}8$



The annual yield [%] of stock investment can be approximated with a normally distributed random variable with μ=10 and σ=2.

a) What is the probability that the yield is within 8% und 14% liegt?
b) Assume that the yields oft wo different years are statistically independent.
   i. What is the probability that the yields in two following years is at least 8%?
   ii. What is the probability that the yields in the next 10 years will be exactly three times less than 11%?
c) Which yield can be maximally expected with a probability of 99%?

a)



$$z = \frac{x-\mu}{\sigma}$$

$1-\Phi(2)$

$$\Rightarrow P(8 \le x \le 12)$$

$$= 1 - \left[1 - \Phi(1) + 1 - \Phi(2)\right]$$

$$= (1 - \Phi(1) - \Phi(2))$$

$$= \Phi(1) + \Phi(2) - 1 \approx 81{,}8\%$$

$1 - \Phi(1)$

$\Phi(1) = 84{,}1\%$ $\Phi(2) = 97{,}7\%$

Binomial distribution

5) The annual yield [%] of stock investment can be approximated with a normally distributed random variable with μ=10 and σ=2.

a) What is the probability that the yield is within 8% und 14% liegt?
b) Assume that the yields oft wo different years are statistically independent.
   i. What is the probability that the yields in two following years is at least 8%?
   ii. What is the probability that the yields in the next 10 years will be exactly three times less than 11%?
c) Which yield can be maximally expected with a probability of 99%?

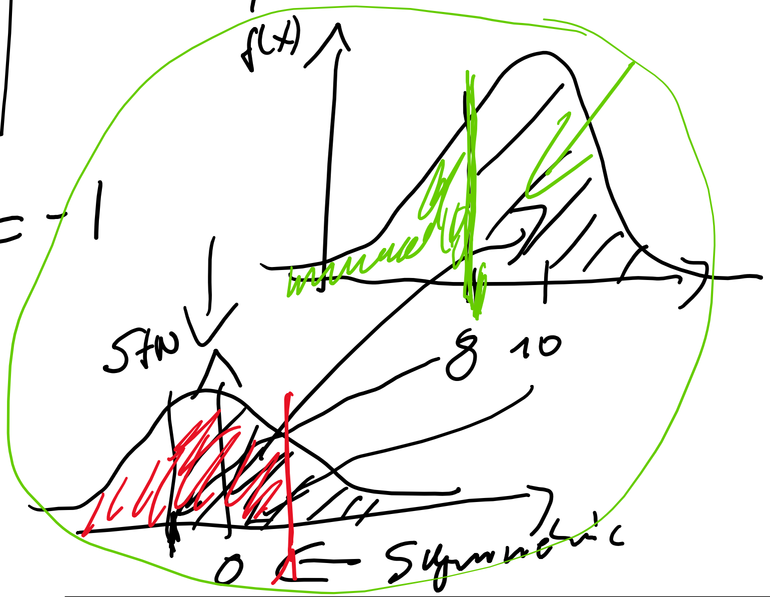
$X \sim N(10, 4) \quad N(\mu, \sigma^2)$

b(i) $P(X \geq 8)$ for two following years

$$\boxed{z = \frac{x - \mu}{2}}$$

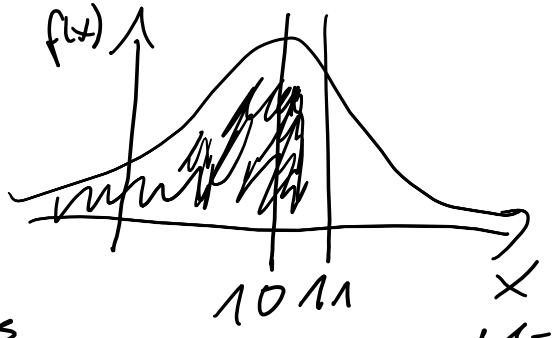
$$z = \frac{8-10}{2} = -1$$

$z >$ 11



$\Rightarrow \phi(1) = 84\% = P(X \geq 8)$

$\Rightarrow [P(X \geq 8)]^2$ because the events are statistically independent $= 0{,}707 \approx 71\%$



(ii) $P(X \leq 11) = P(Z \leq 0{,}5) = 69\%$

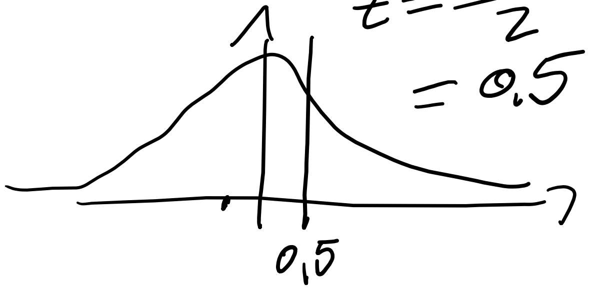
$\Rightarrow$ there $\binom{10}{3}$ possibilities that 3 times the yield is $X \leq 11\%$ out ten events

$$\frac{10!}{7! \cdot 3!} [P(X \leq 11)]^3 \cdot [1 - P(X \leq 11)]^7$$

$$\approx 120\,(0{,}69)^3\,(0{,}31)^7 \approx 0{,}0105$$

$$\approx 1{,}05\%$$

$$z = \frac{11-10}{2} = 0{,}5$$



5) The annual yield [%] of stock investment can be approximated with a normally distributed random variable with μ=10 and σ=2.

a) What is the probability that the yield is within 8% und 14% liegt?
b) Assume that the yields oft wo different years are statistically independent.
   i. What is the probability that the yields in two following years is at least 8%?
   ii. What is the probability that the yields in the next 10 years will be exactly three times less than 11%?
c) Which yield can be maximally expected with a probability of 99%?

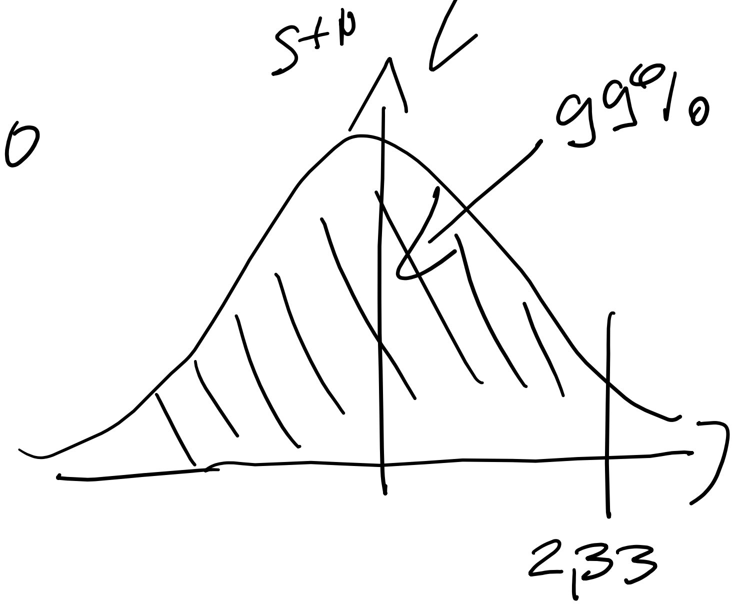
=> which x coincides with die area below the density function of 0,99

$P(z \leq 2{,}33) = 0{,}99$

$\Rightarrow x = ?$ $\qquad z = \frac{x-\mu}{\sigma}$

$= \sigma z + \mu = 2 \cdot 2{,}33 + 10$

$= 14{,}66$



maximum yield is

$14{,}66\,\%$

Exercise

weighted average

1. $T(X_1, ... X_n) = \sum_{i=1}^{n} a_i X_i$ with $\sum_{i=1}^{n} a_i = 1$ ($a_i > 0$) is an unbiased estimator for μ.
2. $\frac{K}{n}$ is an unbiased estimator for p of the binomial distributio $B(n, p)$.
3. $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator for the variance, if μ is unknown.
4. $\hat{s}^2(\mu) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is an unbiased estimator for the variance, if μ is known.
5. $\hat{s}^2(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an asymtotically unbiased estimator for the variance, if μ is unknown.
6. Within the linear unbiased estimators for μ, the arithmetic mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an efficient estimator.

1) $\boxed{E\left(\sum_{i=1}^{n} a_i X_i\right)} = \sum_{i=1}^{n} E(a_i X_i)$

$= \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \cdot \mu$ ← no index i anymore

$= \mu \underbrace{\sum_{i=1}^{n} a_i}_{1} = \boxed{\mu}$

$E(X+Y) \overset{r.i.i}{=} E(X) + E(Y)$

$E(aX) = a E(X)$

2) $\frac{K}{n}$ is unbiased Est. für p für binomial distribution

number of successes → K is just in arithmetic mean of X_i binomial distributed

n ← number of events

$X_i \to 1$, $X_i \to 0$

but $\frac{K}{n}$ is an estimator out of the class of estimators of question 1

$\Rightarrow \frac{K}{n}$ is also unbiased!

Exercise

$\bar X = \frac{1}{n}\sum X_i \Rightarrow \sum X_i = n\bar X$

1. $T(X_1, \dots X_n) = \sum_{i=1}^n a_i X_i$ with $\sum_{i=1}^n a_i = 1$ $(a_i > 0)$ is an unbiased estimator for μ.
2. $\frac{K}{X}$ is an unbiased estimator for p of the binomial distributio $B(n, p)$.
3. $\hat\sigma^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar X)^2$ is an unbiased estimator for the variance, if μ is unknown.
4. $\hat s^2(\mu) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is an unbiased estimator for the variance, if μ is known.
5. $\hat s^2(\bar X) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar X)^2$ is an asymtotically unbiased estimator for the variance, if μ is unknown.
6. Within the linear unbiased estimators for μ, the arithmetic mean $\bar X = \frac{1}{n}\sum_{i=1}^n X_i$ is an efficient estimator.

$(a-b)^2 = a^2 - 2ab + b^2$

$$\underbrace{1 + 1 + 1 + 1 + \dots + 1}_{10 \text{ times}} = \underbrace{\mu + \mu + \dots + \mu}_{10 \text{ times}}$$

$\hat\sigma^2$ ← 3) $E(\hat\sigma^2) = E\left(\frac{1}{n-1}\sum (X_i - \bar X)^2\right) = \frac{1}{n-1} E\left(\sum (X_i - \bar X)^2\right)$

$$= \frac{1}{n-1} E\left(\sum (X_i - \mu + \mu - \bar X)^2\right) = \frac{1}{n-1} E\left(\sum \left((X_i - \mu) - (\bar X - \mu)\right)^2\right)$$

$$= \frac{1}{n-1} E\left(\sum \left((X_i - \mu)^2 - 2(X_i - \mu)(\bar X - \mu) + (\bar X - \mu)^2\right)\right)$$

$$= \frac{1}{n-1}\left[\sum E((X_i - \mu)^2) - 2\sum E(X_i - \mu)(\bar X - \mu) + \sum (\bar X - \mu)^2\right)$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^n E(X_i - \mu)^2 - 2(\bar X - \mu)\left[\underbrace{\sum_{i=1}^n X_i}_{n\bar X} - \underbrace{\sum_{i=1}^n \mu}_{n\mu}\right] + (\bar X - \mu)^2 \underbrace{\sum_{i=1}^n 1}_{n}\right)$$

$n\bar X - n\mu = n(\bar X - \mu)$

$$= \frac{1}{n-1}\left[\underbrace{\sum E(X_i - \mu)^2}_{\sigma^2} - 2n E(\bar X - \mu)^2 + n(\bar X - \mu)^2\right]$$

$$= \frac{1}{n-1}\left[\sigma^2 - n\sigma^2_{\bar X}\right]$$

⇒ we ~~Know~~ the Variance of the arithmetic mean scale with $\frac{1}{n}$

$$= \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \frac{1}{n-1}\left[(n-1)\sigma^2\right] = \sigma^2$$

Exercise

1. $T(X_1, \dots X_n) = \sum_{i=1}^{n} a_i X_i$ with $\sum_{i=1}^{n} a_i = 1$ $(a_i > 0)$ is an unbiased estimator for μ.
2. $\frac{K}{X}$ is an unbiased estimator for p of the binomial distributio $B(n, p)$.
3. $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator for the variance, if μ is unknown.
4. $\hat{s}^2(\mu) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is an unbiased estimator for the variance, if μ is known.
5. $\hat{s}^2(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an asymtotically unbiased estimator for the variance, if μ is unknown.
6. Within the linear unbiased estimators for μ, the arithmetic mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an efficient estimator.

$$Var(aX_i) = a^2 Var(X_i)$$

$$\max x^{\frac{1}{2}} y^{\frac{1}{2}} \quad s.t. \; 1000 = 2x + 3y$$

$$T = \sum a_i x_i \qquad \boxed{\sum a_i = 1} \qquad a_i = \frac{1}{n} \text{ for efficiency}$$

$$Var(T) = Var(\sum a_i x_i) = \sum Var(a_i x_i) = \sum a_i^2 Var(x_i)$$

x_i are statistically identically distributed and independent

$$= \sum a_i^2 \sigma^2$$

$$\Rightarrow \min \sum a_i^2 \sigma^2 \quad s.t. \sum a_i = 1$$

$$L = \sum a_i^2 \sigma^2 + \lambda(1 - \sum a_i) = L(a_1, a_2, a_3, a_n, \lambda)$$

$$\Rightarrow \frac{\partial L}{\partial a_i} = 2 a_i \sigma^2 - \lambda = 0 \qquad \boxed{\text{for all } a_i}$$

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

$$\Rightarrow a_i = a_j \text{ for } i \neq j$$

$\Rightarrow$ all parameters have to be the same

$$\Rightarrow a_i = \frac{1}{n} \quad \text{because} \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n} = 1$$

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Exercise

1. $T(X_1, ...X_n) = \sum_{i=1}^{n} a_i X_i$ with $\sum_{i=1}^{n} a_i = 1$ $(a_i > 0)$ is an unbiased estimator for μ.
2. $\frac{K}{X}$ is an unbiased estimator for p of the binomial distributio $B(n, p)$.
3. $\boxed{\hat{\sigma}^2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator for the variance, if μ is unknown. σ^2
4. $\hat{s}^2(\mu) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is an unbiased estimator for the variance, if μ is known.
5. $\hat{s}^2(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an asymtotically unbiased estimator for the variance, if μ is unknown.
6. Within the linear unbiased estimators for μ, the arithmetic mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an efficient estimator.

unbiased gest by using in the definition of Variance

$\sigma^2 = \frac{1}{n} \sum (X_1 - \mu)^2$

$= E(\hat{s}(\mu))$

5) $E(\hat{\sigma}^2(X)) = E\left(\frac{1}{n}\sum(x_i-\bar{x})^2\right) = \frac{1}{n}\underbrace{E\left(\sum(x_i-\bar{x})^2\right)}_{(n-1)\sigma^2}$

$E(aX) = aE(X)$

$= \frac{n-1}{n}\sigma^2 \quad = \lim_{n\to\infty}\left(\frac{n-1}{n}\sigma^2\right) = \sigma^2 \lim_{n\to\infty}\left(\frac{n-1}{n}\right)$

result from (3)

$= \sigma^2 \lim_{n\to\infty}\left(1-\frac{1}{n}\right) = \sigma^2(1-0) = \sigma^2$ ✓

Exercise

7. Suppose $X \sim B(n, p)$: There is no unbiased estimator of the standard deviation of the binomial distribution.
8. Show $MSE := E([\hat{\theta}_n - \theta]^2) = Var(\hat{\theta}_n) - (E(\hat{\theta}_n) - \theta)^2$
9. Suppose $X_1, ..., X_n$ are independently Bernoulli-distributed, then $T(X) = \boxed{X_1 + ... + X_n}$ is a sufficient estimator for p (T is the total number of successes).

* Suppose $X_1, ..., X_n$ are independently normally distributed and σ^2 is known. Then $\bar{X}$ is a sufficent estimator for μ.

7) Suppose there is an unbiased estimator $\hat{\Theta}$

$$E(\hat{\Theta}) = \sum_{k=0}^{n} \Theta(k)\binom{n}{k}p^{k}(1-p)^{n-k}$$ $\leftarrow$ is a Polynom in p $\rightarrow$ This is differentiable

$p \geq 0$

8) $E((\hat{\Theta}-\Theta)^2) = E(\hat{\Theta}^2 - 2\hat{\Theta}\Theta + \Theta^2) = E(\hat{\Theta}^2) - 2E(\hat{\Theta}\Theta) + E(\Theta^2)$

$Var(\hat{\Theta}) = E((\hat{\Theta} - E(\hat{\Theta}))^2) = E(\hat{\Theta}^2 - 2\underline{E(\hat{\Theta})E(\hat{\Theta})} + \underline{(E(\hat{\Theta}))^2}) = E(\hat{\Theta}^2) - (E(\hat{\Theta}))^2$

$\rightarrow$ next lecture

Confidence level

The given probability of the **confidence interval** is called **confidence level**

confidence interval: $\hat{\theta}_a \leq \theta \leq \hat{\theta}_b$ with the lower level a and the upper level b

limits are also random Variables

Estimating the interval: $P(\hat{\theta}_a \leq \hat{\theta} \leq \hat{\theta}_b) = 1 - \alpha$

90%
95%
99%
most common probabilities

confidence level $1-\alpha$: With probability $1-\alpha$ the true parameter θ will be covered by the interval with the limits $[\hat{\theta}_a, \hat{\theta}_b]$. With probability α the true value is not covered.

Note the limits $hat\theta_a, \hat{\theta}_b$ are also random variables!

Keep in mind because in many statistics book the definition of confidence levels confidence intervals do not take this into account

One- and two-sided confidence intervals

One-sided confidence interval

Either the lower limit is $-\infty$ or the upper limit is $+\infty$. The Parameter θ has with probability $1-\alpha$ at most the value $\hat{\theta}_b$ or at least $\hat{\theta}_a$.

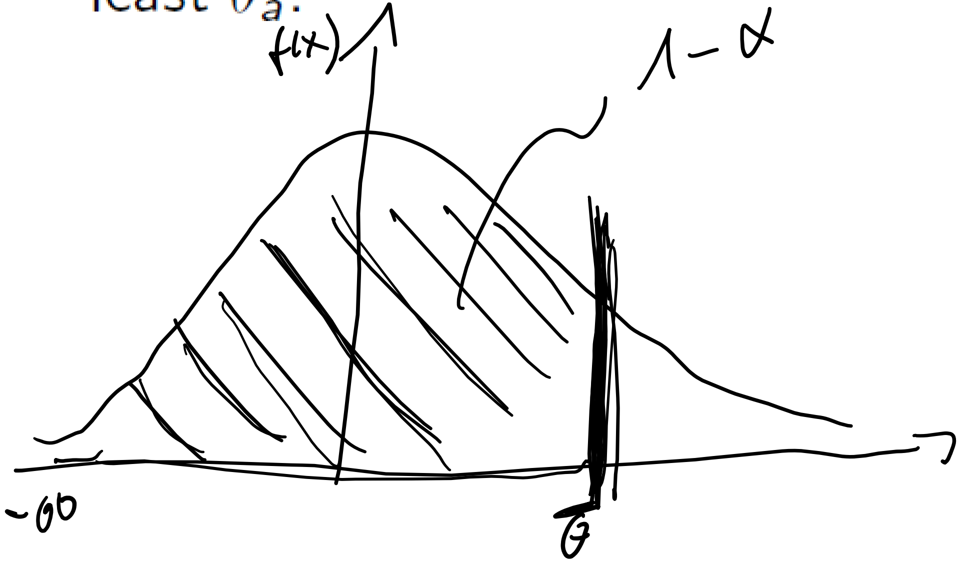
Two-sided confidence interval

We calculate the upper and lower limit ($\hat{\theta}_b$ und $\hat{\theta}_a$) of the interval, which should cover the unknown true parameter.

One- and two-sided confidence intervals

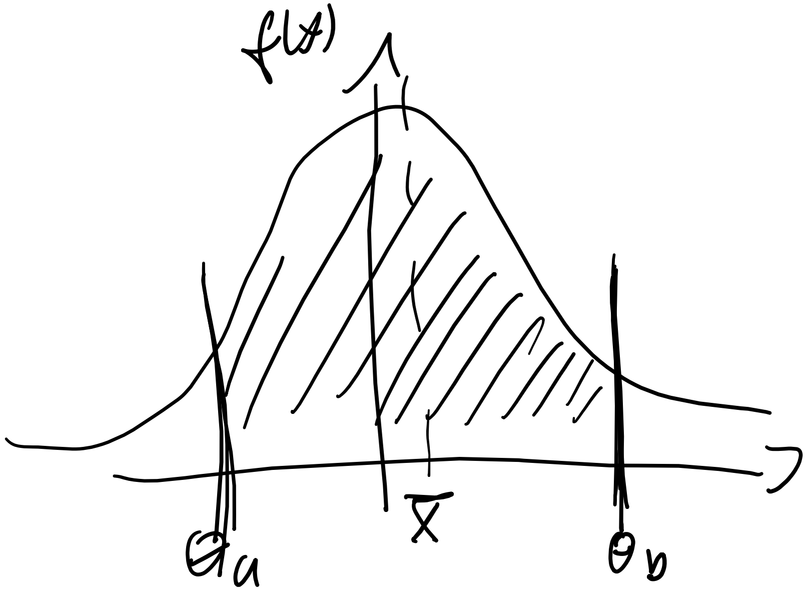
One-sided confidence interval

Either the lower limit is $-\infty$ or the upper limit is $+\infty$. The Parameter θ has with probability $1-\alpha$ at most the value $\hat{\theta}_b$ or at least $\hat{\theta}_a$.



Two-sided confidence interval

We calculate the upper and lower limit ($\hat{\theta}_b$ und $\hat{\theta}_a$) of the interval, which should cover the unknown true parameter.



Calculating confidence intervals

In order to calculate the confidence interval we need a knowledge about the distribution and the variance of the variables. If we do not know the **variance of the population**, we have to estimate this value via the random sample:

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

- For large samples ($\boldsymbol{n \geq 30}$) we can use the central limit theorem and after the standardization we can apply the **Standard-Normal distribution**.
- For small samples **($\boldsymbol{n < 30}$)** we use the **Student-t-Vedistribution** if we know, that the population is normally distributed. Doing this, we also standardize the variable.
- Additionally the sample size n has to be small compared to the size of the population N: $\frac{n}{N} < 5\%$ or $N = \infty$.

Confidence intervals for large samples $n \geq 30$

$$n \geq 30 \implies P\left(\underbrace{\bar{x} - z\frac{\hat{\sigma}}{\sqrt{n}}}_{\theta_u} \leq \mu \leq \underbrace{\bar{x} + z\frac{\hat{\sigma}}{\sqrt{n}}}_{\theta_b}\right) = 1 - \alpha$$

- $1 - \alpha$ confidence level
- $\bar{x}$ (sample mean) point estimator for μ
- z Value of the standard normal distributionfor a probability of $1 - \frac{\alpha}{2}$
- $\hat{\sigma}$ estimated standard deviation (if the variance is known,, we can use σ for $\hat{\sigma}$)
- n der Stichprobenumfang
- Since $n \geq 30$ we can use the normal distribution

(If the population is normally distributed ($X \sim N(\mu, \sigma^2)$) and the variance is known, we have without the restriction ($n \geq 30$) $\frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0, 1)$.)

$z = \frac{x - \mu}{\sigma}$

$\Rightarrow z\sigma + \mu = x$

standard deviation of X

$\Rightarrow$ we have for the standard deviation of $\bar{X}$ $\sigma_{\bar{x}}^2 = \frac{1}{n}\sigma_x$

$\Rightarrow \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

z is first number of the standard normal distribution for confidence level $1-\alpha$

Confidence intervals for samll samples $n < 30$

$n \geqslant 30 \Rightarrow z$

$$\boldsymbol{n < 30} \quad \Longrightarrow \quad P\left(\bar{x} - t_{n-1}\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1}\frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

if the population is normally distributed ($X \sim N(\mu, \sigma^2)$) and the variance is not known.

t_{n-1}: value of the Student-t-distribution with $n - 1$ number of degrees of freedom

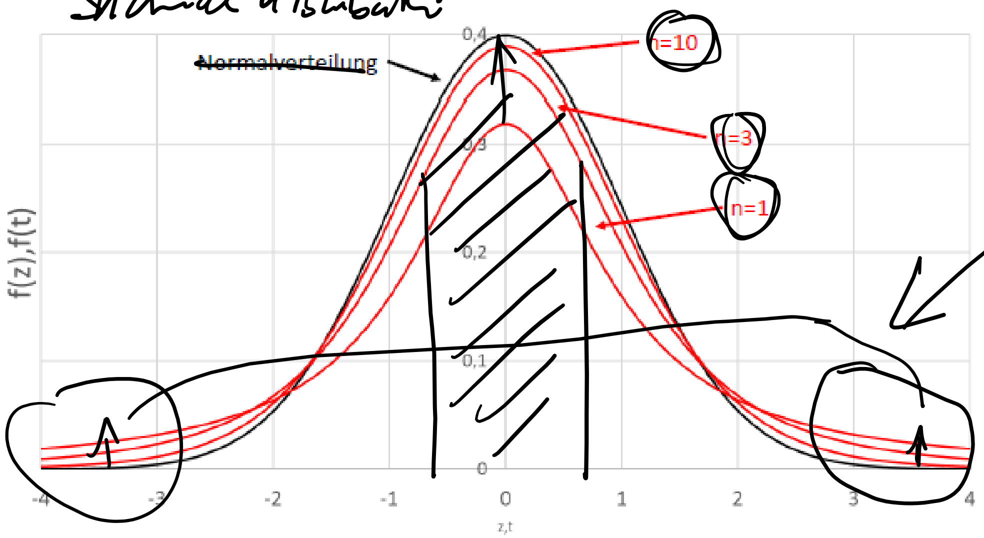
- depends of the degrees of freedom
- converges for large n to the normal distribution
- can be approximated by the normal distribution, if $n \geq 30$

Student-t-distribution

If the variance of a normally distributed population is no known, we replace within the standardization the variance σ with the sample variance $\hat{\sigma}$. The new variable Student-t distributed

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}} \sim t_{n-1}$$

Normal distribution



t-distribution has larger tales than the normal-distribution
⇒ "extreme" events are more probable

The t-dsitribution is similar to the normal distribution but broader. For increasing n the t-distribution reaches the normal distribution.

Student-t-distribution

$N(0,1) = f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2}$

The standardized estimator of the sample mean of a normally distributed data is not normally distributed but t distributed

Density function: $f_n(X = x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{2}\right)^{-\frac{n+1}{2}}$.

Erwartungswert $E(X) = 0$

Varianz $Var(X) = \frac{n}{n-2}$

$\Gamma\left(\frac{n+1}{2}\right)$ → so called gamma function

$\Gamma(x) = \int_0^\infty t^x x^{n-1} dx$

Examples:

- Calculating confidence intervals of expected values of normally distributed data not knowing the variance.
- Comparision of different random samples.
- Testing the coefficients within the linear regression assuming that the error term is normally distributed.

Example

Expected value, unkown variance, normally distributed population $\boldsymbol{n \geq 30}$

The mileage of wheels with a population (N=100.000) should be estimated via a random sample with $n = 625$. Within the sample, we obtain an average mileage of 36.000km and $\hat{\sigma}$=4.800km.

1. Which interval covers the true mileage of the population with probability 95%?
2. What changes, if the sample size increases to 2.500 (confidence level 95%)?
3. What changes, if the confidence level increases to 99% ($n = 625$)?

Example

Expected value, unkown variance, normally distributed population and $\boldsymbol{n < 30}$

What are the average working hours of BA-students of the Jade Hochschule (N=6000) per week in order to finance their studies? In order to answer this question a random sample out of all students is drawn: Within this sample we have:
$\bar{x} = 18,36h$ und $\hat{\sigma} = 3,42$.

Calulate the confidence interval to a confidence level of 90%− if the size of the random sample in WHV is $n = 16$. Assume, that the population normally distributed

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Confidence interval for a proportion

$$\boxed{n\hat{\pi}(1-\hat{\pi}) \geq 9} \Longrightarrow P\left(\hat{\pi} - z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1-\alpha$$

- Dichotomous attribute with a probability of success of π (i.e. $X = 0, 1$ with $P(X = 1) = \pi$).
- $\hat{\pi}$ proportion of successes in the random sample
- For $n\hat{\pi}(1-\hat{\pi}) \geq 9$ Z is approximately normally distributed:

$$\boxed{Z = \frac{\hat{\pi} - \pi}{\sqrt{\hat{\pi}(1-\hat{\pi})}}\sqrt{n} \sim \boxed{N(0,1)}}$$

$X_i \begin{matrix} \nearrow 1 \\ \searrow 0 \end{matrix} \Rightarrow$ proportion of successes

$\hat{\pi} = \bar{x}$

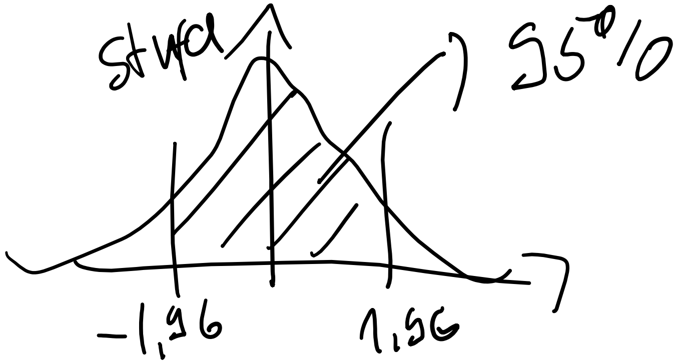
$X_i \sim B(1, \pi)$

$\Rightarrow \sigma^2 = \pi(1-\pi)$

$\hat{\pi}$ is the estimator for the true probability π

Example

Proportion



$$P\left(\hat{\pi} - z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \le \pi \le \hat{\pi} + z\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1-\alpha$$

$\hat{\pi} = \frac{96}{1200} = 8\%$

$z = 1,96$

$n = 1200$

In advance of the uk poll, we want to estimated the proportion of the liberal party in UK. The random sample (asked people) has a size of $n = 1200$ and 96 person answered to vote for the liberal party. Calculate the confidence interval to the confidence level 95%.

$$\Rightarrow 8\% \pm \sqrt{\frac{8\% \cdot 92\%}{1200}}$$

$$\Rightarrow 8\% \pm 1,5\%$$

$I_d \approx 6,5\%$ $\quad I_u \approx 9,5\%$

Confidence intervals for the Variance

Given the parent population is normally distributed ($X \sim N(\mu, \sigma^2)$), then

this is known from Data

$$Y = \boxed{(n-1)S^2/\sigma^2} \sim \chi^2(n-1)$$

Chi 2 - Distributed

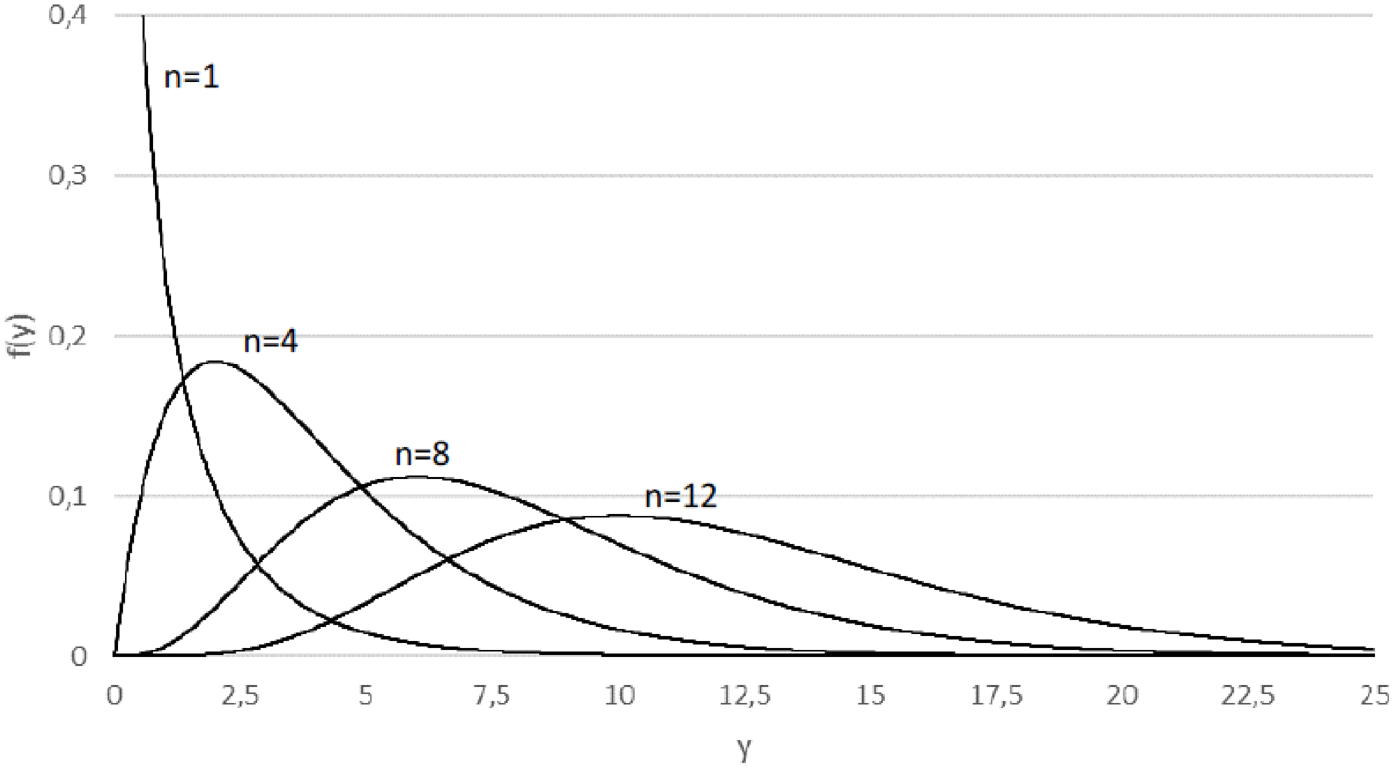
χ^2-distributed with $n-1$ degrees of freedom (S^2 random variable corresponding to the variance $\hat{\sigma}$ of the random sample). For values $\chi^2_{\frac{\alpha}{2}}$ and $\chi^2_{1-\frac{\alpha}{2}}$ for which the χ^2-distribution with $n-1$ degrees of freedom reaches the values $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$, we have:

$$P(\chi^2_{\frac{\alpha}{2};n-1} \leq Y \leq \chi^2_{1-\frac{\alpha}{2};n-1}) = 1-\alpha$$

$$\Longrightarrow$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha$$

χ^2-Distribution



In opposite to the normal and t-distribtion the χ^2-distributionis not symmetric. This means both limits of a two-sided confidence interval has to be calculated separately.

χ^2-distribution

Suppose $X_1, ..., X_n$ are independently identical standard normally distributed, then the distribution of the random variable $X = \sum_{i=1}^{n} X_i^2$ is called χ^2-distribution

Density function:	$f_n(X = x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}}, x > 0\,, 0 \text{ sonst}$
Expected value	$E(X) = n$
Variance	$Var(X) = 2n$

← this distribution also depends on the sample size

Examples:

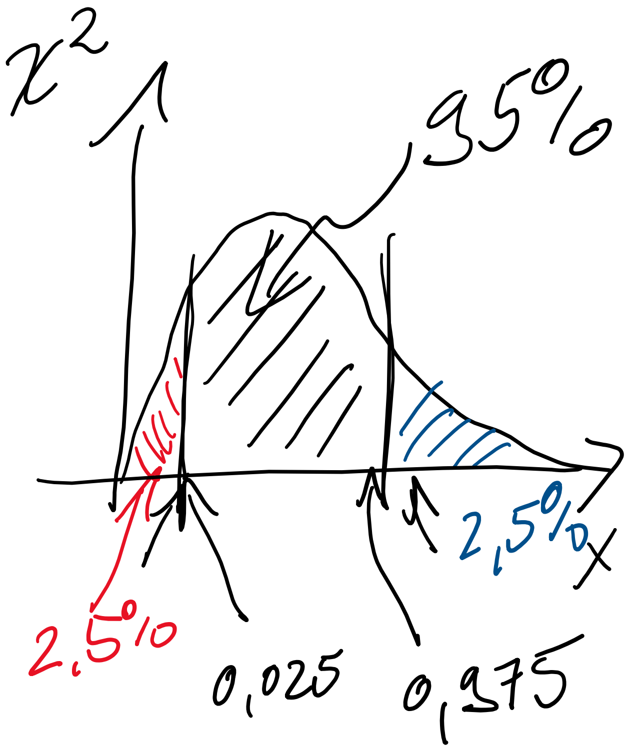
- Variance of a random sample of size n given the parent distribution is normally distributed
- χ^2-goodness of fit test
- χ^2-test of independence (contigency table!)

Example

Variance given a normally distributed parent population

$/\!/\ n$

A stock of the Oldtech company has within the last 19 years an average profit of $\bar{x} = 15$ US-Dollar. The uncertainty of the profit, estimated with the standard deviation is $\hat{\sigma} = 30$ US-Dollar. We assume that the profits are normally distributed. Given a confidence level of 95% calculate the confidence interval for the unknown variance σ^2 of the parent distribution.



$\chi_{0{,}025;18} = 8{,}231$

$\chi_{0{,}975;18} = 31{,}53$

$$\Longrightarrow$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha$$

$s^2 \hat{=} \hat{\sigma}^2 = 900 \qquad n-1 = 18$

$I_u = \dfrac{18 \cdot 900}{8{,}231} \approx 513$

$I_d = \dfrac{18 \cdot 900}{31{,}53} \approx 1968$

Confidence Intervall for the standard deviation $\hat{\sigma}$

$I_d \approx \sqrt{513} \approx 22{,}7$

$I_u \approx \sqrt{1968} \approx 44{,}4$

E2 1) $\bar{x} = 600.000$ $\hat{\sigma} = 90.000$ $n = 225$

(a) Calculate the standard deviation of the average profit?

(b) Calculate a confidence interval of μ for probability of error of $\alpha = 5\%$

~~(c) Extrapolate for a pro~~bability of error $\alpha = 1\%$ a confidence intervall for the whole sector.

a) $\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{90.000}{\sqrt{225}} = 6.000$

b) $\bar{X} \pm z \cdot \hat{\sigma}_{\bar{X}} \Rightarrow \bar{x} \pm z \cdot \frac{\hat{\sigma}}{\sqrt{n}}$

6.000

$z \mathrel{\hat{=}} z_{stud}$

c

0,975

$\Rightarrow 600.000 - 1{,}96 \cdot 6000 = I_d$

$600.000 + 1{,}96 \cdot 6.000 = I_u$

$I_d \approx 588240$

$I_u \approx 611759$

c) $\Rightarrow$ just multiply the limits with 12.100

$I_d = 12.100 \cdot 588240 \approx 7{,}07$ Mrd. €

$I_u = 12.100 \cdot 611759 \approx 7{,}44$ Mrd. €

Bill.

$\hat{\sigma}^2_{\bar{x}} = Var\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2} Var\left(\sum x_i\right) = \frac{1}{n^2} \sum Var(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} \hat{\sigma}^2 = \frac{1}{n^2} \cdot n\hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n}$

$\Rightarrow \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}}$

95%

2. A machine is cutting steel rods of a specific length. Out of the total production of $N = 150$, we take a random sample of $n = 9$. Measuring their length, we obtain:
184,2 mm, 182,6 mm, 185,3 mm, 184,5 mm, 186,2 mm, 183,9 mm, 185,0 mm, 187,1 mm, 184,4 mm.
Due to experience, we know that the parent distribution is normally distributed.

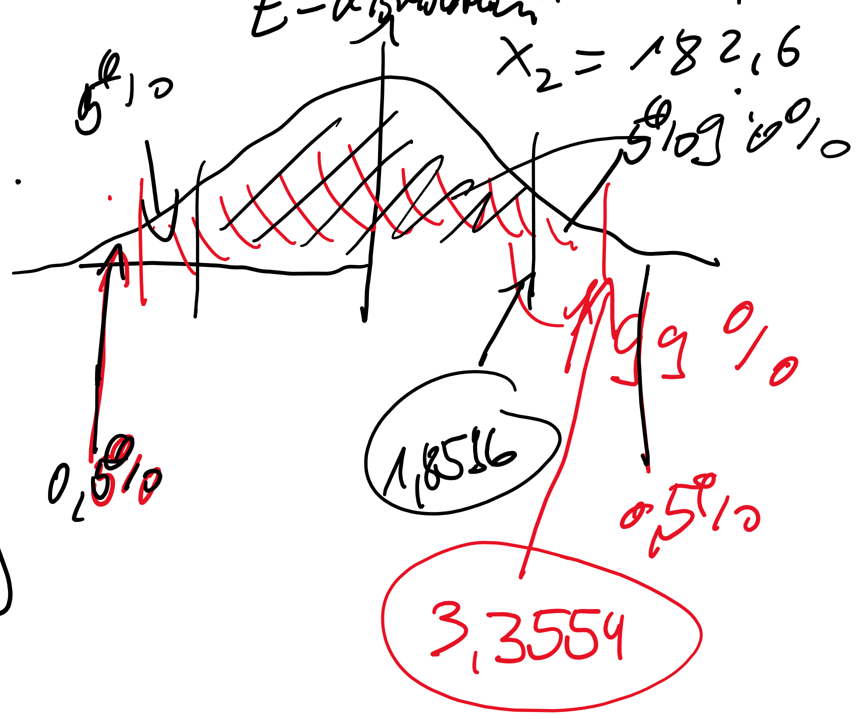
(a) Calculate unbiased estimators for mean and variance of the parent distribution.

(b) Calculate for μ a cofidence interval for the confidence leveles 0,9 and 0,99.

t - distribution

$x_1 = 184,2$

$x_2 = 182,6$



Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 184,8$ $\quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 1,725$

b)

90%: $\bar{x} \pm \frac{\hat{\sigma}}{\sqrt{n}} t_{0,95;n-1}$

$= 184,8 \pm \frac{1,75}{\sqrt{9}} \cdot 1,8596 \Rightarrow [183,99; 185,61]$

99%: $184,8 \pm \frac{1,75}{\sqrt{9}} \cdot 3,3554 \Rightarrow [183,33; 186,27]$

3. Within a random sample of 144 persons in WHV 75% answered, they would like to shop on sundays.

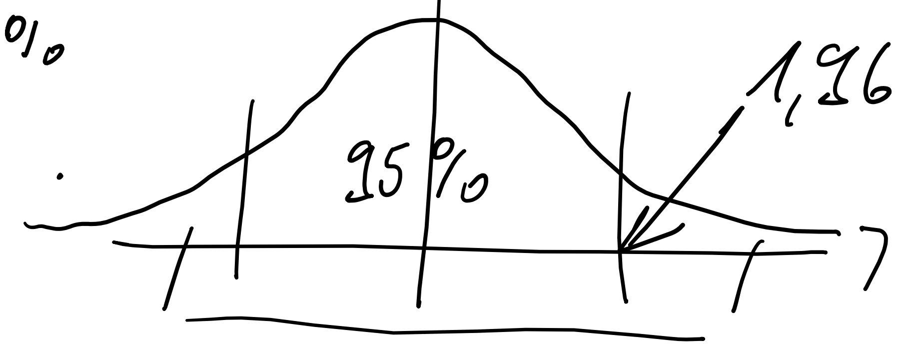
   (a) Calculate a confidence interval for the proportion of inhabitants of WHV, who would prefer a general shop opening on sundays for a probability of error of $\alpha = 0,05$ and $\alpha = 0,1$).

   (b) Calculated the confidence level, if the confidence interval would be represented by 75%±10%.

$\to$ normal distribution

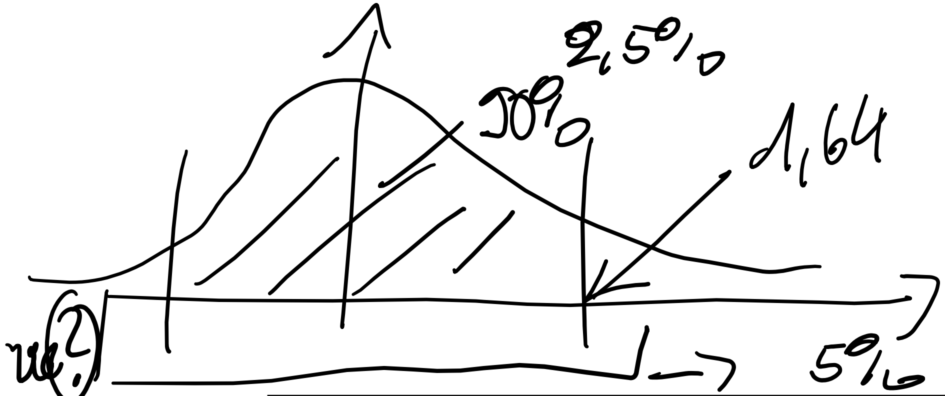
$\hat{\pi} = 0,75 \qquad \hat{\sigma} = \sqrt{\hat{\pi}(1-\hat{\pi})}$

corresponding to 97,5%



$\Rightarrow \left|\; \hat{\pi} \pm \sqrt{\frac{0,75(1-0,75)}{144}} \cdot 1,96 \Rightarrow [0,68 ; 0,82]\right.$

$\alpha = 5\%$



$\Rightarrow \left|\; \hat{\pi} \pm \sqrt{\frac{0,75(1-0,75)}{144}} \cdot 1,64 \Rightarrow [0,69 ; 0,81]\right.$

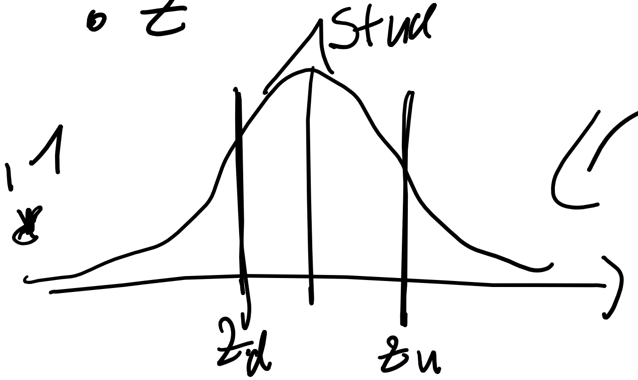
$\alpha = 1\%$

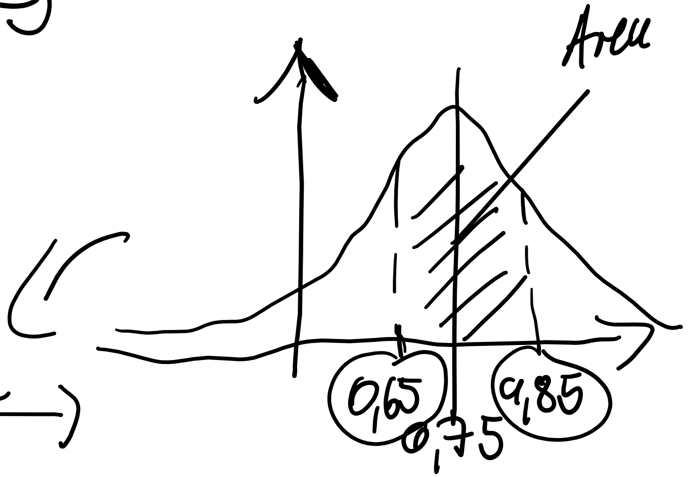
b) $0,75 \pm 0,1 \qquad [0,65 ; 0,85] \Rightarrow$ Confidence Level?

$\hat{\pi} \pm \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \cdot z$

$\Rightarrow \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \cdot z = 0,1$

$\sqrt{\frac{0,75(1-0,75)}{144}} \cdot z = 0,1$



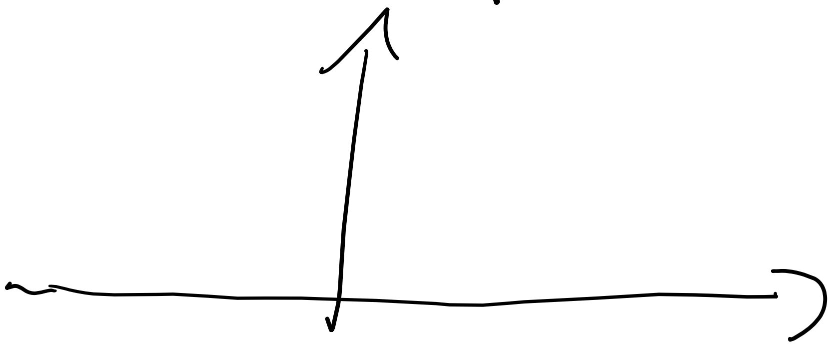


* $\sqrt{\frac{0{,}75(1-0{,}75)}{144}} \cdot z = 0{,}1 \Rightarrow z = \frac{0{,}1}{\sqrt{\frac{0{,}75(1-0{,}75)}{144}}} = 2{,}77$

→ corresponds to 99,72 %

$1 - 99{,}72\,\% = 0{,}28\,\%$

this we have to multiply by 2 in order to get the α-error because we a two-sided confidence interval



$\Rightarrow \boxed{\alpha = 0{,}56\,\%}$

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Hypothesis tests

Assume, you have a random sample and you want to test some hypothesis about for example some parameter of the distribution

Given a specific probability of error, we consider this probability as the probability of making a wrong decision. Thus, given a Hypothesis, the probability of error is the probability that the Hypothesis is not confiremed.

Example:

In your company, you want to know if your marketing activity has increased the awareness of a product. You know, that before the activity, the awareness was roughly 40% (i.e. 40% of the population have known the product). Now you make a survey with $n = 1000$ people asking for the knowledge after the marketing acitivity. Within this sample 420 People answered, that they know the product.

$\Rightarrow \frac{420}{1000} = 42\%$

$\Rightarrow 42\% > 40\%$

ok, more people are knowing my product

Question:

Has the awareness of the product increased within the total population??

$\Rightarrow$Beause of random fluctuations an answer via a point estimation is not possible!

Statistical hypothesistesting is asking:

Is it possible to reject the null hypothesis H_0 that the degree of awareness has not changed, because of the result of the survey with the random sample of $n = 1000$? Similar to the confidence intervalls one has to specify a degree error probability α, for example 5%.

Hypothesis testing

1. Step:

Create a questioning with two hypotheses contradicting each other:

H_0: Null hypothesis (incorporates always the equal sign) H_1: Alternative hypothesis (no equal sign)

Possibilities:

$H_0 : \theta \geq c$ $\qquad$ $H_1 : \theta < c$ left-sided

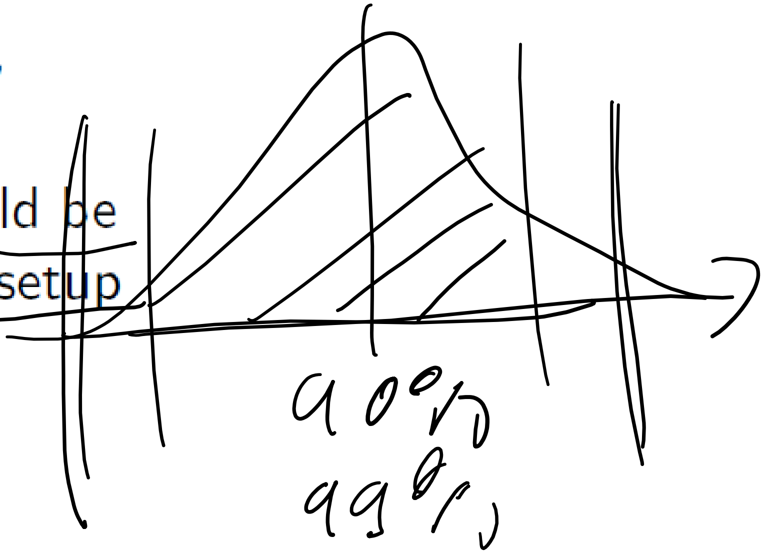
$H_0 : \theta \leq c$ $\qquad$ $H_1 : \theta > c$ right-sided

$H_0 : \boxed{\theta = c}$ $\qquad$ $H_1 : \theta \neq c$ two-sided

Error of 1. and 2. kind

- Error of 1. kind (or α-error): rejecting the null hypothesis, although it is true.
- Error of 2. kind (oder β-error): not rejecting the null hypothesis, although it is not true.

Mostly the error of 1. kind is used. Thus the error probability α should be "small" (but still the "right" *alpha* has always to be chosen from the setup of your experiment!):



$$P(H_0 \,\text{rejecting} | H_0 \,\text{true}) = \alpha$$

Problem:

If one choses a small α the probability not rejecting the null hypothesis is increasing, although it is not true. Generally, we choose an α-error of 1%, 5% or 10%.

Statistics A

Wilhelmshaven

This lecture will be recorded and Subsequently uploaded in the world-wide-web

Function translator (webpage)

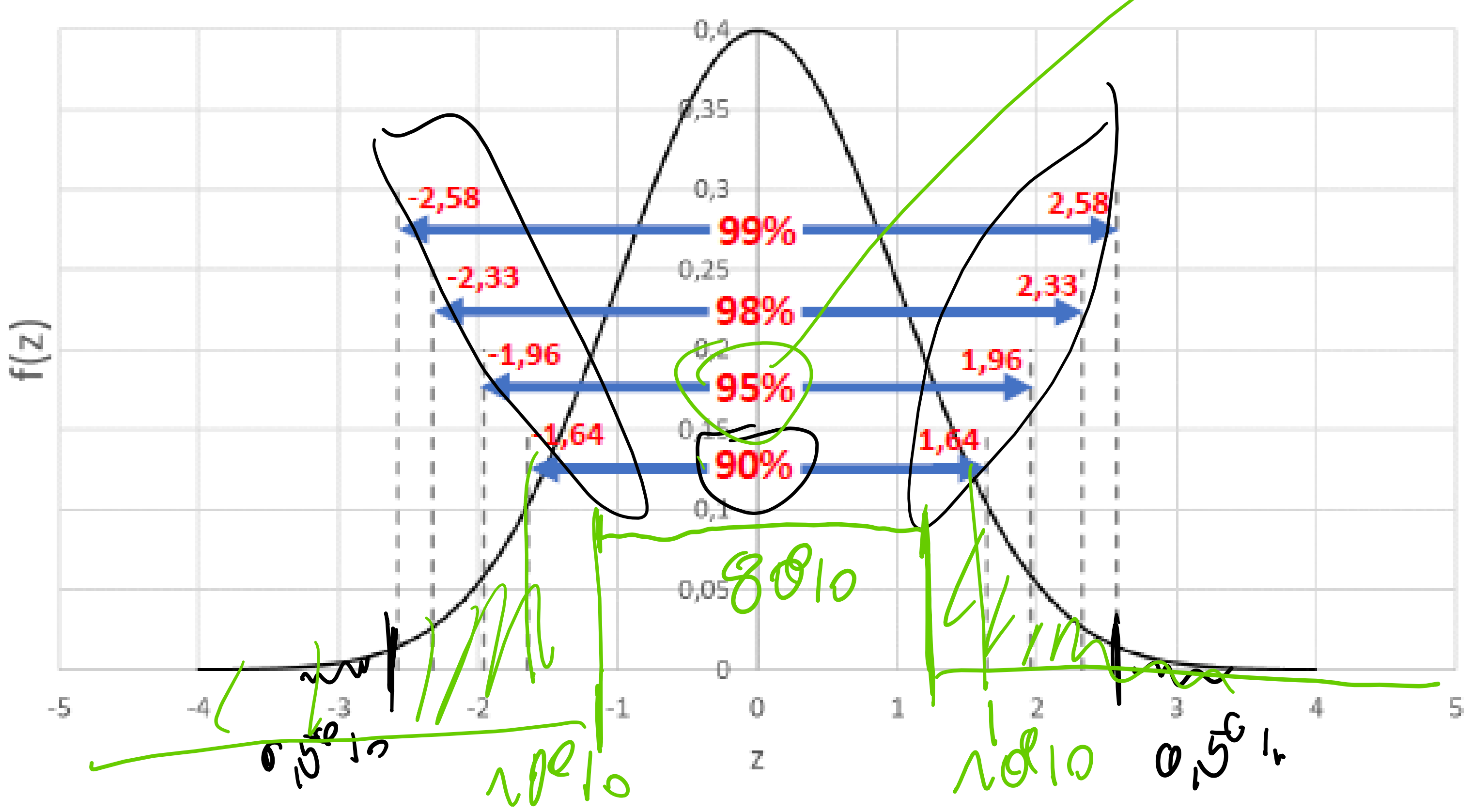
Function translator Excel 1 (add in)

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Relevant Values of the standard normal distribution

z	F(z)
-2.58	0.5 %
-2.33	1 %
-1.96	2.5 %
-1.64	5 %
-1.23	10 %

$\alpha = 10\%$ but
One-sided



Test-statistic

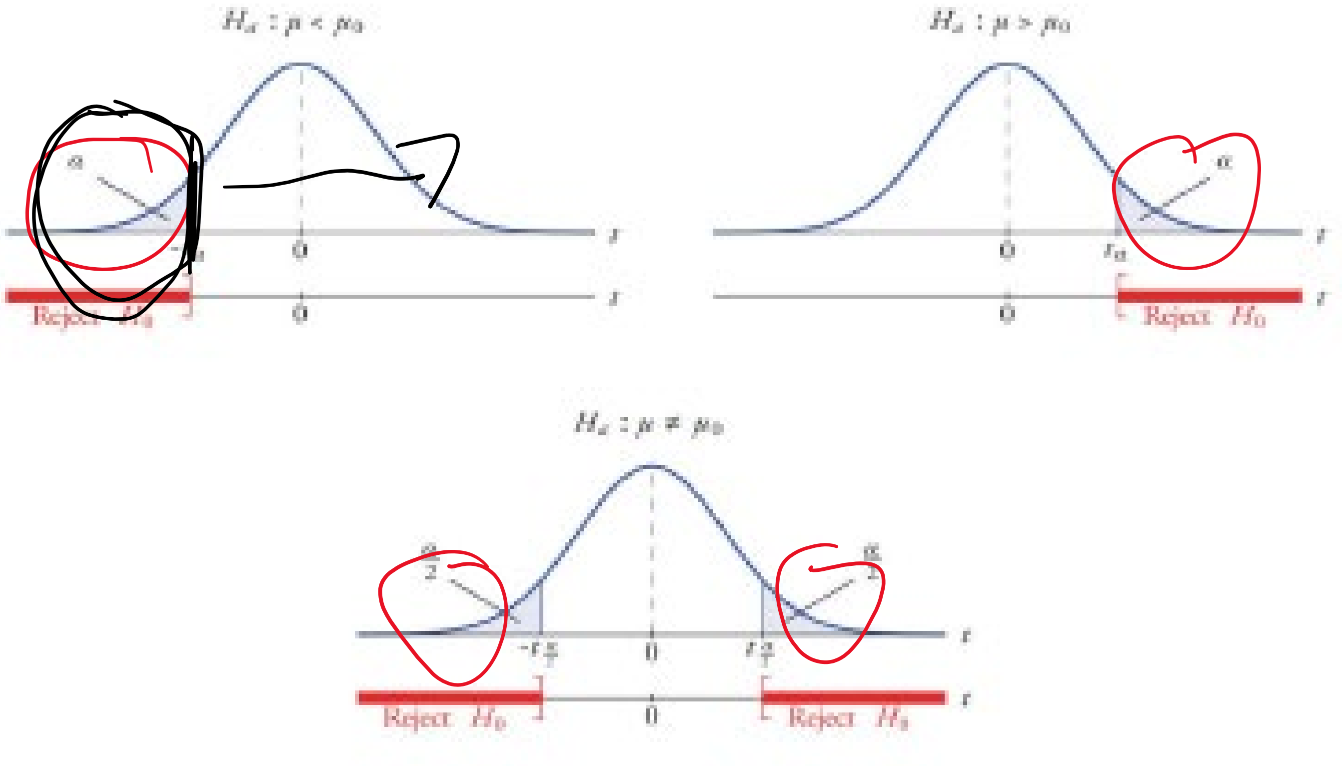
In order to calculated the error probability, we a random sample out of a probability distribution

$\Longrightarrow$ For every test, we need a test-statistic, where know the corresponding distribution or we make an assumption about the distribution.

Calculting the critical value given the niveau α:

- **left sided test**: reject H_0 if the calculated test statistic is lower than the critical value CV is obtained from the assumption α
- **right sided test**: reject H_0 if the calculated test statistic is larger than the critical value CV.
- **two sided test**: reject H_0 if the calculated test statistic is larger than the critical value $+CV$ oder lower than the critical value $-CV$.

Example: Test for the mean



Rejecting the null hypothesis for one- and two sided tests

Hypothesis testing for the mean

Test for μ:

H_0: $\mu \geq, \leq, = \mu_0$
H_1: $\mu <, >, \neq \mu_0$

(a) If $n \geq 30$ and unkown standard deviation

Test-statistic: $Z = \frac{\bar{X}-\mu}{\hat{\sigma}}\sqrt{n}$ and Z is standard normally distributed

(b) if $n < 30$ and unkown standard deviation

Test-statistic: $T = \frac{\bar{X}-\mu}{\hat{\sigma}}\sqrt{n}$ and T is t-distributed with $n-1$ degrees of freedom

Given a random sample we want to check some hypothesis.

Given a specific error probability the hypothesis will be rejected or not.

Example 1

Some machine produces boards with a minnimum width of 10mm. Boards with less than the minimum width cannot be used. The machine has an accurancy of 0,75mm (= standard deviation). Within a random sample of $n = 100$ we have an average width of 9,85mm. Is this shortfall signifikant to a level of ($\alpha = 10\%$) or only a random fluctuation?

1) H_0 : $\mu \geq \mu_0$ $\mu \geq 10\,mm$

H_1 : $\mu < 10\,mm$

2) Test-Statistic $\Rightarrow$ $Z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \frac{9{,}85 - 10}{0{,}75}\sqrt{100} \approx -2$ $\rightarrow$ has to be compared with the critical value CV

3) CV is obtained from α and the assumption on the distribution

$\alpha = 10\%$ Normal distribution

$\Rightarrow$ from a table or statistics software we have

$CV \approx -1{,}23$

4) Since $-2 < -1{,}23$

$\Rightarrow H_0$ is rejected

Example 2

Within the distribution of some product, the sellers obtain a fixed income and a provision depending on their revenue. On average a provision of 600 Euro is payed. In the current month the framework changed due to higher selling-prices (may be a reason for higher revenue) and the holiday season (may be a reason for lower revenue). Within a random sample of $n = 100$ we obtained an average provision of 585 Euro. Assuming the standard deviation of the provision is 80 and choosing $\alpha = 10\%$, do we have a significant lowering of the provision?

$\bar{x} = 585$

$\sigma = 80$

$n = 100 \Leftarrow$ SN-D

1) $H_0 : \mu \geq \mu_0$

$H_1 : \mu < \mu_0$

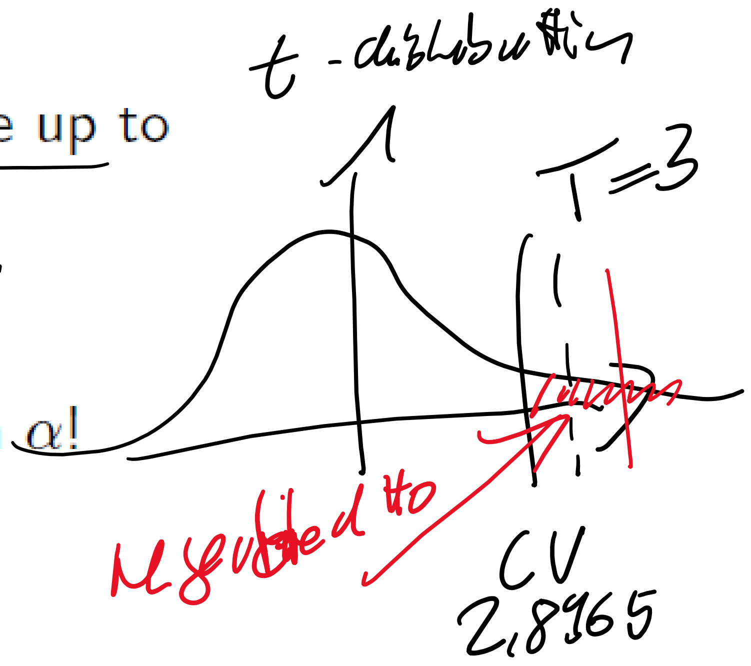
2) $Z = \frac{585 - 600}{80} \sqrt{100} \approx -1{,}875$

3) $-1{,}23$ CV $\Rightarrow$ 4) $-1{,}875 \leq 1{,}23$

Example 3

Some new pharmaceutical product should lower the blood pressure up to 20 points. This we know from the advertisement is this true?

small simple size $\Rightarrow$ T-distribution



Assume the parent distribution ist normally distributed. Choose $(\alpha = 0,01; n = 9; \bar{x} = 21,5; \hat{\sigma} = 1,5)$ Discuss the dependance on α!

1) $H_0: \mu \le 20$

$H_1: \mu > 20$

2) $T = \frac{\bar{x} - \mu}{\hat{\sigma}}\sqrt{n} = \frac{21,5-20}{1,5}\sqrt{9} = 3$ CV = 2,8965

3) CV comes from t-distribution 2,8965

4) $3 > 2,8965$

Test comparing two means

Assume we have two random samples with μ_1, μ_2 und $n_1, n_2 \geq 30$ and unknown variances:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

Test-statistic:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \quad \text{mit} \quad \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

is approximately normally distributed.

Example

Some insurance company has a telephone hotline. The phone calls are evaluated via some new software. Without having the software the phone calls during one month had a connecting on average of 22 minutes within 1500 calls. With the softwarethe connecting time was 18 minutes out of 1200 calls. The unbiased estimation of the standard deviations are given by:

$$\hat{\sigma}_1 = 15 \quad \hat{\sigma}_2 = 12$$

Has the connecting time significantly changed for $\alpha = 5\%$?

4. A provider of mobile games collected the playing time of $n = 120$ randomly drawn users. We suppose, that the playing times x_i are realizations of a normally distributed variable and obtain the following data:

$$\sum_{i=1}^{120} x_i = 21840 \qquad \sum_{i=1}^{120} x_i^2 = 4868856$$

(a) Calculate a confidence interval for the variance of the playing time given a confidence level of 99 %.

(b) Calculate the confidence interval of the standard deviation for a probability of error of $\alpha = 1\%$.

$x_1, x_2, \ldots, x_{120}$

$\sum x_i = 21840 \qquad \sum x_i^2 = 4868856$

$$\hat\sigma^2 = \frac{1}{n-1}\left(\sum_{i=1}^{120}(x_i - \bar{x})^2\right) \Leftarrow$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{120}(x_i^2 - 2x_i\bar{x} - \bar{x}^2)\right)$$

$+(\mu-\mu)$

$$\Rightarrow \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\left(\sum x_i\right)^2\right)$$

$$\Rightarrow \hat\sigma^2 \approx 7512 \quad \Big| \quad \hat\sigma \approx 86{,}67$$

χ^2

99% $\leftarrow$ χ^2 is not symmetrically

$$Y = \frac{(n-1)\hat\sigma^2}{\sigma^2} \sim \chi^2\text{-distributed}$$

$$I_l = \frac{(n-1)\hat\sigma^2}{\chi^2_{99{,}5\%;119}} \approx 5502 \qquad (\hat\sigma^2 = 7512)$$

$\chi^2_{99{,}5\%;119} \approx 162{,}5$

$\chi^2_{0{,}5\%;119} \approx 83$

from table or statistic software

$$I_u = \frac{(n-1)\hat\sigma^2}{\chi^2_{0{,}5\%;119}} \approx 10768$$

just take the root of the limits

$\Rightarrow I_l \approx 74{,}2$

$I_u \approx 103{,}8$

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Test comparing two means

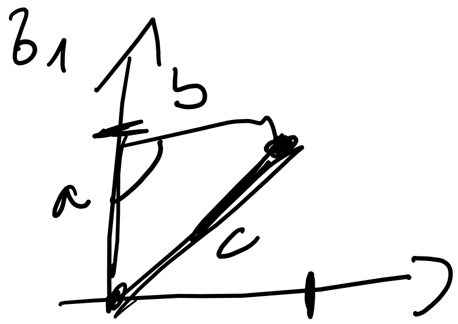
Assume we have two random samples with μ_1, μ_2 und $n_1, n_2 \geq 30$ and unknown variances:

H_0: $\boxed{\mu_1 = \mu_2}$
H_1: $\mu_1 \neq \mu_2$

Test-statistic:

$$\boxed{Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}}} \quad \text{mit} \quad \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

is approximately normally distributed.



$a^2 + b^2 = c^2$ σ_2

$c = \sqrt{a^2 + b^2}$

Example

Some insurance company has a telephone hotline. The phone calls are evaluated via some new software. Without having the software the phone calls during one month had a connecting on average of 22 minutes within 1500 calls. With the softwarethe connecting time was 18 minutes out of 1200 calls. The unbiased estimation of the standard deviations are given by:

$\bar{x}_1 = 22 \quad \bar{x}_2 = 18$

$n_1 = 1500 \quad n_2 = 1200$

$$\hat{\sigma}_1 = 15 \quad \hat{\sigma}_2 = 12$$

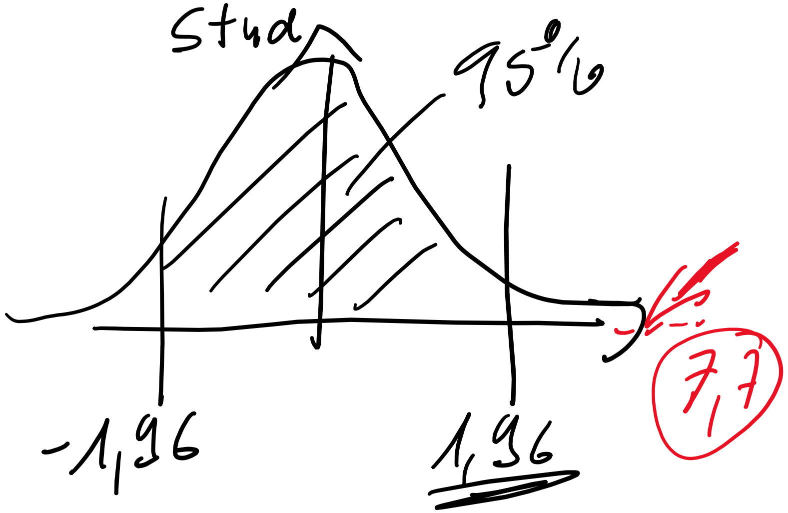
Has the connecting time significantly changed for $\alpha = 5\%$?

two sided

1) $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

2) $z = \dfrac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{x}_1, \bar{x}_2}} \approx \dfrac{22 - 18 - 0}{0{,}52} \approx 7{,}7$



$\hat{\sigma}_{\bar{x}_1, \bar{x}_2} = \sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}$

$= \sqrt{\dfrac{15^2}{1500} + \dfrac{12^2}{1200}} \approx 0{,}52$

3) CV $\Rightarrow$ we reject the null hypothesis

Test for proportion

Remember this is only a special case for testing for the mean

Given a random sample of $X_1, ...X_n$ with X_i are Bernoulli distributed

H_0: $\pi \geq, \leq, = \pi_0$
H_1: $\pi <, >, \neq \pi_0$

Teststatistic:

$$Z = \boxed{\frac{\hat{\pi} - \pi_0}{\hat{\sigma}_\pi}} \text{ mit } \hat{\sigma}_\pi = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$$

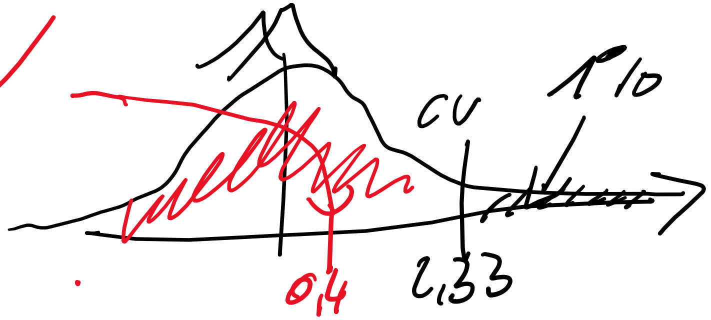
is approximately normally distributed if $\boxed{n\pi_0(1 - \pi_0) \geq 9}$

Example

one-sided

$400 \cdot 0{,}5 \cdot 0{,}5 = 100 \geq 9$ ✓

$\Rightarrow$ we can use the normal distribution



We will test the hypothesis, that a party A has obtained more than 50% of the votes for $\alpha = 0,01$. Within a random sample of $n = 400$ 204 persons, claimed that they have voted for party A.

(1) $H_0: \pi \leq 0{,}5 = \pi_0$

$H_1: \pi > 0{,}5$

(2) Teststatistic

$z = \frac{\hat{\pi} - \pi_0}{\hat{\sigma}_\pi}$

$\hat{\sigma}_\pi = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{0{,}5^2}{400}} = \frac{0{,}5}{20} = 2{,}5\% = 0{,}025$

$\hat{\pi} = \frac{204}{400} = 0{,}51 \Rightarrow z = \frac{0{,}51 - 0{,}5}{0{,}025} = 0{,}4$

(3) $CV = 2{,}33$ since $z < CV \Rightarrow H_0$ is not rejected

χ^2-test of independence

Remember the test for variance also followed the χ^2 - distribution

The χ^2-test of independence checks whether two nominal scaled variables are likely to be related or not.

- H_0: Two variables are independent.
- H_1: Two variables are not independent.

The χ^2-test of independence is based on the contingency table. With the following test statistic, we descide if we reject the Null hypothesis or not for a given level of α

empirical probability

theoretical probability assuming both variables are independent!

$$\chi^2_{emp} = n \sum_{i=1} \sum_{j=1} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

approximately χ^2-distributed with $f = (k-1)(l-1)$ degrees of freedom:

Example

An insurance company wants to increase the number of sold household insurances and is planning a marketing activity. In order to address the right people, it is assumed, that customers who have already a life insurance policy are risk averse and therefore are interested to buy also a household insurance. This is evaluated via the own data of the insurance company. Test the two variables (X: having a household insurance or not/Y: having a life insurance or not) for independence 1% $\boxed{\alpha = 1\%}$

two-sided →

1. H_0: X, Y are independent
   H_1: X, Y are not "
2. Test-statistic $\chi^2_{emp} \approx 107$
3. CV = 7,88 ⇒ H_0 is rejected

$\chi^2(1)$

0,5%

0,5%

	no household	household	
no Life	436	511	547
Life	121	481	602
	557	992	1549

p-emp	no h	h	
no	28%	33%	61%
l	8%	31%	39%
	36%	64%	100%

p-theo	no h	h	
no	22%	39%	61%
l	14%	25%	39%
	36%	64%	100%

$\Rightarrow \chi^2_{emp} \approx 107$

$f = (k-1)(l-1) = (2-1)(2-1) = 1$

$\chi(f=1)$

Statistics A

Wilhelmshaven

This lecture will be recorded and Subsequently uploaded in the world-wide-web

Function translator (webpage)

Function translator Excel 1 (add in)

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Scatterplot

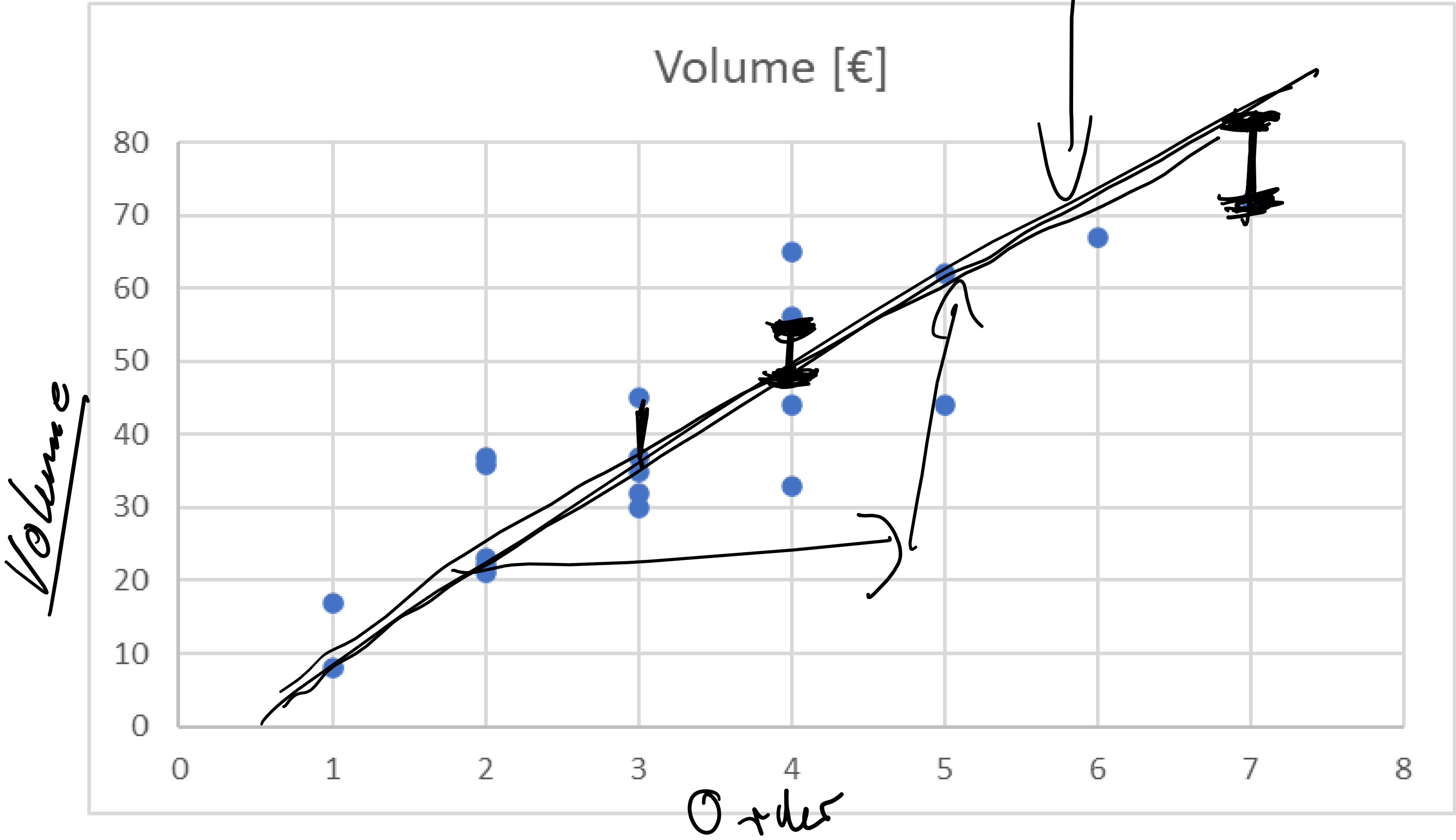
Graphical representation of the attributes of two variables within a two-dimensional coordinate-system.

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Order	2	4	3	6	2	1	5	3	4	2	2	7	4	3	2	1	5	4	3	3	1	6
Volume [€]	36	44	30	67	21	8	62	35	33	37	22	72	65	45	23	17	44	56	32	37	22	48

Scatterplot

How should we fit this line into the data?

X-coordinate: order
Y-coordinate: Volume [Euro]



Assumption:
The higher the number of orders, the higher is the volume

(4-9)

Linear Regression

Example 2:

A company wants to know how the sales volume depends on visits of representatives.

The company has collected the following data

	Visits	Sales volume [€]
year	x	y
2008	9	24
2009	11	33
2010	5	10
2011	13	29
2012	20	42
2013	12	24

We also want to answer the question, from a descriptive point of view, which sales volume we can expect, if we increase the visits up to $x=30$

Linear Regression

- Via a linear regression we want to analyse the dependence between two variables
- We distinguish between a dependent variable (y) and an independent variable (x)
- (x) is also called the explanatory variable
- A change of (x) implicates a change (y)
- The linear regression defines an empirical relation via a linear function
- i.e. $y = f(x) = 3 + \cancel{2}x$ or generally $y = a + bx$

point on the vertical axes

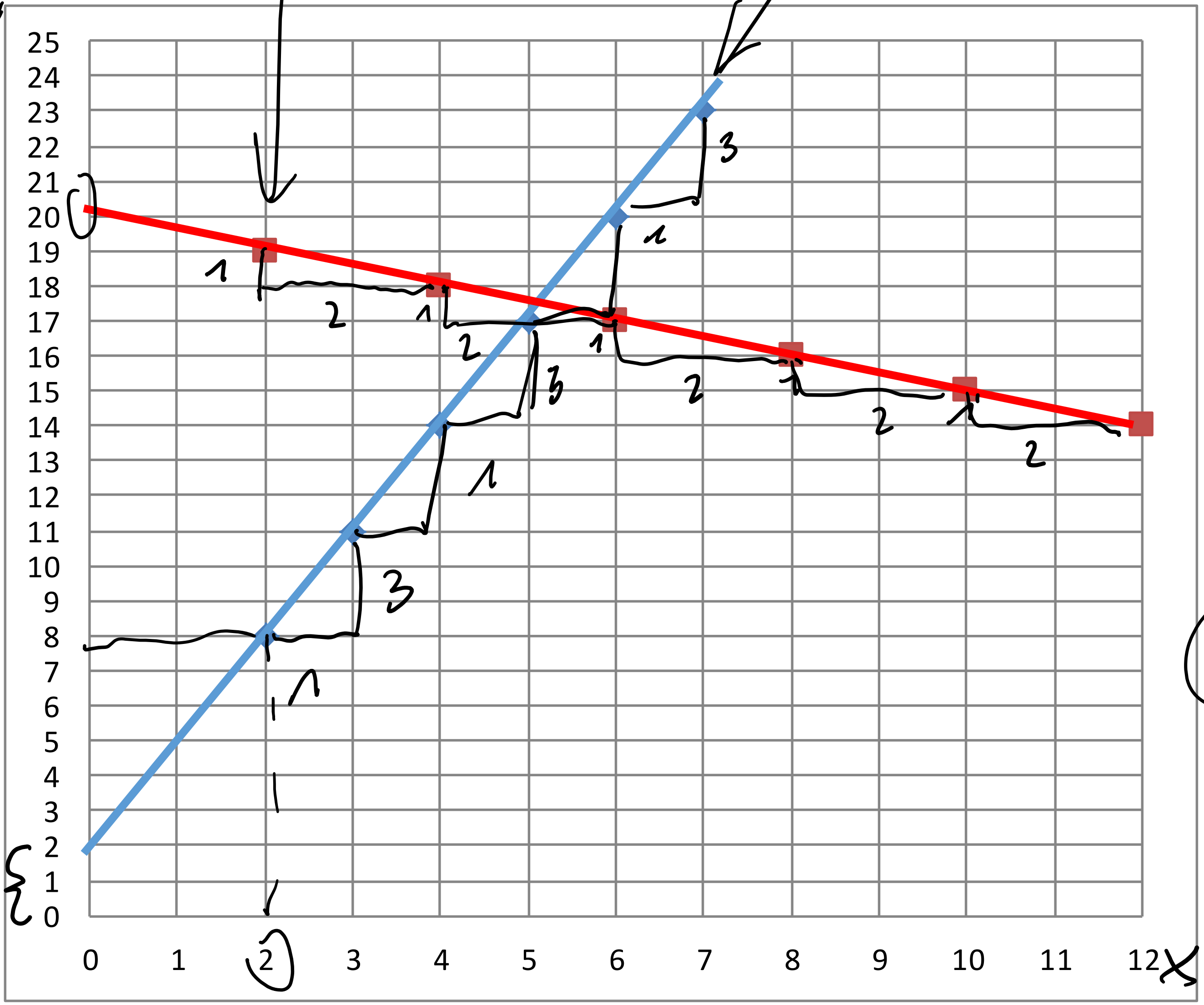
vert

slope of the line

$19 = a + (-\frac{1}{2})\cdot 2$
$= a - 1$
$\Rightarrow a = 20$

$x = 2 \quad y = 19$

Simple Examples



blue data means 3

$\Delta x = 1 \Rightarrow \Delta y = 3$
true for all data points

$y = a + bx$

$\Rightarrow b = \frac{\Delta y}{\Delta x} = 3$

and $a = 2$

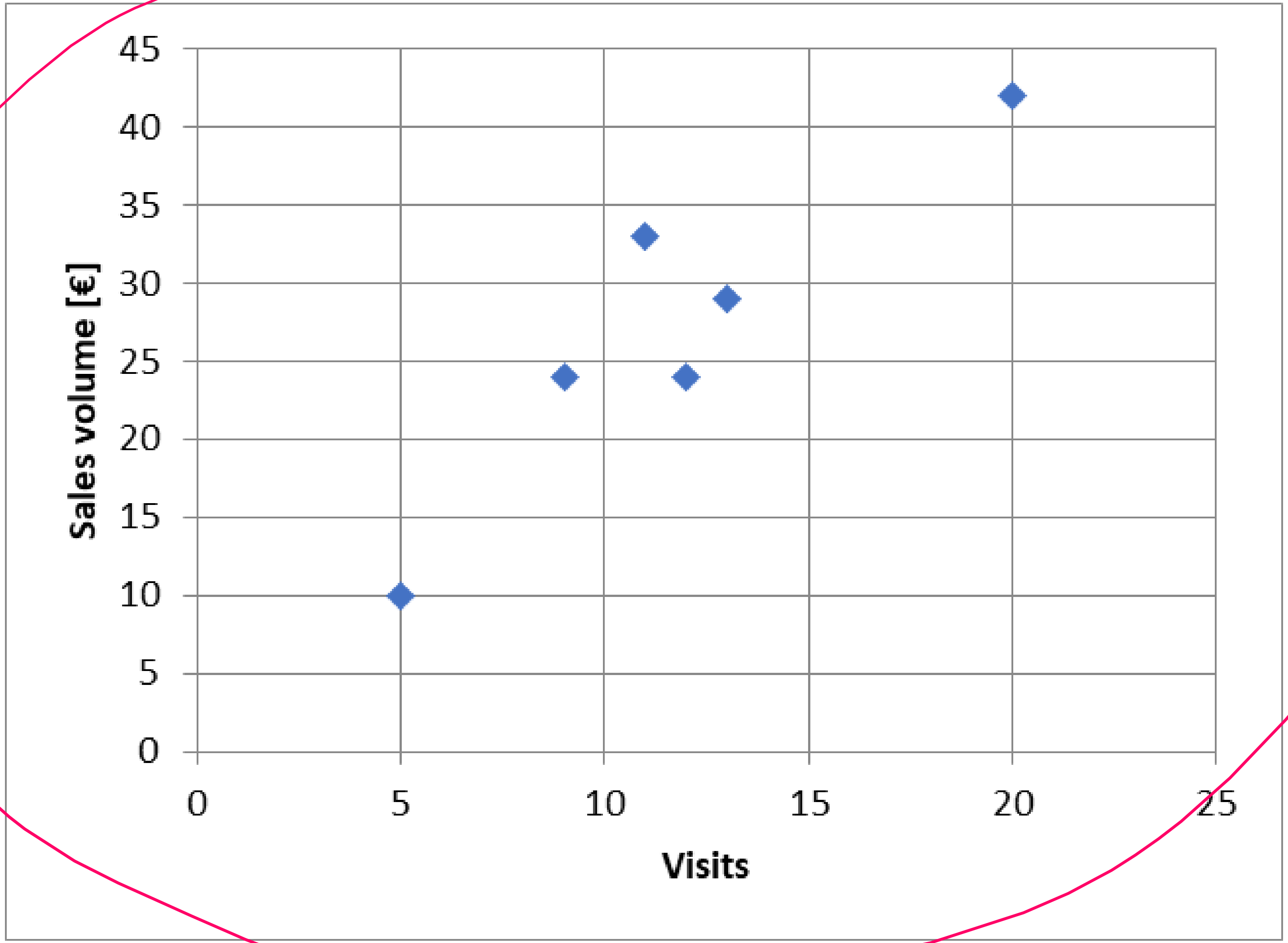
same data point $(x = 2, y = 8)$

$\Rightarrow 8 = a + 3 \cdot 2 \Rightarrow \quad a = 2$

red points
$a = 20$
$b = -\frac{1}{2}$

Determine graphically a regression line for the blue and red data points. What kind of dependence do we have?

Example



Try to fit graphically (by eye) a regression line through the data points

Optimizing problem

How to calculate a regression line $\hat{y}_i = a + bx_i$?

The coefficients a and b has be calculated from the collected data.

$\hat{y}_i$	Value on the regression line
y_i	collected data point
$\varepsilon_i = y_i - \hat{y}_i$	Difference between the data point and the theoretical value on the regression line

error, which we make by the assumption $\hat{y}_i = a + bx_i$

Determine a and b such, that the sum of all quadratic differences ε_i^2 is minimized

aim: $\sum_{i=1} \varepsilon_i^2 \to \min$

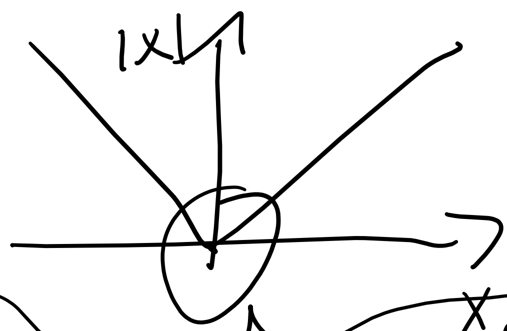
Least squared method

The formal optimizing problem

$$F(a,b) = \sum_{i}^{n}(y_i - (a + bx_i))^2 = \sum_{i}^{n}\epsilon_i^2 \to min$$

4

$\sum \epsilon_i$



not differentiable =) no derivative =) no $f'(x)$

Connected with the normal distribution

why this square

1) no problems with $\pm$ is dev difference

$(y-\bar{y})$

$|y-\bar{y}|$

2) $(y-\bar{y})^2$ is differentiable and $|y-\bar{y}|$ is not

3) Normal distribution

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

Least squared method

$f(b) = 5b \Rightarrow f'(b) = 5$

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$f(b) = bx_i$

$\Rightarrow f'(b) = x_i$

The formal optimizing problem $F(a, b) = \sum_i^n (y_i - (a + bx_i))^2 = \sum_i^n \epsilon_i^2 \rightarrow min$

$\frac{\partial F}{\partial a} = \left[\sum_{i=1}^{n} 2(y_i - (a + bx_i))\right](-1) = 0$

$\Rightarrow \sum_{i=1}^{n} (y_i - (a + bx_i)) = 0 \quad (1)$

FOC

$\frac{\partial F}{\partial b} = \left[\sum_{i=1}^{n} 2(y_i - (a + bx_i))x_i\right](-1)$

$\Rightarrow \sum_{i=1}^{n} (y_i - (a + bx_i))x_i = 0 \quad (2)$

$(1) \Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i \Rightarrow \cancel{n}\bar{Y} = \cancel{n}a + \cancel{n}b\bar{x}$

$\Rightarrow \bar{Y} = a + b\bar{x}$

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

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

$\Rightarrow$ means, the data point $(\bar{x}, \bar{y})$ is on the optimal regression line

Least squared method

$$\frac{\partial F}{\partial a} = -2\left[\sum_i^n (y_i - (a + bx_i))\right] = 0$$

The formal optimizing problem

$$\frac{\partial F}{\partial b} = -2\left[\sum_i^n x_i(y_i - (a + bx_i))\right] = 0 \quad (2)$$

(2) $\Rightarrow \sum_{i=1}^{n} (x_i y_i - a x_i - b x_i^2) = 0$

$\Rightarrow \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} x_i^2 = 0$

since (1) $\bar{y} = a + b\bar{x}$

$\Rightarrow a = \bar{y} - b\bar{x}$ (1)

$n\,\overline{x \cdot y} - n a \bar{x} - n b \overline{x^2} =$

$\Rightarrow \overline{x \cdot y} - \bar{x}(\bar{y} - b\bar{x}) - b\overline{x^2} = 0$

$\Rightarrow \overline{x \cdot y} - \bar{x}\bar{y} + b\,\bar{x}\cdot\bar{x} - b\,\overline{x^2} = 0$

$\overline{x \cdot y} - \bar{x}\cdot\bar{y} = b(\overline{x^2} - \bar{x}^2) \Rightarrow b = \dfrac{\overline{x \cdot y} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}$

Known data

put this in (1) and we obtain for a:

$$a = \bar{y} - \frac{\overline{x \cdot y} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}\,\bar{x}$$

Least squared method

The formal optimizing problem

$$F(a,b) = \sum_i^n (y_i - (a + bx_i))^2 = \sum_i^n \epsilon_i^2 \to min$$

FOC:

$$\frac{\partial F}{\partial a} = -2\left[\sum_i^n (y_i - (a + bx_i))\right] = 0$$

$$\frac{\partial F}{\partial b} = -2\left[\sum_i^n x_i (y_i - (a + bx_i))\right] = 0$$

2 dimensional linear system which can directly be solved

$$\frac{\partial F}{\partial a} = -2\left[\sum_{i}^{n}(y_i - (a + bx_i))\right] = 0$$

Least squared method

$$\frac{\partial F}{\partial b} = -2\left[\sum_{i}^{n}x_i(y_i - (a + bx_i))\right] = 0$$

Formulas

:

$\bar{x}$ $\bar{y}$

$\overline{x \cdot y}$

$\overline{x^2}$ $\bar{x}^2$

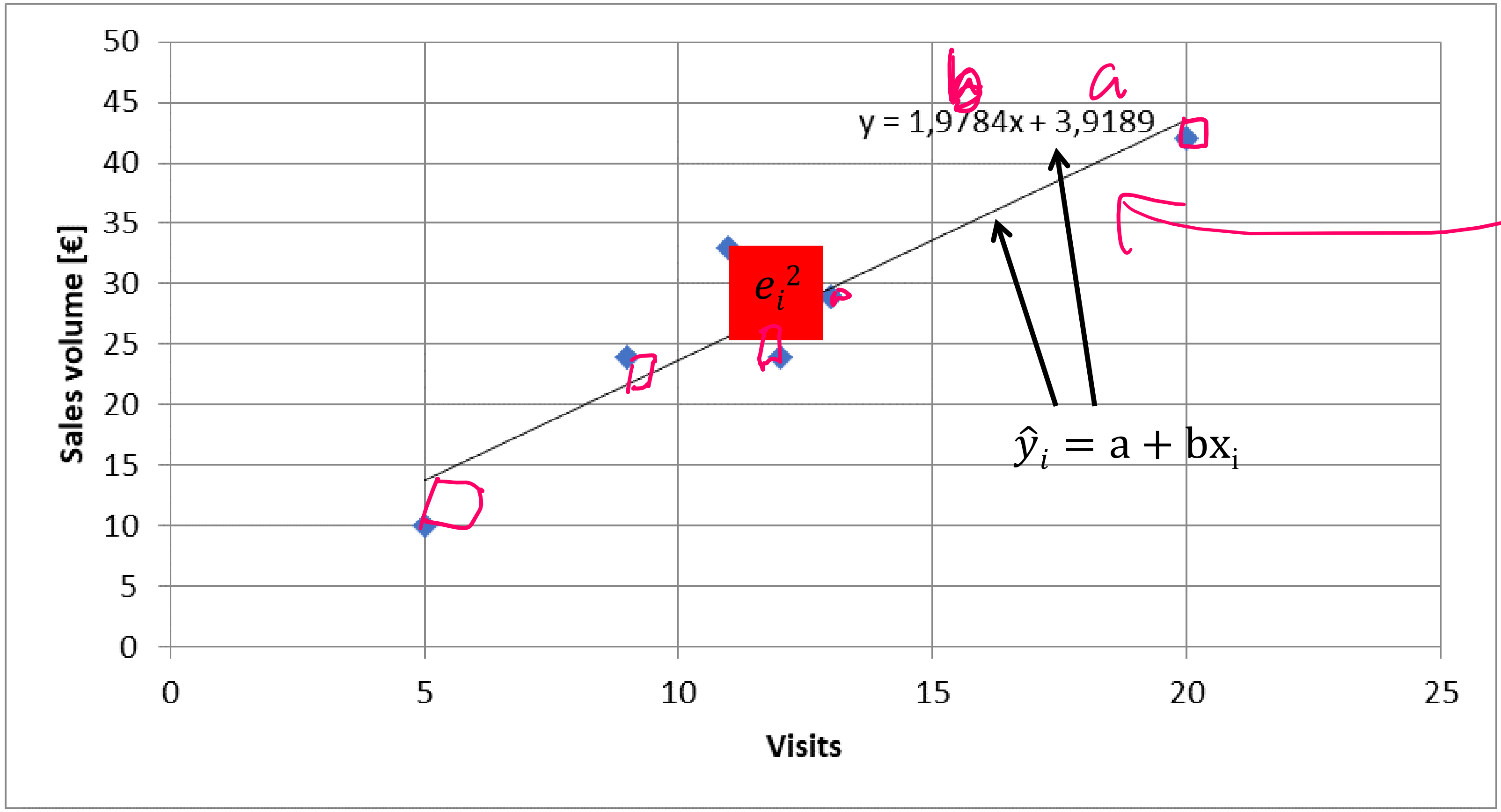
$$a = \frac{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \bar{y} - b\bar{x}$$

$$b = \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{Cov(x, y)}{Var(x)}$$

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Covariance

Calulating the regression line from empirical data



This line minimizes

$\sum_{i=1}^{n} \square_i$

Lineare Regression

	Visits	Sales volume [€]			
year	x	y	xy	x^2	
2008	9	24			
2009	11	33			
2010	5	10			
2011	13	29			
2012	20	42			
2013	12	24			

$\bar{x} \cdot \bar{y} = 315$

$\bar{x}$ $\bar{y}$

$11{,}67$ 27

$\overline{x \cdot y} = 355{,}67$

$\overline{x^2} = 156{,}67$

$\bar{x}^2 = 136{,}11$

$$b = \frac{\underset{\overline{x \cdot y}}{355{,}67} - \underset{\bar{x} \cdot \bar{y}}{315}}{\underset{\overline{x^2}}{156{,}67} - \underset{\bar{x}^2}{136{,}11}} \approx 1{,}97$$

$$a = \bar{y} - b\bar{x} = 27 - 1{,}97 \cdot 11{,}67 \approx 3{,}92$$

Extrapolation for x=30

	Visits	Sales volume [€]			
year	x	y	xy	x^2	yhat
2008	9	24			
2009	11	33			
2010	5	10			
2011	13	29			
2012	20	42			
2013	12	24			

yhat(30)=

Regressions coefficients

- Regressions coefficient b
  - Slope of the regression line.
  - Determines the marginal effect of a change of one unit of the independent variable x onto the dependet variable y
- Regressions coefficient a
  - The value of the dependent variable if the independent variable x=0
    → intercept of the vertical axes

Linear regression

- Measures the linear dependence between two variables

- This dependence can be interpreted, that one variable is quantitatively influencing another variable

- The linear regression is an instrument for forecasting

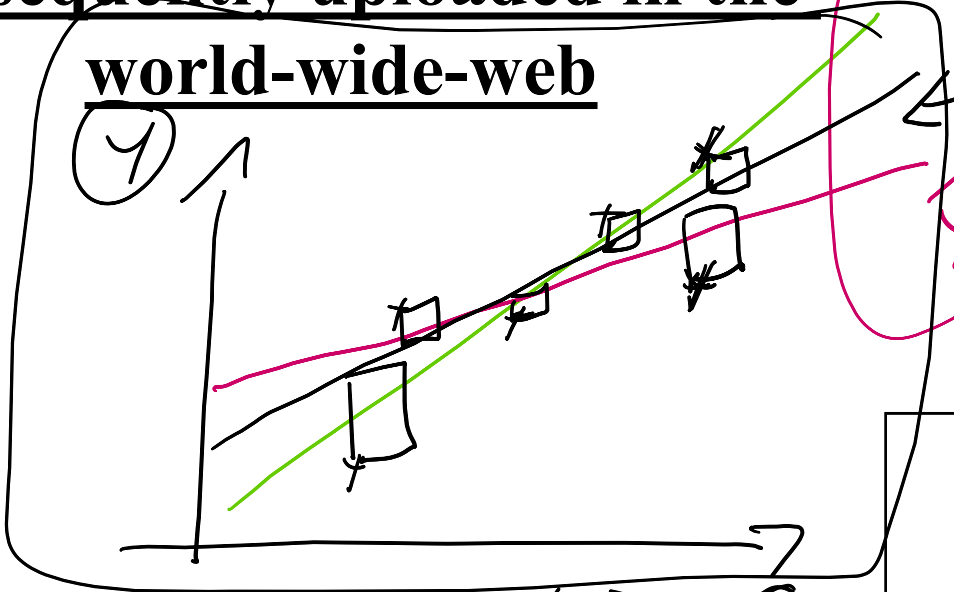
Least squares

Statistics A

was developed by C.F. Gauss, when he measured the distances in northern Germany

Wilhelmshaven

This lecture will be recorded and Subsequently uploaded in the world-wide-web



$\hat{y}_1 = a_1 + b_1 x$

$\hat{y}_0 = a_0 + b_0 x$

$\hat{y}_2 = a_2 + b_2 x$

$\Rightarrow$ choose this linear line for which we have

$\sum \square \to \min$

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Correlation analysis

- In principle for all data sets, we can calculate a regression line.

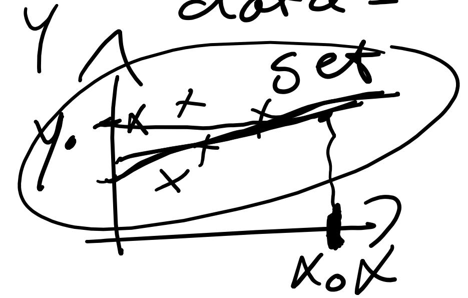
- But we are also interested in the question meaningful is this calculated dependence?.

- For this, we use the correlation analysis, wich gives the possibility to measure the strength of the dependence

- For this, we use the correlation coefficient of Bravais-Pearson which is lying between -1 and +1

$$\hat{y}_i = a + \boxed{b}\, x_i$$

can be fitted to every data-set



Covariance

- The covariance is a measure of the joint variability of two random variables (X,Y), defined by:

$$\mathrm{cov}(X,Y) = E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right]$$

with the unbiased estimator

$$\hat{\sigma}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_j - \bar{x}\right)\left(y_i - \bar{y}\right)$$

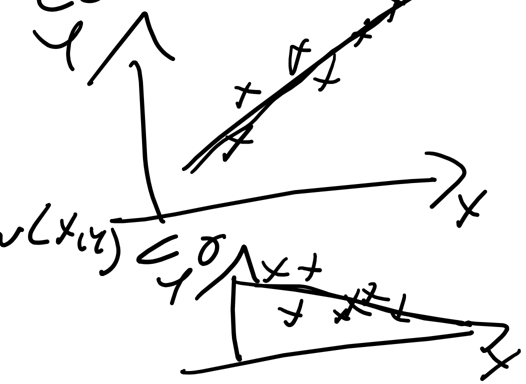
$\hat{\sigma}_{xx}$

Variance $= E\left[\left(x - E(x)\right)^2\right]$

$\mathrm{cov}(x,x) = \mathrm{Var}(x,x)$

if $\mathrm{cov}(X,Y) > 0$

$\mathrm{cov}(x,y) < 0$



- If the greater values of X correspond with the greater values of Y the covariance is positive. In the opposite case, when the greater values of X correspond to the lesser values Y, the covariance is negative. The sign of the covariance shows the tendency in the linear relationship between the variables.

- The magnitude of the covariance can hardly be interpreted, since it is not normalized.

- Therefore, we use the normalized version, called correlation coefficient, showing the strength of the linear relation.

Correlation coefficient

How do we do normalization?

$$R = \frac{N\sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{\sqrt{\left(N\sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i\right)^2\right)\left(N\sum_{i=1}^N y_i^2 - \left(\sum_{i=1}^N y_i\right)^2\right)}}$$

$$= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sum_{i=1}^N (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^N (x_i-\bar{x})^2 \sum_{i=1}^N (y_i-\bar{y})^2}}$$

$$\frac{\frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2 \cdot \frac{1}{n-1}\sum(y_i-\bar{y})^2}}$$

$$= b\sqrt{\frac{Var(X)}{Var(Y)}}$$

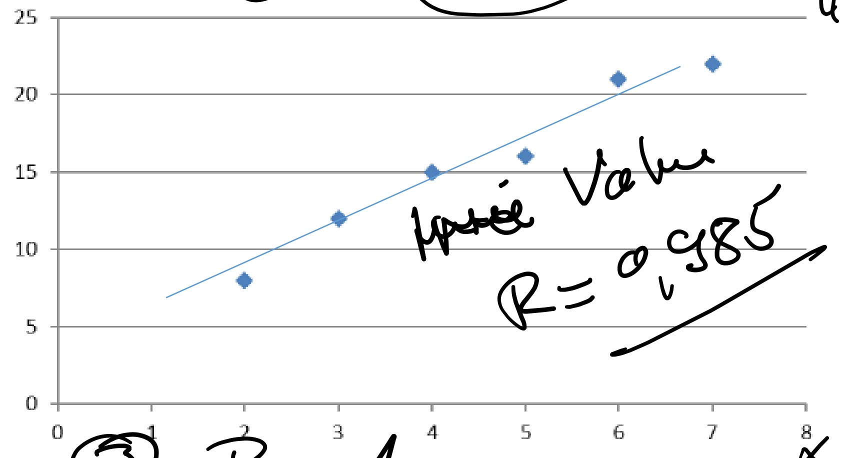
in general the covariance is normalized by the product of the standard deviations of X, Y

General intervals for **R** $\in [-1, 1]$

$\pm$ (0,0;0,2)	→	almost no dependence
$\pm$ [0,2;0,4)	→	low dependence
$\pm$ [0,4;0,6)	→	medium dependence
$\pm$ [0,6;0,8)	→	high dependence
$\pm$ [0,8;1,0)	→	almost full dependence

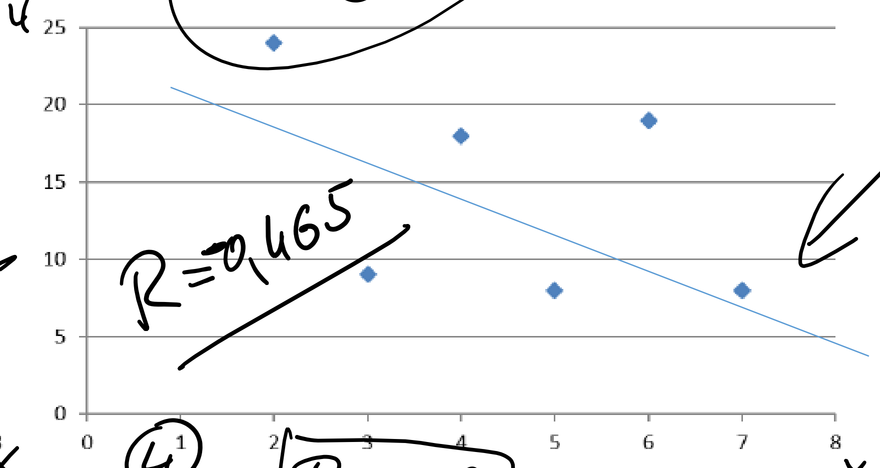
Correlation coefficiant examples

① $R \approx 0,8$



~~true~~ Value $R = 0,985$

$R \approx -0,2$

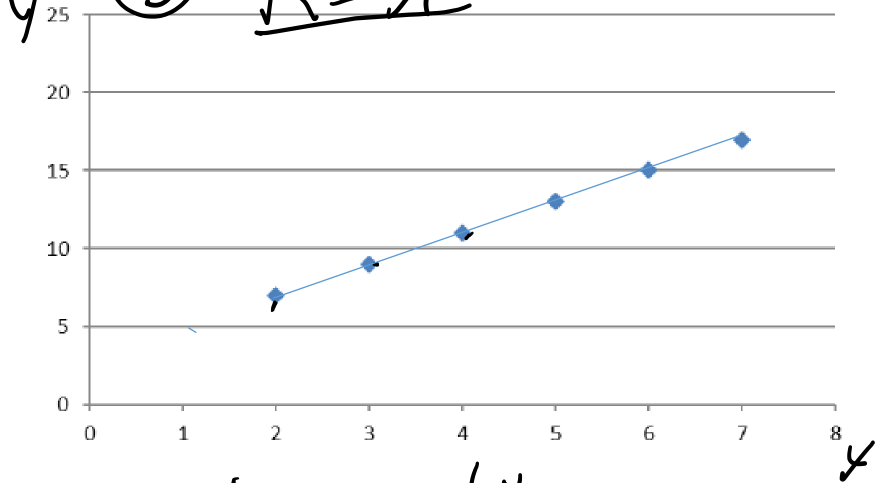


$R = 0,465$

Y is more or less decreasing if X is increasing

$\Rightarrow$ R should negative!

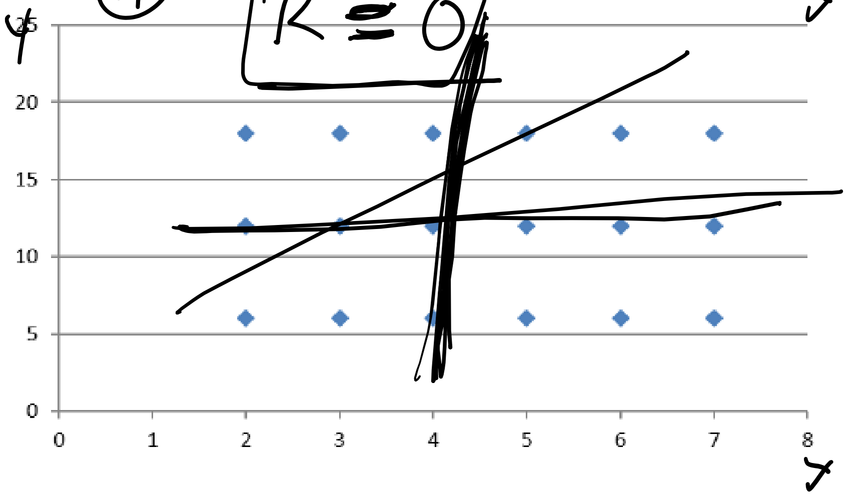
③ $R = 1$



$y = a + bx$

$\Rightarrow$ all data points are on the line

④ $\boxed{R = 0}$



Coefficient of determination R^2

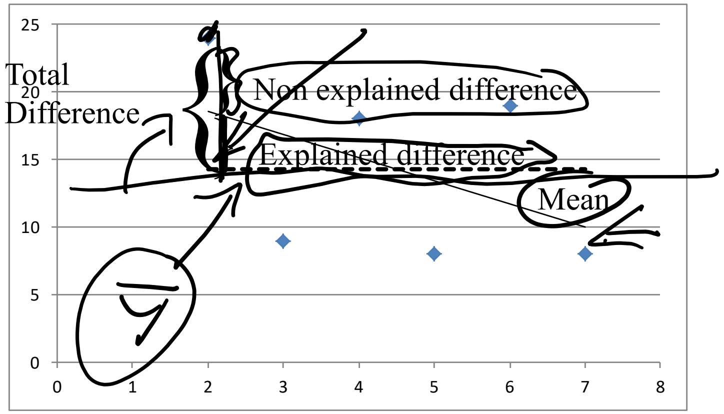
- The squared correlation coefficient R^2 is called coefficient of determination.

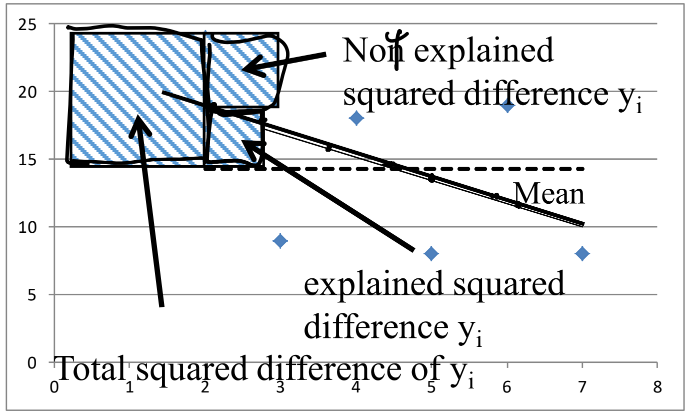
- In quantitative analysis, we often use the coefficient of determination R^2, because it can be intuitively interpreted

- But R^2 does not distinguish between + and – anymore.

- R^2 is in the interval [0,1] since $R \in [-1,1]$

- R^2 grafical interpretation:
  - R^2 equals the proportion of variance explained by the model in relation to the total variance, i.e. how much percent is explained by the regression line
  - (1-R^2) equals the proportion of Variance not explained by the model in relation to the total variance, and has to be explained by other influencing factors





$\hat{y}_i - \bar{y}$ explained difference

$y_i - \hat{y}_i$ not explained difference

$y_i - \bar{y}$ total difference

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

total squared difference

not explained squared difference

explained squared difference

$$\frac{\text{Sum of the squared distances}}{\text{Sum of the squared total differences}}$$

Coefficiant of determination $R^2 = \frac{\sum_{i=1}^{n}\boxed{(\hat{y}_i-\bar{y})^2}}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$

Correlation Coefficiant R^{2C}

$(21{,}72-27)^2$

$3{,}92+1{,}98\cdot 9$

$\hat{y}_i = 3{,}92 + 1{,}98\,x_i$

$\bar{Y} = 27$

N	x	y	$\hat{y}$	$(\hat{y}-\bar{y})^2$	$(y-\bar{y})^2$
1	9	24	21,72		
2	11	33			
3	5	10			
4	13	29			
5	20	42			
6	12	24			
Gesamt					

$R^2=$

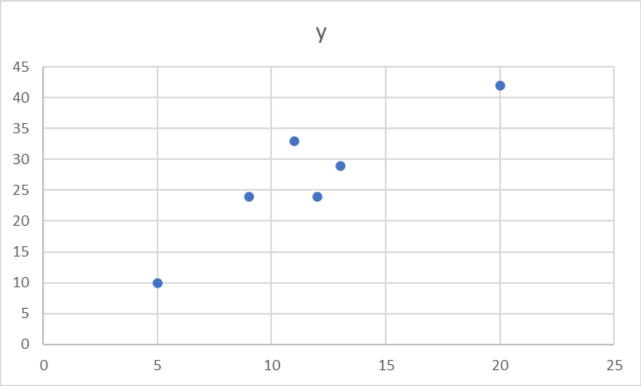
$R=$

Correlation Coefficient R

$\hat{y}_i = 3{,}92 + 1{,}98\,x_i$

$\bar{y} = 27$

	Visits	Sales volume [€]				explained	total
year	x	y	xy	x^2	yhat	(yhat-ybar)^2	(y-ybar)^2
2008	9	24	216	81	21,72	27,83	9,00
2009	11	33	363	121	25,68	1,74	36,00
2010	5	10	50	25	13,81	173,95	289,00
2011	13	29	377	169	29,64	6,96	4,00
2012	20	42	840	400	43,49	271,80	225,00
2013	12	24	288	144	27,66	0,43	9,00
sum	70	162	2134	940		482,72	572,00



$R^2 = \frac{487{,}2}{572}$

← positive dependence

⇒ taking the root for calculating R we need the "+" sign

$R^2 = 0{,}84$

$R = +0{,}92$

Interpretation of R^2

- R^2 measures the strengh of the dependence between two variables X and Y.

- R^2 can be interpreted as the proportion expained by the linear regression.

- $R^2 = 0$, the linear regression model is given just by the constant a und $b=0$

  $\rightarrow$ The change of the independent variable X has no influence on the dependent variable Y

- $R^2 = 1$, then the regression line fully explains the dependence between X and Y and we have $\hat{y}_i = a + bx_i = y_i$

- Make a forecast for inflation for Germany
- First take the oil price the last 5 years (monthly data)
- Exchange rate
- Calculate the dependence between the energy component of the HCPI and the oilprice
- And calculate the total inflation coming from the guessed increase of oil prices in the future due to a possible embargo of Russian oil

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Case study I: Inflation

Forecast of the inflation ind Germany for 2023 based on the historical development of oil prices

- Data (monthly data):
    - Oil price of the last 5 years
    - Exchange rate Euro/USD
    - HCPI (Harmonized Consumer Price Index) Germany
    - HCPI Energy Component Germany

- Calculate the historical dependence between the energy component of the HCPI and the oil price via a linear regression
- Make an assumption about the future exchange rate and Oil price of 2023
- Assume the non-energy component of HCPI stays constant on the last known level.

- Calculate total inflation for 2023 based on the former assumptions and dependences

Case study I: Inflation

https://fred.stlouisfed.org/

Link for Oil price of sort Brent

https://www.bundesbank.de/dynamic/action/de/statistiken/zeitreihen-datenbanken/zeitreihen-datenbank/759778/759778?listId=www_sdks_b01012_2

Link exchange rate USD-Euro

https://ec.europa.eu/eurostat/de/

Link HCPI

Case study I: Inflation

Result:

Dependence between Oil-price and Energy component HCPI

HCPI-Energy=76,415+0,4987*Oilprice[Euros]

-> with the assumptions, that Exchangerates USD:Euro will be at 1:1 in Summer and Oilprice will be at 130 USD per Barrel in Summer until December 2023, we obtain, that Oil prices will nove up Inflation by roughly 0,5 Percentagepoints.

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

1. Properties of the linear regression.

(a) Show $\bar{y} = a + b\bar{x}$

(b) Show $b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$ and $a = \frac{\overline{x^2} \cdot \bar{y} - \bar{x} \cdot \overline{xy}}{\overline{x^2} - \bar{x}^2}$

(c) Show $\sum_{i=1}^{n} \epsilon_i = 0$

(d) Show $\sum_{i=1}^{n} \epsilon_i x_i = 0$

(e) Show $\sum_{i=1}^{N} (y_i - \bar{y})^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$

this follows from FOC (1) (2)

total variance

explained variance

$\hat{Y}_i = a + bx_i \qquad \hat{Y}_i - Y_i = \varepsilon_i$

$Y_i = a + bx_i + \varepsilon_i$

$\Rightarrow \frac{1}{n}\sum x_i = \bar{x}$ and

$\frac{1}{n}\sum Y_i = \bar{Y}$ are on the regression line

$$(1) \quad \frac{\partial F}{\partial a} = -2\left[\sum_i^n (y_i - \underbrace{(a + bx_i)}_{\hat{Y}_i})\right] = 0$$

$$(2) \quad \frac{\partial F}{\partial b} = -2\left[\sum_i^n x_i \underbrace{(y_i - (a + bx_i))}_{\varepsilon_i}\right] = 0$$

a) $\Rightarrow \underbrace{\frac{1}{n}\sum Y_i}_{\bar{Y}} - \underbrace{\frac{1}{n}\sum a}_{=a} \underbrace{\frac{1}{n}\sum bx_i}_{b\bar{x}} = 0 \quad | \frac{1}{n}$

$\bar{Y} - a - b\bar{x} \Rightarrow \bar{Y} = a + b\bar{x} \Rightarrow$ thus $\bar{Y}$ and $\bar{x}$ are on the regression line

c) Take (1) $\Rightarrow \sum \underbrace{Y_i - \hat{Y}_i} = 0 \Rightarrow \boxed{\sum_{i=1}^{n} \varepsilon_i = 0}$ q.e.d.

d) Take (2) $\overset{\text{from (c)}}{\Rightarrow} \boxed{\sum x_i \varepsilon_i = 0}$ q.e.d

1. Properties of the linear regression.

(a) Show $\bar{y} = a + b\bar{x}$

(b) Show $b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$ and $a = \frac{\overline{x^2}\cdot\bar{y} - \bar{x}\cdot\overline{xy}}{\overline{x^2} - \bar{x}^2}$

(c) Show $\sum_{i=1}^{n} \epsilon_i = 0$

(d) Show $\sum_{i=1}^{n} \epsilon_i x_i = 0$

(e) Show $\boxed{\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}$

e) $\sum(y_i - \bar{y})^2 = \cancel{\sum y_i^2} - 2\bar{y}\sum y_i + \cancel{\sum \bar{y}^2}$

$+\Big(\sum(y_i - \hat{y}_i)^2 = \cancel{\sum y_i^2} - 2\sum y_i\hat{y}_i + \sum \hat{y}_i^2$ $\Rightarrow 2\sum\hat{y}_i^2$

$\sum(\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - 2\bar{y}\sum\hat{y}_i + \cancel{\sum\bar{y}^2}$

$\Rightarrow -\cancel{2}\bar{y}\sum y_i = \cancel{2}\sum\hat{y}_i^2 - \cancel{2}\sum y_i\hat{y}_i - \cancel{2}\bar{y}\sum\hat{y}_i \;\Big|\; + \bar{y}\sum\hat{y}_i$

$\bar{y}\sum\underbrace{(\hat{y}_i - y_i)}_{\varepsilon_i} = \sum(\hat{y}_i^2 - y_i\hat{y}_i)$ $\Rightarrow$ by definition $a + bx_i$

$= \sum\hat{y}_i\underbrace{(\hat{y}_i - y_i)}_{\varepsilon_i}$

$\Rightarrow \underbrace{\sum\varepsilon_i}_{0} = \sum(a + bx_i)\varepsilon_i = a\underbrace{\sum\varepsilon_i}_{0} + b\underbrace{\sum x_i\varepsilon_i}_{0}$ q.e.d

$\Rightarrow$ qualitatively we have total variance = unexplained variance + explained variance

2. A housing agent sells last April 26 flats with the following data regarding size G in $[m^2]$ and rental M in [€]. We know $Cov(G,M) = 5760$ $\sigma_M = 223,61$ $\sigma_G = 21,9$ $\mu_M = 1100$ $\mu_G = 80$

(a) Calculate and draw the regression line.

(b) Calculate the average and marginal price per m^2.

(c) Estimate the rental of flat of size 100 m^2.

$Y \hat{=} M \qquad x \hat{=} G$

$\Rightarrow Var(X) = \sigma_G^2$

a) $b = \frac{Cov(x,y)}{Var(x)} = \frac{5760}{21,9^2} \approx 12,01 \qquad a = \overline{y} - b\overline{x} = 1100 - 12,01 \cdot 80 \approx 139,22$

$\hat{y}_i = a + b x_i$

c) $\hat{M}(100 m^2) = 139,22 + 12,01 \cdot 100$

$\approx 1340,20$ €

3. Within a random sample of 100 data points (X,Y) we have $R = -0,93$. Decide which statement is right or wrong?

(1) There is no dependence between X and Y since R is negative. ← wrong means no dependence $\Rightarrow R = 0$

(2) The data points scattering very near around the regression line. ← true since R is very near at -1 and $R \in [-1,1]$

(3) The linear regression has positive slope. ← wrong because R is negative

(4) The dependence of X and Y is inversely proportional. ← true same questions in (3) only what we say if y is going down x is going up

(5) Which proportion of the total variance is explained by the linear regression with a R=-0,93?

$\Rightarrow$ coefficient of determination

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = (-0,93)^2 \approx 86,5\%$$

← explained variation

← total variation

4. In 2021 a brewery has the following output and costs:

Month	Output [Hektoliter] x	Costs [€] y
Jan	600	6500
Feb	680	8200
Mrz	720	7300
Apr	1010	8900
Mai	900	9900
Jun	990	10000
Jul	1270	10300
Aug	1440	12500
Sep	1380	11500
Okt	1010	9200
Nov	830	8200
Dez	1070	9300

(a) Calculate the Cost function via a linear regression.

(b) Interprete economically the parameters.

(c) Calculate the correlation between Output and costs?

(d) Estimate the total costs and average variable costs of an output of 1100 Hektoliter.

	x	y	x^2	xy
mean	991,666667	9316,66667	1049650	9627583,33
	$\overline{x}$	$\overline{y}$	$\overline{x^2}$	$\overline{xy}$

we are interested into the cost function of our brewery $\Rightarrow$ $c(x)$ $\to$ function form!

a) $$b = \boxed{\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}} = \frac{9627583{,}\bar{3} - 991{,}\bar{6} \cdot 9316{,}\bar{6}}{1049650 - (991{,}\bar{6})^2} = 5{,}865$$

$$a = \bar{y} - b\bar{x} = 9316{,}\bar{6} - 5{,}865 \cdot 991{,}\bar{6} = 3500{,}308$$

$\Rightarrow$ costfunction $c(x) = \underbrace{3500{,}308}_{\text{fixed costs}} + \overbrace{5{,}865 \cdot x}^{\text{variable costs}}$

$c(x) = \alpha + \beta x^2$

b) $a :=$ fix cost $\qquad b :=$ variable costs per unit because we have a linear dependence

variable costs $:= b \cdot x$

$:=$ marginal costs $= c'(x) = b$

c) $R =$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} = 0{,}87 \Rightarrow R = 0{,}93$$

$c(1100) = 3500{,}308 + 5{,}865 \cdot 1100 = 9952{,}067$

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>

[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Markowitz and CAPM

Bernhard Köster

Summer Term 2022

Example: Minimum-Variance-Portfolio

A Portfolio constists of two assets a and b with investment weight $\alpha \in [0, 1]$. Prices of the assets are random variables with means μ_a, μ_b and variances σ_a^2, σ_b^2. Calculate the portfolio with the minimal variance, if assets are correlated with $\rho \in (-1, 1)$.

$\rho = R$

Optimizing problems with many variables

Suppose we have $\vec{x} = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$. For continously twice differentiable functions, we have:

for example $u(\mu_1, \mu_2, \sigma_1, \sigma_1, \rho)$

$\frac{\partial f}{\partial x_1} = 0 \quad \frac{\partial f}{\partial x_2} = 0 \quad \dots \rightarrow$ Investmentparameter α

$\rightsquigarrow$ 20% in Asset 1
80% " " 2

$c'(x) = 0$ FOC

Maximum $c''(x^*) < 0$

$$f(\vec{x}) \longrightarrow \max \implies \underbrace{\frac{\partial f}{\partial \vec{x}}(\vec{x}^*) = 0}_{\text{FOC}} \text{ and } Hess\, f(\vec{x}^*) \text{ is negative definite}$$

or

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{pmatrix}$$

$$f(\vec{x}) \longrightarrow \min \implies \frac{\partial f}{\partial \vec{x}}(\vec{x}^*) = 0 \text{ and } Hess\, f(\vec{x}^*) \text{ ist positive definite}$$

$\Rightarrow$ there exists a global min/max

Very often in economics, we are confronted not with global optimization but with optimizing problems subject to some restrictions. Then we have

microeconomics

$$f(\vec{x}) \quad \longrightarrow \quad \max\ (\min) \quad \text{s.t. } g_1(\vec{x}) = 0, g_2(\vec{x}) = 0, \ldots, g_m(\vec{x}) = 0\ (n > m)$$

$\boxed{m - p_1x_1 - p_2x_2 = 0}$

Budget restriction

Lagrange function   i.e. utility

with FOC

$$\mathcal{L} = f(\vec{x}) + \lambda_1 g_1 + \lambda_2 g_2 + \ldots + \lambda_m g_m$$

$\mathcal{L}(x_1 \ldots x_n, \lambda_1 \ldots \lambda_m)$

FOC

$$\boxed{\begin{array}{ccc} \frac{\partial \mathcal{L}}{\partial x_1} = 0 & \ldots & \frac{\partial \mathcal{L}}{\partial x_n} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda_1} = 0 & \ldots & \frac{\partial \mathcal{L}}{\partial \lambda_n} = 0 \end{array}}$$

For SOC including the rendered hessian matrix and regularity conditions for the existence of optima (i.e. implicit functions theorem is valid!) are beyond this lecture and we refer to the literature.

Markowitz Portfolio Modell

Assumptions:

1. An Investor has different risky investment opportunities with finite means and variances.
2. Investors minimize one-period-utility with dereasing marginal utility in wealth and a risk avers structure.
3. Risk ist measured via the volatility of the expected yield.
4. Decision varibales are only expected yield and variances.
5. Given the Risk investors prefer higher yield to lower yield. Given the yield. Investors prefer lower risk to higher risk.

$\Rightarrow$ in general risk avers investors

Notation

random variables

- R_1, R_2: Yield asset 1 and 2
- $\mu_1 = E(R_1), \mu_2 = E(R_2)$: expected yield asset 1 and 2
- $\sigma_1 = \sigma(R_1), \sigma_2 = \sigma(R_2)$: standard deviation asset 1 and 2
- $\rho = \rho(R_1, R_2) = \boxed{Cov(R_1, R_2)/(\sigma_1\sigma_2)}$: correlation coefficiant of the yields of asset 1 and 2
- R: yield of a Portfolio constisting of asset 1 and 2
- $\mu = \mu(R)$: expected yield of a portfolio
- $\sigma = \sigma(R)$: standard deviation of a portfolios

Expected yield and Variance of a Portfolios

$$\rho = \frac{cov(R_1, R_2)}{\sigma_1 \sigma_2}$$

$E(X + \overset{\text{number}}{a} Y) = E(X) + aE(Y)$

X, Y are random variables

$$R = \alpha R_1 + (1-\alpha) R_2 \quad (1)$$

$$\mu = E(R) = \mu = \alpha\mu_1 + (1-\alpha)\mu_2 \quad (2)$$

$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \quad (3)$$

- W.l.o.g. $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$ $\alpha \in [0,1]$: proportion invested in asset 1 $\Rightarrow$ $(1-\alpha)$ investment in asset 2

$Var(\alpha R_1 + (1-\alpha)R_2) = E([\alpha R_1 + (1-\alpha)R_2 - E(\alpha R_1 + (1-\alpha)R_2)]^2)$

$= E([\alpha R_1 + (1-\alpha)R_2 - \alpha\mu_1 - (1-\alpha)\mu_2]^2)$

$= E([\alpha(R_1 - \mu_1) + (1-\alpha)(R_2 - \mu_2)]^2)$

$= E(\alpha^2(R_1-\mu_1)^2 + (1-\alpha)^2(R_2-\mu_2)^2 + 2\alpha(1-\alpha)(R_1-\mu_1)(R_2-\mu_2))$

$= \alpha^2 \underbrace{E((R_1-\mu_1)^2)}_{\sigma_1^2} + (1-\alpha)^2 \underbrace{E((R_2-\mu_2)^2)}_{\sigma_2^2} + 2\alpha(1-\alpha) \underbrace{E((R_1-\mu_1)(R_2-\mu_2))}_{cov(R_1, R_2)}$

$$E(R) = \mu = \alpha\mu_1 + (1-\alpha)\mu_2 \qquad (2)$$

$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \qquad (3)$$

What kind of portfolios are accessible for an investor?

$\rho = 1$ perfect correlation

$$\Rightarrow \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\sigma_1\sigma_2 \overset{1}{=} x^2 + y^2 + 2xy = (x+y)^2$$

see $\alpha^2\sigma_1^2 = (\alpha\sigma_1)^2 \mid \alpha\sigma_1 = x$

$(1-\alpha)\sigma_2 = y$

$$= (\alpha\sigma_1 + (1-\alpha)\sigma_2)^2$$

$$\Rightarrow \sigma = \alpha\sigma_1 + \sigma_2 - \alpha\sigma_2 = \sigma_2 + \alpha(\sigma_1 - \sigma_2) \Rightarrow \alpha = \frac{\sigma - \sigma_2}{\sigma_1 - \sigma_2} = \alpha(\sigma)$$

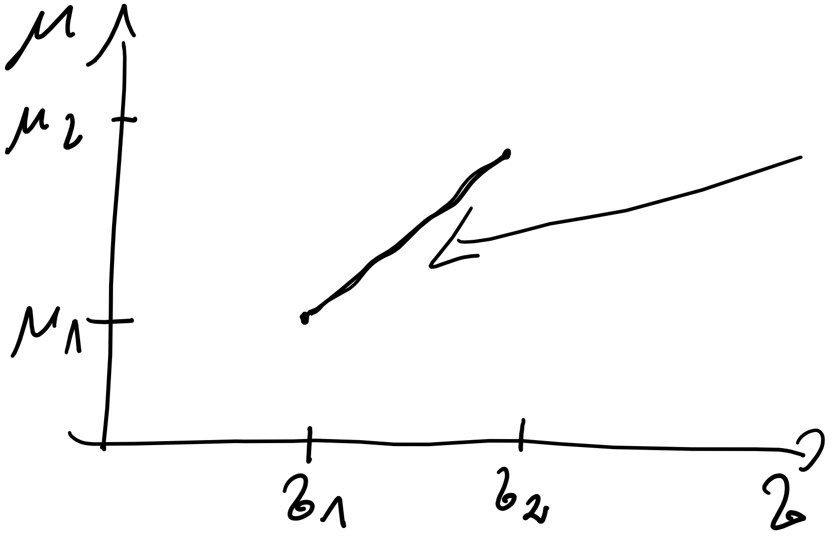
variable

are all parameters

$$\Rightarrow \mu = \alpha(\mu_1 - \mu_2) + \mu_2 = \frac{\sigma - \sigma_2}{\sigma_1 - \sigma_2}(\mu_1 - \mu_2) + \mu_2$$

this is just a linear function

$\mu(\sigma) \rightarrow$ which kind of dependence?



all these combinations of (μ, σ) are accessible

$$E(R) = \mu = \alpha\mu_1 + (1-\alpha)\mu_2 = \alpha(\mu_1-\mu_2)-\mu_2 \neq 0 \qquad (2)$$

$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \qquad (3)$$

since we have a quadratic dependence, we can ask for the portfolio and the parameter α which creates the minimum risk i.e. minimum variance

no correlation $\rho = 0$

$$\Rightarrow \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 = \sigma^2(\alpha) \Rightarrow$$

$$\sigma^{2\prime}(\alpha) = 2\alpha\sigma_1^2 - 2(1-\alpha)\sigma_2^2 = 0$$

$$\Rightarrow \alpha\sigma_1^2 - \sigma_2^2 + \alpha\sigma_2^2 = 0 \Rightarrow \boxed{\alpha = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}} \leftarrow$$

Investment parameter for the Minimum-Variance-Portfolio

$$1 = \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

$$f(x) = (1-x)^2$$

$$f'(\alpha) = 2(1-\alpha)(-1)$$

σ_{MVP} = Minimum Variance

Inserting α into $\sigma^2(\alpha)$

$$\Rightarrow \sigma^2 = \left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2\sigma_1^2 + \left(\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2} - \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2\sigma_2^2$$

$$= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \quad \begin{matrix} < \sigma_1^2 \\ < \sigma_2^2 \end{matrix}$$

$\Rightarrow$ investing in a portfolio with a proportion of α in asset 1 you can reduce your risk!

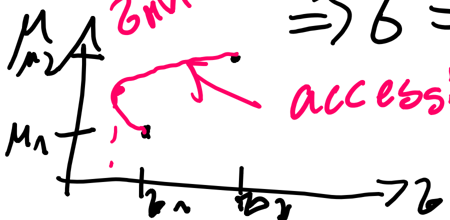
What are the accessible portfolios?

from (2) we obtain $\alpha = \frac{\mu-\mu_2}{\mu_1-\mu_2}$ $\qquad 1 = \frac{\mu_1-\mu_2}{\mu_1-\mu_2}$

insert in (3)

$$\Rightarrow \sigma^2 = \left(\frac{\mu-\mu_2}{\mu_1-\mu_2}\right)^2\sigma_1^2 + \underbrace{\left(1 - \frac{\mu-\mu_2}{\mu_1-\mu_2}\right)^2}_{\left(\frac{\mu-\mu_1}{\mu_1-\mu_2}\right)^2}\sigma_2^2$$

$$\Rightarrow \sigma^2(\mu) \rightsquigarrow \mu(\sigma)$$



accessible (μ,σ) combinations with $\rho=0$

$$E(R) = \mu = \alpha\mu_1 + (1-\alpha)\mu_2 \quad (2)$$
$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \quad (3)$$

$x = \alpha\sigma_1 \quad y = (1-\alpha)\sigma_2$

$\rho = -1 \Rightarrow \sigma^2 = (\alpha\sigma_1)^2 + ((1-\alpha)\sigma_2)^2 - 2\alpha\sigma_1(1-\alpha)\sigma_1 \stackrel{1}{=} x^2 + y^2 - 2xy = (x-y)^2$

$\Rightarrow \sigma^2 = (\alpha\sigma_1 - (1-\alpha)\sigma_2)^2$

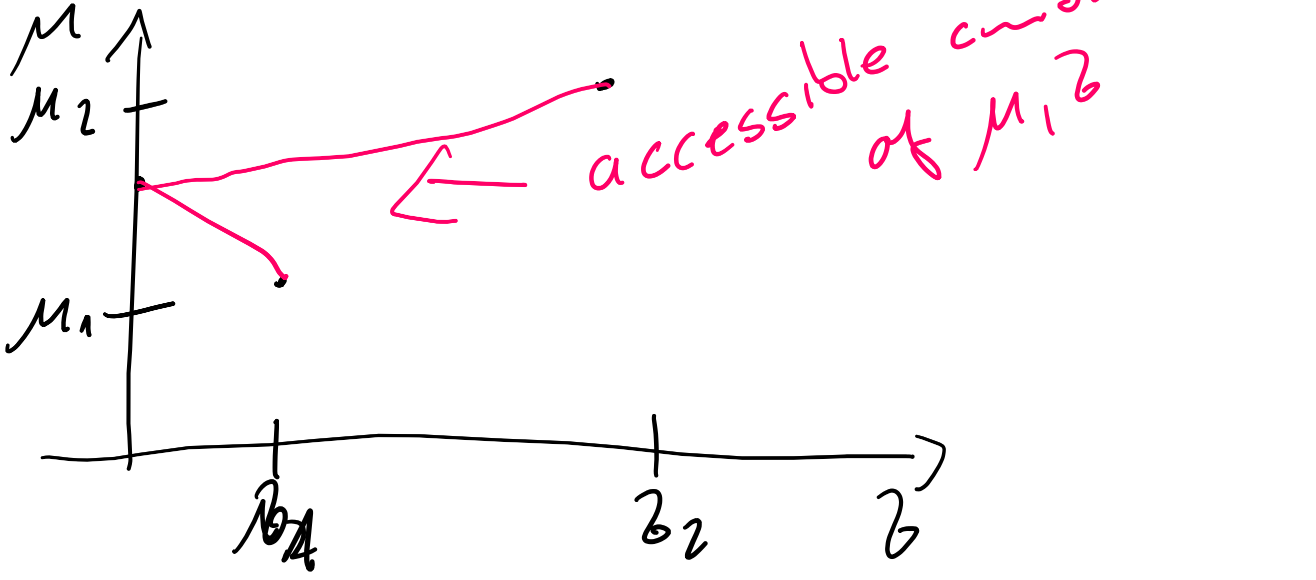
$\Rightarrow \sigma = |\alpha\sigma_1 + \alpha\sigma_2 - \sigma_2| \Rightarrow$ can be zero $\Rightarrow \alpha(\sigma_1+\sigma_2) - \sigma_2 = 0$

$\Rightarrow \boxed{\alpha^* = \dfrac{\sigma_2}{\sigma_1+\sigma_2}}$ where the portfolio has no risk

$\sigma^2(\alpha^*) = 0$

$\Rightarrow$ insert this dependence in μ

$\Rightarrow$ again you obtain a linear dependence between μ and σ



Accessible portfolios

We obtain all accessible portfolios (μ, σ) inserting (2) in (3) solving for σ.

$$\sigma^2 = \frac{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}{(\mu_1-\mu_2)^2}\left[\mu - \frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2-(\mu_1+\mu_2)\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}\right]^2 \tag{4}$$

$= \sigma^2(\mu)$

$$+\frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2} \tag{5}$$

$$\sigma^2 = VarK^2(R_1, R_2)[\mu - \mu_{MVP}]^2 + \sigma_{MVP}^2 \tag{6}$$

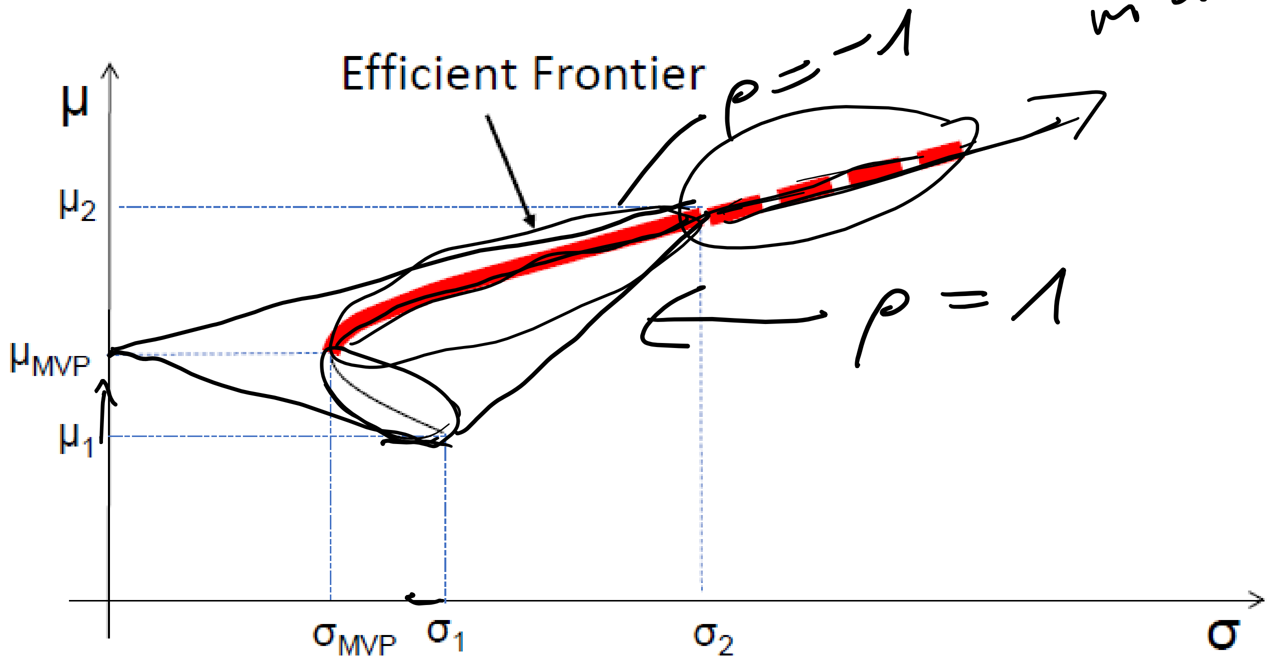
Accessible portfolios

- $VarK^2(R_1, R_2) = \frac{Var(R_1.R_2)}{[E(R_1.R_2)]^2} = \frac{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}{(\mu_1-\mu_2)^2}$ (Coefficient of variation)
- $\mu_{MVP} = \frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2-(\mu_1+\mu_2)\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$ and $\sigma^2_{MVP} = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$

$(\mu_{MVP}, \sigma_{MVP})$: Expected value and standard deviation of the global minimum-variance-portfolios. (6) represents a parabola with the vertex $(\mu_{MVP}, \sigma_{MVP})$. For the extrema $\rho = \pm 1$ we obtain separate solutions.

Efficiant frontier

Since we assume risk avers investors (higher risk implicates higher yields) only the upper part of the parabola is economically relevant. This is called the *efficiant frontier.*

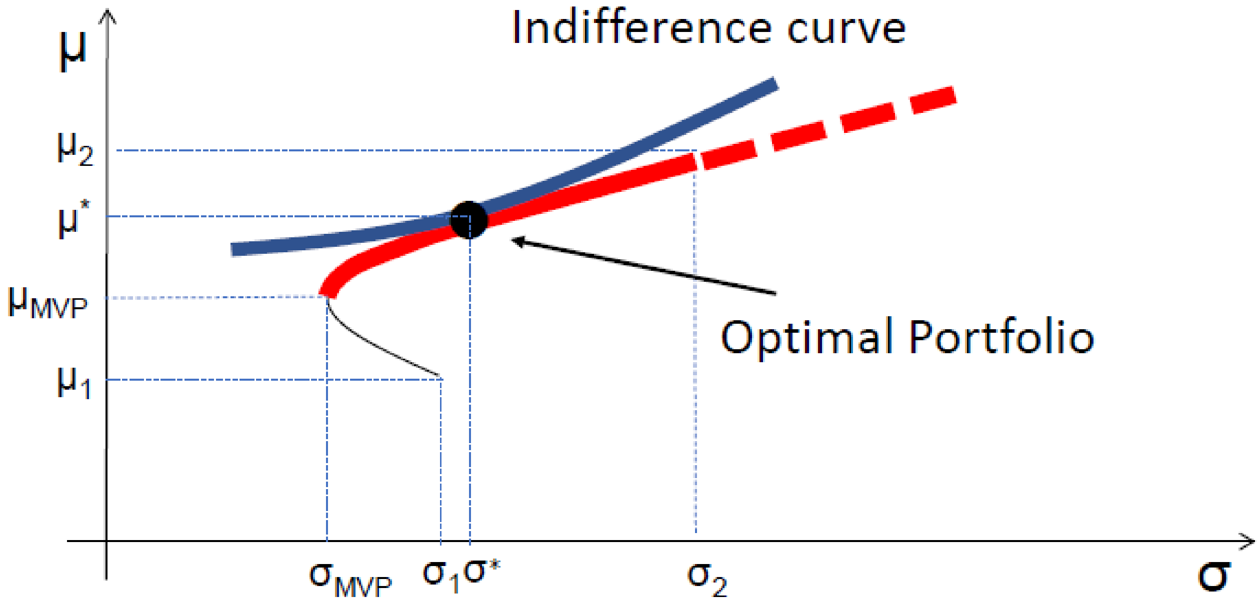


in a region with $\alpha < 0$
$\alpha > 1$
is also possible since in financial markets you can short sell an asset

Optimal Portfolio

In general a risk avers utility function is given by $U(\mu, \sigma)$ with $\frac{\partial U}{\partial \mu} > 0$ and $\frac{\partial U}{\partial \sigma} < 0$. The optimizing problem is then given by:

$$\max_{\mu,\sigma} U(\mu, \sigma) \quad \text{N.B. } \sigma^2 = A[\mu - \mu_{MVP}]^2 + \sigma^2_{MVP} \quad (A = VarK^2(R_1, R_2)) \quad (7)$$



Example

The explicit utility is given by $U(\mu, \sigma) = \mu - \frac{a}{2}\sigma^2$ with $(a > 0)$. $\Longrightarrow$

$$\mathcal{L}(\mu, \sigma, \lambda) = \mu - \frac{a}{2}\sigma^2 + \frac{\lambda}{2}(\sigma^2 - (A[\mu - \mu_{MVP}]^2 + \sigma^2_{MVP}))$$

$\Longrightarrow$ FOC

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= 1 - A\lambda(\mu - \mu_{MVP}) = 0 \\
\frac{\partial \mathcal{L}}{\partial \sigma} &= -a\sigma + \lambda\sigma = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \sigma^2 - (A[\mu - \mu_{MVP}]^2 + \sigma^2_{MVP}) = 0
\end{aligned}$$

$$\Longrightarrow \quad \mu^* = \mu_{MVP} + \frac{1}{aA} \quad \sigma^* = \sqrt{\frac{1}{Aa^2} + \sigma^2_{MVP}} \quad \alpha^* = \frac{\mu^* - \mu_2}{\mu_1 - \mu_2}$$

Statistics A

Wilhelmshaven

This lecture will be recorded and Subsequently uploaded in the world-wide-web

Function translator (webpage)

Function translator Excel 1 (add in)

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

Markowitz Portfolio Modell

Assumptions:

1. An Investor has different risky investment opportunities with finite means and variances.
2. Investors minimize one-period-utility with dereasing marginal utility in wealth and a risk avers structure.
3. Risk ist measured via the volatility of the expected yield.
4. Decision varibales are only expected yield and variances.
5. Given the Risk investors prefer higher yield to lower yield. Given the yield. Investors prefer lower risk to higher risk.

$\Rightarrow$ in general risk avers investors

Notation

random variables

- R_1, R_2: Yield asset 1 and 2
- $\mu_1 = E(R_1), \mu_2 = E(R_2)$: expected yield asset 1 and 2
- $\sigma_1 = \sigma(R_1), \sigma_2 = \sigma(R_2)$: standard deviation asset 1 and 2
- $\rho = \rho(R_1, R_2) = \boxed{Cov(R_1, R_2)/(\sigma_1\sigma_2)}$: correlation coefficiant of the yields of asset 1 and 2
- $\boxed{R}$: yield of a Portfolio constisting of asset 1 and 2
- $\mu = \mu(R)$: expected yield of a portfolio
- $\sigma = \sigma(R)$: standard deviation of a portfolios

Expected yield and Variance of a Portfolios

$Var(\alpha X + Y)$

$\rho = \frac{cov(R_1, R_2)}{\sigma_1 \sigma_2}$

$\boxed{E(X + \alpha Y) = E(X) + \alpha E(Y)}$

X, Y are random variables

$$R = \alpha R_1 + (1 - \alpha) R_2 \qquad (1)$$

$$\mu = E(R) = \mu = \alpha\mu_1 + (1 - \alpha)\mu_2 \qquad (2)$$

$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1 - \alpha)^2\sigma_2^2 + 2\alpha(1 - \alpha)\rho\sigma_1\sigma_2 \qquad (3)$$

- W.l.o.g. $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$ $\alpha \in [0, 1]$: proportion invested in asset 1

$\Rightarrow (1-\alpha)$ investment in asset 2

$Var(\alpha R_1 + (1-\alpha)R_2) = E([\alpha R_1 + (1-\alpha)R_2 - E(\alpha R_1 + (1-\alpha)R_2)]^2)$

$= E([\alpha R_1 + (1-\alpha)R_2 - \alpha\mu_1 - (1-\alpha)\mu_2]^2)$

$= E([\alpha(R_1 - \mu_1) + (1-\alpha)(R_2 - \mu_2)]^2)$

$= E(\alpha^2(R_1 - \mu_1)^2 + (1-\alpha)^2(R_2 - \mu_2)^2 + 2\alpha(1-\alpha)(R_1 - \mu_1)(R_2 - \mu_2))$

$= \alpha^2 \underbrace{E((R_1 - \mu_1)^2)}_{\sigma_1^2} + (1-\alpha)^2 \underbrace{E((R_2 - \mu_2)^2)}_{\sigma_2^2} + 2\alpha(1-\alpha) \underbrace{E((R_1 - \mu_1)(R_2 - \mu_2))}_{cov(R_1, R_2)}$

$$E(R) = \mu = \alpha\mu_1 + (1-\alpha)\mu_2 \quad (2)$$

$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \quad (3)$$

what kind of portfolios are accesible for an investor?

$\rho = 1$ perfect correlation

see $\alpha^2 \sigma_1^2 = (\alpha\sigma_1)^2 \mid \alpha\sigma_1 = x$

$(1-\alpha)\sigma_2 = y$

$$\Rightarrow \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\sigma_1\sigma_2 \overset{1}{=} x^2 + y^2 + 2xy$$

$$= (x+y)^2$$

$$= (\alpha\sigma_1 + (1-\alpha)\sigma_2)^2$$

$$\Rightarrow \sigma = \alpha\sigma_1 + \sigma_2 - \alpha\sigma_2 = \sigma_2 + \alpha(\sigma_1 - \sigma_2) \Rightarrow \alpha = \frac{\sigma - \sigma_2}{\sigma_1 - \sigma_2} = \alpha(\sigma)$$

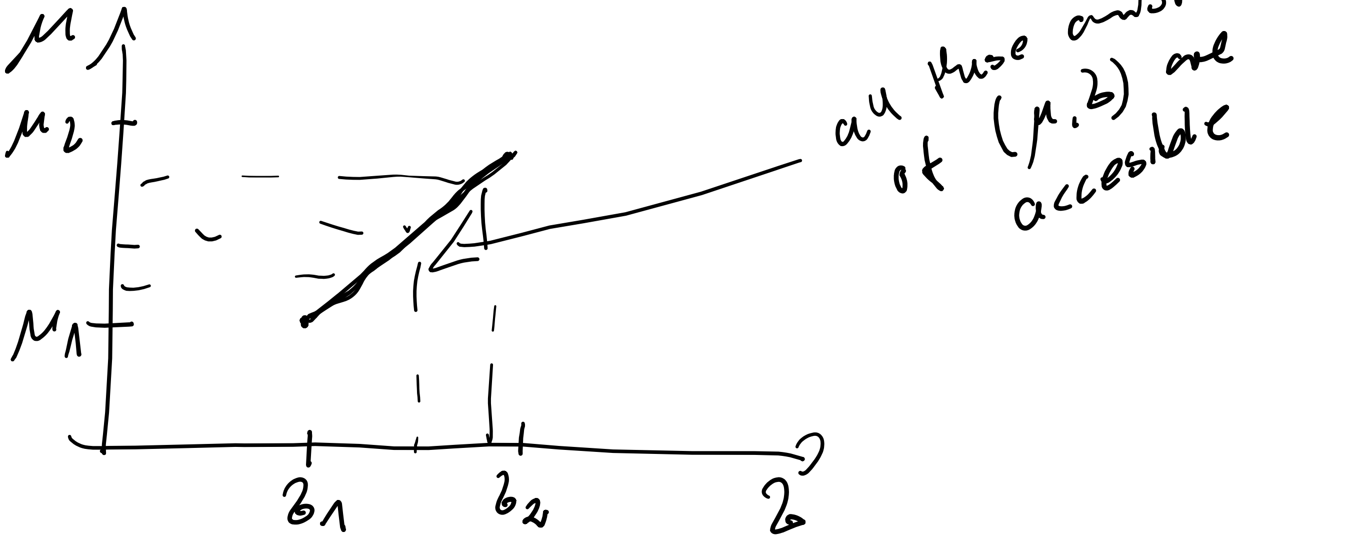
variable

are all parameters

$$\Rightarrow \mu = \alpha(\mu_1 - \mu_2) + \mu_2 = \frac{\sigma - \sigma_2}{\sigma_1 - \sigma_2}(\mu_1 - \mu_2) + \mu_2$$

this is just a linear function

$\mu(\sigma) \rightarrow$ which kind of dependence?



$$E(R) = \mu = \alpha\mu_1 + (1-\alpha)\mu_2 = \alpha(\mu_1-\mu_2) - \mu_2 \neq 0 \qquad (2)$$

$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \qquad (3)$$

since we have a quadratic dependence, we can ask for the portfolio or the parameter α which creates the minimum risk i.e. minimum variance

no correlation $\rho = 0$

$$\Rightarrow \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 = \sigma^2(\alpha) \Rightarrow$$

$$\sigma^{2\prime}(\alpha) = 2\alpha\sigma_1^2 - 2(1-\alpha)\sigma_2^2 = 0$$

$$1 = \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2}$$

$$f(x) = (1-x)^2$$

$$f'(\alpha) = 2(1-\alpha)(-1)$$

$$\Rightarrow \alpha\sigma_1^2 - \sigma_2^2 + \alpha\sigma_2^2 = 0 \Rightarrow \boxed{\alpha = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}}$$

Investment parameter for the Minimum-Variance-Portfolio

Inserting α into $\sigma^2(\alpha)$

$$\Rightarrow \sigma^2 = \left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2\sigma_1^2 + \left(\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2} - \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2\sigma_2^2$$

$$= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} \quad < \sigma_1^2, \quad < \sigma_2^2$$

$\Rightarrow$ investing in a portfolio with a proportion of α in asset 1 you can reduce your risk!

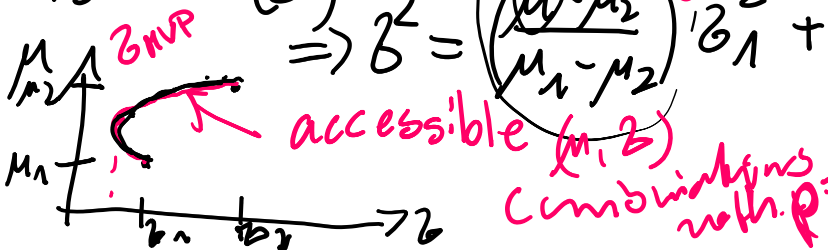
σ_{MVP} = Minimum Variance

What are the accessible portfolios?

from (2) we obtain $\alpha = \frac{\mu-\mu_2}{\mu_1-\mu_2}$ $\quad \frac{\mu_1-\mu_2}{\mu_1-\mu_2}$

insert in (3)

$$\Rightarrow \sigma^2 = \left(\frac{\mu-\mu_2}{\mu_1-\mu_2}\right)^2\sigma_1^2 + \underbrace{\left(1 - \frac{\mu-\mu_2}{\mu_1-\mu_2}\right)^2}_{\left(\frac{\mu-\mu_1}{\mu_1-\mu_2}\right)^2}\sigma_2^2 \Rightarrow \sigma^2(\mu) \rightsquigarrow \mu(\sigma)$$



$$E(R) = \mu = \alpha\mu_1 + (1-\alpha)\mu_2 \qquad (2)$$

$$Var(R) = \sigma^2 = \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\alpha(1-\alpha)\rho\sigma_1\sigma_2 \qquad (3)$$

$x = \alpha\sigma_1 \quad y = (1-\alpha)\sigma_2$

$\boxed{\rho = -1} \Rightarrow \sigma^2 = (\alpha\sigma_1)^2 + ((1-\alpha)\sigma_2)^2 - 2\alpha\sigma_1(1-\alpha)\sigma_1 \overset{!}{=} x^2 + y^2 - 2xy = (x-y)^2$

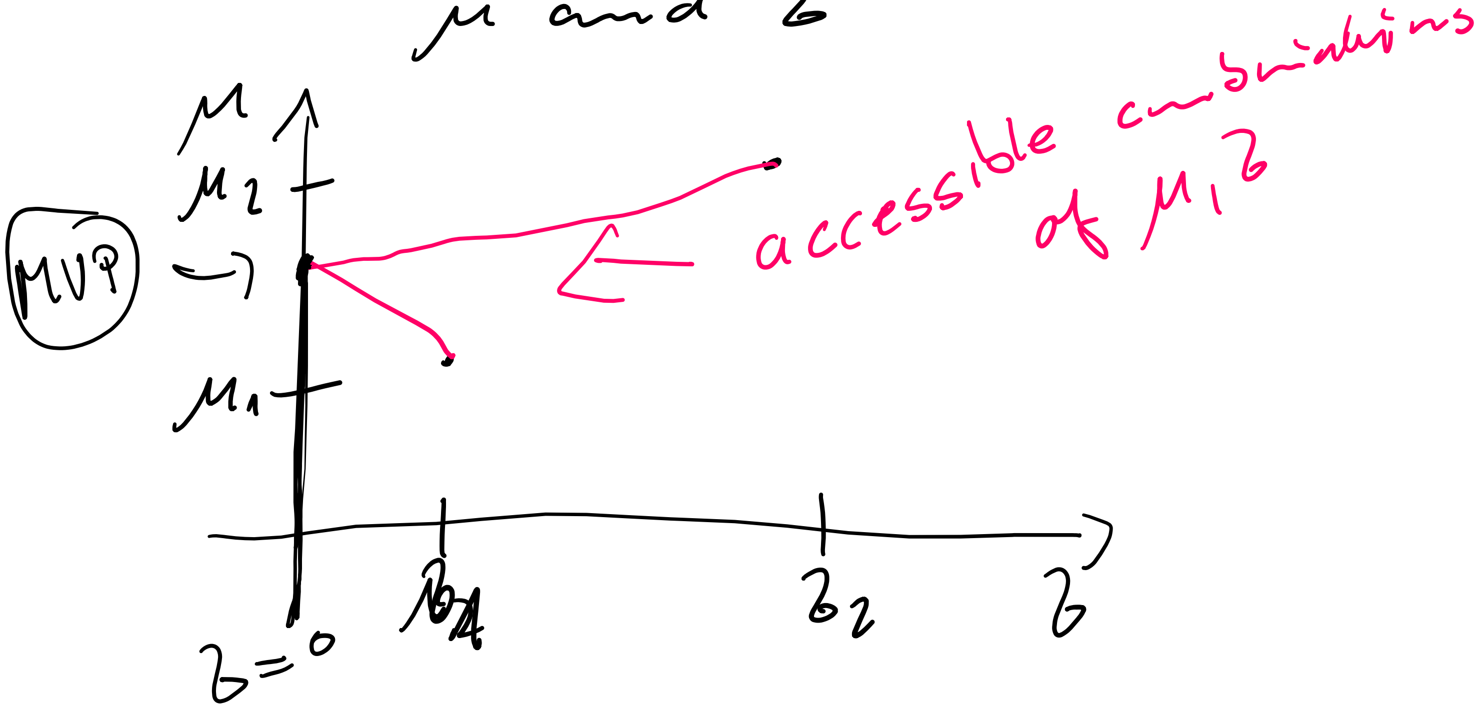
$\Rightarrow \sigma^2 = (\alpha\sigma_1 - (1-\alpha)\sigma_2)^2$

$\Rightarrow \boxed{\sigma = |\alpha\sigma_1 + \alpha\sigma_2 - \sigma_2|} \Rightarrow$ can be zero $\Rightarrow \alpha(\sigma_1+\sigma_2) - \sigma_2 = 0$

$\Rightarrow \boxed{\alpha^* = \dfrac{\sigma_2}{\sigma_1+\sigma_2}}$ where the portfolio has no risk $\rightarrow \sigma^2(\alpha^*) = 0$

$\Rightarrow$ insert this dependence in μ

$\Rightarrow$ again you obtain a linear dependence between μ and σ



Accessible portfolios

$\rho \in (-1, 1)$

We obtain all accessible portfolios (μ, σ) inserting (2) in (3) solving for σ.

$$\sigma^2 = \frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{(\mu_1 - \mu_2)^2}\left[\mu - \frac{\mu_2\sigma_1^2 + \mu_1\sigma_2^2 - (\mu_1 + \mu_2)\rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}\right]^2 \tag{4}$$

$$+ \frac{(1 - \rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{5}$$

$$\sigma^2 = VarK^2(R_1, R_2)\underbrace{[\mu - \mu_{MVP}]^2}_{} + \sigma^2_{MVP} \tag{6}$$

$= \sigma^2(\mu)$

Function for the efficient frontier!

Difference between the Expected value of the Portfolio $\mu(\alpha)$ relative to the MVP

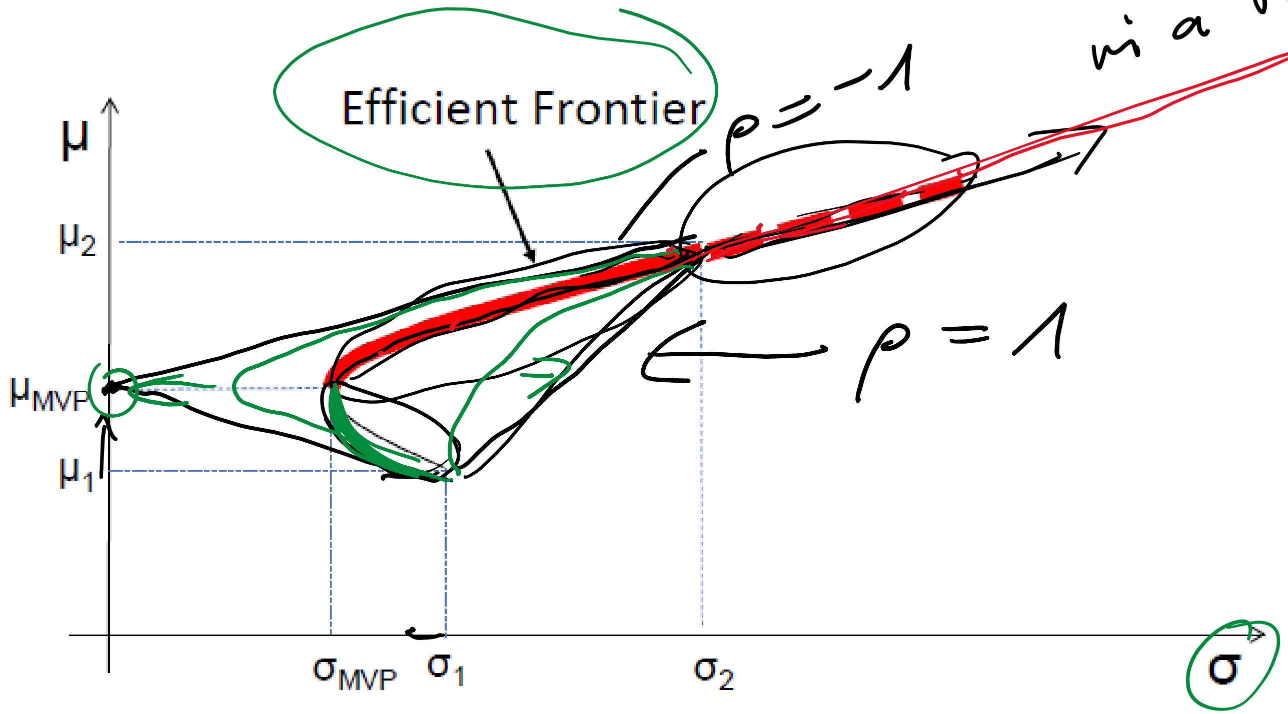
Accessible portfolios

- $VarK^2(R_1, R_2) = \frac{Var(R_1.R_2)}{[E(R_1.R_2)]^2} = \frac{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}{(\mu_1-\mu_2)^2}$ (Coefficient of variation)
- $\mu_{MVP} = \frac{\mu_2\sigma_1^2+\mu_1\sigma_2^2-(\mu_1+\mu_2)\rho\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$ and $\sigma_{MVP}^2 = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2-2\rho\sigma_1\sigma_2}$

$(\mu_{MVP}, \sigma_{MVP})$: Expected value and standard deviation of the global minimum-variance-portfolios. (6) represents a parabola with the vertex $(\mu_{MVP}, \sigma_{MVP})$. For the extrema $\rho = \pm 1$ we obtain separate solutions.

Efficiant frontier

Since we assume risk avers investors (higher risk implicates higher yields) only the upper part of the parabola is economically relevant. This is called the *efficiant frontier*.



$\alpha \in [0,1]$

α could negative or positive

in a region with $\alpha < 0$, $\alpha > 1$ is also possible since in financial markets you can short sell an asset

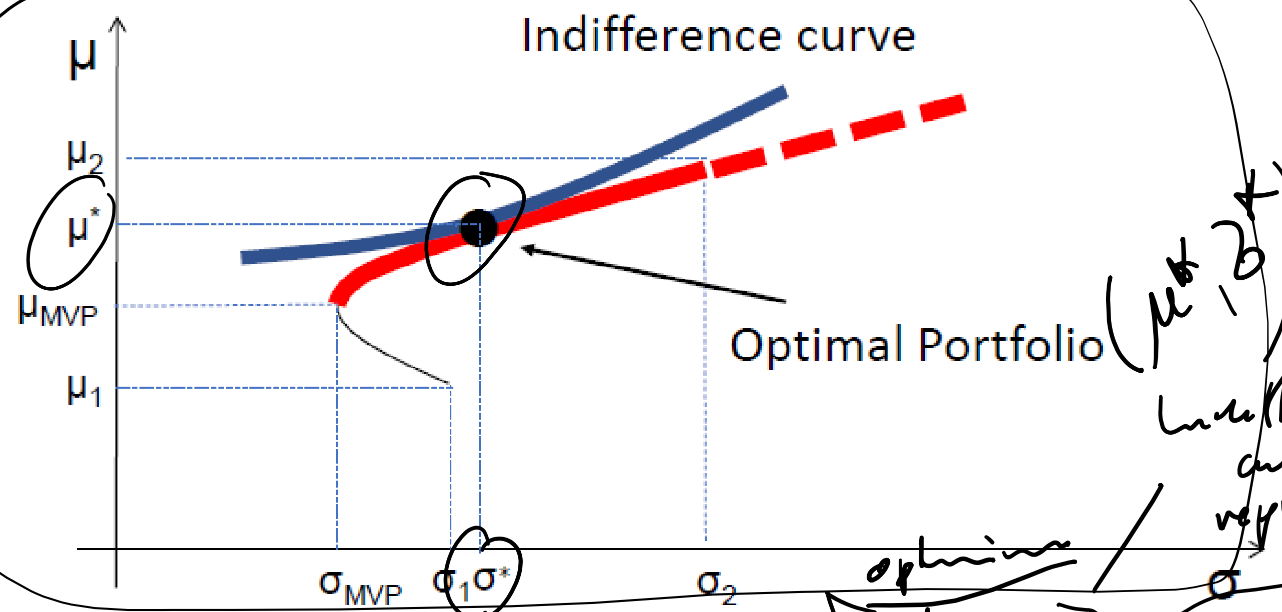
Optimal Portfolio

"good" "bad"
+ −

In general a risk avers utility function is given by $U(\mu,\sigma)$ with $\frac{\partial U}{\partial \mu} > 0$ and $\frac{\partial U}{\partial \sigma} < 0$. The optimizing problem is then given by:

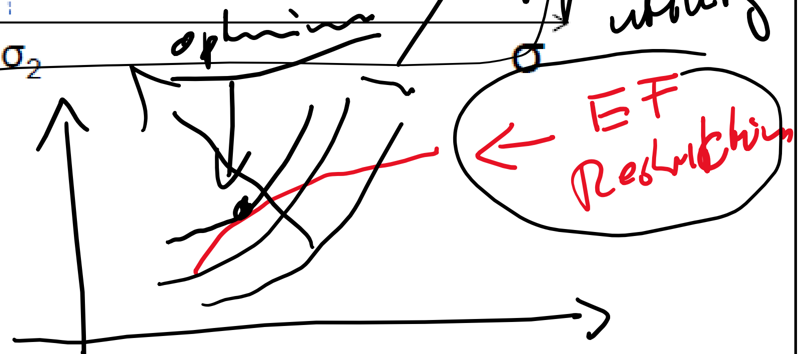
$$\max_{\mu,\sigma} U(\mu,\sigma) \quad \text{N.B.} \boxed{\sigma^2 = A[\mu-\mu_{MVP}]^2 + \sigma^2_{MVP}} \quad (A = VarK^2(R_1,R_2)) \quad (7)$$

$\max(u(x_1, x_2))$ $(m = p_1x_1 + p_2x_2)$



(μ^*, σ^*)

Indifference curves representing utility



Restriction is just the efficient frontier

Example

linear in μ — quadratic dependence of σ

The explicit utility is given by $U(\mu, \sigma) = \mu - \frac{a}{2}\sigma^2$ with $(a > 0)$. $\Longrightarrow$

micro $u = \sqrt{x_1 x_2}$

$L = \sqrt{x_1 x_2} + \lambda(m - p_1 x_1 - p_2 x_2)$

$\to \sigma A(\mu - \mu_{MVP})$

$$\mathcal{L}(\mu, \sigma, \lambda) = \underbrace{\mu - \frac{a}{2}\sigma^2}_{\text{utility}} + \frac{\lambda}{2}\underbrace{(\sigma^2 - (A[\mu - \mu_{MVP}]^2 + \sigma^2_{MVP}))}_{\text{Restriction}}$$

$\Longrightarrow$ FOC

$$\frac{\partial \mathcal{L}}{\partial \mu} = 1 - A\lambda(\mu - \mu_{MVP}) = 0 \Rightarrow 1 = A\lambda(\mu - \mu_{MVP}) \Rightarrow \frac{1}{\lambda} = \mu - \mu_{MVP}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma} = -a\sigma + \lambda\sigma = 0 \Rightarrow a\sigma = \lambda\sigma \Rightarrow a = \lambda$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \frac{1}{2}(\sigma^2 - (A\underbrace{[\mu - \mu_{MVP}]^2}_{[\frac{1}{A}]^2} + \sigma^2_{MVP})) = 0 \Rightarrow$$

solve for σ from μ and calculate μ

$$\Longrightarrow \quad \mu^* = \mu_{MVP} + \frac{1}{aA} \quad \sigma^* = \sqrt{\frac{1}{Aa^2} + \sigma^2_{MVP}} \quad \boxed{\alpha^* = \frac{\mu^* - \mu_2}{\mu_1 - \mu_2}}$$

$\mu^* = \alpha\mu_1 + (1 - \alpha)\mu_2$

my investment plan

Portfolio with an additional riskless asset

Assumptions:

1. It is possible to invest any amount in the riskless asset with yield R_0.
2. Any investment can be financed via a riskless credit with interest rate R_0.

Notation

in general non zero standard deviation of the risky investment

- R_p, R_0: Yield of the risky portfolio, riskless interest rate
- $\mu_p = E(R_p), \mu_0 = R_0$: Expected yield of the risky portfolio, expected yield of the riskless investment
- $\sigma_p = \sigma(R_p), \sigma_0 = 0$: Standard deviation of the risky portfolio and the riskless investment
- R yield of the total investment consisting of the risky portfolio and the riskless investment
- $\mu = \mu(R)$: Expected yield of the total investment
- $\sigma = \sigma(R)$: Standard deviation of the total investment

Expected yield and standard deviation of the total investment

risky part   riskless part

put on the vertical axes   slope   how does this function look like

$Var(aX) = a^2 \underbrace{Var(X)}_{\sigma^2}$

from numbers

$\mu = A + 3\sigma$

Ex: $y = 3 + 6x$

$$R = \beta R_p + (1-\beta)R_0 \quad (8)$$

$$\mu(\beta) = E(R) = \mu = \beta\mu_p + (1-\beta)R_0 = R_0 + \beta(\mu_p - R_0) \quad (9)$$

$$\sigma(\beta) = Var(R) = \sigma^2 = \beta^2\sigma_p^2 \quad (10)$$

$Var(\beta R_p + \underbrace{(1-\beta)R_0}_{})$

$$\Rightarrow \underbrace{\boxed{\mu = R_0 + \frac{\mu_p - R_0}{\sigma_p}\cdot\sigma}}_{} \quad (11)$$

Sharpe Ratio SR_p

therefore β drops out

we have σ_0 of the riskless investment is $\sigma_0 = 0$

Difference between risky investment and riskless investment

found from data and statistic estimations

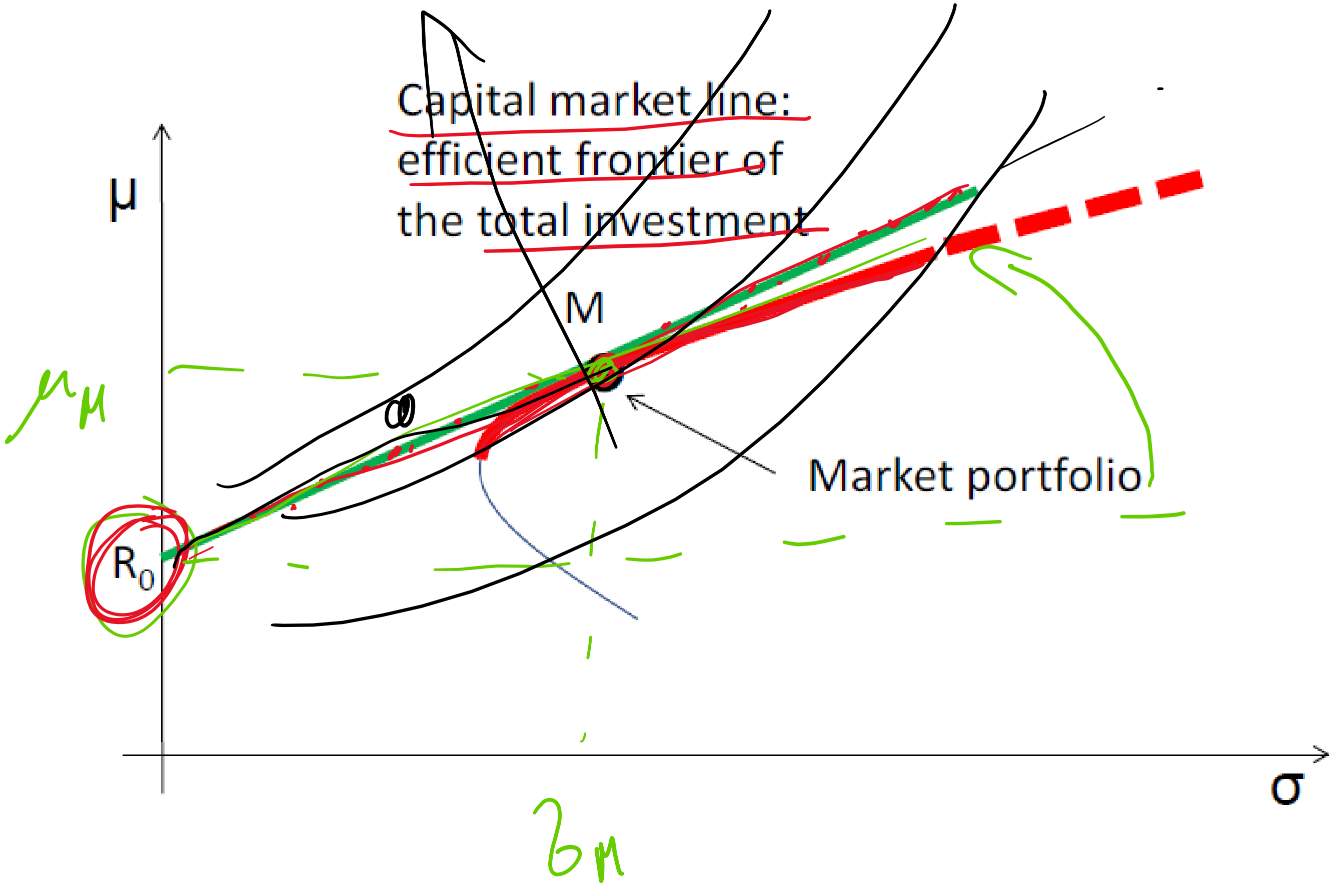
are ~~the~~ all accessible risk profiles (μ, σ) given the investment opportunities

in A_1, A_2 ← risky

R_0 ← riskless

Capital market line

The accessible portfolios of the total investment are obetained via equation (11). The efficient total investments are calculated via the tangent including the point $(0, R_0)$ touching the efficient frontier of the risky portfolio:



Capital market line and the market portfolio

In general we obtain the capital market line via

solve for σ^2

$$\mu = R_0 + B\sigma \quad \text{oder} \quad \mu = R_0 + \frac{\sigma}{\sqrt{\tilde{B}}} \quad \text{mit} \quad \tilde{B} := \frac{1}{B^2} \tag{12}$$

the function to be optimized

$$\Longrightarrow \sigma^2 = \tilde{B}(\mu - R_0)^2 = g(\tilde{B}, \mu) \tag{13}$$

For the efficiant frontier of the risky portfolio we have:

Restriction

$$\sigma^2 = A[\mu - \mu_{MVP}]^2 + \sigma^2_{MVP} = h(\mu) \tag{14}$$

The capital market line is (equation (12)) the line with maximum slope B touching the efficient frontier of risky profolio (equation (14)). Thus, we have to ~~minize~~ minimize the parameter $\tilde{B}$ subject to the capital market line and efficient frontier are touching each other.

Capital market line and the market portfolio

$$\min_{\mu,\tilde{B}} \tilde{B} \quad \text{N.B.} \quad g(\tilde{B},\mu) = h(\mu) \tag{15}$$

$$\implies \quad \mathcal{L}(\tilde{B},\mu,\lambda) = \tilde{B} + \lambda(\underbrace{\tilde{B}(\mu - R_0)^2 - ((A[\mu-\mu_{MVP}]^2 + \sigma^2_{MVP})}_{\text{Restriction}}) \tag{16}$$

FOC:

$$\frac{\partial \mathcal{L}}{\partial \tilde{B}} = 1 + \lambda(\mu - R_0)^2 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = 2\lambda\left(\tilde{B}(\mu - R_0) - A(\mu - \mu_{MVP})\right) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \tilde{B}(\mu - R_0)^2 - (A[\mu - \mu_{MVP}]^2 + \sigma^2_{MVP}) = 0$$

$\Rightarrow \lambda = -\frac{1}{(\mu - R_0)^2}$

$\Rightarrow$ this can be solved for $\mu(\tilde{B})$

$\tilde{B}(\mu)$

$\Rightarrow$ we can solve for $\tilde{B}$

Capital market line and the market portfolio

Solving FOC, we obtain:

$$B_{KML} = \sqrt{\frac{1}{A} + \left[\frac{\mu_{MVP} - R_0}{\sigma_{MVP}}\right]^2} \quad (17)$$

$$\mu_M = \mu_{MVP} + \frac{\sigma^2_{MVP}}{A(\mu_{MVP} - R_0)} \quad (18)$$

$$\sigma^2_M = \sigma^2_{MVP}\left[\frac{\sigma^2_{MVP}}{A(\mu_{MVP} - R_0)^2} + 1\right] \quad (19)$$

with B_{KML}: slope of the capital market line (μ_M, σ_M): Expected yield and stadard deviation of the market portfolio. The market portfolio consists of $\alpha_M = \frac{\mu_M - \mu_2}{\mu_1 - \mu_2}$ of asset 1 and $1 - \alpha_M = \frac{\mu_M - \mu_1}{\mu_2 - \mu_1}$ of asset 2.

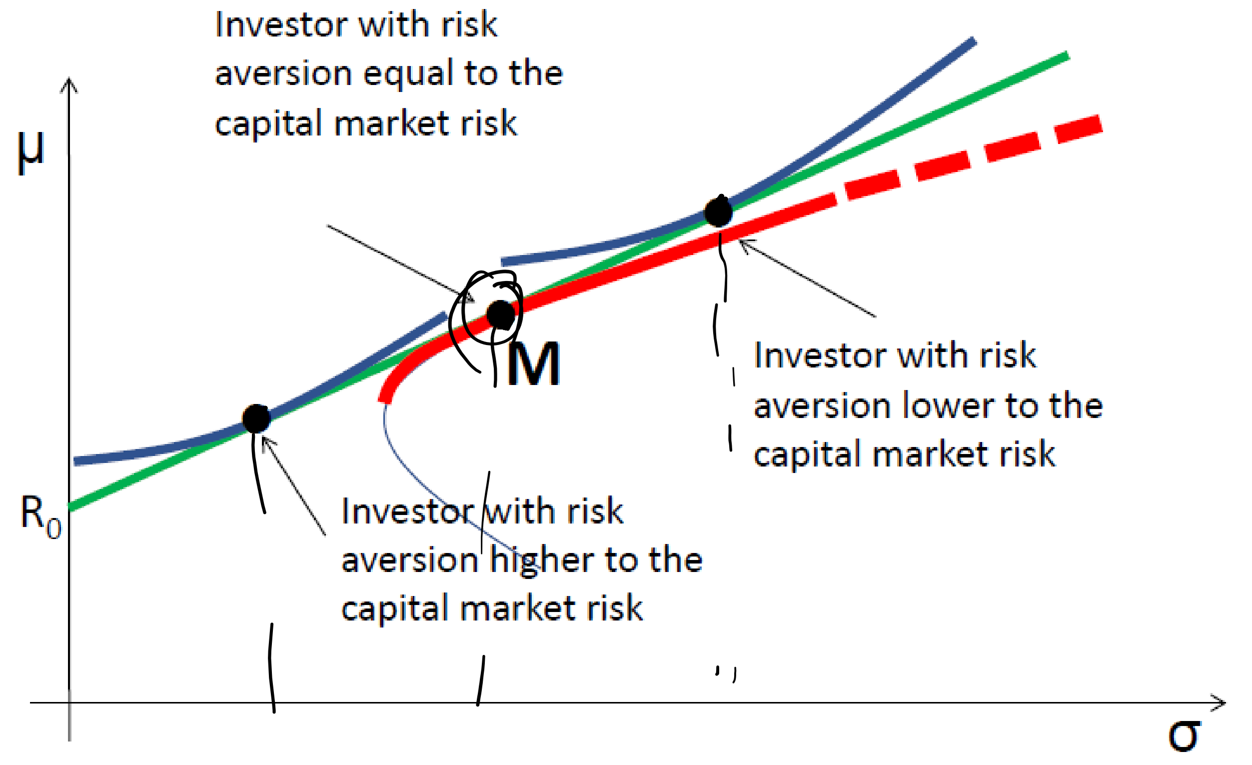
$\Longrightarrow$ It is possible to replicate every risk profile via investing in the riskless asset and the risky market portfolio, starting with 100% investment in the riskless asset $(\mu, \sigma) = (R_0, 0)$.

represents in general the ~~risk of~~ average risk of the capital. market including the risky investment possibilities

Optimal Investment of the riskless asset

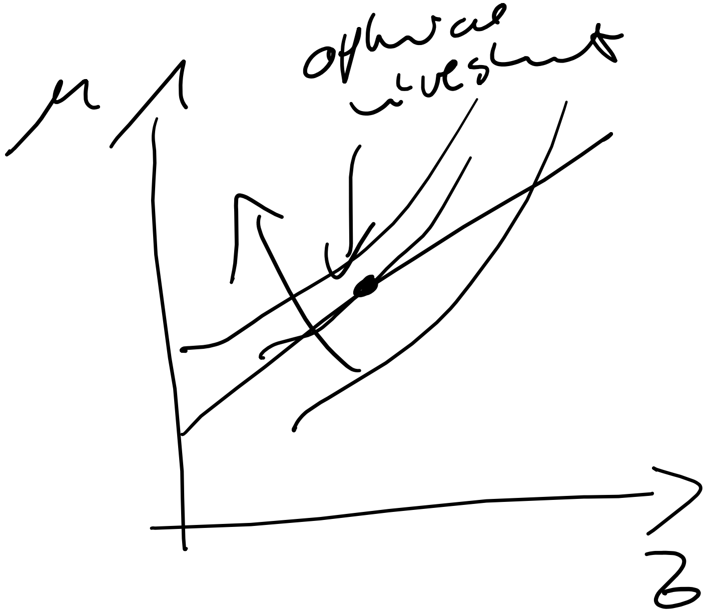
The condition for the optimal investment is now given by the tangent point of capital market line and the indifference curve.

$$\max_{\mu,\sigma} U(\mu,\sigma) \quad \text{N.B. } \mu = R_0 + \sigma B_{KML} \tag{20}$$



Example

Again utility is given by $U(\mu, \sigma) = \mu - \frac{a}{2}\sigma^2$ mit $(a > 0)$. $\Longrightarrow$



$$\mathcal{L}(\mu, \sigma, \lambda) = \mu - \frac{a}{2}\sigma^2 + \lambda(R_0 + \sigma B_{KML} - \mu)$$

$\Longrightarrow$ FOC

$$\left.\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= 1 - \lambda = 0 \\
\frac{\partial \mathcal{L}}{\partial \sigma} &= -a\sigma + \lambda B_{KML} = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= R_0 + \sigma B_{KML} - \mu = 0
\end{aligned}\right\}$$

homework

$$\Longrightarrow \quad \mu^* = R_0 + \frac{B_{KML}^2}{a} \quad \sigma^* = \frac{B_{KML}}{a} \quad \beta^* = \frac{\mu^* - R_0}{\mu_M - R_0}$$

CAPM and linear regression

Over or under performance of security i in relation to the market portfolio M is given by (due to the assumptions of the Makowitz model)e:

$$\boxed{\mu_i - R_0 = \underbrace{\frac{COV(R_i, R_M)}{\sigma_M^2}}_{\beta_i}(\mu_M - R_0)} \quad \text{CAPM-equation}$$

famous β - factors

In analogy to the capital market line this condition related to a specific security is named security market line. compare with the capital market line

$\mu_i - R_0$ is the risk premium of security i related to the riskless interes rate R_0 equaling β times the risk premium of the market portfolio related to R_0. Therfore β can be interpreted as the individual risk of security i relativ to the market risk.

estimation via linear regression

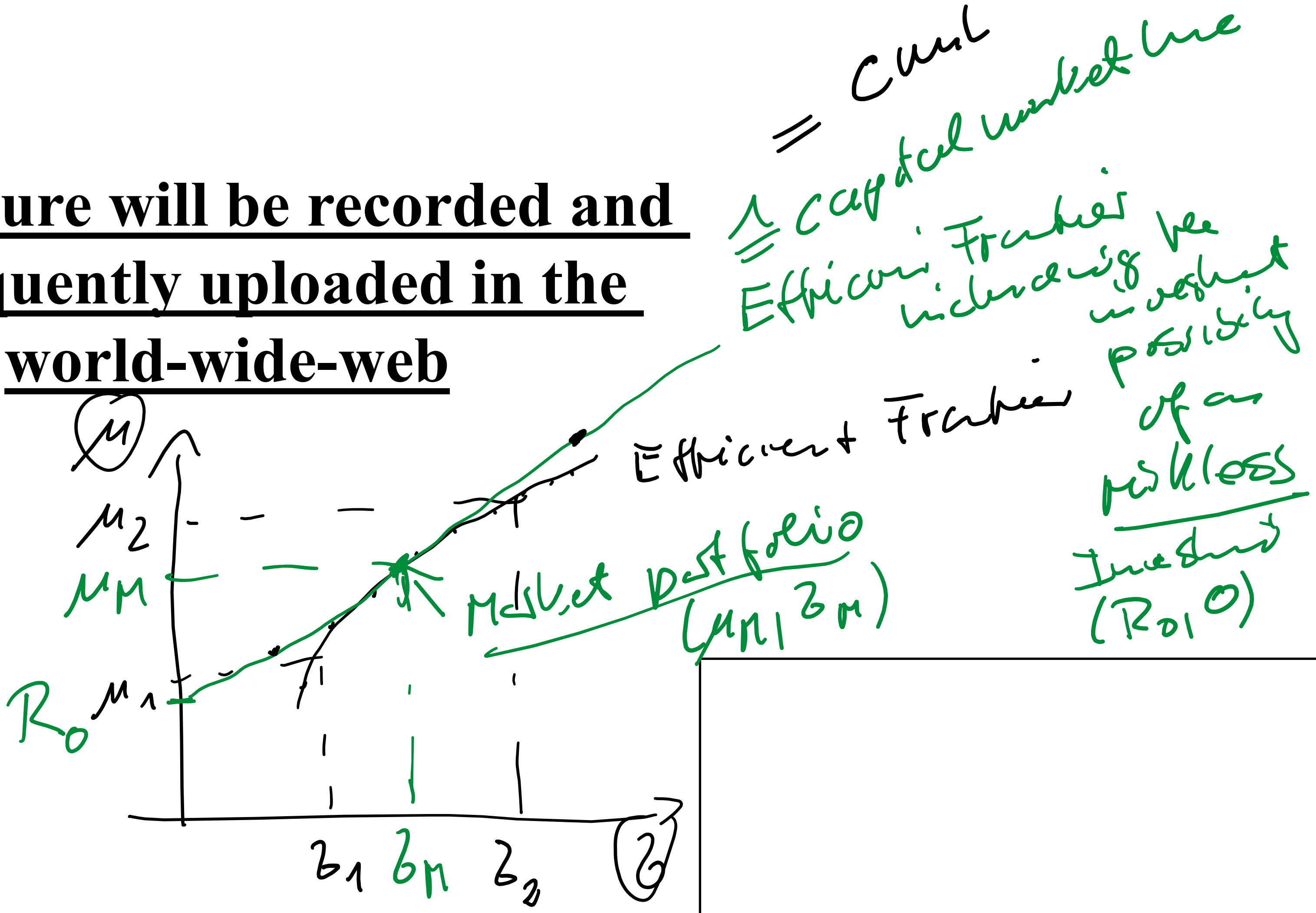
all DAX securities
DAX itself

⤳ google finance
yahoo finance } → direct plugin into excel

Statistics A

Wilhelmshaven

<u>This lecture will be recorded and Subsequently uploaded in the world-wide-web</u>



[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

CAPM and linear regression → Capital Asset Pricing-Modell

Over or under performance of security i in relation to the market portfolio M ⟹ Equilibrium-model
is given by (due to the assumptions of the Makowitz model)e: yield of the market portfolio

SML

$$\mu_i - R_0 = \underbrace{\frac{COV(R_i, R_M)}{\sigma_M^2}}_{\beta_i}(\mu_M - R_0) \qquad \text{CAPM-equation}$$

famous β-factors

In analogy to the capital market line this condition related to a specific security is named security market line. compare with the capital market line

$\mu_i - R_0$ is the risk premium of security i related to the riskless interes rate R_0 equaling β times the risk premium of the market portfolio related to R_0. Therfore β can be interpreted as the individual risk of security i relativ to the market risk.

estimation via linear regression

all DAX securities
DAX itself

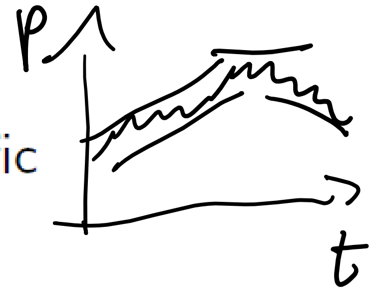
⇝ google finance
yahoo finance } → direct plugin into excel

CAPM and linear regression

Over or under performance of security i in relation to the market portfolio M is given by (due to the assumptions of the Makowitz model)e:

$$\mu_i - R_0 = \underbrace{\frac{COV(R_i, R_M)}{\sigma_M^2}}_{\beta_i}(\mu_M - R_0) \quad \text{CAPM-equation}$$

$\mu_M \mathrel{\hat{=}}$ large stock-market index
S&P 500
MSCI-World
DAX
Nikkei
Dow Jones



In analogy to the capital market line this condition related to a specific security is named security market line.

$\mu_i - R_0$ is the risk premium of security i related to the riskless interes rate R_0 equaling β times the risk premium of the market portfolio related to R_0.
Therfore β can be interpreted as the individual risk of security i relativ to the market risk. What data should we take in order to estimate the β-factors?

$\mu_i \mathrel{\hat{=}}$ historical yields of single securities ← individual DAX securities

$R_0 \mathrel{\hat{=}}$ Interest rates of governments bonds or we can take the reference interest rates of the financial markets Libor oder $\boxed{\text{Euribor}}$ ← 3-month

Which Period and which frequency should we take

Why not taking daily data which is accissible?

-> but Daily can be in general highly correlated

-> the general calculated risk of a single security is mainly driven by a short subperiod

Why not taking yearly data?

-> but the die sample size goes down

-> we should increase the whole period, we are looking at

-> i.e we got back the last 30 years

~> dot com ~2001

. financial crisis 2008/09

. Corona crisis 2020

-> but this means that the framework of financial markets was highly changed in this period

-> reasonable choice of frequency monthly

Period length 5 - 10 years

Where do I get the data

-> - DAX

https://de.finance.yahoo.com/quote/ADS.DE/history?period1=1262304000&period2=1653264000&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true

- DAX-titles

Addis 1. Mai 160

1. April 150

$R_i = \frac{160}{150} - 1$

- Euribor

https://sdw.ecb.europa.eu/

https://fred.stlouisfed.org/

you have also a excel add-in

p.a. $\Rightarrow \frac{\text{Euribor}}{12}$

https://www.bundesbank.de/dynamic/action/de/statistiken/zeitreihen-datenbanken/zeitreihen-datenbank/759778/759778?listId=www_sgeldmkt_mb03_neu

Dax-Performance

https://query1.finance.yahoo.com/v7/finance/download/NFLX?period1=1022112000&period2=1589241600&interval=1d&events=history

$\rightarrow \Delta R_i = R_{it} - R_{0t} = \beta(R_{Mt} - R_{0t})$

$\Rightarrow$ linear Regression with no absolute Parameter

$y = a + bx$

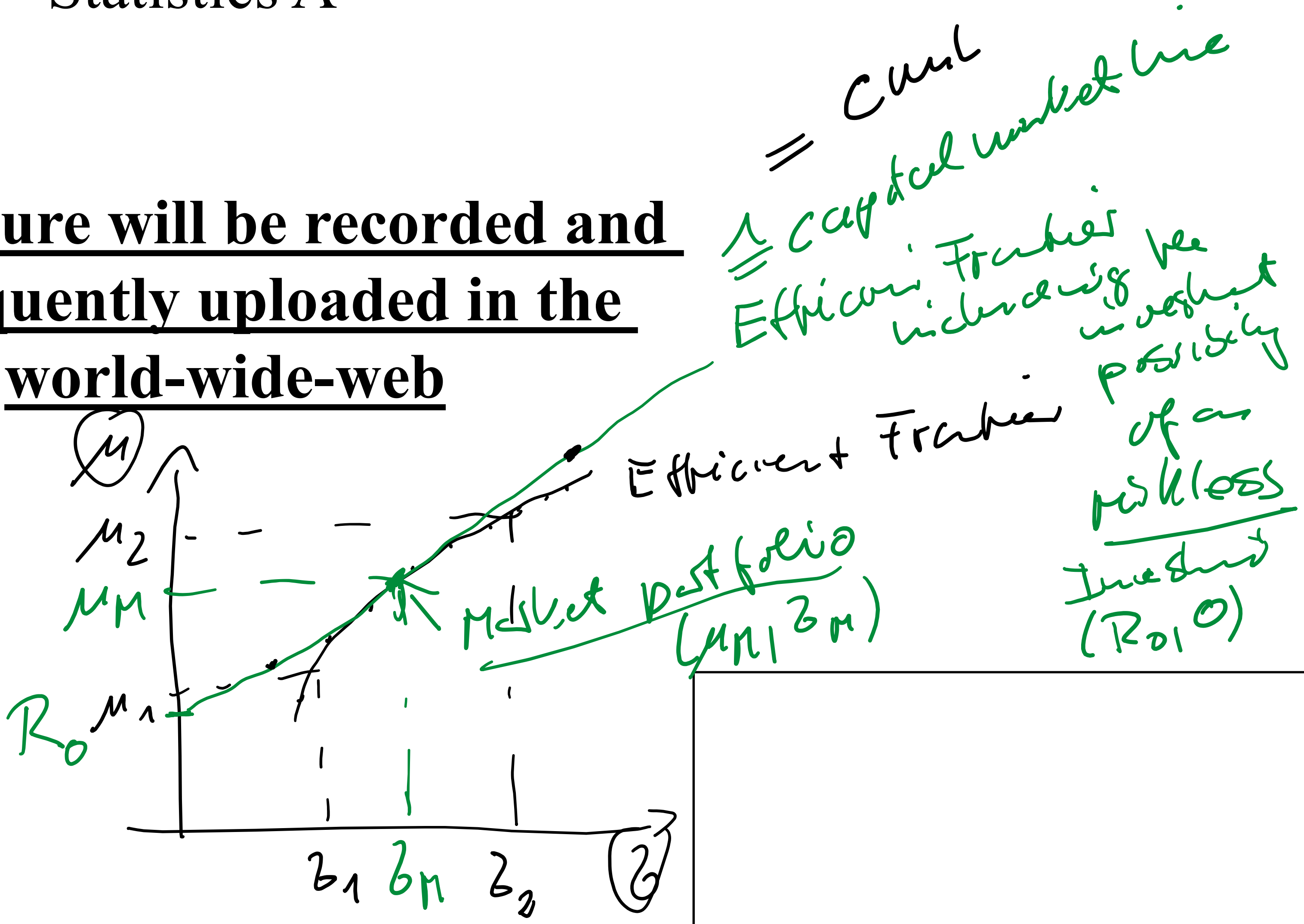
$\Rightarrow a = 0$ in our case

$\leadsto \Delta \quad R_{it} + R_{0t} = \beta(R_{Mt} - R_{0t}) + \varepsilon_t$

Statistics A

Wilhelmshaven

This lecture will be recorded and Subsequently uploaded in the world-wide-web



[Function translator (webpage)]

[Function translator Excel 1 (add in)]

Prof. Dr. Bernhard Köster
Jade-Hochschule Wilhelmshaven
http://www.bernhardkoester.de/vorlesungen/inhalt.html

CAPM and linear regression

Over or under performance of security i in relation to the market portfolio M is given by (due to the assumptions of the Makowitz model):

$$\mu_i - R_0 = \underbrace{\frac{COV(R_i, R_M)}{\sigma_M^2}}_{\beta_i} (\mu_M - R_0) \quad \text{CAPM-equation}$$

$>/< 1$

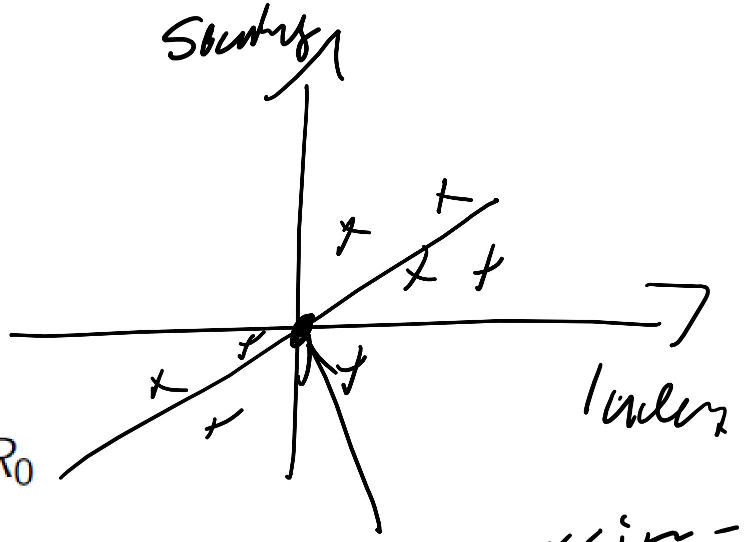
$+ \alpha = 0$

R_i guaranteed Interest rate DAX guaranteed interest rate

SML

CML (Markowitz)

In analogy to the capital market line this condition related to a specific security is named security market line.



$\mu_i - R_0$ is the risk premium of security i related to the riskless interes rate R_0 equaling β times the risk premium of the market portfolio related to R_0. Therfore β can be interpreted as the individual risk of security i relativ to the market risk.

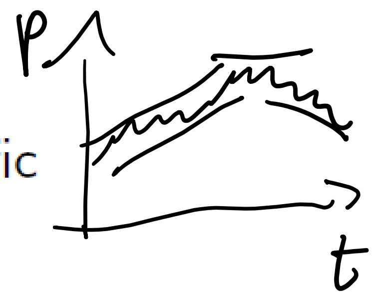
linear regression line has to pass through the origin

CAPM and linear regression

Over or under performance of security i in relation to the market portfolio M is given by (due to the assumptions of the Makowitz model)e:

$$\mu_i - R_0 = \underbrace{\frac{COV(R_i, R_M)}{\sigma_M^2}}_{\beta_i}(\mu_M - R_0) \quad \text{CAPM-equation}$$

$\mu_M \hat{=}$ large stock-market index
S&P 500
MSCI-World
DAX
Nikkei
Dow Jones



In analogy to the capital market line this condition related to a specific security is named security market line.

$\mu_i - R_0$ is the risk premium of security i related to the riskless interes rate R_0 equaling β times the risk premium of the market portfolio related to R_0. Therfore β can be interpreted as the individual risk of security i relativ to the market risk. What data should we take in order to estimate the β-factors?

$\mu_i \hat{=}$ historical yields of single securities ← individual DAX securities

$R_0 \hat{=}$ Interest rates of governments bonds or we can take the reference interest rates of the financial markets Libor oder $\boxed{\text{Euribor}}$ ← 3-month

Which Period and which frequency should we take

Why not taking daily data which is accissible?

-> but Daily can be in general highly correlated

-> the general calculated risk of a single security is mainly driven by a short subperiod

Why not taking yearly data?

-> but the sample size goes down

-> we should increase the whole period, we are looking at

-> i.e we got back the last 30 years

~> .dot com ~2001

. financial crisis 2008/09

. Corona crisis 2020

-> but this means that the framework of financial markets was highly changed in this period

-> reasonable choice of frequency monthly

Period length 5 - 10 years

Where do I get the data

-> - DAX

https://de.finance.yahoo.com/quote/ADS.DE/history?period1=1262304000&period2=1653264000&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true

- DAX-titles

Addis 1. Mai 160

1. April 150

$R_i = \frac{160}{150} - 1$

- Euribor

https://sdw.ecb.europa.eu/

https://fred.stlouisfed.org/

you have also a excel add-in

p.a. => $\frac{\text{Euribor}}{12}$

https://www.bundesbank.de/dynamic/action/de/statistiken/zeitreihen-datenbanken/zeitreihen-datenbank/759778/759778?listId=www_sgeldmkt_mb03_neu

Dax-Performance

https://query1.finance.yahoo.com/v7/finance/download/NFLX?period1=1022112000&period2=1589241600&interval=1d&events=history

$\rightarrow \Delta R_i = R_{it} - R_{ot} = \beta (R_{Mt} - R_{ot})$

=> linear Regression with no absolute Parameter

$y = a + bx$

=> $a = 0$ in our case

$\sim) \quad \Delta \; R_{it} + R_{ot} = \beta (R_{Mt} - R_{ot}) + \varepsilon_t$