

## Exercises Tutorial 1

### Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

## Introduction

**Next Tutorial:**

**Group V: HPH G3**

**Group H: HCI H2.1**

**Group E: HCI D2**

**Group K: HCI D8**

What we will do are to:

- ✓ see applications of the topics we have learned in the lecture.
- ✓ calculate probabilities by ourselves.
- ✓ prepare for the examination.

What is required?

- ✓ preparation for lecture and exercise
- ✓ open-mind: not hesitate to ask
- ✓ do exercise by yourself
- ✓ bring with you script and exercises!

## Organization

➤ Office hours:

Vicky-Kazu-Hari-Eva

Monday 11:30-12:30

Thursday 13:30-14:30

Matthias Schubert:

Thursday 14.00-16.00

➤ Materials: all the material is available from

[http://www.ibk.ethz.ch/fa/education/ss\\_statistics/index](http://www.ibk.ethz.ch/fa/education/ss_statistics/index)

- Script
- Exercises and their solutions (for exercise tutorials)
- Past exam paper (for self study)
- Presentations (uploaded day before)
- Glossary

## Examination

- Two assessments are held during semester (3<sup>rd</sup> May and 14<sup>th</sup> June).
- Assessments count for 1/3 of final mark
- Examination (e.g. October, March...) counts for 2/3 of final mark
- Information can be found in the Preamble of the script.
- If you have questions, **do not hesitate to ask the assistants !!!!!!!!!!!!!!!!!!!!!!!**

**Assessments and examination are in English**

**Assessments: Multiple choice questions and 1 exercise.**

## Examination

### Students with first subscription in 2006 or before:

*If you already have the Testat required:*

#### Option 1:

Go directly to the final exam (October/March)-  
the mark you manage in the exam is  
your final mark for the course.

#### Option 2:

Repeat BOTH assessments during the SS07.  
Final mark will be:  
(1/3 from assessments)+(2/3 from final exam)

Inform assistants till 20<sup>th</sup> of April

*If you do not have the Testat :*

Must do the two assessments.

No matter what your mark is in the  
assessments you get the Testat😊 and so  
can go to the final exam

## Exercise tutorials

- At least 2 new exercises shown in steps (in the content of the last lecture)
- At least 1 full solution of an exercise of the previous tutorial
- Group presentation of 1 exercise (25 min including questions)
  - not obligatory
  - helpful for you to do
  - not marked
  - present using any means you choose (pc, board etc.)
  - show to assistants the solution (use Monday's office hours)
- Information can be found in the Preamble of the script.
- If you have questions, **do not hesitate to ask the assistants !!!!!!!!!!!!!!!!!!!!!**

## Exercise 1.1 (Earthquake)

In spite of a small seismic activity, the risk of a large earthquake with significant consequences always exists. A large earthquake may occur once in 1000 years. In a given region, 300 years have passed without a significant earthquake occurring. How large is the probability that a significant earthquake will occur in this region, in the current year?

1. The probability has increased.
2. The probability has remained the same.
3. The probability has decreased.





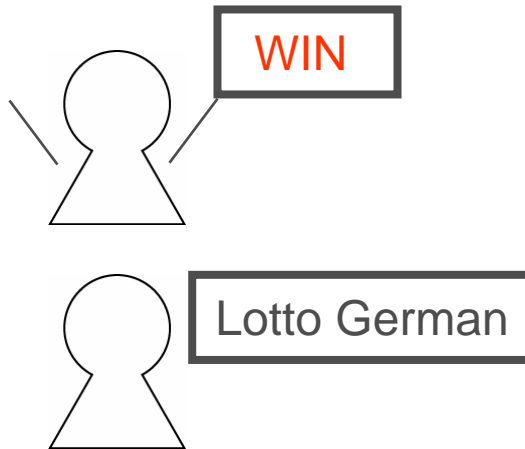
Let's think in daily life...

You go to a lottery shop with your friend, then your friend buys a lottery "Swiss" and you buy a lottery "German". (These lotteries have only one winner respectively!!!)



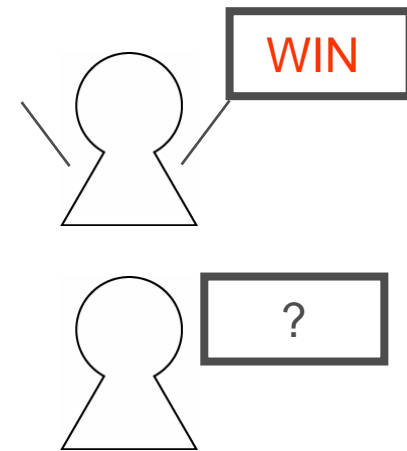
Let's think in daily life...

Then your friend won the lottery.



Let's think in daily life...

Does the probability that you win the lottery "German" change?



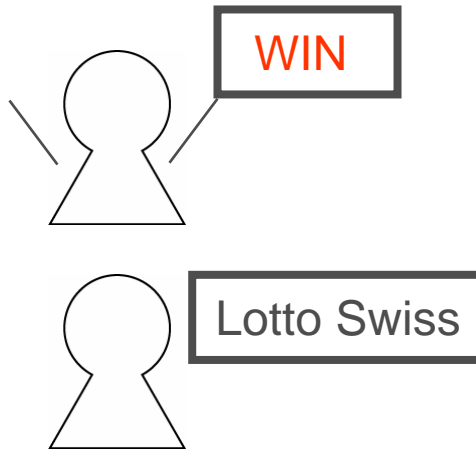
Let's think in daily life...

Then another day, you go to a lottery shop with your friend, then you and your friend buy one lottery "Swiss" respectively. (Again, this lottery has only one winner!!!)



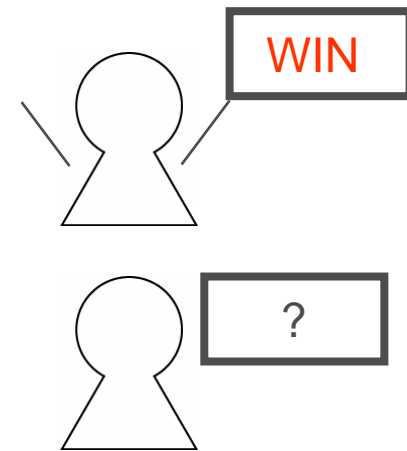
Let's think in daily life...

Then your friend won the lottery.



Let's think in daily life...

Does the probability that you win the lottery change?



## Answer 1.1 (Earthquake)

In spite of a small seismic activity, the risk of a large earthquake with significant consequences always exists. A large earthquake may occur once in 1000 years. In a given region, 300 years have passed without a significant earthquake occurring. How large is the probability that a significant earthquake will occur in this region, in the current year?

1. The probability has increased.
2. The probability has remained the same.
3. The probability has decreased.



## Exercise 1.2 (Risk definition)

Considering an activity with only one event with potential consequences, the risk is that probability that this event will occur multiplied with the consequences given the event occurs.

Which one of the following events is the riskiest?

Event	1	2	3
Event probability	10%	1%	20%
Consequences	100 SFr	500 SFr	100 SFr



## Answer 1.2 (Risk definition)

Considering an activity with only one event with potential consequences, the risk is that probability that this event will occur multiplied with the consequences given the event occurs.

Which one of the following events is the riskiest?

$$\text{Risk} = (\text{Probability}) \times (\text{Consequences})$$

Event	1	2	3
Event probability	10%	1%	20%
Consequences	100 SFr	500 SFr	100 SFr
Risk	10 SFr	5 SFr	20 SFr

## Exercise 1.3 (Risk of different activities)

Following a number of different activities is given, which involve death as a possible consequence. Which is the riskiest one?

1. Crossing a bridge
2. Smoking 20 cigarettes per day
3. Traveling 100000 km by train

### Answer 1.3 (Risk of different activities)

Following a number of different activities is given, which involve death as a possible consequence. Which is the riskiest one?

1. Crossing a bridge
- 2. Smoking 20 cigarettes per day**
3. Traveling 100000 km by train

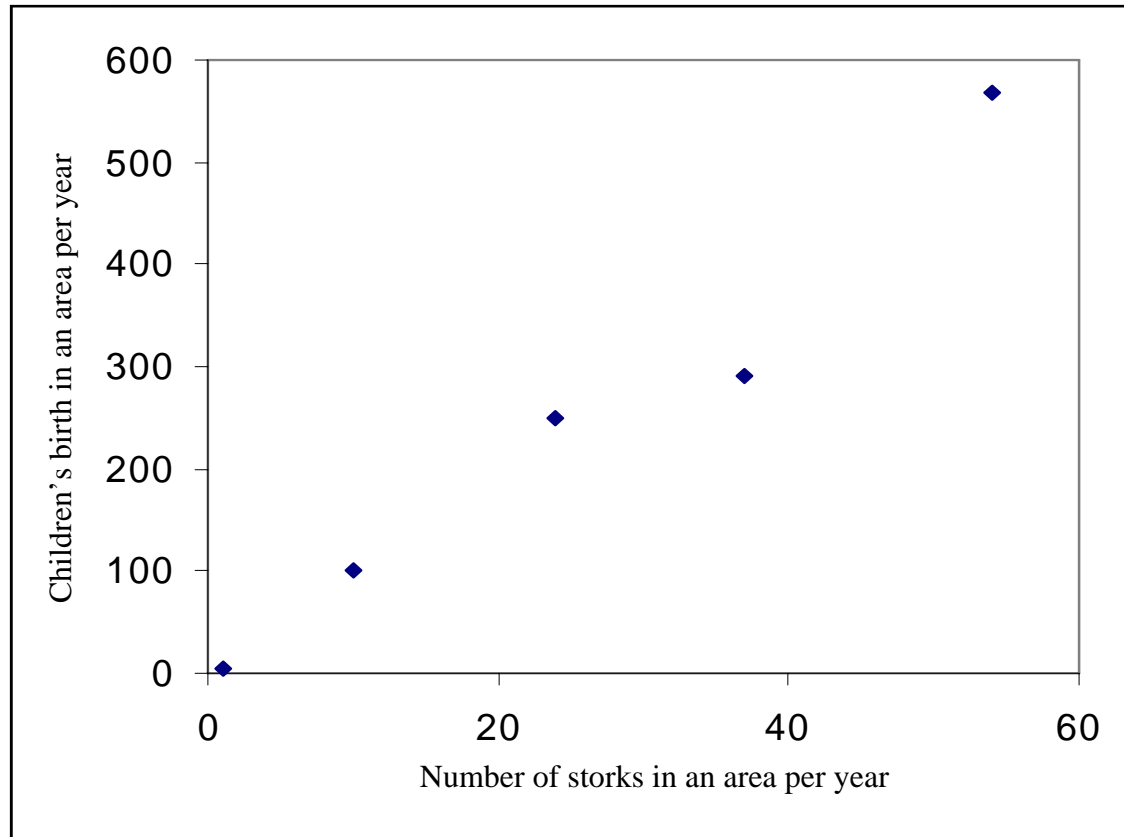
<b>Mean death risk</b> Per year and per 100.000 persons	
Overall	
100	Wood cutting, wood transport
90	Forest enterprise
50	Worker on a construction site
15	Chemistry industry
10	Mechanical factory
5	Office work
Miscellaneous risks	
400	20 cigarettes a day
300	1 bottle of wine per day
150	Motor bicycling
100	Wing aircraft as a hobby
20	Driving a car (20-24 years)
10	Pedestrian, Houseworker
10	10000 km by personal car
5	Hiking
3	10000 km on the highway
1	Plane crash per flight
1	Fire in a building
1	10000 km by train
0.2	Death due to earthquake
0.1	Death due to lightning

## Exercise 1.4 (storks)

In a region, an investigation was carried out of the number of storks and births. It was figured out that when the number of storks is high then the amount of births is also high and vice versa. The statistics indicate that these events – the number of births and the number of storks – are correlated. What do you think?

1. It has been proved statistically that the storks bring the children.
2. They are in fact correlated, but it does not necessarily mean there is a causal relation.
3. The statistical analysis has shown that the stork is a protected bird.

Let's see...



Ref.: Uni Heidelberg

## Answer 1.4 (storks)

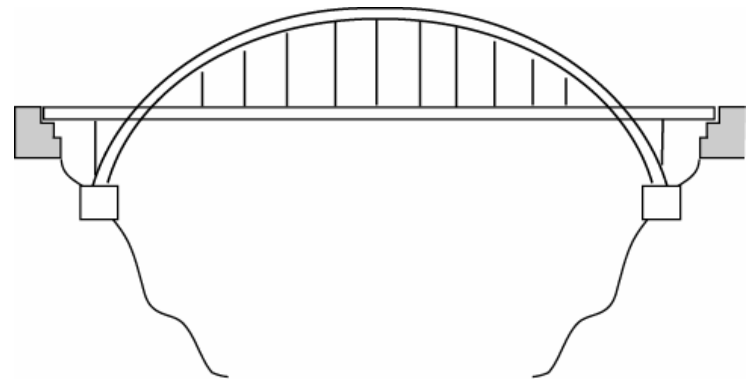
In a region, an investigation was carried out of the number of storks and births. It was figured out that when the number of storks is high then the amount of births is also high and vice versa. The statistics indicate that these events – the number of births and the number of storks – are correlated. What do you think?

1. It has been proved statistically that the storks bring the children.
2. There is no direct connection between the two events so we cannot speak about correlation.
3. The statistical analysis has shown that the stork is a protected bird.

## Exercise 1.5 (Bridge collapse)

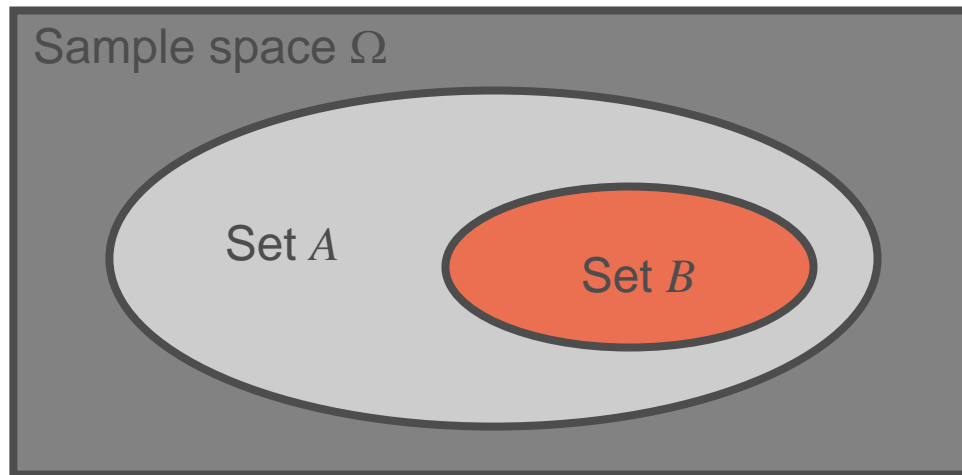
A reinforced concrete bridge shows large cracks at mid span. As a result water can reach the reinforcement and eventually corrosion will initiate. What is more probable?

1. A failure of the bridge at mid span under the action of an abnormal load.
2. A failure of the bridge at any section under the action of an abnormal load.



Let's see...

The definition of “probability measure”



$$0 \leq P(E_i) \leq 1$$

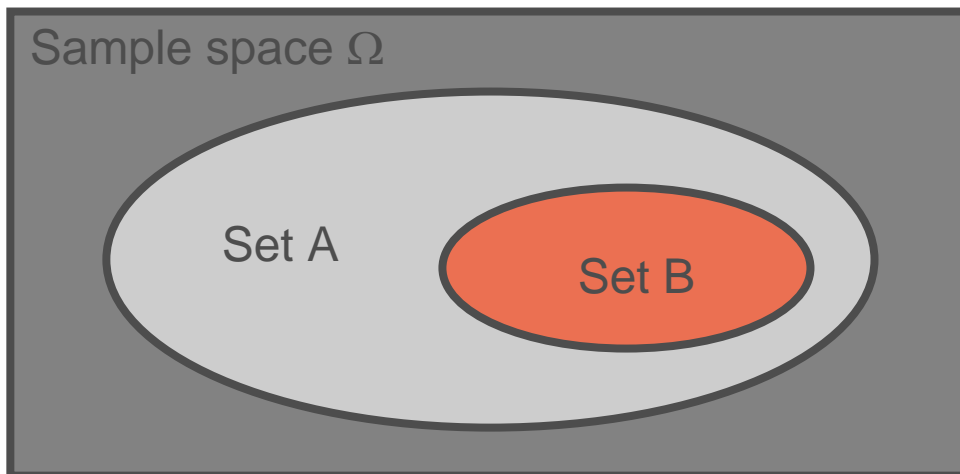
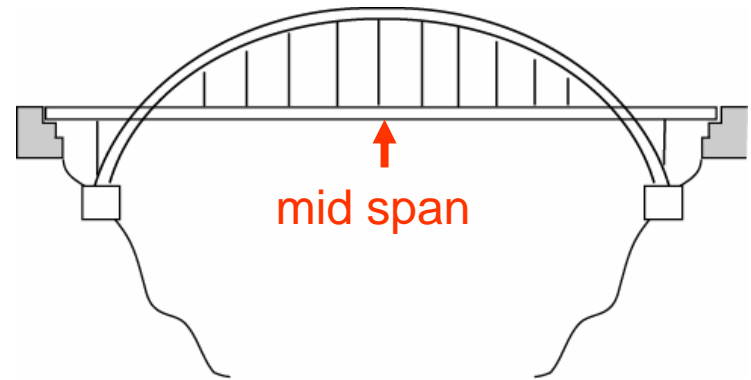
$$P(\Omega) = 1$$

$$P(A) \geq P(B)$$



Let's think...

1. A failure of the bridge at mid span under the action of an abnormal load.
2. A failure of the bridge under the action of an abnormal load.

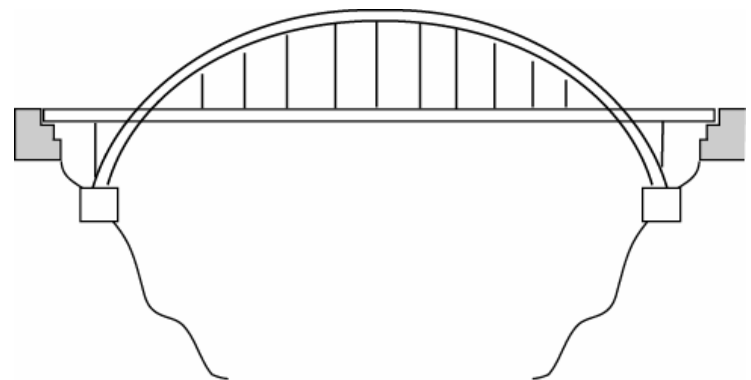


$$0 \leq P(B) \leq P(A) \leq P(\Omega) = 1$$

## Answer 1.5 (Bridge collapse)

A reinforced concrete bridge shows large cracks at mid span. As a result water can reach the reinforcement and eventually corrosion will initiate. What is more probable?

1. A failure of the bridge at mid span under the action of an abnormal load.
- 2.** A failure of the bridge under the action of an abnormal load.



## Exercise 1.6 (1000% safety)

Engineer Meier is “1000%” certain that the pedestrian bridge constructed by him is capable to withstand the load resulting from the bike racers taking part in the “Tour de Suisse”. Which statement is correct?

1. Mr. Meier has made a wrong evaluation.  
A “200%” certainty would be enough.
2. If Mr. Meier made no miscalculations, he is right.
3. There is neither 1000% certainty nor absolute safety in civil engineering.

## Answer 1.6 (1000% safety)

Engineer Meier is “1000%” certain that the pedestrian bridge constructed by him is capable to withstand the load resulting from the bike racers taking part in the “Tour de Suisse”. Which statement is correct?

1. Mr. Meier has made a wrong evaluation.  
A “200%” certainty would be enough.
2. If Mr. Meier made no miscalculations, he is right.
3. There is neither 1000% certainty nor absolute safety in civil engineering.

## Exercise 1.7

In an Alp region, there are 25 very high summits. These are covered with snow over the entire year and each day there is the same probability of occurrence of an avalanche. This amounts to  $1/40$ .

How large is the probability in this region of at least two avalanches occurring at the same day?

It is assumed that only one avalanche may occur on the same summit at the same day.

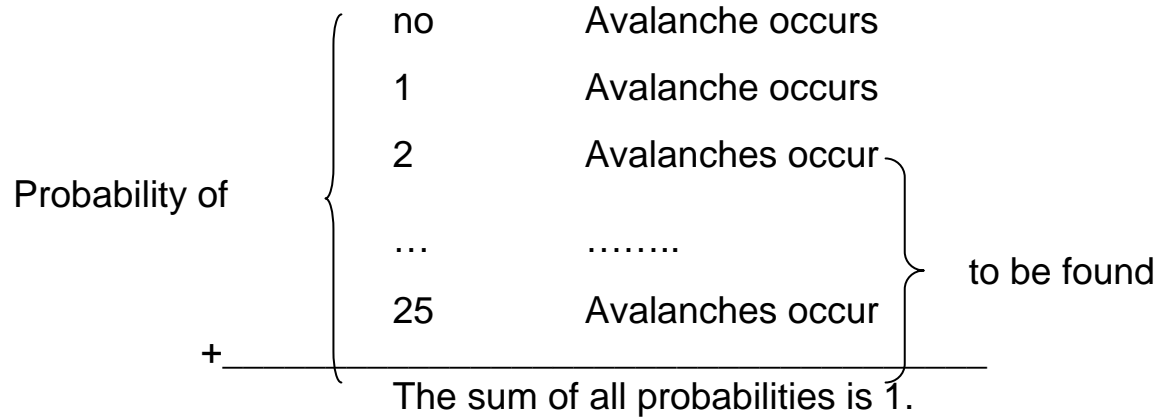


## Exercise 1.7

Probability of	{	no	Avalanche occurs	}	to be found
		1	Avalanche occurs		
		2	Avalanches occur		
		...	.....		
		25	Avalanches occur		
+					
	The sum of all probabilities is 1.				

Therefore the probabilities that no avalanche occurs and that only 1 avalanche occurs need to be determined and be subtracted from the sum of all probabilities.

## Exercise 1.7



Therefore the probabilities that no avalanche occurs and that only 1 avalanche occurs need to be determined and be subtracted from the sum of all probabilities.

The probability of occurrence of an avalanche at one summit is:  $P_j(avalanche) = \frac{1}{40} = 0.025$

The probability that no avalanche occurs at one summit is:  $P_j(no\ avalanche) = 1 - \frac{1}{40} = 0.975$

## Exercise 1.7

The probability of occurrence of an avalanche at one summit is:

The probability that no avalanche occurs at one summit is:

The probability of an avalanche only at one summit and at no other summit is calculated as:

The probability that *no avalanche occurs at any summit* (event  $A$ ) is calculated as:

The probability that *only one avalanche occurs in 25 summits* (event  $B$ ) is

The probability that *at least two avalanches occur* (event  $C$ ) can be calculated as:



## Exercise 1.8

A non destructive test method is carried out to determine whether the reinforcement of a component is corroded or not. From a number of past tests, it is known that the probability of the reinforcement being corroded is 1%. If the reinforcement is corroded, this will be indicated by the test. However there is also a 10% probability that the test will indicate that the reinforcement is corroded although this is not true (false indication).

How large is the probability that corrosion is present, if the non destructive test indicates corrosion?



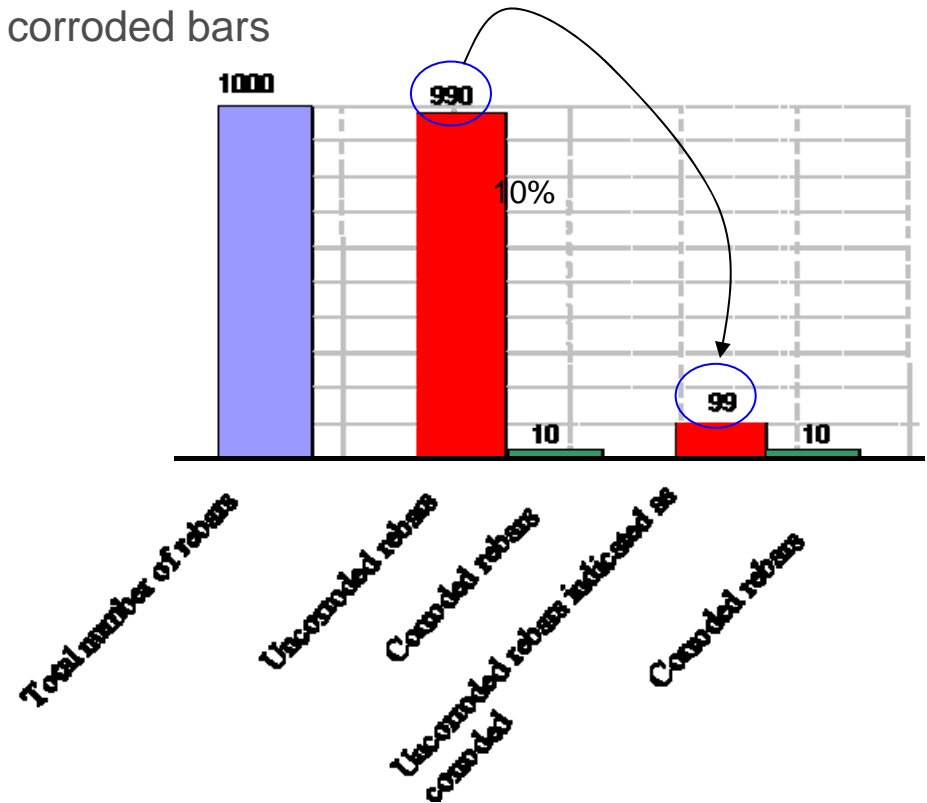
## Exercise 1.8

Let us assume that we have 1000 reinforcement bars (rebars).

According to tests: 1% of these rebars are corroded;  
 10 corroded, (990 not corroded)

Test will indicate the corroded bars: 10 corroded bars

10% probability of false indication:



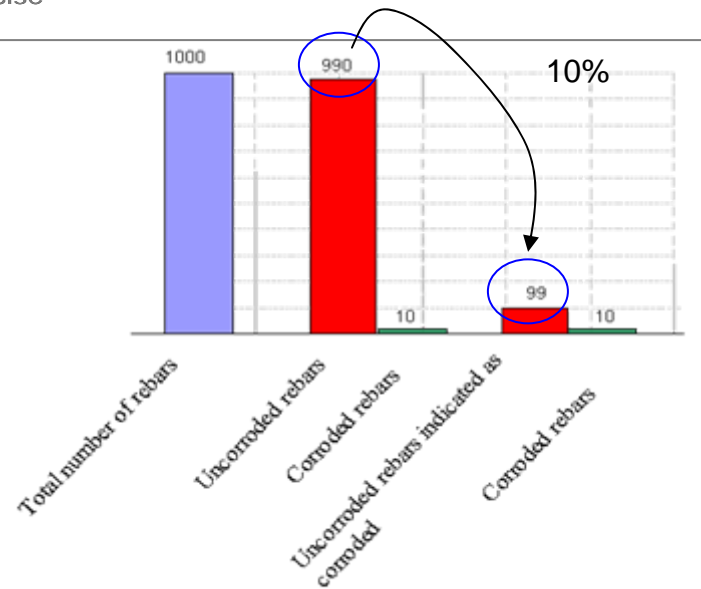
## Exercise 1.8

Let us assume that we have 1000 reinforcement bars (rebars).

According to tests:            1% of these rebars are corroded;  
    10 corroded,    (990 not corroded)

Test will indicate the corroded bars:        10 corroded bars

10% probability of false indication:



How many are indicated as corroded?

$$99+10=109$$

BUT...only 10 are truly corroded

So the probability that corrosion is present provided that the test indicates corrosion is:

$$P(\text{corrosion}) = \frac{10}{10 + 99} = 0.0917$$

No group exercise next tutorial 😊

Oups.....

Please correct the following:

Annex A:

Pages A.3 to A.5 – Module B:

different sequence of questions –  
Answer to B.3 should be corrected  
no need to re print-can change by hand

Pages A.6 to A.8- Module C:

answer stated as: C.3 is actually continuation of answer C.2 and eventually there are 9 answers- not 10. Again no need to re-print, correct by hand

If you have downloaded and printed the exercises after the 19<sup>th</sup> of March then you have the corrected version----but still pls check☺

Part B – Self Assessment questions Module B

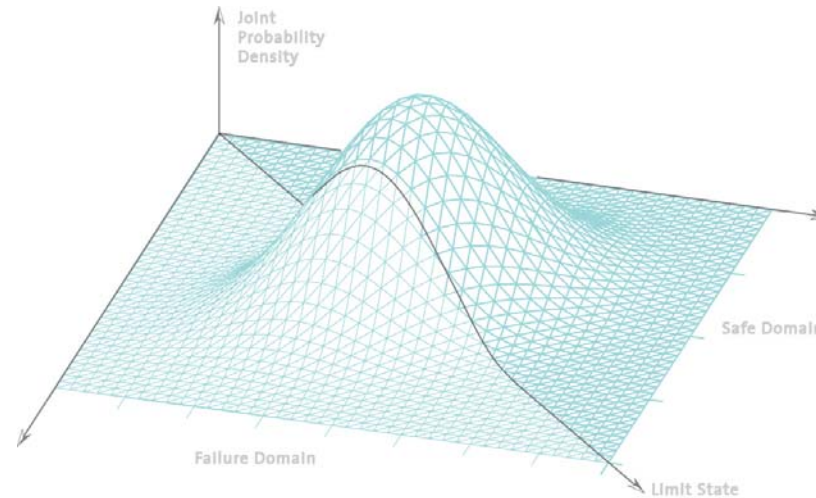
B.1 A person is asked what is the probability for achieving a “head” when flipping a coin. The person after 1000 experiments (flips with the coin) observes that “head” has occurred 333 times and hence answers that the probability for “head” is 0.333.

On which interpretation of probability is this estimation based on?

**Frequentistic definition**  $P(A) = \lim_{n_{\text{exp}} \rightarrow \infty} \frac{N_A}{n_{\text{exp}}}$  for  $n_{\text{exp}} \rightarrow \infty$   $N_A$  = number of experiments where A occurred  
 $n_{\text{exp}}$  = total number of experiments.

**Classical definition**  $P(A) = \frac{n_A}{n_{\text{tot}}}$   $n_A$  = number of equally likely ways by which an experiment may lead to A  
 $n_{\text{tot}}$  = total number of equally likely ways in the experiment.

**Bayesian definition**  $P(A)$  = degree of belief that A will occur



## Exercises Tutorial 2

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETH Zurich

## Introduction

### Function and operation

In the school we have learned about “normal” functions:

		input	output	
	f	2	4	$f(x) = x^2$
Now	P	“1 comes out”	1/6	ideal dice

In the school we have learned about “normal” operations:

	$+, -, \times, \div$	for real numbers	:	$2 + 4$
Now	$\cap, \cup, \bar{\phantom{x}}$	for set	:	$A \cap B$



## Exercise 2.1 a)

Which of the following expressions are meaningful in the way they are written?

$$P[A \cup [B \cap C]]$$

$$P[A] + P[B]$$

$$P[\bar{A}] \cap P[B]$$

$$\overline{P[B]}$$

Answer 2.1 a)

Which of the following expressions are meaningful in the way they are written?

$P[A \cup [B \cap C]]$

$P[A] + P[B]$

$P[\bar{A}] \cap P[B]$        $P[ ]$  is a real number, while  $\cap$  is an operation for set.

$\overline{P[B]}$        $P[ ]$  is a real number, while  $\bar{\phantom{x}}$  is an operation for set.

Probabilities cannot be separated and complementary events describe quantities and not the probability.

## Exercise 2.1 b)

Assume  $A$ ,  $B$  and  $C$  represent different events. Explain in words the meaning of the following expressions and what do they represent in mathematical terms (i.e. numbers, vectors, functions, sets,...)

$$A \cup B$$

$$\bar{B} \cap C$$

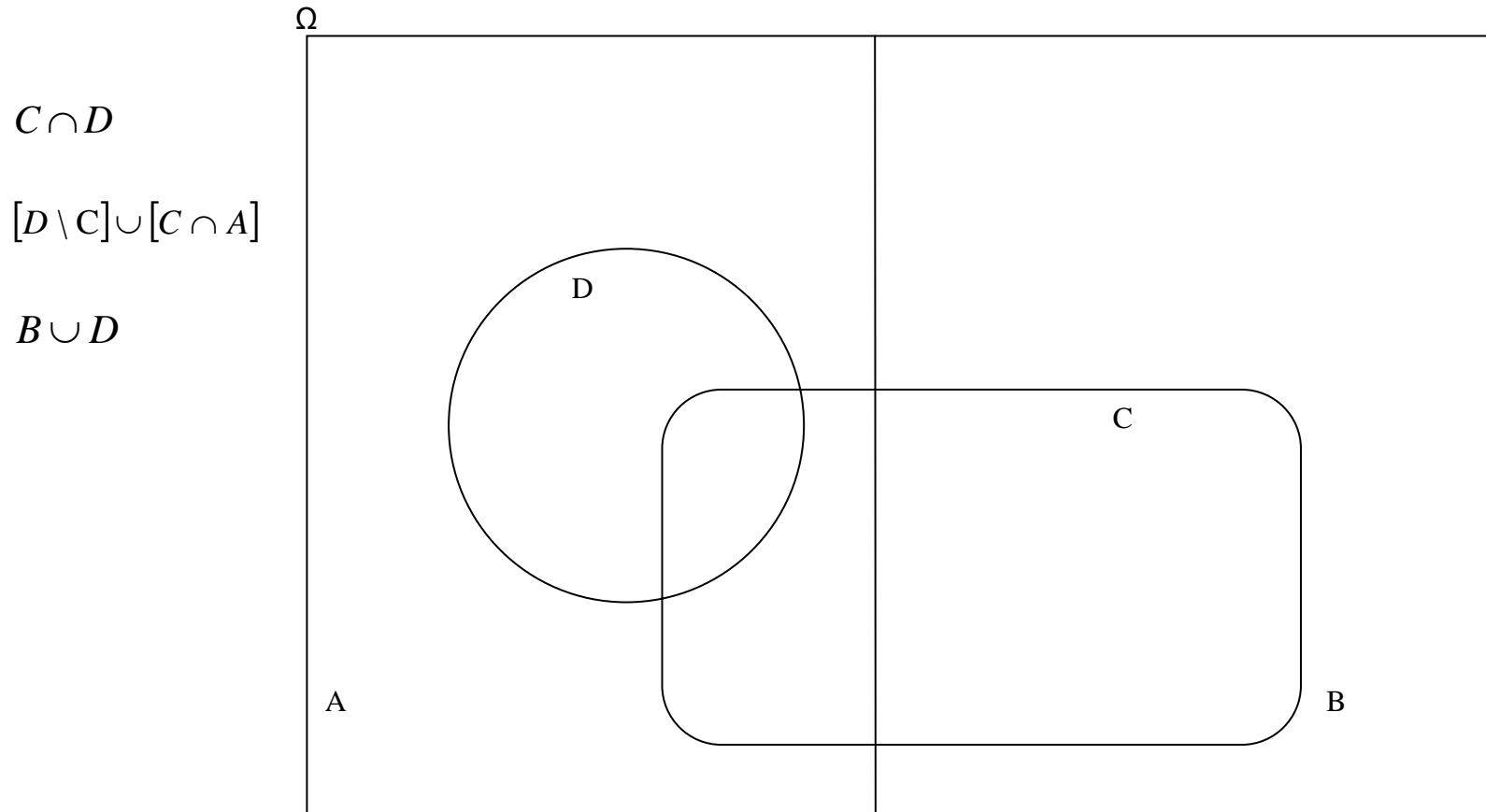
$$P[A]$$

$$P\left[[A \cap B \cap C] \cup [\bar{A} \cap \bar{B} \cap \bar{C}]\right]$$

$$\emptyset$$

Exercise 2.1 c)

Using the diagram provided below show the following events.



$C \cap D$

$[D \setminus C] \cup [C \cap A]$

$B \cup D$

## Exercise 2.2

We are throwing an ideal dice and considering the following events:

1.  $A$ : “An even number comes.”  
 $B$ : “A number dividable by 3 comes.”
2.  $A$ : “An even number comes.”  
 $B$ : “A prime number comes.”

Calculate the probability that the both events occur simultaneously for each case.



## Exercise 2.2

We are throwing an ideal dice and considering the following events:

- A*: “An even number comes.”**  
***B*: “A number dividable by 3 comes.”**

Calculate the probability that both events occur simultaneously

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$1. \quad A = \{2, 4, 6\}, \quad B = \{3, 6\}, \quad A \cap B = \{6\}$$

$$P(A) = 1/2, \quad P(B) = 1/3, \quad P(A \cap B) = 1/6$$



## Exercise 2.2

We are throwing an ideal dice and considering the following events:

1.  $A$ : “An even number comes.”  
 $B$ : “A number dividable by 3 comes.”
2.  $A$ : “An even number comes.”  
 $B$ : “A prime number comes.”

Calculate the probability that both events occur simultaneously

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$1. \quad A = \{2, 4, 6\}, \quad B = \{3, 6\}, \quad A \cap B = \{6\}$$

$$P(A) = 1/2, \quad P(B) = 1/3, \quad P(A \cap B) = 1/6$$

$$2. \quad A = \{2, 4, 6\}, \quad B = \{2, 3, 5\}, \quad A \cap B = \{2\}$$

$$P(A) = 1/2, \quad P(B) = 1/2, \quad P(A \cap B) = 1/6$$



Definition of “independent”

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

but, what does it really mean??

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

What is the probability of  $B$   
when you already know  $A$ ?

=

What is the probability of  $B$   
without any information?

The information on  $A$  cannot increase or decrease the probability of  $B$ .

More frankly, **information on  $A$  is helpless, when you want to know about  $B$**  in a probabilistic sense.



Why it is important to know about events being “independent” or not.

The story on an election **some decades ago...**

President election – two candidates A and B.

- TV station wanted to estimate the result of the election in advance.
- They asked those who voted which candidate they in fact voted, **by telephone.**
- Based on many many evidences, they concluded that: **A won the election.**
- However after counting the votes: **B won the election.**

## Review Exercises

---

What happened in reality,

- Those having a telephone were rich at that period.
- Rich people tended to vote candidate A.
- Poor people tended to vote candidate B.
- There were much more poor people than rich people.
- TV stations asked only by telephone!

Review Exercises

---

What happened in reality,

- Those having a telephone were rich at that period.
- Rich people tended to vote candidate A.
- Poor people tended to vote candidate B.
- There were much more poor people than rich people.
- TV stations asked only by telephone!

Let's assume that

$T$  is the event that a person has a telephone.

$V$  is the event representing which candidate he/she voted.

What they wanted to know is the probability of  $V$  :  $P(V)$

What they knew actually is the conditional probability of  $V$  given  $T$  :  $P(V|T)$

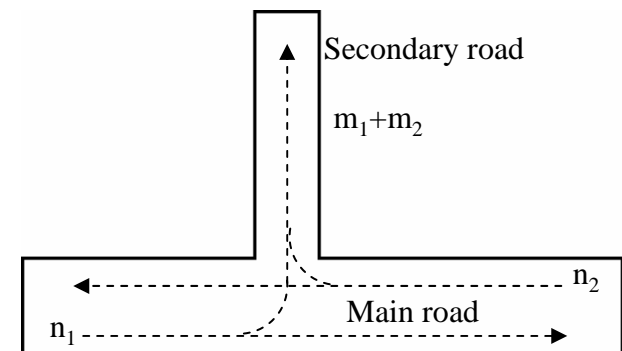
$P(V) \neq P(V|T)$  meaning  $V$  and  $T$  are not independent!

### Exercise 2.3

The observation of the traffic flow at a crossing, see the following figure, shows that  $n_1 = 50$  vehicles move on the main road in direction 1.

From those  $m_1 = 25$  vehicles turn to the secondary road.  $n_2 = 200$  vehicles move on the main road in direction 2 and  $m_2 = 40$  vehicles turn to the secondary road.

How large is the probability that a vehicle moving on the main road will turn to the secondary road?



## Answer 2.3

$n_1 = 50$  vehicles move on the main road in direction 1.  
 $m_1 = 25$  vehicles turn to the secondary road.  
 $n_2 = 200$  vehicles move on the main road in direction 2.  
 $m_2 = 40$  vehicles turn to the secondary road.

Easy!! (Classical definition of probability)

$$P(B) = \frac{m_1 + m_2}{n_1 + n_2} = \frac{25 + 40}{50 + 200} = 0.26$$

Number of equally likely ways by which an experiment leads to  $B$

Total number of equally likely ways of the experiment

where  $B$  is the event that a vehicle turns to the secondary road.

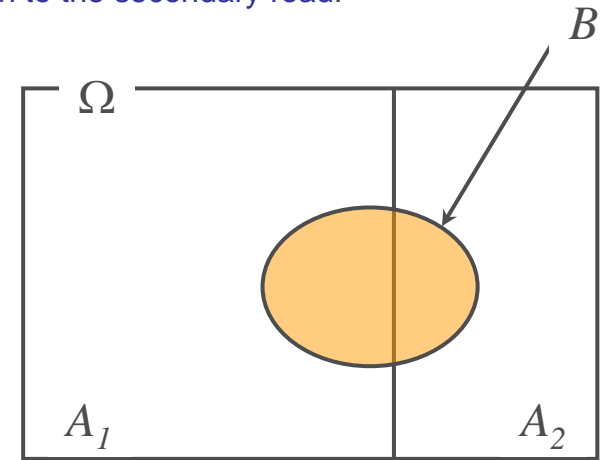
BUT, think it a little more.....

## Answer 2.3

$B$  is the event that a vehicle turns to the secondary road.

$$P(B) = P(B \cap A_1) + P(B \cap A_2) \Rightarrow$$

$n_1 = 50$  vehicles move on the main road in direction 1.  
 $m_1 = 25$  vehicles turn to the secondary road.  
 $n_2 = 200$  vehicles move on the main road in direction 2.  
 $m_2 = 40$  vehicles turn to the secondary road.



### Answer 2.3

$n_1 = 50$  vehicles move on the main road in direction 1.  
 $m_1 = 25$  vehicles turn to the secondary road.  
 $n_2 = 200$  vehicles move on the main road in direction 2  
 $m_2 = 40$  vehicles turn to the secondary road.

$B$  is the event that a vehicle turns to the secondary road.

$$P(B) = P(B \cap A_1) + P(B \cap A_2) \Rightarrow$$

$$P(B) = P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2)$$

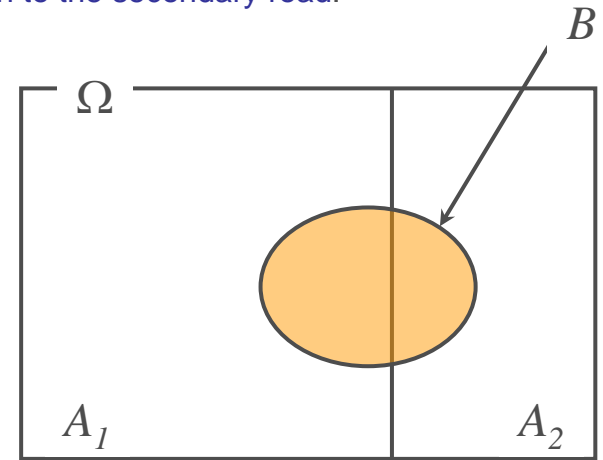
Probability that a vehicle is moving in one direction

$$P(A_1) = \frac{n_1}{(n_1+n_2)} = \frac{50}{(50+200)} = 0.2 \qquad P(A_2) = \frac{n_2}{(n_1+n_2)} = \frac{200}{(50+200)} = 0.8$$

Probability that a vehicle will turn to the secondary road

$$P(B | A_1) = \frac{m_1}{n_1} = \frac{25}{50} = 0.5 \qquad P(B | A_2) = \frac{m_2}{n_2} = \frac{40}{200} = 0.2$$

$$P(B) = P(A_1) \cdot P(B | A_1) + P(A_2) \cdot P(B | A_2) = 0.2 \cdot 0.5 + 0.8 \cdot 0.2 = 0.26$$



## Exercise 2.4

Measurements are to be carried out with measurement devices.

Since a large number of devices is required, 20% of them will be provided by IAC (Institute for the Atmosphere and Climate) and 80% will be provided by IHW (Institute for Hydraulics and Water management).

5% of the devices provided by IAC do not fulfill the required accuracy, while 2% of the devices provided by IHW do not fulfill the required accuracy.

A student carried out a measurement using a device without knowing from which institute the device was provided. Thereby, she found the inaccuracy involved in the measurement.

How large is the probability that the measurement was carried out with a device provided by IAC?



20% of the devices will be provided by IAC  
 80% will be provided by IHW

5% of the devices provided by IAC do not fulfill the required accuracy  
 2% of the devices provided by IHW do not fulfill the required accuracy.

Simplify the statement in the exercise!

$A$  = Device of Institute A (IAC)  
 $B$  = Device of Institute B (IHW)  
 $D$  = Inaccurate device

We do not know from which institute the device was provided.

How large is the probability that the measurement was carried out with a device provided by IAC?

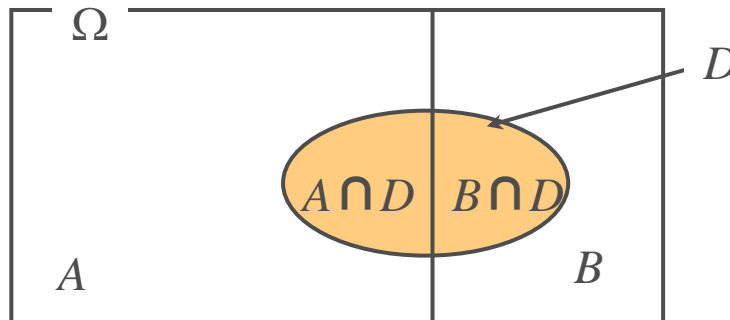
What is known???

$P(A) = 0.2$        $P(B) = 0.8$

What is required???

Probability of using an inaccurate device:

$P(D | A) = 0.05$        $P(D | B) = 0.02$



## Exercise 2.5

A non destructive test method is carried out to determine whether the reinforcement of a component is corroded or not. From a number of past experiments, it is known that the probability that the reinforcement is corroded is 1%. If the reinforcement is corroded, this will be always indicated by the test. However, there is a 10% probability that the test will indicate that the reinforcement is corroded even if there is no corrosion.

How large is the probability that corrosion is present, if the non destructive test indicates corrosion? Calculate the probability using the Bayes' theorem.

## Exercise 2.5 – Steps for solving

### Identify the events:

$K$  reinforcement is corroded

$A$  test indication

### What is known???

the probability that the reinforcement is corroded is 1%.

If the reinforcement is corroded, this will be always indicated by the test.

there is a 10% probability that the test will indicate that the reinforcement is corroded even if there is no corrosion

## Exercise 2.5 – Steps for solving

### Identify the events:

$K$  reinforcement is corroded

$A$  test indication

### What is known???

the probability that the reinforcement is corroded is 1%.

If the reinforcement is corroded, this will be always indicated by the test.

there is a 10% probability that the test will indicate that the reinforcement is corroded even if there is no corrosion

### What is required???

the probability that corrosion is present, if the non destructive test indicates corrosion

## Exercise 2.6

The failure of a building in the city of Tokyo may be caused by two independent events:

$F_1$  : A big earthquake.

$F_2$  : A strong typhoon.

The annual probabilities of occurrence of the above events are:

$$P(F_1) = 0.04$$

$$P(F_2) = 0.08$$

Calculate the annual failure probability of the building.

## Exercise 2.6

### Identify the events:

$F_1$	:	A big earthquake.
$F_2$	:	A strong typhoon.

### What is known?

$$P(F_1) = 0.04$$

$$P(F_2) = 0.08$$

### What is required?

Calculate the annual failure probability of the building.

## Exercise 2.7- (Group Exercise to be presented on 05.04.07)

Due to the increasing demand on drinking and processing water, the groundwater discharge flow has to be discussed. The hazard of long-term ground-lowering is analysed, whereby it is assumed that the ground-lowering depends on the thickness of the clay layer,  $h$ .

The thickness of the clay layer is classified in the following:

$$C_1: 0 \leq h \leq 20 \text{ cm}$$

$$C_2: 20 \text{ cm} < h \leq 40 \text{ cm}$$

$$C_3: 40 \text{ cm} < h$$

Based on experience a geologist estimates the following prior probabilities that the thickness of the clay layer at a site belongs to one of the above cases:

$$P(C_1) = 0.2 \quad P(C_2) = 0.47$$

A geo-electrical test may be useful to update the prior probability on the ground category, although the test result may not always be correct. From past experience, the probabilities of the correct/false indication of the test are known as are listed in the (uncompleted) table below:

Category of thickness of clay layer $C_i$	Indication of the category of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84		0.03
$C_2$	0	0.77	
$C_3$		0.02	0.89

## Exercise 2.7- (Group Exercise to be presented on 05.04.07)

### Lots of information----Simplify!!!

- the ground-lowering depends on the thickness of the clay layer,  $h$ .

- Classification of thickness layer: (**Events**)

$$C_1: 0 \leq h \leq 20\text{cm}$$

$$C_2: 20\text{cm} < h \leq 40\text{cm}$$

$$C_3: 40\text{cm} < h$$

- prior probabilities: (**Known probabilities**)

$$P(C_1) = 0.2 \quad P(C_2) = 0.47 \quad P(C_3) = 0.33$$

- test to update the prior probability on the ground category, the test result may not always be correct.

- probabilities of the correct/false indication of the test:

Category of thickness of clay layer $C_i$	Indication of the category of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84		0.03
$C_2$	0	0.77	
$C_3$		0.02	0.89



## Exercise 2.7- (Group Exercise to be presented on 05.04.07)

### Lots of information----Simplify!!!

- the ground-lowering depends on the thickness of the clay layer,  $h$ .
- Classification of thickness layer: (**Events**)  $C_1: 0 \leq h \leq 20\text{cm}$      $C_2: 20\text{cm} < h \leq 40\text{cm}$      $C_3: 40\text{cm} < h$
- prior probabilities: (**Known probabilities**)  $P(C_1) = 0.2$      $P(C_2) = 0.47$      $P(C_3) = 0.33$
- test to update the prior probability on the ground category, the test result may not always be correct.
- probabilities of the correct/false indication of the test:

Category of thickness of clay layer $C_i$	Indication of the category of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84		0.03
$C_2$	0	0.77	
$C_3$		0.02	0.89

### What is required???

- Complete the table
- A geo-electrical test was carried out and indicated  $C_3$  as the thickness of the clay layer. What is the probability that the thickness of the clay layer belongs to  $C_1$ ,  $C_2$ ,  $C_3$  ?

## Exercise 2.7- (Group Exercise to be presented on 05.04.07)

### What is required???

- Complete the table
- A geo-electrical test was carried out and indicated  $C_3$  as the thickness of the clay layer. What is the probability that the thickness of the clay layer belongs to  $C_1$ ,  $C_2$ ,  $C_3$  ?

### a. Complete the table

Category of thickness of clay layer $C_i$	Indication of the category of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84		0.03
$C_2$	0	0.77	
$C_3$		0.02	0.89

$$P(I = C_1 | C_1) + P(I = C_2 | C_1) + P(I = C_3 | C_1) = 1$$

---

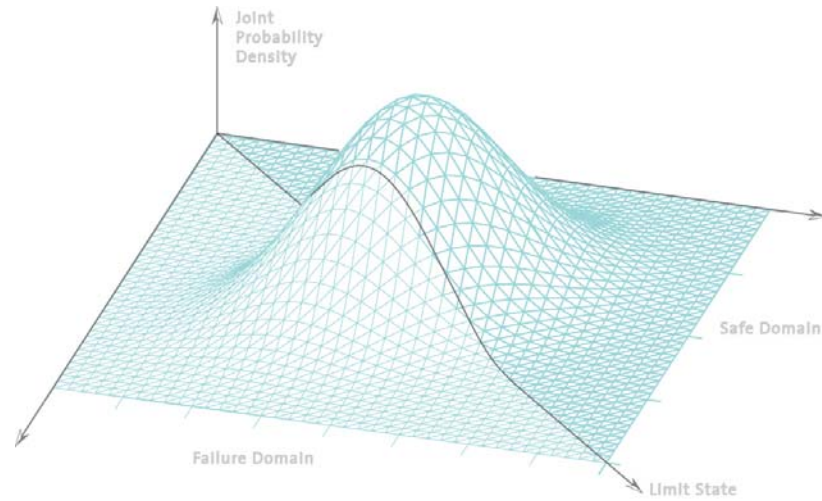
## Exercise 2.7- (Group Exercise to be presented on 05.04.07)

---

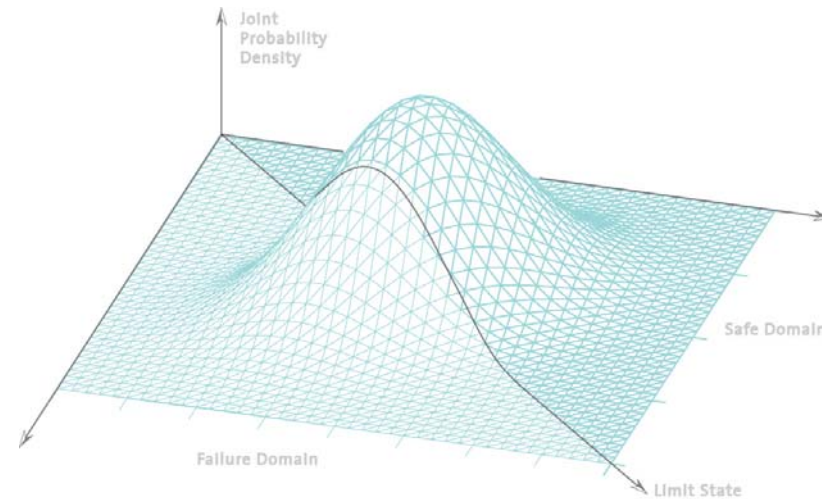
### What is required???

- a. Complete the table
  - b. A geo-electrical test was carried out and indicated  $C_3$  as the thickness of the clay layer. What is the probability that the thickness of the clay layer belongs to  $C_1$ ,  $C_2$ ,  $C_3$  ?
- 
- b. **How can we express this??**

Use Bayes' theorem (script section B.5)



**QUESTIONS ? ? ? ? ? ?**



## Exercises Tutorial 3

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

## Exercise 2.7- (Group Exercise)

### Lots of information----Simplify!!!

- the ground-lowering depends on the thickness of the clay layer,  $h$ .
- Classification of thickness layer: **(Events)**  $C_1: 0 \leq h \leq 20cm$      $C_2: 20cm < h \leq 40cm$      $C_3: 40cm < h$
- prior probabilities: **(Known probabilities)**  $P(C_1) = 0.2$      $P(C_2) = 0.47$      $P(C_3) = 0.33$
- test to update the prior probability on the ground category, the test result may not always be correct.
- probabilities of the correct/false indication of the test:

Category of thickness of clay layer $C_i$	Indication of the category of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84		0.03
$C_2$	0	0.77	
$C_3$		0.02	0.89

$$P(I = C_1 | C_1) = 0.84$$

$$P(I = C_2 | C_2) = 0.77$$

$$P(I = C_2 | C_3) = 0.02$$

### What is required???

- Complete the table
- A geo-electrical test was carried out and indicated  $C_3$  as the thickness of the clay layer. What is the probability that the thickness of the clay layer belongs to  $C_1$ ,  $C_2$ ,  $C_3$  ?

## Exercise 2.7- (Group Exercise)

- the ground-lowering depends on the thickness of the clay layer,  $h$ .
- Classification of thickness layer: (**Events**)  $C_1: 0 \leq h \leq 20cm$      $C_2: 20cm < h \leq 40cm$      $C_3: 40cm < h$
- prior probabilities: (**Known probabilities**)  $P(C_1) = 0.2$      $P(C_2) = 0.47$      $P(C_3) = 0.33$
- test to update the prior probability on the ground category, the test result may not always be correct.
- probabilities of the correct/false indication of the test:

Category of thickness of clay layer $C_i$	Indication of the category of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84		0.03
$C_2$	0	0.77	
$C_3$		0.02	0.89

$$P(I = C_1 | C_1) + P(I = C_2 | C_1) + P(I = C_3 | C_1) = 1$$

$$0.84 + P(I = C_2 | C_1) + 0.03 = 1 \Rightarrow P(I = C_2 | C_1) = 0.13$$

### Exercise 2.7- (Group Exercise)

- the ground-lowering depends on the thickness of the clay layer,  $h$ .
- Classification of thickness layer: **(Events)**  $C_1: 0 \leq h \leq 20cm$      $C_2: 20cm < h \leq 40cm$      $C_3: 40cm < h$
- prior probabilities: **(Known probabilities)**  $P(C_1) = 0.2$      $P(C_2) = 0.47$      $P(C_3) = 0.33$
- test to update the prior probability on the ground category, the test result may not always be correct.
- probabilities of the correct/false indication of the test:

Category of the thickness of the clay layer $C_i$	Indication of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84	<b>0.13</b>	0.03
$C_2$	0	0.77	<b>0.23</b>
$C_3$	<b>0.09</b>	0.02	0.89

$$P(I = C_1 | C_1) + P(I = C_2 | C_1) + P(I = C_3 | C_1) = 1$$

$$0.84 + P(I = C_2 | C_1) + 0.03 = 1 \Rightarrow P(I = C_2 | C_1) = \mathbf{0.13}$$



Exercise 2.7- (Group Exercise to be presented on 05.04.07)

**What is required???**

- a. Complete the table
- b. A geo-electrical test was carried out and indicated  $C_3$  as the thickness of the clay layer. What is the probability that the thickness of the clay layer belongs to  $C_1$ ,  $C_2$ ,  $C_3$  ?

b. Use Bayes' theorem (script section B.5)

$$P(C_1 | I = C_3) = \frac{P(I = C_3 | C_1)P(C_1)}{P(I = C_3 | C_1)P(C_1) + P(I = C_3 | C_2)P(C_2) + P(I = C_3 | C_3)P(C_3)} = 0.015$$

Prior probabilities

$$P(C_1) = 0.2$$

$$P(C_2) = 0.47$$

$$P(C_3) = 0.33$$

Category of the thickness of the clay layer $C_i$	Indication of the thickness of the clay layer		
	$I = C_1$	$I = C_2$	$I = C_3$
$C_1$	0.84	<b>0.13</b>	0.03
$C_2$	0	0.77	<b>0.23</b>
$C_3$	<b>0.09</b>	0.02	0.89

---

 Exercise 2.7- (Group Exercise to be presented on 05.04.07)

**What is required???**

- a. Complete the table
- b. A geo-electrical test was carried out and indicated  $C_3$  as the thickness of the clay layer. What is the probability that the thickness of the clay layer belongs to  $C_1$ ,  $C_2$ ,  $C_3$  ?

- b. Use Bayes' theorem (script section B.5)

$$P(C_1 | I = C_3) = \frac{P(I = C_3 | C_1)P(C_1)}{P(I = C_3 | C_1)P(C_1) + P(I = C_3 | C_2)P(C_2) + P(I = C_3 | C_3)P(C_3)} = 0.015$$

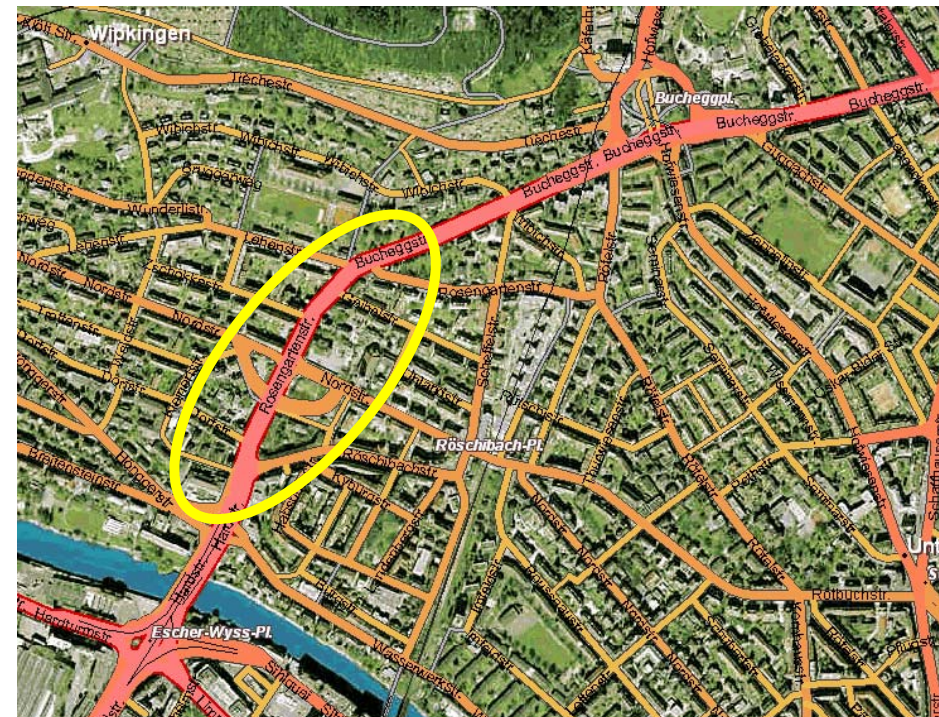
$$P(C_2 | I = C_3) = \frac{P(I = C_3 | C_2)P(C_2)}{P(I = C_3 | C_1)P(C_1) + P(I = C_3 | C_2)P(C_2) + P(I = C_3 | C_3)P(C_3)} = 0.265$$

$$P(C_3 | I = C_3) = \frac{P(I = C_3 | C_3)P(C_3)}{P(I = C_3 | C_1)P(C_1) + P(I = C_3 | C_2)P(C_2) + P(I = C_3 | C_3)P(C_3)} = 0.720$$

### Exercise 3.1 (Descriptive Statistics)

Two sets of data are provided, each of which represents the daily traffic flow in Rosengartenstrasse in Zurich during the month of April 2001  
 Direction 1 corresponds to driving towards Bucheggplatz, while direction 2 corresponds to driving towards Escher Wyss Platz.

Date	Direction 1	Direction 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877



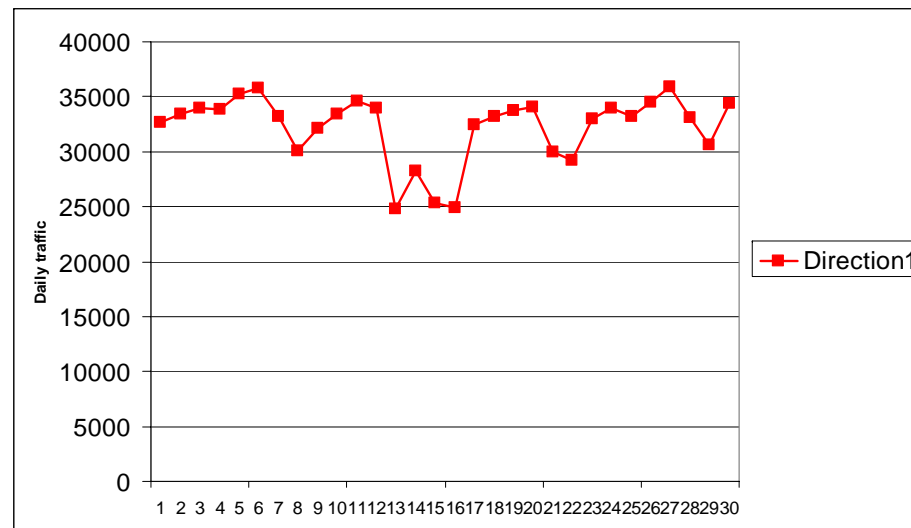
What do we want to know?

Which is the best way to know it? – plot, histogram, statistics etc.

For example,

if you are interested in:

the change in the traffic of direction 1 during the month



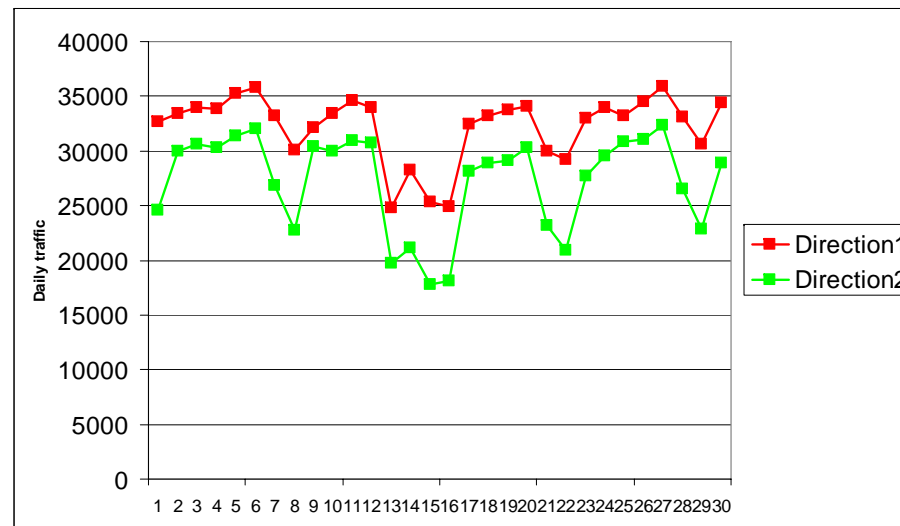
What do we want to know?

Which is the best way to know it? – plot, histogram, statistics etc.

For example,

if you are interested in:

the relation between the traffic of direction 1 and that of direction 2,



What do we want to know?

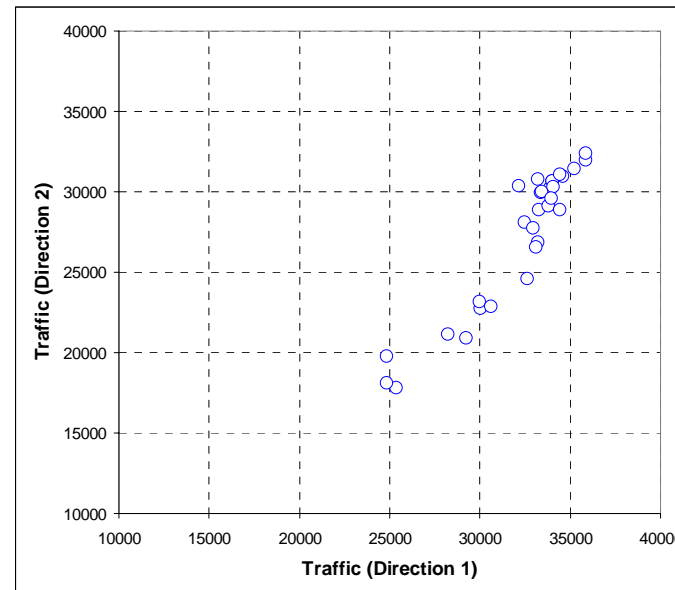
Which is the best way to know it? – plot, histogram, statistics etc.

For example,

if you are interested in:

the relation between the traffic of direction 1 and that of direction 2,

but you are not interested in the time element

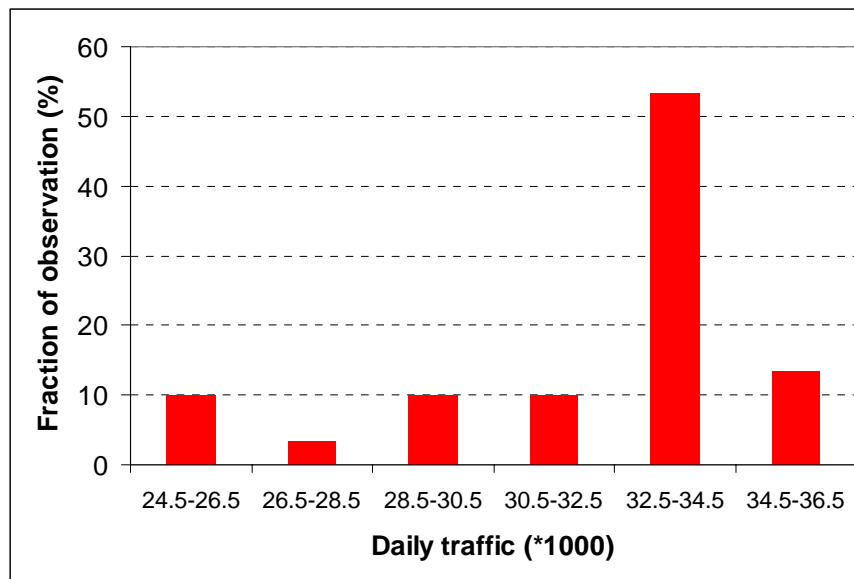


Correlated!

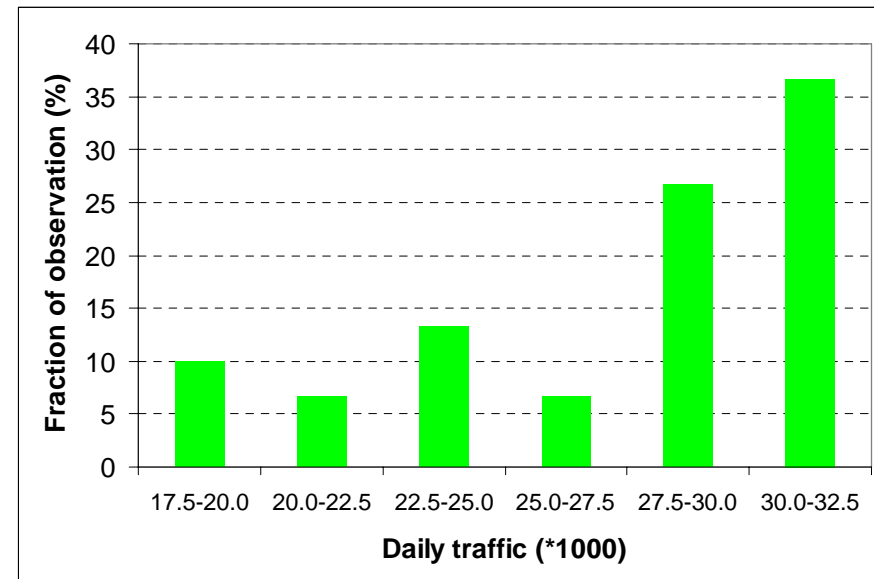
What do we want to know?

Which is the best way to know it? – plot, histogram, statistics etc.

For example,  
if you are interested in:  
traffic volume of each direction



Direction 1



Direction 2

We will look today into...

how to represent and compare the properties of sets of data which you have

- graphically
  - frequency distribution (histogram)
  - cumulative frequency distribution
  
- numerically
  - median
  - quantile
  
- a summary plot
  - Tukey box plot
  
- Correlation between data sets

You can use excel, matlab and/or other programming/statistics software....**BUT**

Make sure **ALWAYS** to insert functions by yourself or check that the functions provided by the used program agree with those of the used script!



## Exercise 3.1

Provide frequency distributions and cumulative frequency distributions of the observed data. What is your first impression of the data? Try to make comparison between the two directions.


Date	Direction 1	Direction 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877

### Steps

1. sort the data
2. select the number of intervals
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

## Step 1 (sort the data)

Date	Direction 1	Direction 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877

sort/order  
  
 in ascending  
 order

### Steps

1. sort the data
2. select the number of intervals
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

Direction 1	Direction 2
24846	17805
24862	18123
25365	19735
28252	20903
29224	21145
29976	22762
30035	22828
30613	23141
32158	24609
32472	26525
32618	26846
32962	27746
33091	28117
33197	28858
33198	28877
33245	29080
33380	29586
33406	29965
33788	29994
33888	30263
33937	30313
34007	30366
34013	30629
34076	30680
34425	30788
34455	30958
34576	31074
35237	31405
35843	31994
35852	32384

**Steps**

1. sort the data
2. **select the number of intervals**
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

## Step 2 (**select the number of intervals**)

No general rule but suggestion - (script Equation (C.8))

$$k = 1 + 3.3 \log_{10} n$$

$k$  is the number of the intervals,  $n$  is the number of the data.

In this case,  $n = 30$

$$k = 1 + 3.3 \log_{10} 30 = 5.87 \approx 6 \text{ intervals}$$

For direction 1,

minimum = 24846

max = 35852

we may select the intervals as follows:

$$[24.5 \ 26.5 \ 28.5 \ 30.5 \ 32.5 \ 34.5 \ 36.5] \ (*1000)$$

**Steps**

1. sort the data
2. select the number of intervals
3. **count the data in each interval**
4. draw the frequency distribution
5. draw the cumulative frequency distribution

**Step 3 (count the data in each interval)**

Direction 1

24846  
24862  
25365  
28252  
29224  
29976  
30035  
30613  
32158  
32472  
32618  
32962  
33091  
33197  
33198  
33245  
33380  
33406  
33788  
33888  
33937  
34007  
34013  
34076  
34425  
34455  
34576  
35237  
35843  
35852

**Count**



Direction 1	Interval (Number of cars *10 <sup>3</sup> )	Interval Midpoint (Number of cars *10 <sup>3</sup> )	Number of observations
	24.5-26.5	25.5	3
	26.5-28.5	27.5	1
	28.5-30.5	29.5	3
	30.5-32.5	31.5	3
	32.5-34.5	33.5	16
	34.5-36.5	35.5	4

**Steps**

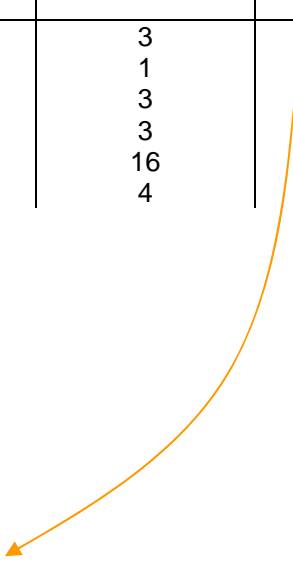
1. sort the data
2. select the number of intervals
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

Step 4 (draw the frequency distribution)

But first some calculations....

	Interval (Number of cars *10 <sup>3</sup> )	Interval Midpoint (Number of cars *10 <sup>3</sup> )	Number of observations	Frequency %
<b>Direction 1</b>	24.5-26.5	25.5	3	10.000
	26.5-28.5	27.5	1	3.333
	28.5-30.5	29.5	3	10.000
	30.5-32.5	31.5	3	10.000
	32.5-34.5	33.5	16	53.333
	34.5-36.5	35.5	4	13.333

$$\begin{aligned} \text{Frequency\%} &= \frac{n_o}{n} 100 \\ &= \frac{3}{30} 100 = 10 \end{aligned}$$



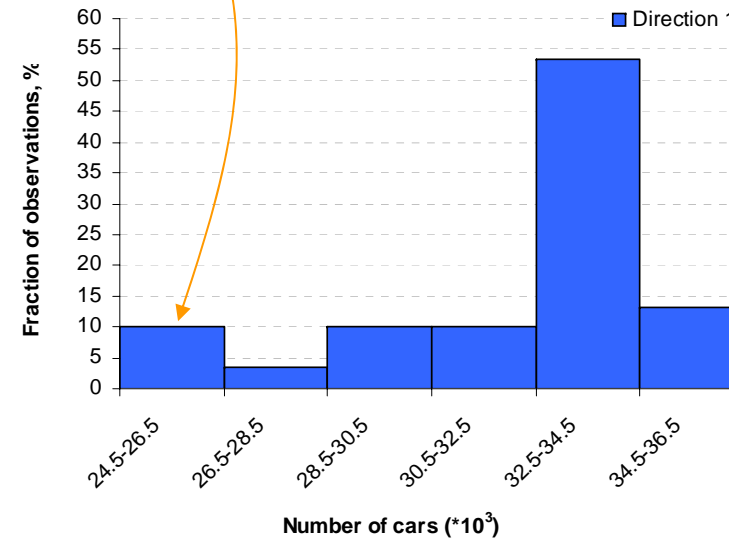
Step 4 (draw the frequency distribution)

Steps

1. sort the data
2. select the number of intervals
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

Direction 1	Interval (Number of cars *10 <sup>3</sup> )	Interval Midpoint (Number of cars *10 <sup>3</sup> )	Number of observations	Frequency %
	24.5-26.5	25.5	3	10.000
	26.5-28.5	27.5	1	3.333
	28.5-30.5	29.5	3	10.000
	30.5-32.5	31.5	3	10.000
	32.5-34.5	33.5	16	53.333
	34.5-36.5	35.5	4	13.333

Draw 



**Steps**

1. sort the data
2. select the number of intervals
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

**Step 5 (draw the cumulative frequency distribution)**

Direction 1	Interval (Number of cars *10 <sup>3</sup> )	Interval Midpoint (Number of cars *10 <sup>3</sup> )	Number of observations	Frequency %	Cumulative frequency
		24.5-26.5	25.5	3	10.000
	26.5-28.5	27.5	1	3.333	0.133
	28.5-30.5	29.5	3	10.000	0.233
	30.5-32.5	31.5	3	10.000	0.333
	32.5-34.5	33.5	16	53.333	0.867
	34.5-36.5	35.5	4	13.333	1.000

Cumulate →

/100

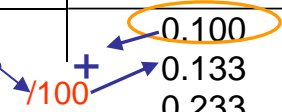
**Steps**

1. sort the data
2. select the number of intervals
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

**Step 5 (draw the cumulative frequency distribution)**

Cumulate →

Direction 1	Interval (Number of cars *10 <sup>3</sup> )	Interval Midpoint (Number of cars *10 <sup>3</sup> )	Number of observations	Frequency %	Cumulative frequency
	24.5-26.5	25.5	3	10.000	0.100
	26.5-28.5	27.5	1	3.333	0.133
	28.5-30.5	29.5	3	10.000	0.233
	30.5-32.5	31.5	3	10.000	0.333
	32.5-34.5	33.5	16	53.333	0.867
	34.5-36.5	35.5	4	13.333	1.000





**Steps**

1. sort the data
2. select the number of intervals
3. count the data in each interval
4. draw the frequency distribution
5. draw the cumulative frequency distribution

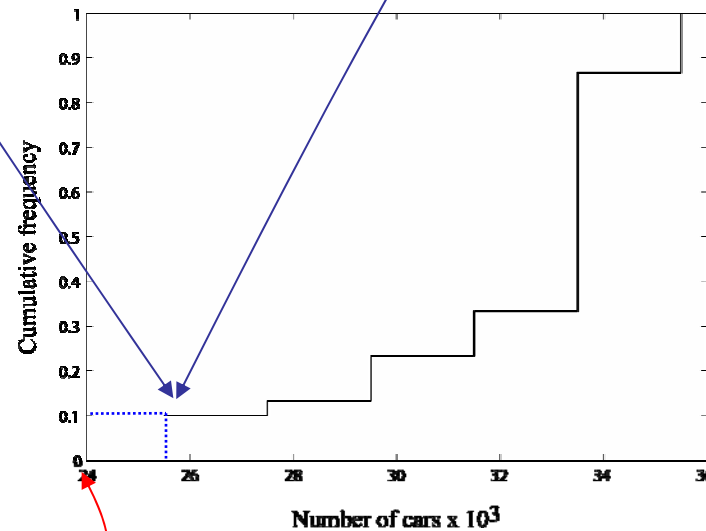
**Step 5 (draw the cumulative frequency distribution)**

Cumulate →

Direction 1	Interval (Number of cars *10 <sup>3</sup> )	Interval Midpoint (Number of cars *10 <sup>3</sup> )	Number of observations	Frequency %	Cumulative frequency
		24.5-26.5	25.5	3	10.000
	26.5-28.5	27.5	1	3.333	0.133
	28.5-30.5	29.5	3	10.000	0.233
	30.5-32.5	31.5	3	10.000	0.333
	32.5-34.5	33.5	16	53.333	0.867
	34.5-36.5	35.5	4	13.333	1.000

**Draw**

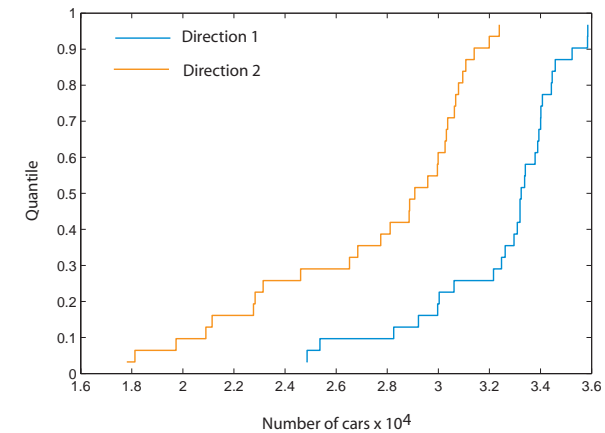
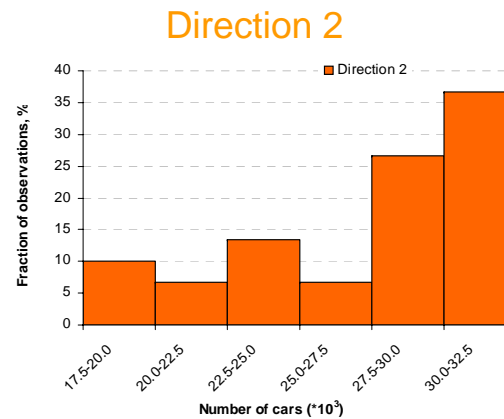
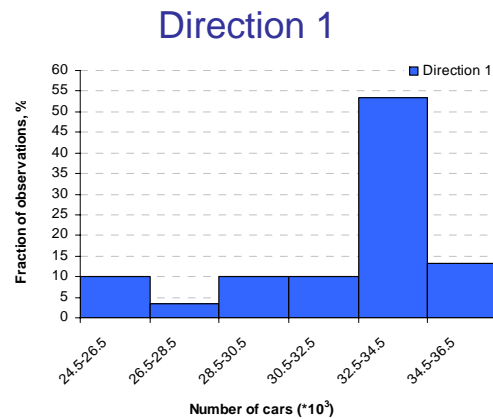
y=0



### Answer 3.1

Do the same for direction 2.

What can we know from these plots?



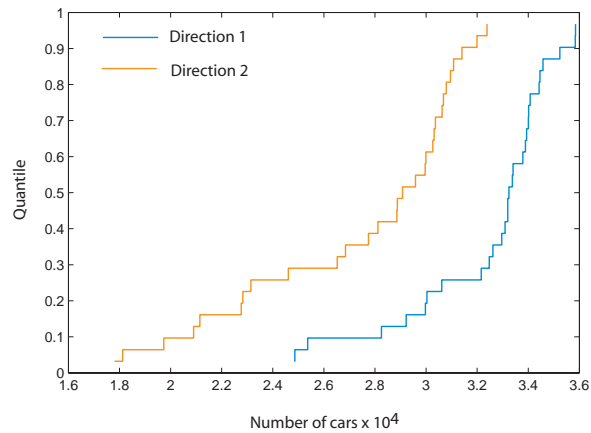
- Traffic is much heavier in direction 1 than direction 2.
- There are large variations in traffic for both directions.
- etc....

These figures give nice overviews of the data!

## Answer 3.1

- a. When we have in hand all observations:  
prefer to plot the cumulative distribution plot using the quantiles of the data!

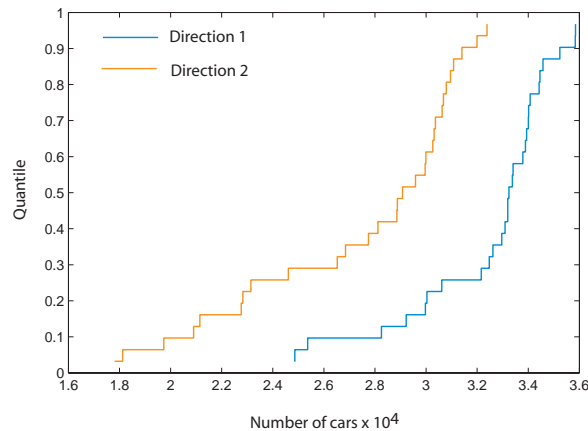
a.



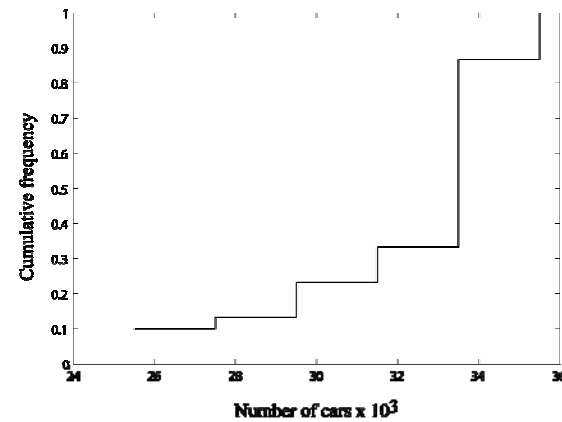
### Answer 3.1

- a. When we have in hand all observations:  
prefer to plot the cumulative distribution plot using the quantiles of the data!
- b. If we have in hand only the intervals observed and the frequency of observations within each interval  
a. is not possible so...plot the cumulative frequency!

a.



b.



---

## Quantiles

A quantile is related to a given percentage  $\alpha$ , for which  $\alpha\%$  of all observations in the data set have smaller values.

e.g. the 0.65 quantile of a given data set of observations corresponds to the observation for which 65% of all observations in the data set have smaller values

## Quantiles

A quantile is related to a given percentage  $\alpha$ , for which  $\alpha\%$  of all observations in the data set have smaller values.

e.g. the 0.74 quantile of a given data set of observations corresponds to the observation for which 74% of all observations in the data set have smaller values

Direction 1	Direction 2
24846	17805
24862	18123
25365	19735
28252	20903
29224	21145
29976	22762
30035	22828
30613	23141
32158	24609
32472	26525
32618	26846
32962	27746
33091	28117
33197	28858
33198	28877
33245	29080
33380	29586
33406	29965
33788	29994
33888	30263
33937	30313
34007	30366
34013	30629
34076	30680
34425	30788
34455	30958
34576	31074
35237	31405
35843	31994
35852	32384

Q=0.74

## Quantiles

A quantile is related to a given percentage  $\alpha$ , for which  $\alpha\%$  of all observations in the data set have smaller values.

e.g. the 0.74 quantile of a given data set of observations corresponds to the observation for which 74% of all observations in the data set have smaller values

Direction 1	Direction 2
24846	17805
24862	18123
25365	19735
28252	20903
29224	21145
29976	22762
30035	22828
30613	23141
32158	24609
32472	26525
32618	26846
32962	27746
33091	28117
33197	28858
33198	28877
33245	29080
33380	29586
33406	29965
33788	29994
33888	30263
33937	30313
34007	30366
34013	30629
34076	30680
34425	30788
34455	30958
34576	31074
35237	31405
35843	31994
35852	32384

**Q=0.74**

**74% of the observations  
Have a smaller value!**

## Quantiles

A quantile is related to a given percentage  $\alpha$ , for which  $\alpha\%$  of all observations in the data set have smaller values.

e.g. the 0.65 quantile of a given data set of observations corresponds to the observation for which 65% of all observations in the data set have smaller values

How to calculate it????

$$Q_i = \frac{i}{n+1}, \quad n : \text{total number of observations}, \quad i=1,2,\dots,n$$



## Exercise 3.2

Use the Tukey box plot to provide a summary of the main features of the distribution of each data set. Plot the Tukey box plots on the same graph so that you are able to compare these features. Do you observe any symmetry in the data sets?

### Steps

1. calculate the median
2. calculate the 75%- and 25%- quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Steps

1. calculate the median
2. calculate the 75%- and 25%- quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Step 1 (calculate the median)

Just take the central value (50%-quantile).

but.....

if the number of data is even, this is not possible!

In that case, take the two values around the center, then take the average.

Direction 1  
 24846  
 24862  
 25365  
 28252  
 29224  
 29976  
 30035  
 30613  
 32158  
 32472  
 32618  
 32962  
 33091  
 33197  
 33198  
 33245  
 33380  
 33406  
 33788  
 33888  
 33937  
 34007  
 34013  
 34076  
 34425  
 34455  
 34576  
 35237  
 35843  
 35852

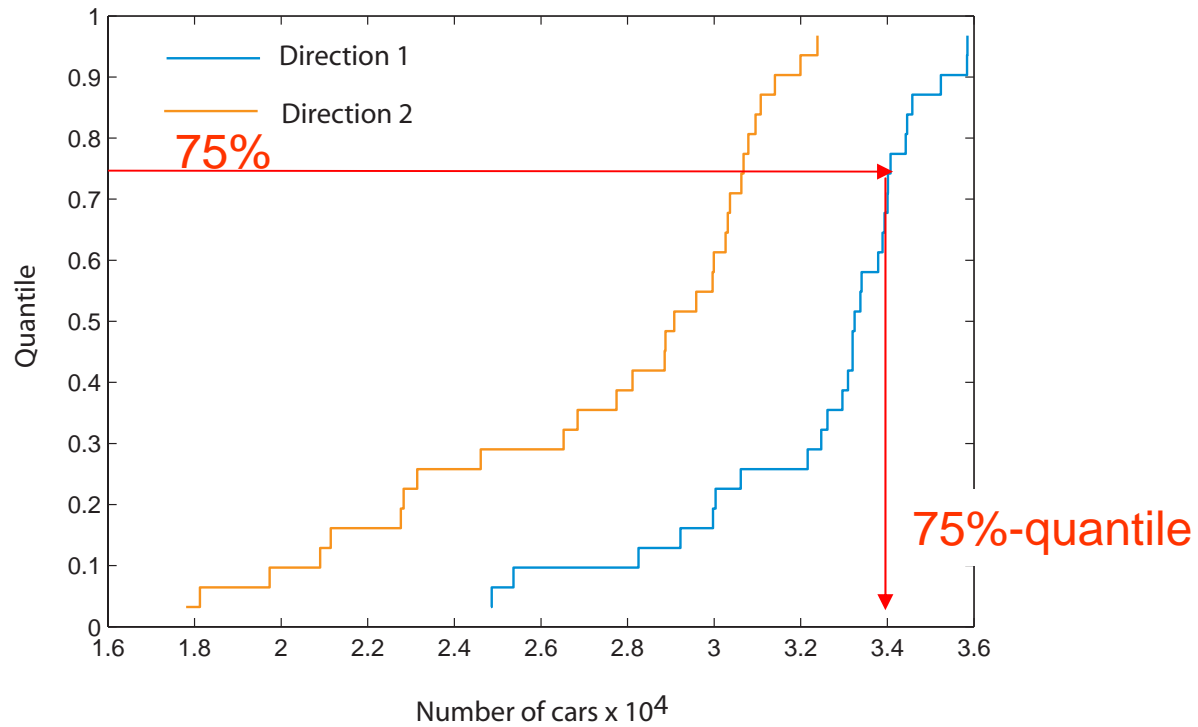
$$\text{Median is } \frac{33198 + 33245}{2} = 33221.5$$

Steps

1. calculate the median
2. calculate the 75%- and 25%- quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Step 2 (calculate the quantiles)

Roughly speaking,



Step 2 (calculate the quantiles)

More strictly speaking,

$$Q_i = \frac{i}{n+1}, \quad n : \text{total number of observations}$$

Direction 1	i	i/31
24846	1	0.03
24862	2	0.06
25365	3	0.10
28252	4	0.13
29224	5	0.16
29976	6	0.19
30035	7	0.23
30613	8	0.26
32158	9	0.29
32472	10	0.32
32618	11	0.35
32962	12	0.39
33091	13	0.42
33197	14	0.45
33198	15	0.48
33245	16	0.52
33380	17	0.55
33406	18	0.58
33788	19	0.61
33888	20	0.65
33937	21	0.68
34007	22	0.71
34013	23	0.74
34076	24	0.77
34425	25	0.81
34455	26	0.84
34576	27	0.87
35237	28	0.90
35843	29	0.94
35852	30	0.97

← 75%

Step 2 (calculate the quantiles)

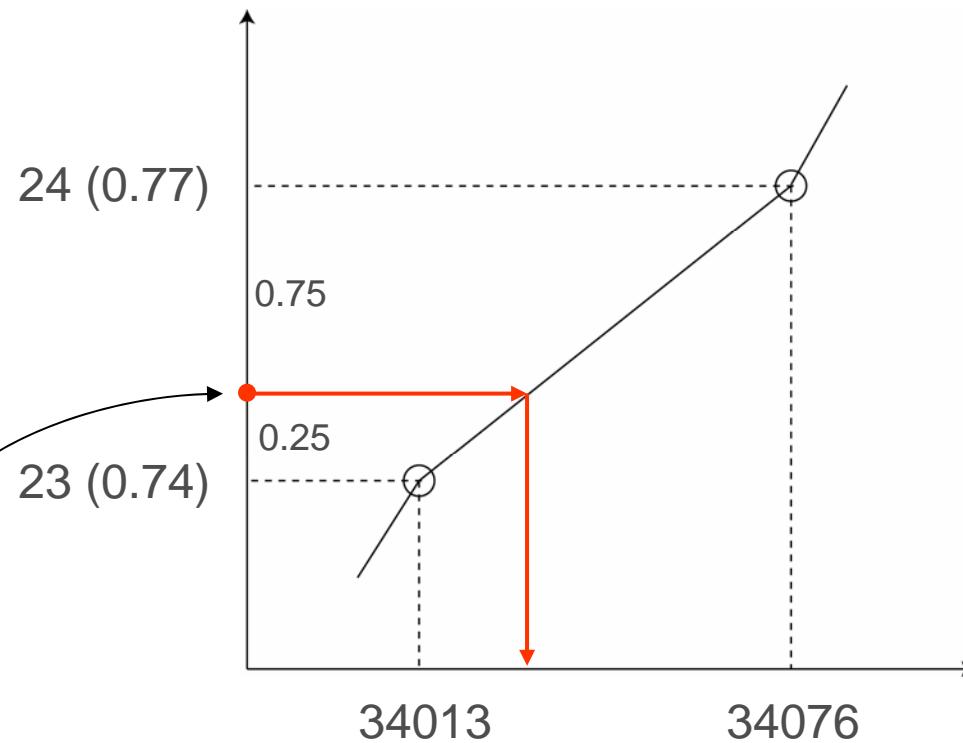
Interpolation

33700	19	0.61
33888	20	0.65
33937	21	0.68
34007	22	0.71
34013	23	0.74
34076	24	0.77
34425	25	0.81
...	...	...

$$v = nQ_v + Q_v$$

$$v = 30 \cdot 0.75 + 0.75 = 23.25$$

$$x_v^o = (1 - p)x_{23}^o + px_{24}^o = (1 - 0.25) \cdot 34013 + 0.25 \cdot 34076 = 34028.75 \approx 34029 \text{ cars}$$



Steps

1. calculate the median
2. calculate the 75% and 25% quantile.
3. **calculate the adjacent values.**
4. check for outside values
5. draw the Tukey box plot

Step 3 (calculate the adjacent values)

$$\begin{array}{l}
 Q_{0.75} = 34029 \\
 Q_{0.25} = 30469
 \end{array}
 \left. \vphantom{\begin{array}{l} Q_{0.75} \\ Q_{0.25} \end{array}} \right\} \begin{array}{l} \text{Interquartile range} \\ \downarrow \\ r \equiv Q_{0.75} - Q_{0.25} = 34029 - 30469 = 3560 \end{array}$$

Upper adjacent value: largest observation  $\leq$  (75% quantile) + 1.5r

In this case, largest value less than  $34029 + 1.5 \times 3560 = 39363$

- 33198
- 33245
- 33380
- 33406
- 33788
- 33888
- 33937
- 34007
- 34013
- 34076
- 34425
- 34455
- 34576
- 35237
- 35843
- 35852**

If the largest observation is less than that value,  
take the largest observation as the upper adjacent value.

Upper adjacent value = 35852

Steps

1. calculate the median
2. calculate the 75% and 25% quantile.
3. **calculate the adjacent values.**
4. check for outside values
5. draw the Tukey box plot

Step 3 (calculate the adjacent values)

$$\left. \begin{array}{l} Q_{0.75} = 34029 \\ Q_{0.25} = 30469 \end{array} \right\} r \equiv Q_{0.75} - Q_{0.25} = 34029 - 30469 = 3560$$

Lower adjacent value: smallest observation  $\geq$  (25% quantile) - 1.5r

In this case, lowest value larger than  $30469 - 1.5 \times 3560 = 25129$

Direction 1	24846
	24862
25129	25365
	28252
	29224
	29976
	30035
	30613
	32158
	32472
	32618
	32962
	33091
	33197
	33198

If the lowest observation is more than that value,  
take the lowest observation as the lower adjacent value.

lower adjacent value = 25365

Direction 1  
 24846  
 24862  
 25365  
 28252  
 29224  
 29976  
 30035  
 30613  
 32158  
 32472  
 32618  
 32962  
 33091  
 33197  
 33198  
 33245  
 33380  
 33406  
 33788  
 33888  
 33937  
 34007  
 34013  
 34076  
 34425  
 34455  
 34576  
 35237  
 35843  
 35852

**Step 4**

(check for outside values)

**Outside values:**

Outside the upper and lower adjacent values

24846

24862

**summary**

Upper adjacent value: 35852

75% quantile : 34029

Median : 33222

25% quantile : 30469

Lower adjacent value: 25365

**Steps**

1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot



Step 4  
(draw the Tukey box plot)

Steps

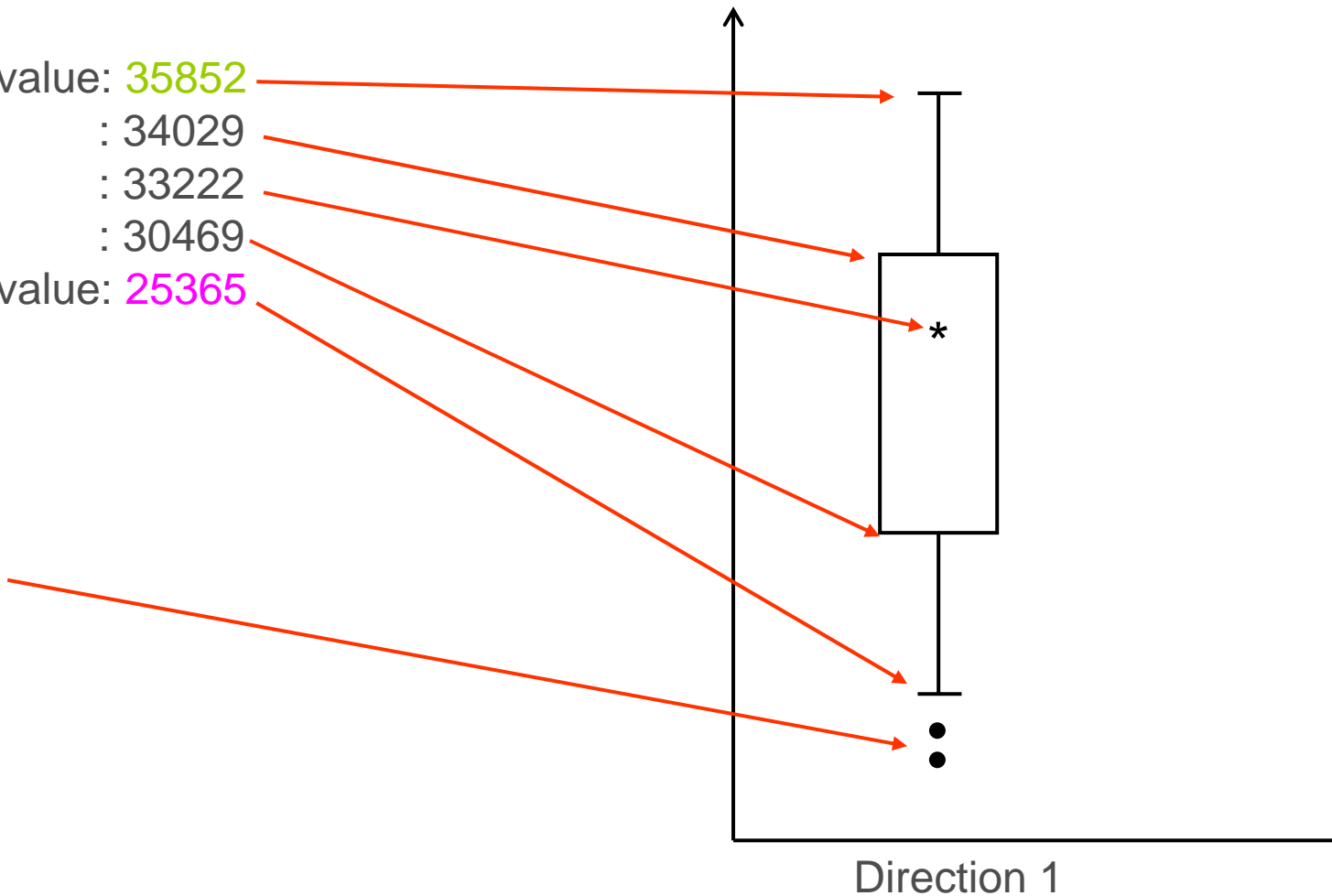
1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

summary

Upper adjacent value: 35852  
 75% quantile : 34029  
 Median : 33222  
 25% quantile : 30469  
 Lower adjacent value: 25365

Outside values:

24846  
 24862



## Answer 3.2

Use the Tukey box plot to provide a summary of the main features of the distribution of each data set.

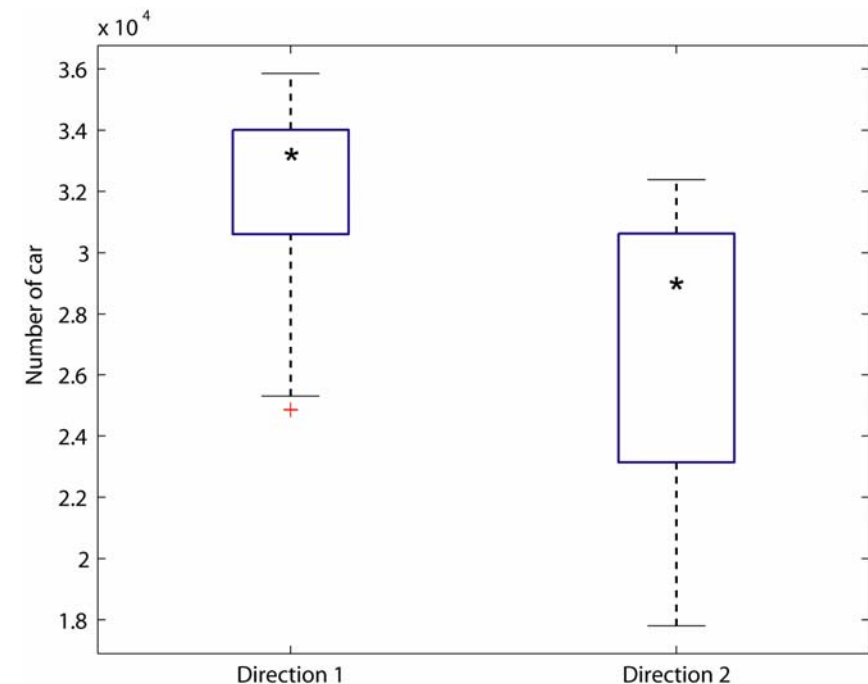
- median
- Adjacent values
- Upper and lower quartiles
- Outside values

Plot the Tukey box plots on the same graph so that you are able to compare these features.

- All features are higher in direction 1. Larger traffic volume in direction 1.
- Larger interquartile range in direction 2: observations more widely dispersed around the median.

Do you observe any symmetry in the data sets?

- No symmetry is observed.
- The median is closer to the upper adjacent value in both directions.
- Left skewed.



### Exercise 3.5

The data sets in Table 3.5.1 show the number of newcomers to the university and the number of total students at the university.

Estimate the correlation of these numbers using the following calculation sheet.

	Univ. A	Univ. B	Univ. C	Univ. D	Univ. E	Univ. F
Newcomer	3970	732	499	1300	3463	2643
Total students	24273	5883	2847	5358	23442	17076

**Table 3.5.1** Number of newcomers to the university and the number of total students at the university.

### Exercise 3.5

Estimate the correlation of these numbers using the following calculation sheet.

	Univ. A	Univ. B	Univ. C	Univ. D	Univ. E	Univ. F
Newcomer	3970	732	499	1300	3463	2643
Total students	24273	5883	2847	5358	23442	17076

**Table 3.5.1** Number of newcomers to the university and the number of total students at the university.

### What is known?

Newcomers:  $X$   
 total students:  $Y$   
 Number of newcomers:  $x_i, i=1, \dots, 6$   
 Number of total students:  $y_i, i=1, \dots, 6$   
 Number of observations/university:  $n=6$

### Exercise 3.5

Estimate the correlation of these numbers using the following calculation sheet.

	Univ. A	Univ. B	Univ. C	Univ. D	Univ. E	Univ. F
Newcomer	3970	732	499	1300	3463	2643
Total students	24273	5883	2847	5358	23442	17076

**Table 3.5.1** Number of newcomers to the university and the number of total students at the university.

#### What is known?

Newcomers:  $X$   
 total students:  $Y$   
 Number of newcomers:  $x_i, i=1, \dots, 6$   
 Number of total students:  $y_i, i=1, \dots, 6$   
 Number of observations/university:  $n=6$

#### What is required?

Correlation:  $r_{XY} = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}$

#### Need to:

Calculate the sample mean values:  $\bar{x}$   $\bar{y}$

Calculate sample standard deviations:  $s_X$   $s_Y$

### Exercise 3.5

Estimate the correlation of these numbers using the following calculation sheet.

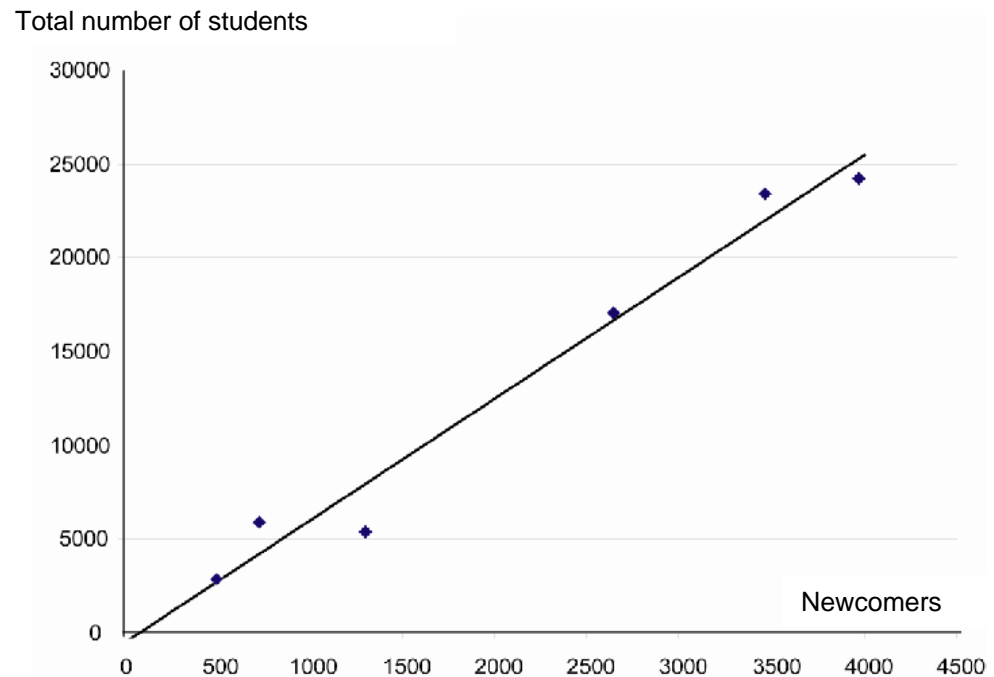
	Univ. A	Univ. B	Univ. C	Univ. D	Univ. E	Univ. F
Newcomer	3970	732	499	1300	3463	2643
Total students	24273	5883	2847	5358	23442	17076

Table 3.5.1 Number of newcomers to the university and the number of total students at the university.

**From a first view are they correlated  
???????**

**Give a rough estimation for the  
correlation coefficient!!!!**

$$-1 \leq r_{XY} \leq 1$$



### Solution 3.5

	Univ. A	Univ. B	Univ. C	Univ. D	Univ. E	Univ. F
Newcomer	3970	732	499	1300	3463	2643
Total students	24273	5883	2847	5358	23442	17076

**Table 3.5.1** Number of newcomers to the university and the number of total students at the university.

#### What is known?

Newcomers:  $X$   
 total students:  $Y$   
 Number of newcomers:  $x_i, i=1, \dots, 6$   
 Number of total students:  $y_i, i=1, \dots, 6$   
 Number of observations/university:  $n=6$

#### What is required?

Correlation: 
$$r_{XY} = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}$$

#### Need to:

Calculate the sample mean values: 
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Calculate sample standard deviations: 
$$s_X = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \quad s_Y = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})$$

### Solution 3.5

	Univ. A	Univ. B	Univ. C	Univ. D	Univ. E	Univ. F
Newcomer	3970	732	499	1300	3463	2643
Total students	24273	5883	2847	5358	23442	17076

**Table 3.5.1** Number of newcomers to the university and the number of total students at the university.

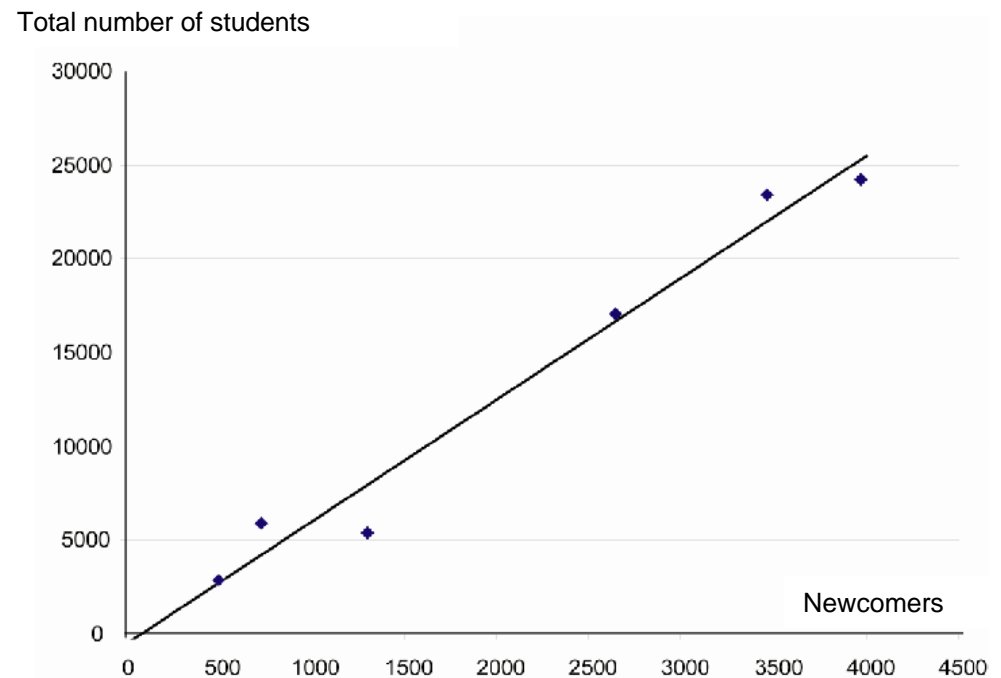
	$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y - \bar{y})^2$	$(x_i - \bar{x})(y - \bar{y})$
A	3970	24273	1868	11126	3493161	123787876	20793574
B	732	5883	-1369	-7264	1874161	52765696	9944942
C	499	2847	-1602	-10300	2566404	106090000	16501516
D	1300	5358	-801	-7789	641601	60668521	6239887
E	3463	23442	1362	10295	1855044	105987025	14020755
F	2643	17076	542	3929	293764	15437041	2129134
$\Sigma$	12607	78879	-	-	10724135	464736159	69629807
$\Sigma/n$	2101	13147	-	-	1787356	77456026.5	11604968
$\sqrt{\Sigma/n}$	-	-	-	-	1337	8801	-



### Solution 3.5

$$r_{XY} = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y} = \frac{11604968}{1337 \cdot 8801} = 0.99$$

as expected high correlation coefficient - positive



### Exercise 3.4 (Group exercise- to be presented on 12.04.07)

Resistivity measurements help to predict the possible corrosion of bridge structures. During a general bridge inspection the data shown in Table 3.2 were obtained from resistivity measurements along the two bridge lanes (direction 1 and 2):

- a. Draw two box plots for the data provided in Table 3.4.1 (direction 1 and direction 2). Show the main features of the box plots and write their values next to the corresponding points on the diagrams. Plot also the outside values, if any.
- b. Tukey box plot is a helpful tool for assessing the symmetry of data sets. Discuss the symmetry/skewness of the resistivity data for both lanes.
- c. Choose a suitable number of intervals and plot the histogram for the resistivity data of direction 1.

### Exercise 3.4 (Group exercise- to be presented on 12.04.07)

- a. Draw two box plots for the data provided in Table 3.4.1 (direction 1 and direction 2). Show the main features of the box plots and write their values next to the corresponding points on the diagrams. Plot also the outside values, if any.
- b. Tukey box plot is a helpful tool for assessing the symmetry of data sets. Discuss the symmetry/skewness of the resistivity data for both lanes.
- c. Choose a suitable number of intervals and plot the histogram for the resistivity data of direction 1.

According to  
exercise 3.2!!!

According to  
exercise 3.1!!!

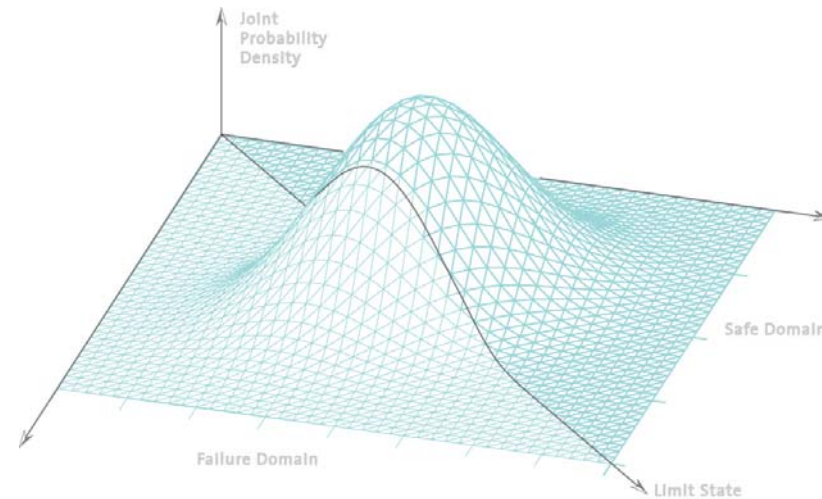
### What should be in the presentation of the solution?

a, b and c!!

An example of calculation where applicable e.g. features of the Tukey box plot etc....

Try to work with a simple calculator, diagrams can be on a transparency made by hand 😊

You can try for yourself to solve in e.g. excel or matlab or other.



## Exercises Tutorial 4

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

- General
  - Correlation plots:  
Plot the UNORDERED observations
  - Quantile estimation:  
Order the available data, calculate then the corresponding quantiles

What do we want to know?

Which is the best way to know it? – plot, histogram, statistics etc.

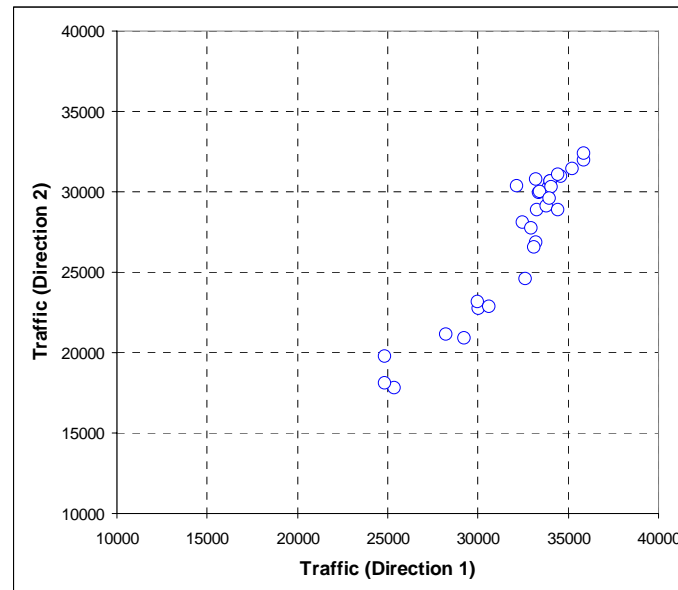
For example,

if you are interested in:

the relation between the traffic of direction 1 and that of direction 2,

but you are not interested in the time element

The graph is correct! Check the unordered pairs of the data!



Correlated!

## Quantiles

A quantile is related to a given percentage  $\alpha$ , for which  $\alpha\%$  of all observations in the data set have smaller values.

e.g. the 0.74 quantile of a given data set of observations corresponds to the observation for which 74% of all observations in the data set have smaller values

Direction 1	Direction 2
24846	17805
24862	18123
25365	19735
28252	20903
29224	21145
29976	22762
30035	22828
30613	23141
32158	24609
32472	26525
32618	26846
32962	27746
33091	28117
33197	28858
33198	28877
33245	29080
33380	29586
33406	29965
33788	29994
33888	30263
33937	30313
34007	30366
34013	30629
34076	30680
34425	30788
34455	30958
34576	31074
35237	31405
35843	31994
35852	32384

**Q=0.74**

**74% of the observations  
Have a smaller value!**

**Correct slide in last week's  
ppt - the unordered data  
were shown.**

### Exercise 3.4 (Group Exercise)

Resistivity measurements help to predict the possible corrosion of bridge structures. During a general bridge inspection the data shown in Table 3.4.1 were obtained from resistivity measurements along the two bridge lanes (direction 1 and 2):

- a. Draw two box plots for the data provided in Table 3.4.1 (direction 1 and direction 2). Show the main features of the box plots and write their values next to the corresponding points on the diagrams. Plot also the outside values, if any.
- b. Tukey box plot is a helpful tool for assessing the symmetry of data sets. Discuss the symmetry/skewness of the resistivity data for both lanes.
- c. Choose a suitable number of intervals and plot the histogram for the resistivity data of direction 1.



## Exercise 3.4 (Group Exercise)

Resistivity measurements help to predict the possible corrosion of bridge structures. During a general bridge inspection the data shown in Table 3.4.1 were obtained from resistivity measurements along the two bridge lanes (direction 1 and 2):

- a. Draw two box plots for the data provided in Table 3.4.1 (direction 1 and direction 2). Show the main features of the box plots and write their values next to the corresponding points on the diagrams. Plot also the outside values, if any.
- b. Tukey box plot is a helpful tool for assessing the symmetry of data sets. Discuss the symmetry/skewness of the resistivity data for both lanes.
- c. Choose a suitable number of intervals and plot the histogram for the resistivity data of direction 1.

### Steps

1. calculate the median
2. calculate the 75%- and 25%- quantile
3. calculate the adjacent values
4. check for outside values
5. draw the Tukey box plot

---

Step 1 (**calculate the median**)

50%-quantile

$$v = nQ_v + Q_v$$

Median is the value at location:

Steps

1. **calculate the median**
2. calculate the 75%- and 25%- quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

---

Step 2 (calculate the 75%- and 25%- quantile)

$$v = nQ_v + Q_v$$

Upper quartile (75% quantile):

Lower quartile (25% quantile):

Steps

1. calculate the median
2. calculate the 75%- and 25%- quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

---

### Step 3 (calculate the adjacent values)

#### Steps

1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Upper adjacent value: largest observation  $\leq (75\% \text{ quantile}) + 1.5r$

In this case, largest value less than

If the largest observation is less than that value,  
take the largest observation as the upper adjacent value.

Upper adjacent value =

---

### Step 3 (calculate the adjacent values)

#### Steps

1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Lower adjacent value: smallest observation  $\geq (25\% \text{ quantile}) - 1.5r$

In this case, lowest value larger than

If the lowest observation is more than that value,  
take the lowest observation as the lower adjacent value.

lower adjacent value :

Try the same steps for Direction 2!

Steps

1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

---

3.4.c

Steps

1. Define number of intervals
2. Count no. of observations within each interval
3. Plot histogram.

### Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.
- What do you observe in regard to the traffic flows in directions 1 and 2?
- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

Date	Direction 1	Direction 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877



### Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.
- What do you observe in regard to the traffic flows in directions 1 and 2?

Direction 2	Direction 1
17805	24846
18123	24862
19735	25365
20903	28252
21145	29224
22762	29976
22828	30035
23141	30613
24609	32158
26525	32472
26846	32618
27746	32962
28117	33091
28858	33197
28877	33198
29080	33245
29586	33380
29965	33406
29994	33788
30263	33888
30313	33937
30366	34007
30629	34013
30680	34076
30788	34425
30958	34455
31074	34576
31405	35237
31994	35843
32384	35852

### Steps

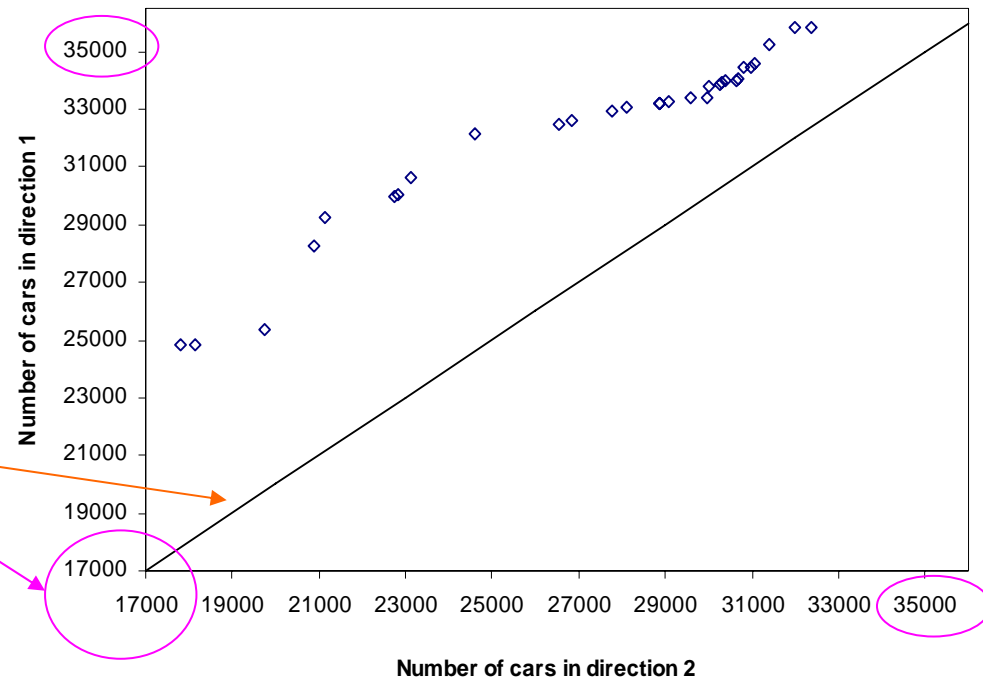
1. sort the data (if not sorted)
2. If  $n_x = n_y$  plot the data in an x-y system using the same scale and origin for x and y
3. Draw the line  $x=y$
4. Compare the two data sets

### Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.

#### Steps

1. sort the data (if not sorted)
2. If  $n_x = n_y$  plot the data in an x-y system using the same scale and origin for x and y
3. Draw the line  $x=y$  (symmetry line)
4. Compare the two data sets

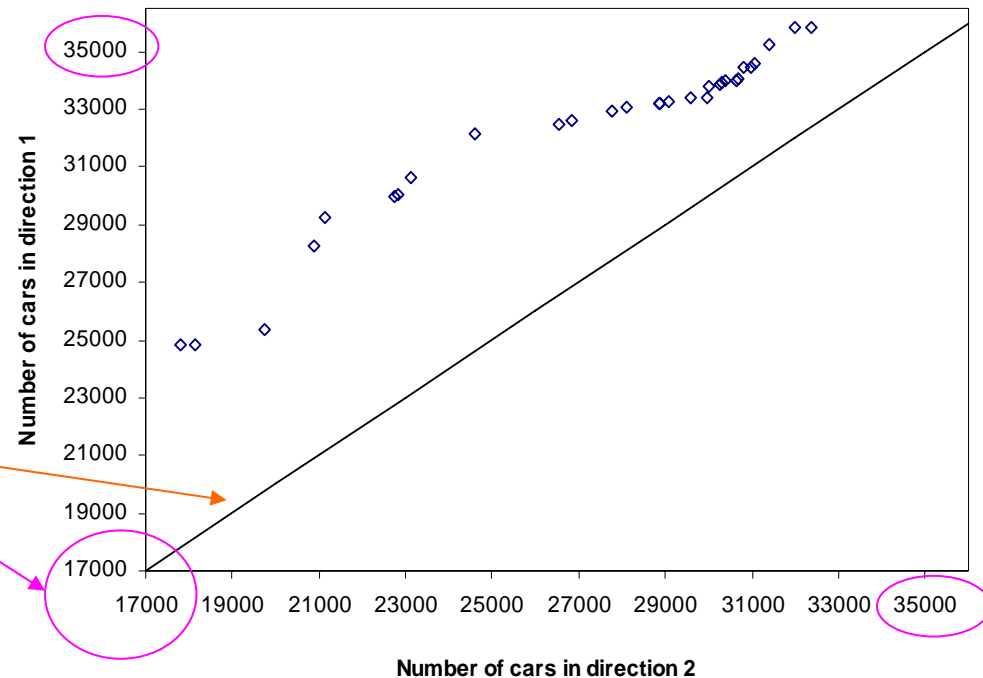


### Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.

#### Steps

1. sort the data (if not sorted)
2. If  $n_x = n_y$  plot the data in an x-y system using the same scale and origin for x and y
3. Draw the line  $x=y$  (symmetry line)
4. Compare the two data sets



The data lie far from the symmetry line

Concentrated on the side of direction 1- higher traffic flow in direction 1

## Exercise 3.3

- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

Date	Direction 1	Direction 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877

### Steps

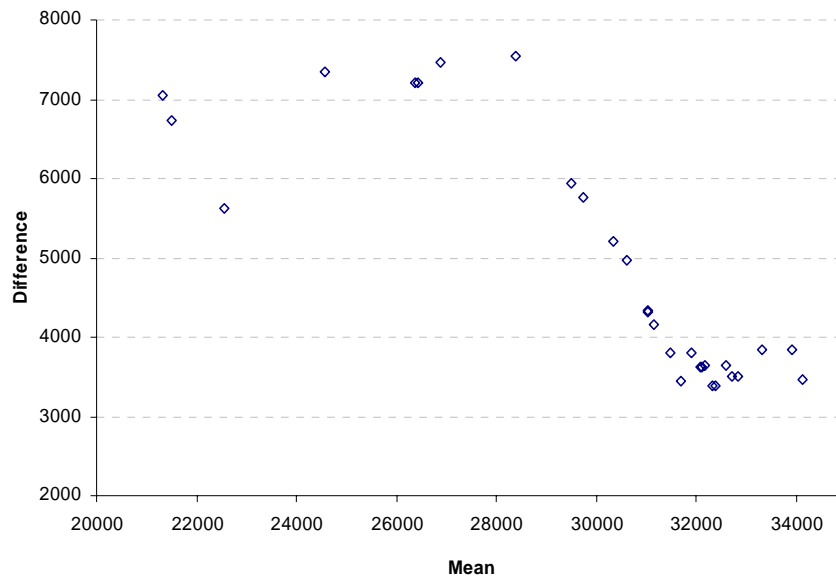
1. sort the data (if not sorted)
2. Calculate  $y_i - x_i$  and plot it on the y-axis
3. Calculate  $(y_i + x_i)/2$  and plot it on the x-axis
4. Discuss...

### Exercise 3.3

- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

#### Steps

1. sort the data (if not sorted)
2. Calculate  $y_i - x_i$  and plot it on the y-axis
3. Calculate  $(y_i + x_i)/2$  and plot it on the x-axis



$x_i$	$y_i$	$y_i - x_i$	$(y_i + x_i)/2$
17805	24846	7041	21325.5
18123	24862	6739	21492.5
19735	25365	5630	22550.0
20903	28252	7349	24577.5
21145	29224	8079	25184.5
22762	29976	7214	26369.0
22828	30035	7207	26431.5
23141	30613	7472	26877.0
24609	32158	7549	28383.5
26525	32472	5947	29498.5
26846	32618	5772	29732.0
27746	32962	5216	30354.0
28117	33091	4974	30604.0
28858	33197	4339	31027.5
28877	33198	4321	31037.5
29080	33245	4165	31162.5
29586	33380	3794	31483.0
29965	33406	3441	31685.5
29994	33788	3794	31891.0
30263	33888	3625	32075.5
30313	33937	3624	32125.0
30366	34007	3641	32186.5
30629	34013	3384	32321.0
30680	34076	3396	32378.0
30788	34425	3637	32606.5
30958	34455	3497	32706.5
31074	34576	3502	32825.0
31405	35237	3832	33321.0
31994	35843	3849	33918.5
32384	35852	3468	34118.0

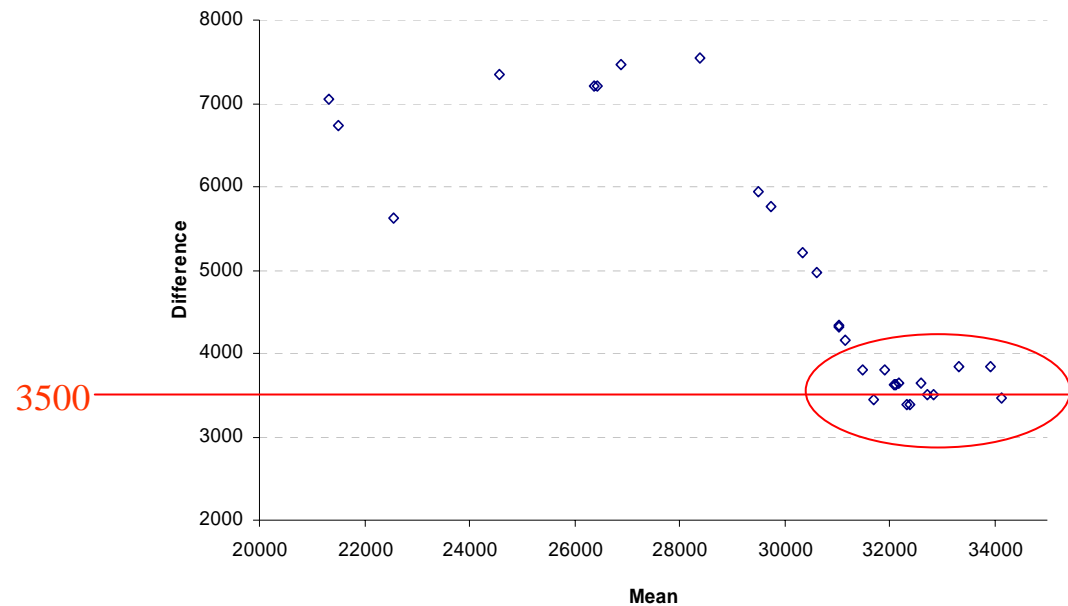
### Exercise 3.3

- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

#### Steps

#### 4. Discuss

for a large part of the data sets the traffic flow in direction 1 is about 3500 cars per day higher than in direction 2



### Exercise 4.1

The monthly expense [CHF] for water consumption including sewage fee for a 2-persons household may be considered as a random variable with the following density function:

$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases} \xrightarrow{\text{Change to}} f_X(x) = \begin{cases} c \cdot x \cdot \left(15 - \frac{x}{4}\right) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

- Which value of  $c$  should be chosen?
- Describe the cumulative distribution function  $F_X(x)$  of the random variable  $X$ .
- Which of the following four values, 30.00 CHF, 40.00 CHF, 50.00 CHF and 60.00 CHF does not exceed the 90%-quantile of the monthly expense?
- How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?

Solution 4.1 a. Which value of  $c$  should be chosen?

Probability density function

$$f_X(x) \geq 0 \quad \longleftarrow \text{Non-negative}$$

$$\int_{\Omega} f_X(x) dx = 1 \quad \longleftarrow \text{Area} = 1$$

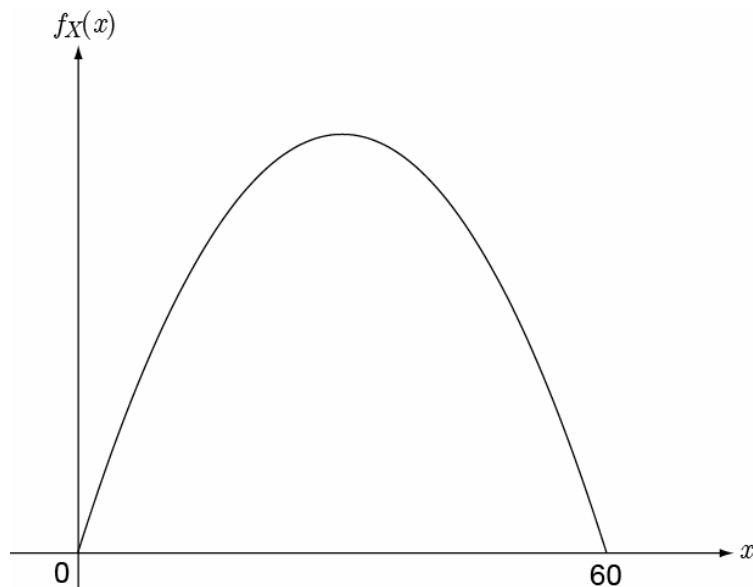


Solution 4.1 a. Which value of  $c$  should be chosen?

Probability density function

$$f_X(x) \geq 0 \quad \leftarrow \text{Non-negative}$$

$$\int_{\Omega} f_X(x) dx = 1 \quad \leftarrow \text{Area} = 1$$



$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

$$\int_0^{60} c \cdot x \cdot (60 - x) dx = 1 \Rightarrow c = \frac{1}{36000}$$

Solution 4.1 b. Describe the cumulative distribution function  $F_X(x)$  of the random variable  $X$ .

Cumulative distribution function

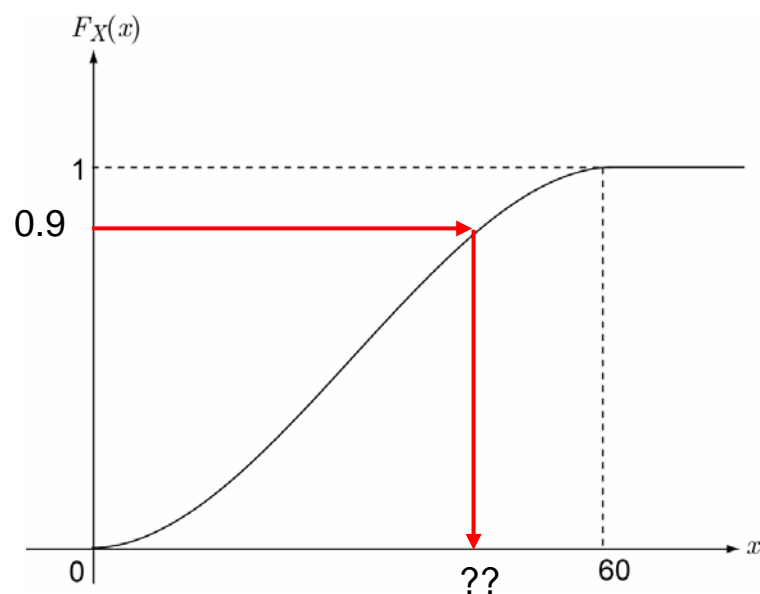
$$F_X(x) = \int_{\Omega} f_X(x) dx$$

$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{36000} \cdot \left( \frac{60}{2} \cdot x^2 - \frac{1}{3} \cdot x^3 \right) & 0 \leq x \leq 60 \\ 1 & 60 < x \end{cases}$$

Solution 4.1 c. Which of the following four values, 30.00 CHF, 40.00 CHF, 50 CHF and 60 CHF does not exceed the 90%-quantile of the monthly expense?

First we need to find the value corresponding to the 90% quantile

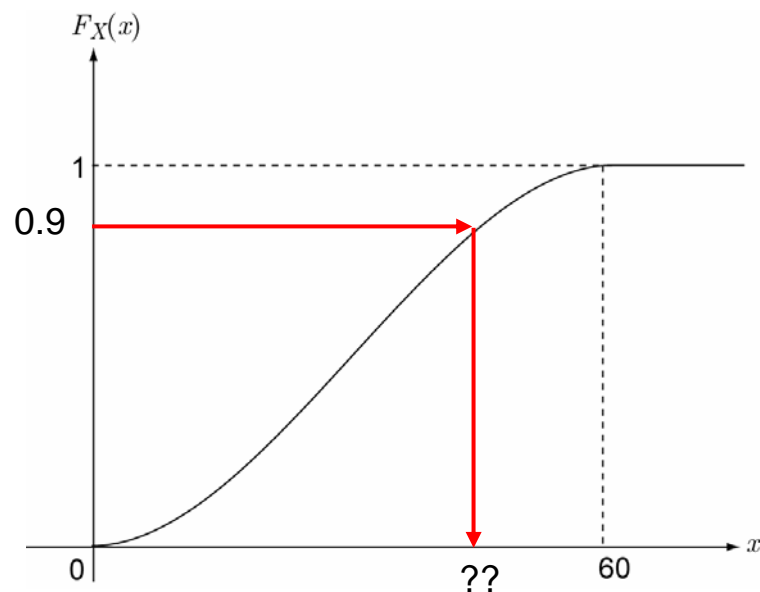


$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{36000} \cdot \left( \frac{60}{2} \cdot x^2 - \frac{1}{3} \cdot x^3 \right) & 0 \leq x \leq 60 \\ 1 & 60 < x \end{cases}$$

Solution 4.1    c. Which of the following four values, 30.00 CHF, 40.00 CHF, 50 CHF and 60 CHF does not exceed the 90%-quantile of the monthly expense?

First we need to find the value corresponding to the 90% quantile

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{36000} \cdot \left( \frac{60}{2} \cdot x^2 - \frac{1}{3} \cdot x^3 \right) & 0 \leq x \leq 60 \\ 1 & 60 < x \end{cases}$$

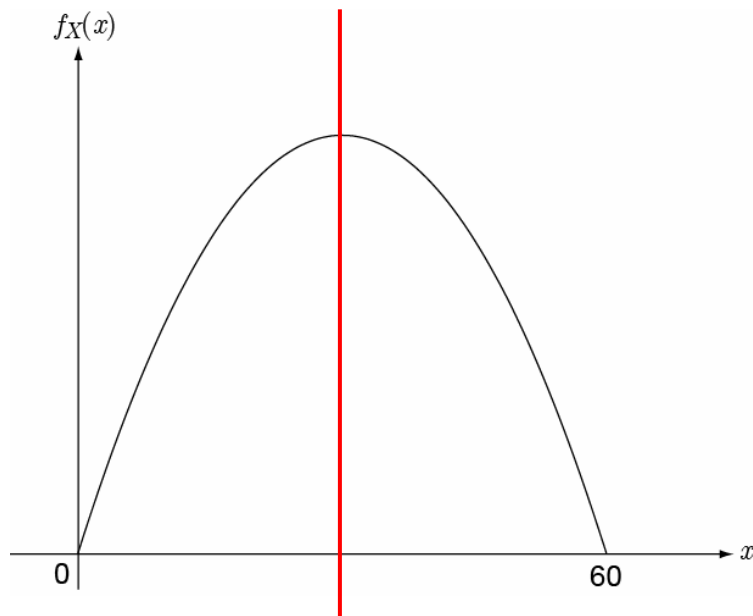


$$P(X \leq a) = F_X(x) = 0.9$$

$$P(X \leq a) = \frac{1}{36000} \cdot \int_0^a x(60-x) dx \Rightarrow a = \dots$$

Solution 4.1

d. How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?



Mean = 30

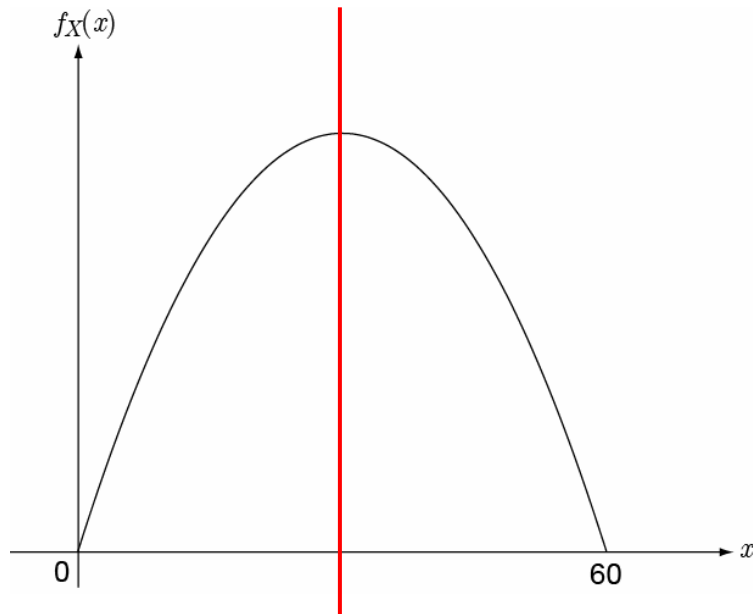
We can say this directly by looking at the Probability density function. WHY???

Mean = 30

Solution 4.1 d. How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?

Mean---First moment

$$\mu = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$



Mean = 30

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) dx = \frac{1}{36000} \cdot \int_0^{60} x^2 \cdot (60 - x) dx$$

## Exercise 4.2

The probability function of a basic variable is shown in the following figure.

a. determine analytically the PDF and the CDF.

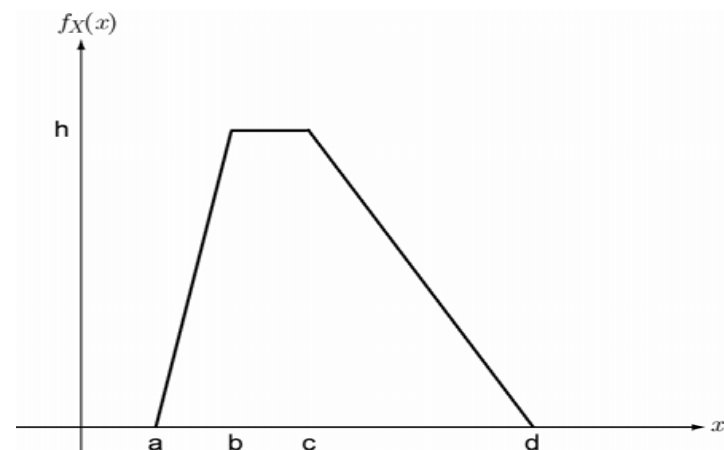
Let  $a=1, b=2, c=3, d=6$ . (Change in the exercise)

b. Define the mode and the parameter  $h$ .

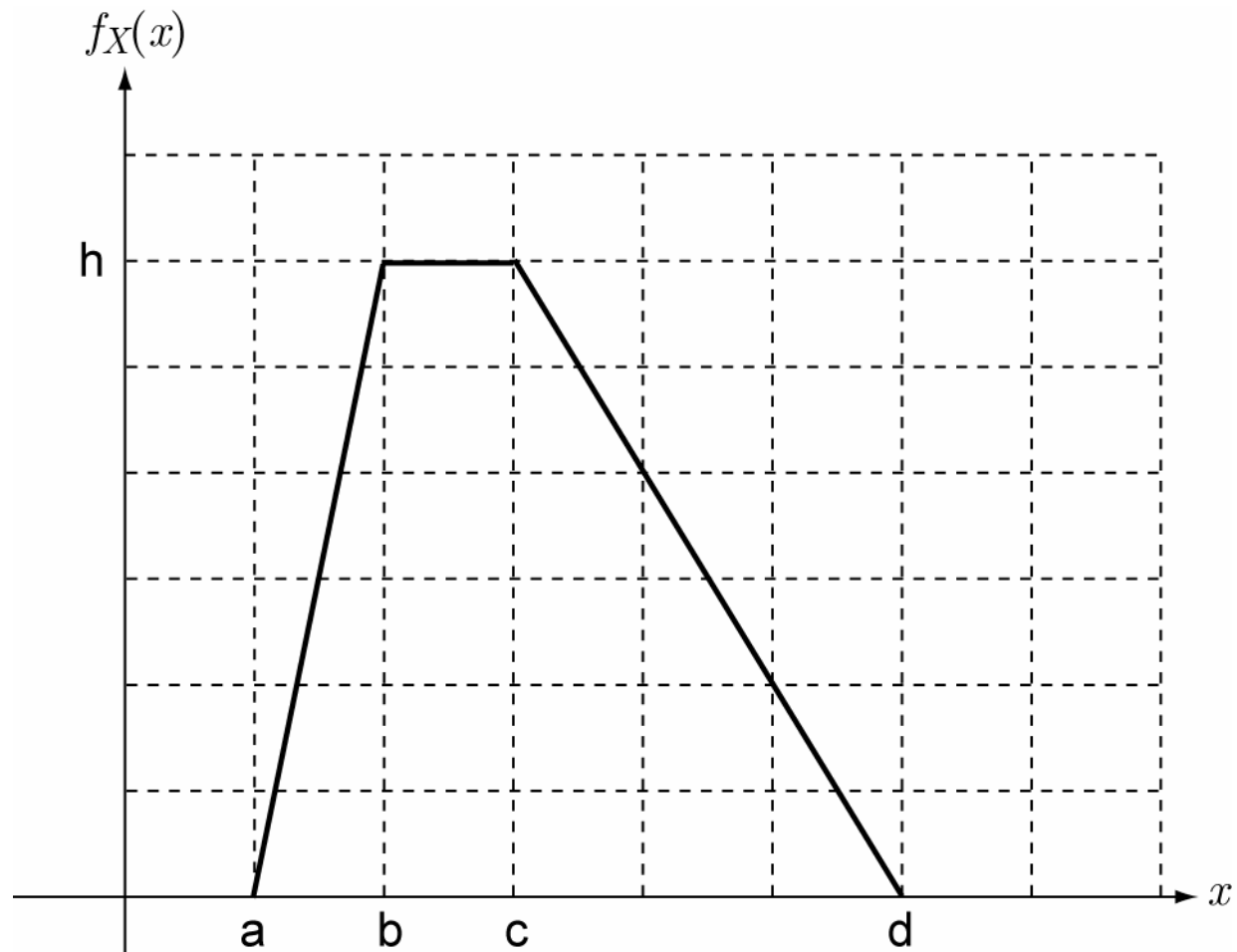
c. Calculate the mean value.

d. Calculate the value of the median.

e. Obtain graphically the median of the pdf. Discuss how the mean value may be evaluated graphically.



First think along with the definition, then think it again graphically.

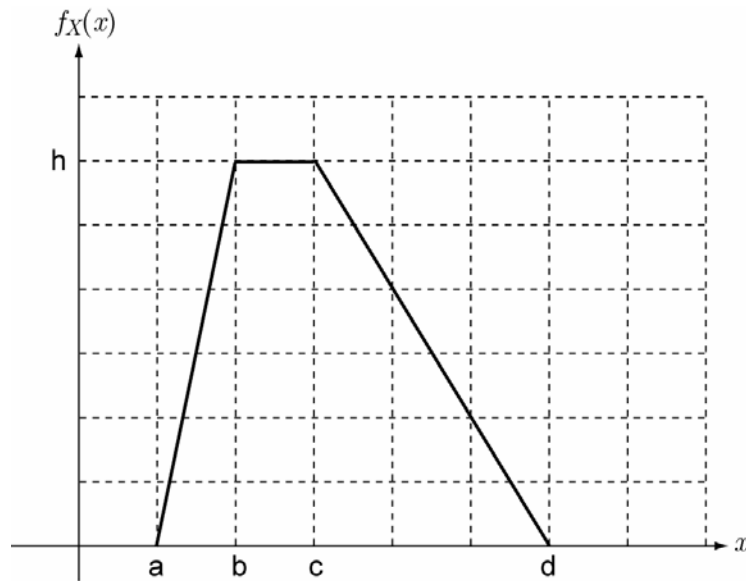




Solution 4.2

a. determine analytically the PDF and the CDF.

**PDF – Probability Density Function**

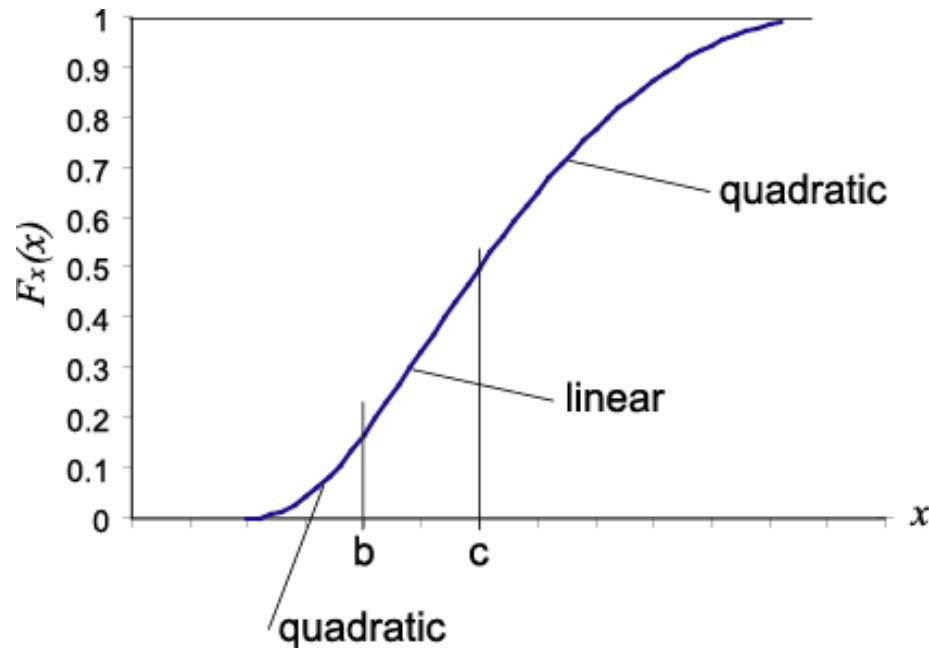


$$f_X(x) = \begin{cases} 0 & x < a \\ h \cdot \frac{(x-a)}{(b-a)} & a \leq x < b \\ h & b \leq x < c \\ h \cdot \frac{(x-d)}{(c-d)} & c \leq x < d \\ 0 & d \leq x \end{cases}$$

Solution 4.2

a. determine analytically the PDF and the CDF.

**CDF – Cumulative Distribution Function**



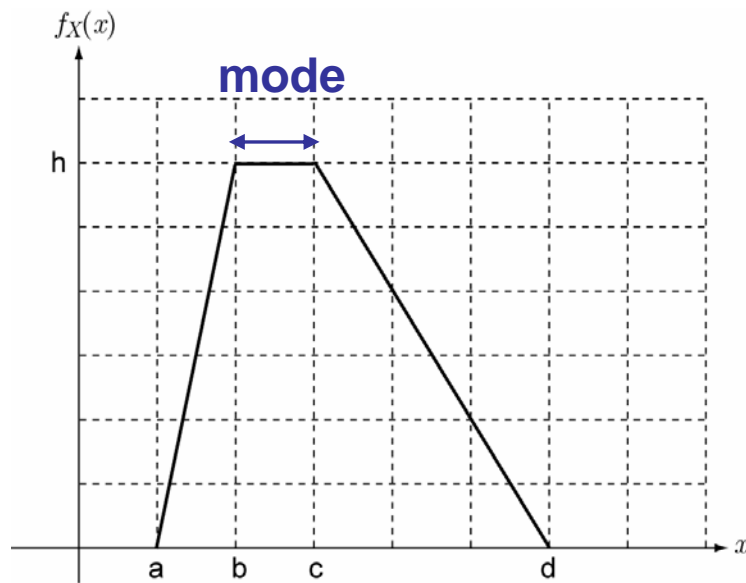
$$F_X(x) = \int_{\Omega} f_X(x) dx$$

$$F_X(x) = \begin{cases} 0 & x < a \\ h \cdot \frac{(x-a)^2}{2 \cdot (b-a)} + C_1 & a \leq x < b \\ h \cdot x + C_2 & b \leq x < c \\ h \cdot \frac{(x-d)^2}{2 \cdot (c-d)} + C_3 & c \leq x < d \\ 1 & d \leq x \end{cases}$$

The four constants can be calculated by using the boundary conditions

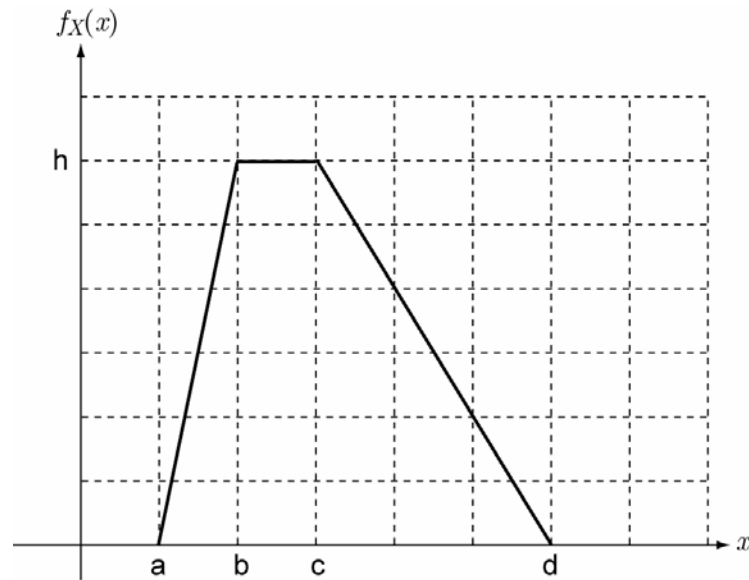
## Solution 4.2

- b. define the **mode** and the parameter  $h$ . ( $a=1, b=2, c=3, d=6$ )



Solution 4.2

b. define the mode and the parameter  $h$ . ( $a=1, b=2, c=3, d=6$ )

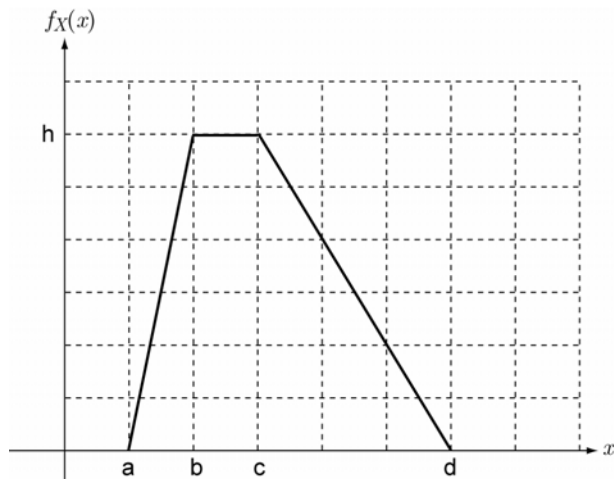


$\int_{-\infty}^{\infty} f_X(x) dx = 1$  Area under the density function

$\frac{(d-a) + (c-b)}{2} \cdot h = 1 \Rightarrow \dots h = \dots$

Solution 4.2

c. Calculate the value of the mean (a=1, b=2, c=3, d=6)



$$f_X(x) = \begin{cases} 0 & x < a \\ h \cdot \frac{(x-a)}{(b-a)} & a \leq x < b \\ h & b \leq x < c \\ h \cdot \frac{(x-d)}{(c-d)} & c \leq x < d \\ 0 & d \leq x \end{cases}$$

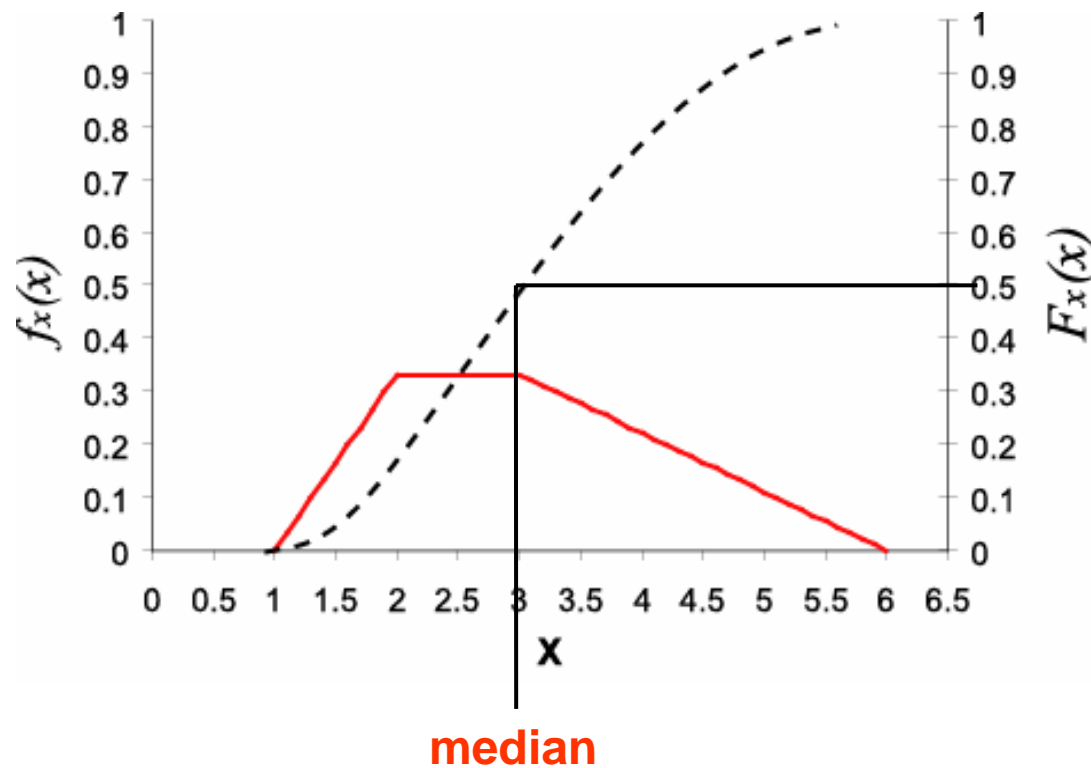


$$f_X(x) = \begin{cases} 0 & x < 1 \\ \frac{(x-1)}{3} & 1 \leq x < 2 \\ \frac{1}{3} & 2 \leq x < 3 \\ -\frac{(x-6)}{9} & 3 \leq x < 6 \\ 0 & 6 \leq x \end{cases}$$

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) \cdot dx = \int_1^2 \frac{x \cdot (x-1)}{3} dx + \int_2^3 \frac{x}{3} \cdot dx + \int_3^6 \frac{-x \cdot (x-6)}{9} dx = \dots$$

## Solution 4.2

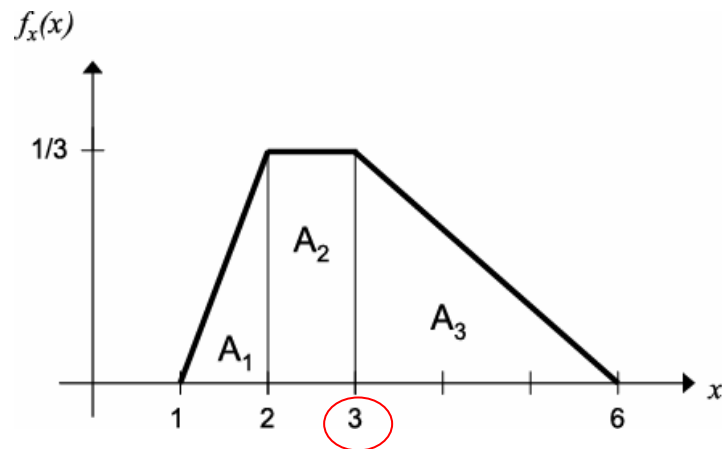
d. Calculate the value of the median.

**Graphically from the CDF****Analytically**

$$P(X \leq x) = \int_1^x f_X(x) dx = 0.5$$

## Solution 4.2

e. Obtain graphically the median of the pdf. Discuss how the mean value may be evaluated graphically.

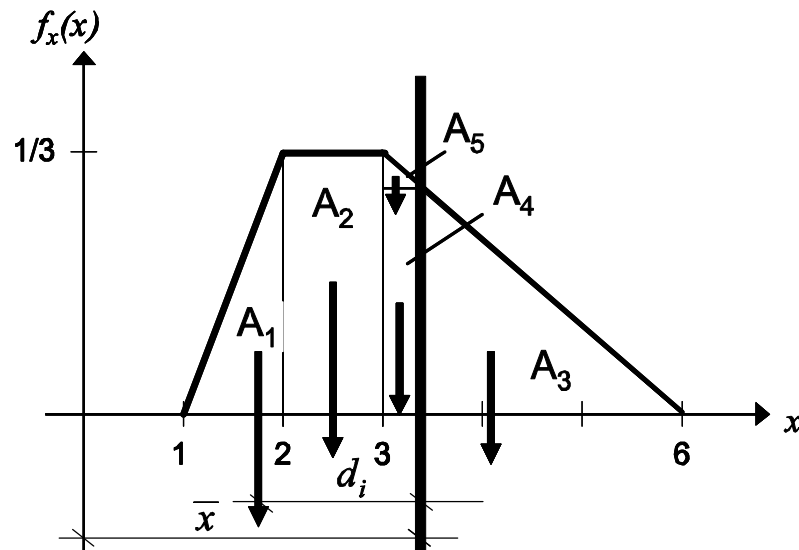
**Graphically from the PDF**

$$\left. \begin{aligned} A_1 &= (2-1) \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} \\ A_2 &= (3-2) \cdot \frac{1}{3} = \frac{1}{3} \\ A_3 &= (6-3) \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{2} \end{aligned} \right\} 0.5$$

**Median:** location at which the area under the density function is equal to 0.5

## Solution 4.2

e. Obtain graphically the median of the pdf. **Discuss how the mean value may be evaluated graphically.**

**Graphically from the PDF**

$$\sum_{i=1}^5 A_i \cdot d_i = 0$$

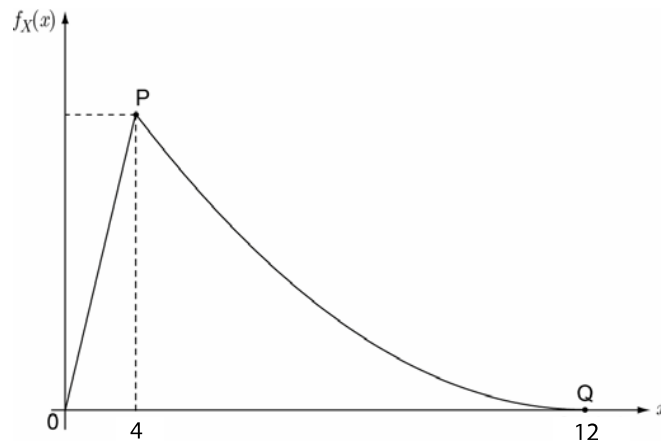
1. **Estimate moments for each shape**
2. **Take equilibrium around the hypothesized location of the center of gravity**

**Mean: center of gravity of the shape of the probability density function.**



**Exercise 4.3 (Group exercise- to be presented on 19.04.07)**

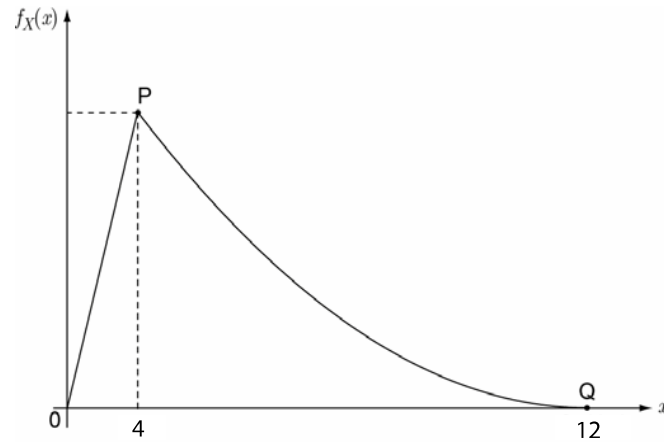
The probability density function of a random variable  $X$  is shown in Figure 4.3.1. In the interval  $[0, 4]$  the function is linear and in the interval  $[4, 12]$  the function is parabolic which is tangent to  $x$ -axis at point  $Q$ .



- Determine the coordinate of point  $P(x,y)$  and then describe the probability density function.
- Describe and draw the cumulative distribution function of  $X$  with some characteristic numbers in the figure.
- Calculate the mean value of  $X$ .
- Calculate  $P[X > 4]$ .

### Exercise 4.3 (Group exercise- to be presented on 19.04.07)

- a. Determine the coordinate of point P(x,y) and describe the probability density function.

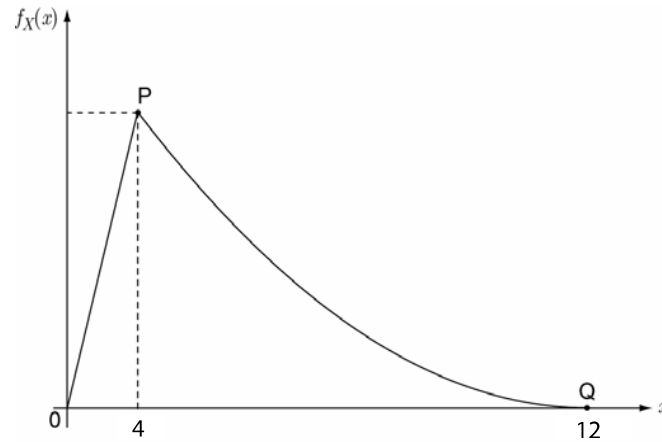


#### Steps:

- Define the pdf in the interval  $[0,12]$
- Find coordinates of P by remembering that the area under the density function is equal to 1!

**Exercise 4.3 (Group exercise- to be presented on 19.04.07)**

- b. Describe and draw the cumulative distribution function of  $X$  with some characteristic numbers in the figure.



**Steps:**

1.  $\int_{\Omega} f_X(x) dx = 1$

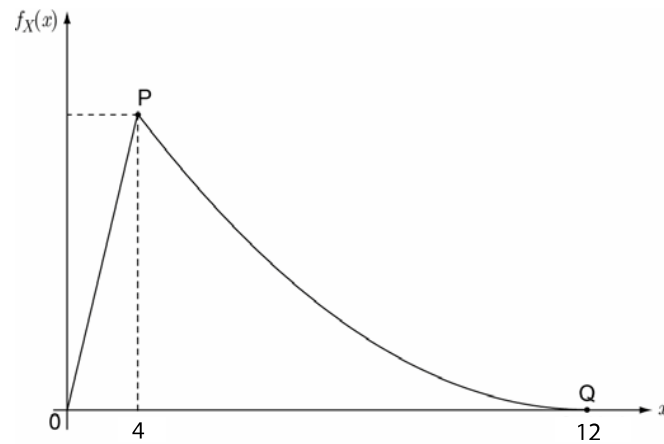
**2. Draw...!**

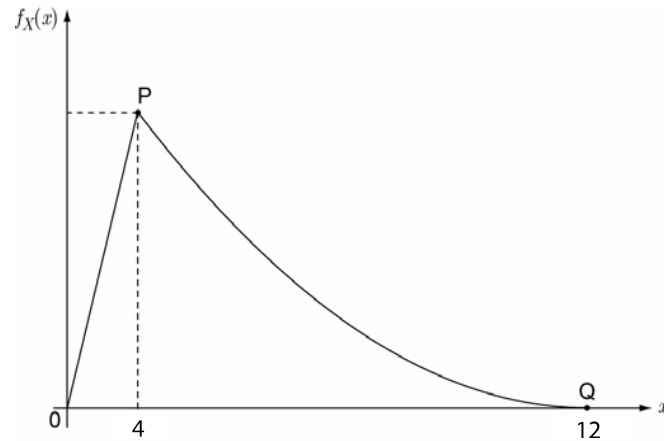
**Exercise 4.3 (Group exercise- to be presented on 19.04.07)**

c. Calculate the mean value

**Steps (Remember Exercise 4.2):**

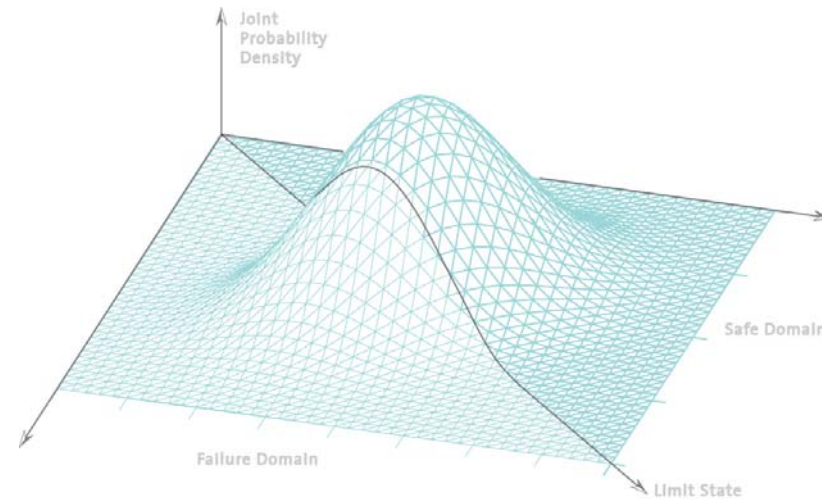
**1.**  $\mu_x = E[x]$



**Exercise 4.3 (Group exercise- to be presented on 19.04.07)****d.** Calculate  $P[X > 4]$ .**Steps (Remember Exercise 4.2):**Exceedance probability  $P[X > \alpha]$  is  $1 - P[X \leq \alpha]$ 

**1.**  $P[X > 4] = 1 - P[X \leq 4]$

How can this be expressed???



## Exercises Tutorial 5

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

### Exercise 4.1

The monthly expense [CHF] for water consumption including sewage fee for a 2-persons household may be considered as a random variable with the following density function:

$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases} \xrightarrow{\text{Change to}} f_X(x) = \begin{cases} c \cdot x \cdot \left(15 - \frac{x}{4}\right) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

### Exercise 4 Solution

- Which value of  $c$  should be chosen?
- Describe the cumulative distribution function  $F_X(x)$  of the random variable  $X$ .
- Which of the following four values, 30.00 CHF, 40.00 CHF, 50.00 CHF and 60.00 CHF does not exceed the 90%-quantile of the monthly expense?
- How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?

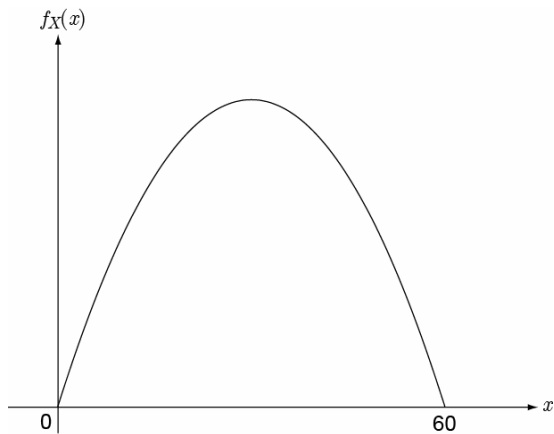
Solution 4.1 a. Which value of  $c$  should be chosen?

Probability density function

$$f_X(x) \geq 0 \quad \leftarrow \text{Non-negative}$$

$$\int_{\Omega} f_X(x) dx = 1 \quad \leftarrow \text{Area} = 1$$

$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$



$$\int_0^{60} c \cdot x \cdot (60 - x) dx = 1 \Rightarrow c \cdot \int_0^{60} x \cdot (60 - x) dx = 1 \Rightarrow$$

$$c \cdot \left[ \frac{60}{2} \cdot x^2 - \frac{1}{3} x^3 \right]_0^{60} = 1 \Rightarrow c \cdot (108000 - 72000) = 1 \Rightarrow$$

$$c = \frac{1}{36000}$$



Solution 4.1 b. Describe the cumulative distribution function  $F_X(x)$  of the random variable  $X$ .

Cumulative distribution function

$$F_X(x) = \int_{-\infty}^x f_X(t) dt \quad f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

Work in the intervals where the pdf is defined

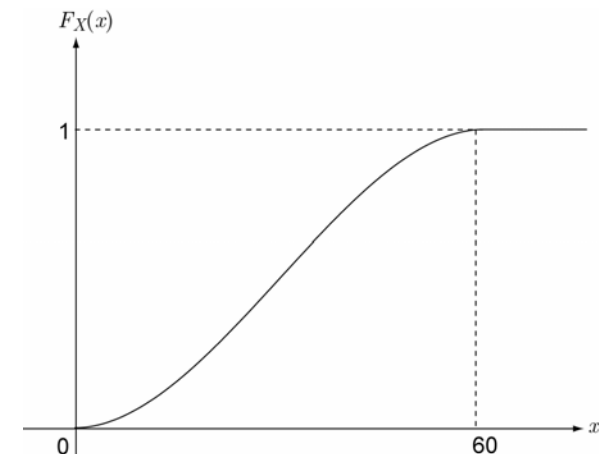
Case of  $0 \leq x \leq 60$

$$F_X(x) = \int_0^x f_X(t) dt = \int_0^x \frac{1}{36000} \cdot t(60-t) dt = \frac{1}{36000} \left[ \frac{60}{2} t^2 - \frac{1}{3} t^3 \right]_0^x = \frac{1}{36000} \left( 30x^2 - \frac{1}{3} x^3 \right)$$

Case of  $60 < x$

$$F_X(x) = 1$$

$$F_X(x) = \begin{cases} \frac{1}{36000} \cdot \left( \frac{60}{2} \cdot x^2 - \frac{1}{3} \cdot x^3 \right) & 0 \leq x \leq 60 \\ 1 & 60 < x \end{cases}$$



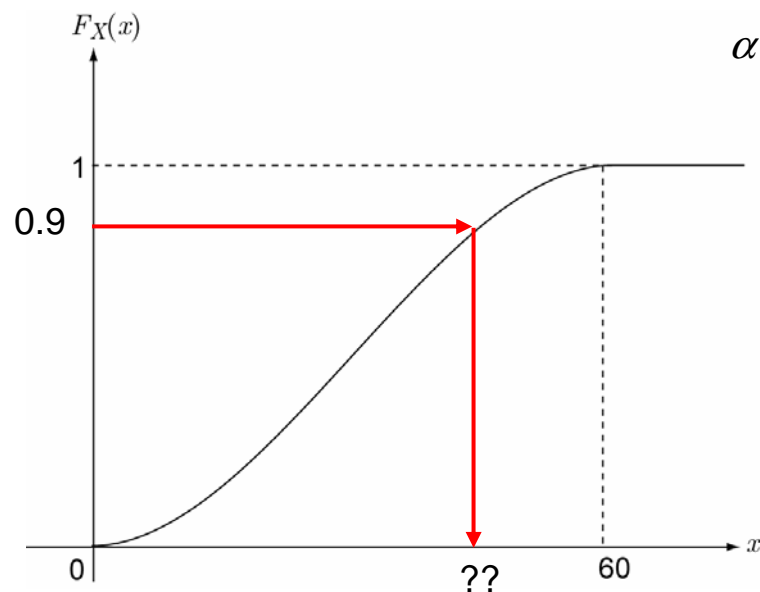
Solution 4.1 c. Which of the following four values, 30.00 CHF, 40.00 CHF, 50 CHF and 60 CHF does not exceed the 90%-quantile of the monthly expense?

First we need to find the value corresponding to the 90% quantile

$$P(X \leq \alpha) = F_X(x) = 0.9$$

$$P(X \leq \alpha) = \frac{1}{36000} \cdot \int_0^{\alpha} t(60-t) dt \Rightarrow 0.9 = \frac{1}{36000} \cdot \int_0^{\alpha} t(60-t) dt \Rightarrow$$

$$\alpha = 48.30$$

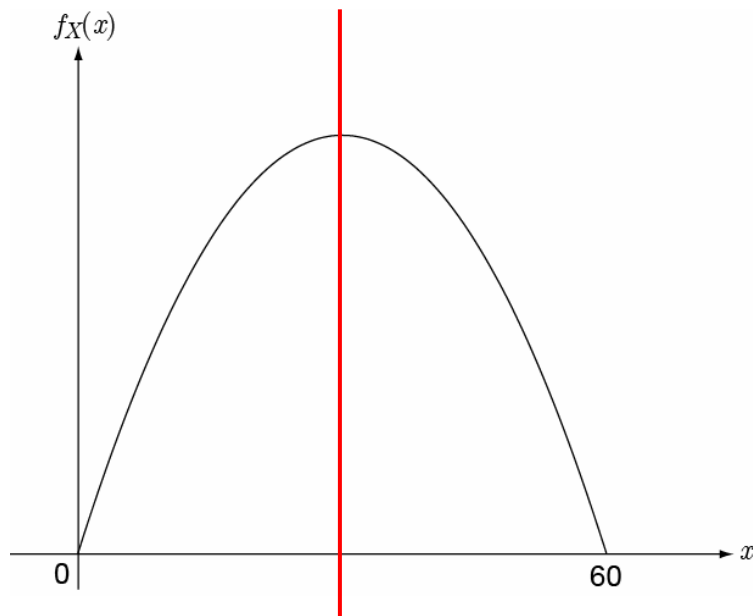


So the values of 30 and 40 CHF do not exceed the 90%- quantile of the monthly expense

Solution 4.1 **d. How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?**

Mean---First moment

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$



Mean = 30

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x) dx = \\ &= \frac{1}{36000} \int_0^{60} x^2 (60 - x) dx = \frac{1}{36000} \left[ 20x^3 - \frac{1}{4}x^4 \right]_0^{60} \\ &= \frac{1}{36000} (4320000 - 3240000) = \frac{1080000}{36000} = 30 \end{aligned}$$

### Exercise 5.1

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to

$$\sqrt{\frac{1}{3}}$$

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov[6X, 4Y]$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

---

Generally: (script Equations D.16 to D.18)

### Expectation operator

$$E[c] = c$$

$$E[cX] = cE[X]$$

$$E[a + bX] = a + bE[X]$$

$$E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$$

### Variance operator

$$\text{Var}[c] = 0$$

$$\text{Var}[cX] = c^2 \text{Var}[X]$$

$$\text{Var}[a + bX] = b^2 \text{Var}[X]$$

## Exercise 5.1

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to  $\sqrt{\frac{1}{3}}$

### Steps:

- Which is the first thing to do??

## Exercise 5.1

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to  $\sqrt{\frac{1}{3}}$

### Steps:

- Which is the first thing to do??

$$E[a + bX] = a + bE[X]$$

$$E[6X - 4Y + 2] = \dots$$

Find the expected value and variance of  $X$  and  $Y$ !

What do these measures express???

## Exercise 5.1

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to  $\sqrt{\frac{1}{3}}$

### Steps:

- Find the expected value and variance of  $X$  and  $Y$ ! (*Script Equations D.5-D11*)

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx = \int_{-1}^1 \frac{1}{2} x dx = \left[ \frac{1}{4} x^2 \right]_{-1}^1 = 0$$



### Exercise 5.1

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to  $\sqrt{\frac{1}{3}}$

#### Steps:

- Find the expected value and variance of  $X$  and  $Y$ ! (*Script Equations D.5-D11*)

$$E[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx = \int_{-1}^1 \frac{1}{2} x dx = \left[ \frac{1}{4} x^2 \right]_{-1}^1 = 0$$

$$Var(X) = E[X^2] - (E[X])^2 \rightarrow E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_{-1}^1 \frac{1}{2} x^2 dx = \left[ \frac{x^3}{6} \right]_{-1}^1 = \frac{1}{3}$$

## Exercise 5.1

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to  $\sqrt{\frac{1}{3}}$

### Steps:

2. After calculating the expected value and variance of both variables use the properties of the respective operator: (*Script Equations D.16-D18*)

a.  $E[6X - 4Y + 2]$

$$E[a + bX] = a + bE[X]$$

$$E[6X - 4Y + 2] = \dots$$

### Exercise 5.1

- a. Calculate the expected value of  $6X - 4Y + 2$
- b. Calculate the covariance of  $Cov(6X; 4Y)$
- c. Calculate the variance of  $6X - 4Y + 2$
- d. Calculate the expected value of  $6X^2 - 4Y^2$

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to  $\sqrt{\frac{1}{3}}$

**Steps:**

2. After calculating the expected value and variance of both variables use the properties of the respective operator: (*Script Equations D.16-D.18*)

b.  $Cov[6X, 4Y]$

$$Var[a + bX] = b^2 Var[X]$$

Script Equation D.23

**Watch this!!!**

$$Cov[6X, 4Y] = 6 \cdot 4 \cdot Cov[X, Y]$$

$$Cov[X, Y] = \rho_{XY} \cdot \sqrt{Var[X] \cdot Var[Y]}$$

## Exercise 5.1

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

The marginal probability density functions of a two dimensional random variable  $Z = (X, Y)^T$  are defined as:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad f_Y(y) = \begin{cases} \frac{3}{4} \cdot (2y - y^2) & \text{for } 0 \leq y \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

The correlation coefficient between  $X$  and  $Y$  equals to

$$\sqrt{\frac{1}{3}}$$

### Steps:

2. After calculating the expected value and variance of both variables use the properties of the respective operator: (*Script Equations D.16-D.18*)

c.  $Var[6X - 4Y + 2] = Var[6X] + Var[4Y] - 2 \cdot Cov[6X, 4Y]$

Watch this!!!

WHY??

## Exercise 5.1

### Steps:

2. After calculating the expected value and variance of both variables use the properties of the respective operator: (*Script Equations D.16-D.18*)

c.  $Var[6X - 4Y + 2] = Var[6X] + Var[4Y] - 2 \cdot Cov[6X, 4Y]$

Watch this!!!

WHY??

### Script Equation D.25

$$Y = a_0 + \sum_{i=1}^n a_i X_i$$

$$E[Y] = a_0 + \sum_{i=1}^n a_i E[X_i]$$

$$Var[Y] = \sum_{i=1}^n a_i^2 Var[X_i] + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^n a_i a_j C_{X_i X_j}$$

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

### Exercise 5.1

- a. Calculate the expected value of  $6X - 4Y + 2$
- b. Calculate the covariance of  $Cov(6X; 4Y)$
- c. Calculate the variance of  $6X - 4Y + 2$
- d. Calculate the expected value of  $6X^2 - 4Y^2$

**Steps:**

2. After calculating the expected value and variance of both variables use the properties of the respective operator: (*Script Equations D.16-D.18*)

c.  $Var[6X - 4Y + 2] = Var[6X] + Var[4Y] - 2 \cdot Cov[6X, 4Y]$



**WHY??**

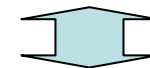
Script Equation D.25

$$Y = a_0 + \sum_{i=1}^n a_i X_i$$

$$E[Y] = a_0 + \sum_{i=1}^n a_i E[X_i]$$

$$Var[Y] = \sum_{i=1}^n a_i^2 Var[X_i] + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^n a_i a_j C_{X_i X_j}$$

$$Var[6X - 4Y + 2] = Var[6X] + Var[4Y] - 2 \cdot Cov[6X, 4Y]$$



$$Var[6X - 4Y + 2] = [6^2 Var[X]] + [(-4)^2 Var[Y]] + 2[6[-4]Cov[X, Y]]$$

---

## Exercise 5.1

### Steps:

2. After calculating the expected value and variance of both variables use the properties of the respective operator: (*Script Equations D.16-D.18*)

d.  $E[6X^2 - 4Y^2] = 6E[X^2] - 4E[Y^2]$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Calculate the expected value of  $6X - 4Y + 2$
- Calculate the covariance of  $Cov(6X; 4Y)$
- Calculate the variance of  $6X - 4Y + 2$
- Calculate the expected value of  $6X^2 - 4Y^2$

## Exercise 5.2

However, in previous years, the measurement has been undertaken with a less accurate device (it is herein called “less accurate device”). The correspondence between the measurement with the accurate device and the measurement with the less accurate device is of interest.

Therefore, the joint probability of the measured wind speeds with both devices is established by calibration in the following next few years.

Table 5.2.1 shows the *joint probability* of the numbers of the days when measured wind speed exceeds the threshold with the accurate device and with the less accurate device.

$N_U$  represents the number of the days when the wind speed measured with the *accurate device exceeds* the threshold, and  $N_G$  represents the number of the days when the wind speed measured with the *less accurate device exceeds* the threshold.

	$N_U = 0$	$N_U = 1$	$N_U = 2$	$N_U = 3$	$P(N_G)$
$N_G = 0$	0.2910	0.0600	0.0000	0.0000	<b>0.3510</b>
$N_G = 1$	0.0400	0.3580	0.0100	0.0000	0.4080
$N_G = 2$	0.0100	0.0250	0.1135	0.0300	0.1785
$N_G = 3$	0.0005	0.0015	0.0100	0.0505	0.0625
$P(N_U)$	0.3415	0.4445	0.1335	0.0805	$\Sigma = 1.00$



---

## Exercise 5.2

Simplify!!

What is given??

Measurements of wind speed are taken with an accurate and a less accurate device.

The joint probability of the measurements by both devices of the number of days in which the wind speed exceeds a threshold.

$N_U$  represents the number of the days when the wind speed measured with the *accurate device exceeds* the threshold, and

$N_G$  represents the number of the days when the wind speed measured with the *less accurate device exceeds* the threshold.

---

## Exercise 5.2

Simplify!!

What is given??

Measurements of wind speed are taken with an accurate and a less accurate device.

The joint probability of the measurements by both devices of the number of days in which the wind speed exceeds a threshold.

$N_U$  represents the number of the days when the wind speed measured with the *accurate device exceeds* the threshold, and

$N_G$  represents the number of the days when the wind speed measured with the *less accurate device exceeds* the threshold.

What is required??

- a. Calculate the probability that the number of days at which the wind speed, measured by each device, exceeds the threshold coincides.
- b. Assume that the accurate device always measures the exact wind speed.  
What are the probabilities that the wind speed which exceeds the threshold prevails 0, 1, 2 and 3 time(s) in a year when the wind speed measured with the less accurate device exceeds the threshold twice?

### Exercise 5.2 a.

$N_U$  represents the number of the days when the wind speed measured with the *accurate device exceeds* the threshold, and

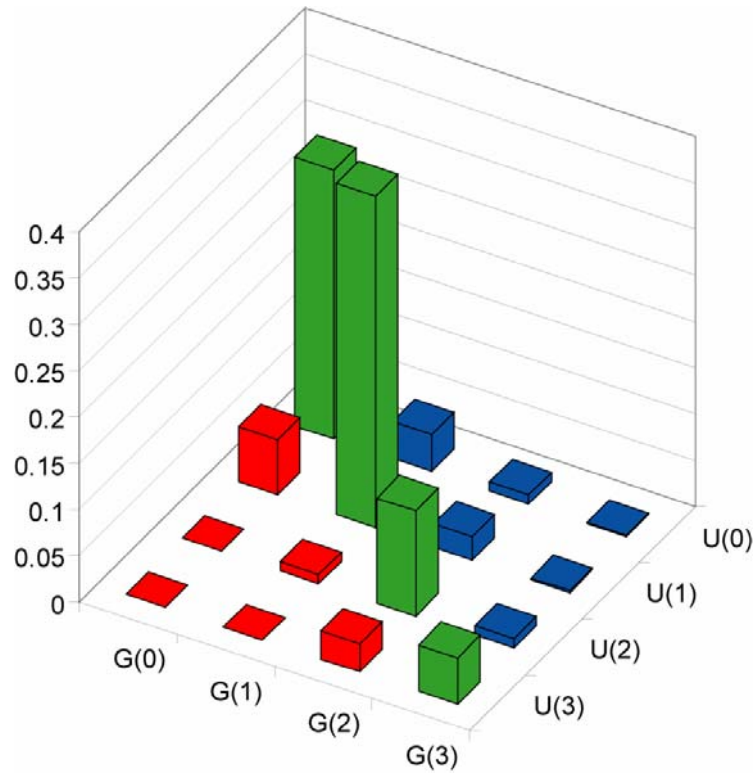
$N_G$  represents the number of the days when the wind speed measured with the *less accurate device exceeds* the threshold.

	$N_U = 0$	$N_U = 1$	$N_U = 2$	$N_U = 3$	$P(N_G)$
$N_G = 0$	0.2910	0.0600	0.0000	0.0000	<b>0.3510</b>
$N_G = 1$	0.0400	0.3580	0.0100	0.0000	0.4080
$N_G = 2$	0.0100	0.0250	0.1135	0.0300	0.1785
$N_G = 3$	0.0005	0.0015	0.0100	0.0505	0.0625
$P(N_U)$	0.3415	0.4445	0.1335	0.0805	$\Sigma = 1.00$

- a. Calculate the probability that the number of days at which the wind speed, measured by each device, exceeds the threshold coincides.

$$P[N_U = N_G] = 0.2910 + 0.3580 + 0.1135 + 0.0505 = 0.813$$

Exercise 5.2



a. Probability that  $N_U = N_G$

$$P[N_U = N_G] = 0.2910 + 0.3580 + 0.1135 + 0.0505 = 0.813$$

Probability that  $N_U < N_G$

Probability that  $N_U > N_G$

	$N_U = 0$	$N_U = 1$	$N_U = 2$	$N_U = 3$	$P(N_G)$
$N_G = 0$	0.2910	0.0600	0.0000	0.0000	<b>0.3510</b>
$N_G = 1$	0.0400	0.3580	0.0100	0.0000	0.4080
$N_G = 2$	0.0100	0.0250	0.1135	0.0300	0.1785
$N_G = 3$	0.0005	0.0015	0.0100	0.0505	0.0625
$P(N_U)$	0.3415	0.4445	0.1335	0.0805	$\Sigma = 1.00$

Exercise 5.2 b.

- b. Assume that the accurate device always measures the exact wind speed.  
 What are the probabilities that the wind speed which exceeds the threshold prevails 0, 1, 2 and 3 time(s) in a year when the wind speed measured with the less accurate device exceeds the threshold twice?

Conditional probability....Baye's rule

	$N_U = 0$	$N_U = 1$	$N_U = 2$	$N_U = 3$	$P(N_G)$
$N_G = 0$	0.2910	0.0600	0.0000	0.0000	<b>0.3510</b>
$N_G = 1$	0.0400	0.3580	0.0100	0.0000	0.4080
$N_G = 2$	0.0100	0.0250	0.1135	0.0300	0.1785
$N_G = 3$	0.0005	0.0015	0.0100	0.0505	0.0625
$P(N_U)$	0.3415	0.4445	0.1335	0.0805	$\Sigma = 1.00$

**How can we write this???**

---

Exercise 5.3 (Group exercise- to be presented on 26.04.07)

Highway bridges may require maintenance in their life time. The duration where no maintenance is required,  $T$ , is assumed exponentially distributed with the mean value of 10 years. The maintenance activity takes some time, which is represented by  $S$ . The time is also assumed exponentially distributed with the mean value of 1/12 year.

- a. Assuming that  $T$  and  $S$  are independent, obtain the distribution of the time between subsequent maintenance activities are initiated,  $Z$ , i.e.,  $Z=S+T$ .
- b. How large is the probability  $P(Z \leq 5)$  ?
- c. Assume that two bridges in a highway system are opened and the times until the bridges require the maintenance are represented by  $T_1$  and  $T_2$  which are independent identically distributed as  $T$ .  
How large is the probability that in the next 5 years no maintenance is required for the two bridges?

### Example with discrete random variables

Consider two dices. The outcome of throwing each one is described by the discrete random variables  $X$  and  $Y$ .

We are looking for the probability that the sum of the numbers that will come out when we will throw the dices is equal to e.g. 10.

The sum is a random variable itself which can be described as:

$$Z=X+Y$$

The probability that the sum is equal to 10 is: ???

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

## Example with discrete random variables

Consider two dices. The outcome of throwing each one is described by the discrete random variables  $X$  and  $Y$ .

We are looking for the probability that the sum of the numbers that will come out when we will throw the dices is equal to e.g. 10.

The sum is a random variable itself which can be described as:

$$Z=X+Y$$

The probability that the sum is equal to 10 is:

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

---

In the case of continuous random variables:

$$Z=X+Y$$



---

Exercise 5.3 (Group exercise- to be presented on 26.04.07)

What is given?

Time where no maintenance is required,  $T$ : exponentially distributed  $\mu_T = 10$  years

Time of maintenance  $S$ : exponentially distributed  $\mu_S = \frac{1}{12}$  years

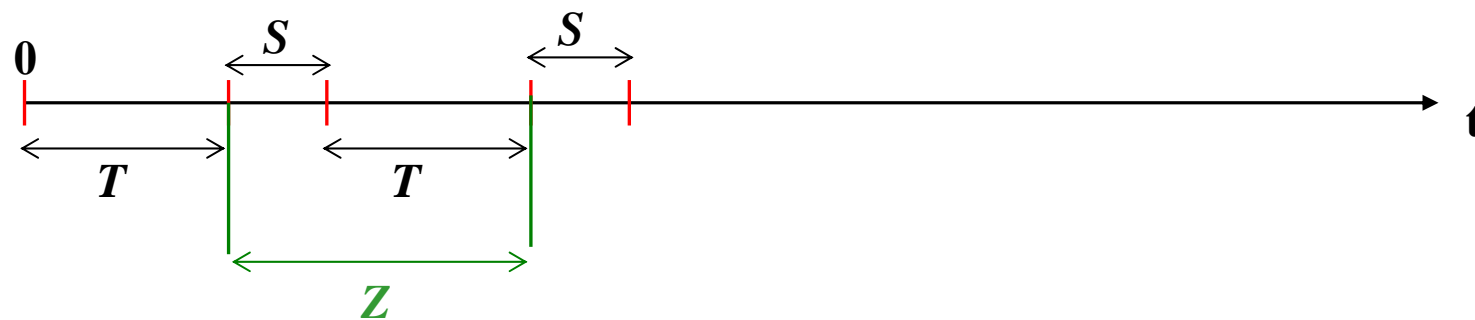
Exercise 5.3 (Group exercise- to be presented on 26.04.07)

What is given?

Time where no maintenance is required,  $T$ : exponentially distributed  $\mu_T = 10$  years

Time of maintenance  $S$ : exponentially distributed  $\mu_S = \frac{1}{12}$  years

- a. Assuming that  $T$  and  $S$  are independent, obtain the distribution of the time between subsequent maintenance activities are initiated,  $Z$ , i.e.,  $Z=S+T$ .



---

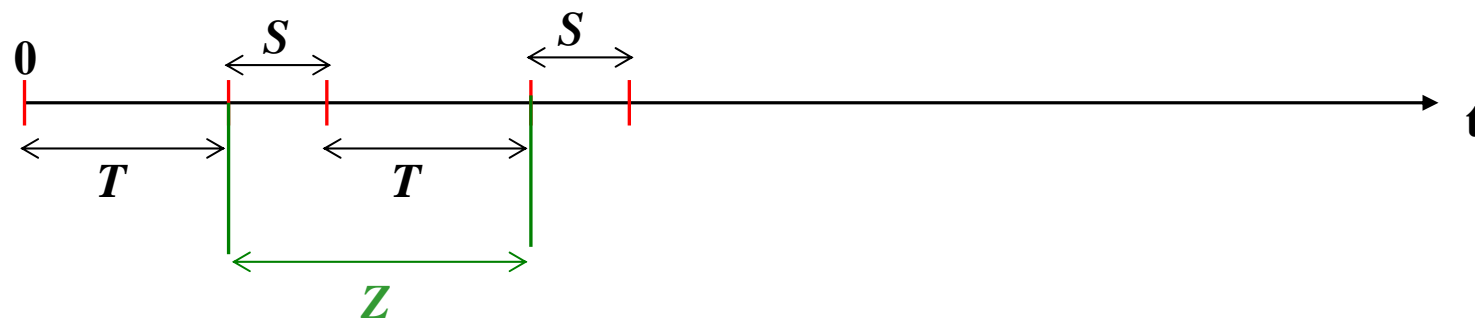
Exercise 5.3 (Group exercise- to be presented on 26.04.07)

What is given?

Time where no maintenance is required,  $T$ : exponentially distributed  $\mu_T = 10$  years

Time of maintenance  $S$ : exponentially distributed  $\mu_S = \frac{1}{12}$  years

- a. Assuming that  $T$  and  $S$  are independent, obtain the distribution of the time between subsequent maintenance activities are initiated,  $Z$ , i.e.,  $Z=S+T$ .



In the case of continuous random variables:  $Z=X+Y$

---

Exercise 5.3 (Group exercise- to be presented on 26.04.07)

What is given?

Time where no maintenance is required,  $T$ : exponentially distributed  $\mu_T = 10$  years

Time of maintenance  $S$ : exponentially distributed  $\mu_S = \frac{1}{12}$  years

- a. Assuming that  $T$  and  $S$  are independent, obtain the distribution of the time between subsequent maintenance activities are initiated,  $Z$ , i.e.,  $Z=S+T$ .

$Z$  is defined within the interval  $[0; z]$

$$f_Z(z) = \int_0^z f_T(t) f_S(z-t) dt = \dots$$

$$f_T(t) = \frac{1}{\mu_T} \cdot e^{\frac{-t}{\mu_T}} \quad f_S(s) = \frac{1}{\mu_S} \cdot e^{\frac{-s}{\mu_S}}$$

---

Exercise 5.3 (Group exercise- to be presented on 26.04.07)

b. How large is the probability  $P(Z \leq 5)$  ?

$$P(Z \leq 5) = F_Z(z) = \dots$$

---

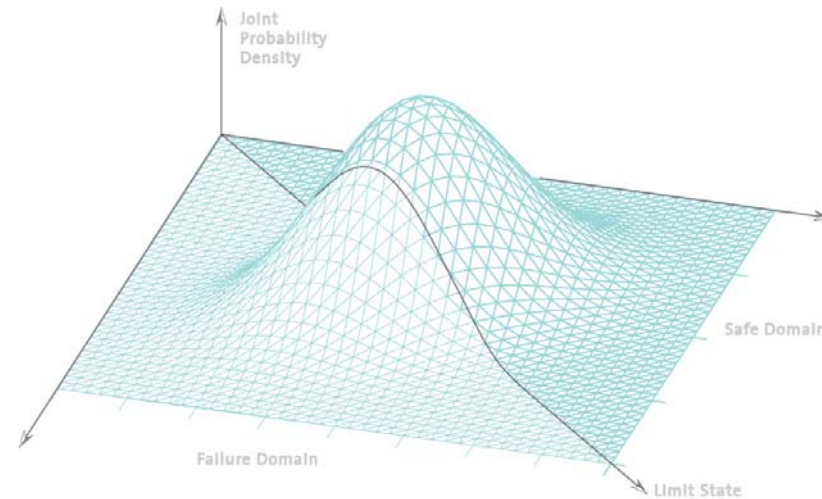
Exercise 5.3 (Group exercise- to be presented on 26.04.07)

b. How large is the probability  $P(Z \leq 5)$  ?

$$P(Z \leq 5) = F_Z(z) = \dots$$

c. Assume that two bridges in a highway system are opened and the times until the bridges require the maintenance are represented by  $T_1$  and  $T_2$  which are independent identically distributed as  $T$ . How large is the probability that in the next 5 years no maintenance is required for the two bridges?

**How can we express the problem??**



## Exercises Tutorial 6

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

---

**Assessment 1:**

**Thursday 3rd of May**

**Where?: HCI G7**

**Time?: 8.00-** try to be exact on your time!

**Allowed?:** Everything (script and the course's material is more than enough)

**Not Allowed? :** mobile phones, pc's etc that provide communication means

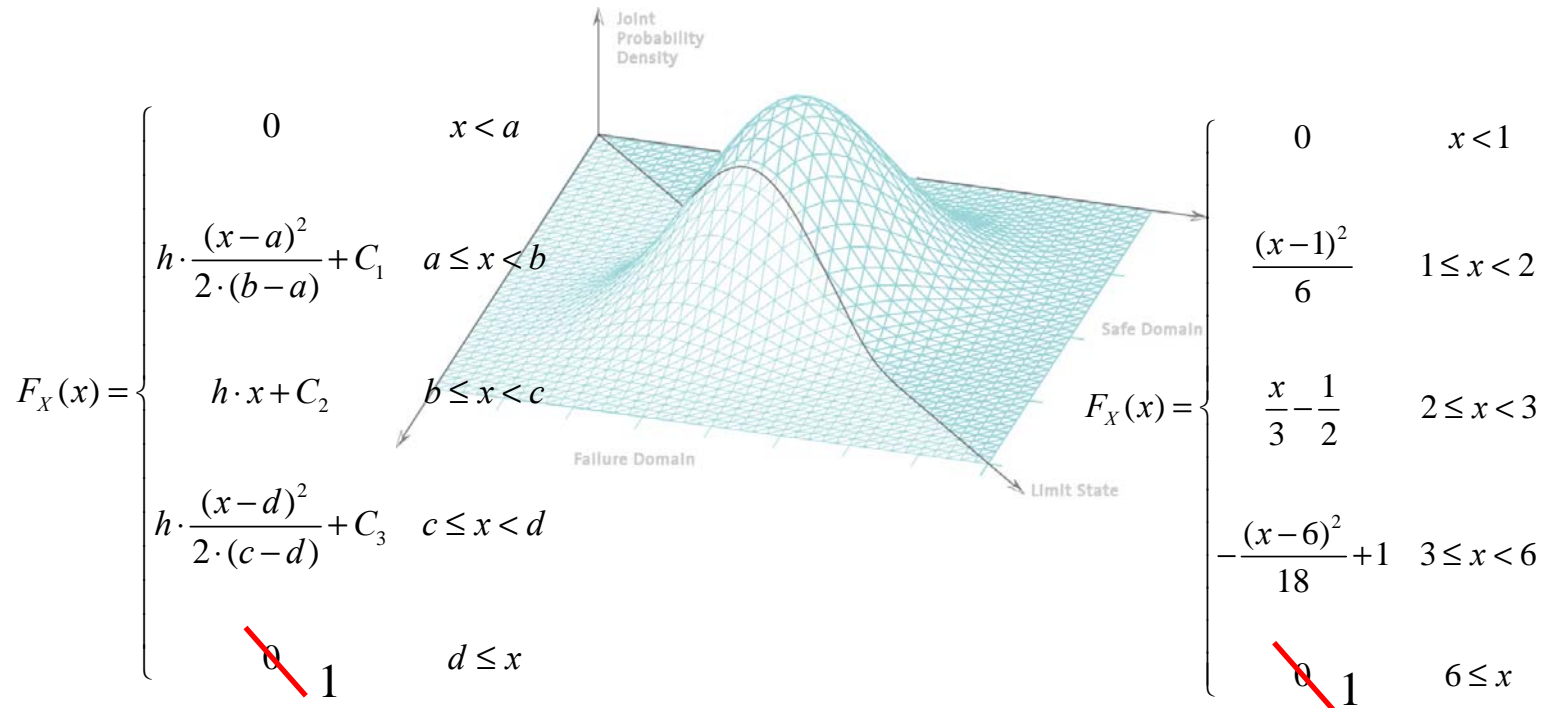
**Up to which lecture? Lecture 6**

**Do not forget:**

- Legi
- Calculator
- Pens!



**Please correct in the solutions of the tutorial exercises:**



On pages 4.2 and 4.4

Correct also the numbering of exercise 6.2

### Exercise 5.3 (Group exercise)

Highway bridges may require maintenance in their life time. The duration where no maintenance is required,  $T$ , is assumed exponentially distributed with the mean value of 10 years. The maintenance activity takes some time, which is represented by  $S$ . The time is also assumed exponentially distributed with the mean value of 1/12 year.

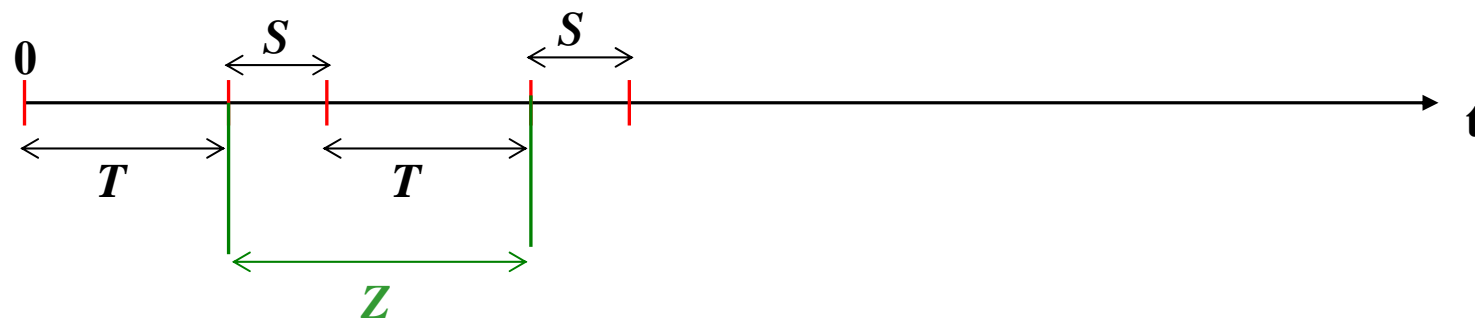
- a. Assuming that  $T$  and  $S$  are independent, obtain the **distribution of the time** between which subsequent maintenance activities are initiated,  $Z$ , i.e.,  $Z=S+T$ .
- b. How large is the probability  $P(Z \leq 5)$ ?
- c. Assume that two bridges in a highway system are opened and the times until the bridges require the maintenance are represented by  $T_1$  and  $T_2$  which are independent identically distributed as  $T$ .  
How large is the probability that in the next 5 years no maintenance is required for the two bridges?

### Exercise 5.3 (Group exercise)

Time where no maintenance is required,  $T$ : exponentially distributed  $\mu_T = 10$  years

Time of maintenance  $S$ : exponentially distributed  $\mu_S = \frac{1}{12}$  years

- a. Assuming that  $T$  and  $S$  are independent, obtain the distribution of the time between subsequent maintenance activities are initiated,  $Z$ , i.e.,  $Z=S+T$ .



In the case of continuous random variables:  $Z=X+Y$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx$$

## Exercise 5.2 a.

Measurements of wind speed are taken with an accurate and a less accurate device.

The joint probability of the measurements by both devices of the number of days in which the wind speed exceeds a threshold.

$N_U$  represents the number of the days when the wind speed measured with the *accurate device exceeds* the threshold, and

$N_G$  represents the number of the days when the wind speed measured with the *less accurate device exceeds* the threshold.

	$N_U = 0$	$N_U = 1$	$N_U = 2$	$N_U = 3$	$P(N_G)$
$N_G = 0$	0.2910	0.0600	0.0000	0.0000	<b>0.3510</b>
$N_G = 1$	0.0400	0.3580	0.0100	0.0000	0.4080
$N_G = 2$	0.0100	0.0250	0.1135	0.0300	0.1785
$N_G = 3$	0.0005	0.0015	0.0100	0.0505	0.0625
$P(N_U)$	0.3415	0.4445	0.1335	0.0805	$\Sigma = 1.00$

- a. Calculate the probability that the number of days at which the wind speed, measured by each device, exceeds the threshold coincides.

$$P[N_U = N_G] = 0.2910 + 0.3580 + 0.1135 + 0.0505 = 0.813$$

Exercise 5.2 b.

- b. Assume that the accurate device always measures the exact wind speed.  
 What are the probabilities that the wind speed which exceeds the threshold prevails  
 0, 1, 2 and 3 time(s) in a year when the wind speed measured with the **less accurate device**  
**exceeds the threshold twice?**

Conditional probability-Bayes's rule  $P[N_U | N_G = 2] = \frac{P[N_U \cap (N_G = 2)]}{P[N_G = 2]}$

	$N_U = 0$	$N_U = 1$	$N_U = 2$	$N_U = 3$	$P(N_G)$
$N_G = 0$	0.2910	0.0600	0.0000	0.0000	<b>0.3510</b>
$N_G = 1$	0.0400	0.3580	0.0100	0.0000	0.4080
$N_G = 2$	0.0100	0.0250	0.1135	0.0300	0.1785
$N_G = 3$	0.0005	0.0015	0.0100	0.0505	0.0625
$P(N_U)$	0.3415	0.4445	0.1335	0.0805	$\Sigma = 1.00$

$$P[N_U = 0 | N_G = 2] = \frac{P[(N_U = 0) \cap (N_G = 2)]}{P[N_G = 2]} = \frac{0.01}{0.1785} = 0.056$$

$$P[N_U = 1 | N_G = 2] = \frac{P[(N_U = 1) \cap (N_G = 2)]}{P[N_G = 2]} = \frac{0.025}{0.1785} = 0.1401$$

$$P[N_U = 2 | N_G = 2] = \frac{P[(N_U = 2) \cap (N_G = 2)]}{P[N_G = 2]} = \frac{0.1135}{0.1785} = 0.6359$$

$$P[N_U = 3 | N_G = 2] = \frac{P[(N_U = 3) \cap (N_G = 2)]}{P[N_G = 2]} = \frac{0.03}{0.1785} = 0.1681$$

## Exercise 6.1

Let  $\{X_i\}_{1 \leq i \leq 50}$  be independent, identically Normal distributed with mean value of  $\mu = 1$  and standard deviation of  $\sigma = 2$ . Define:

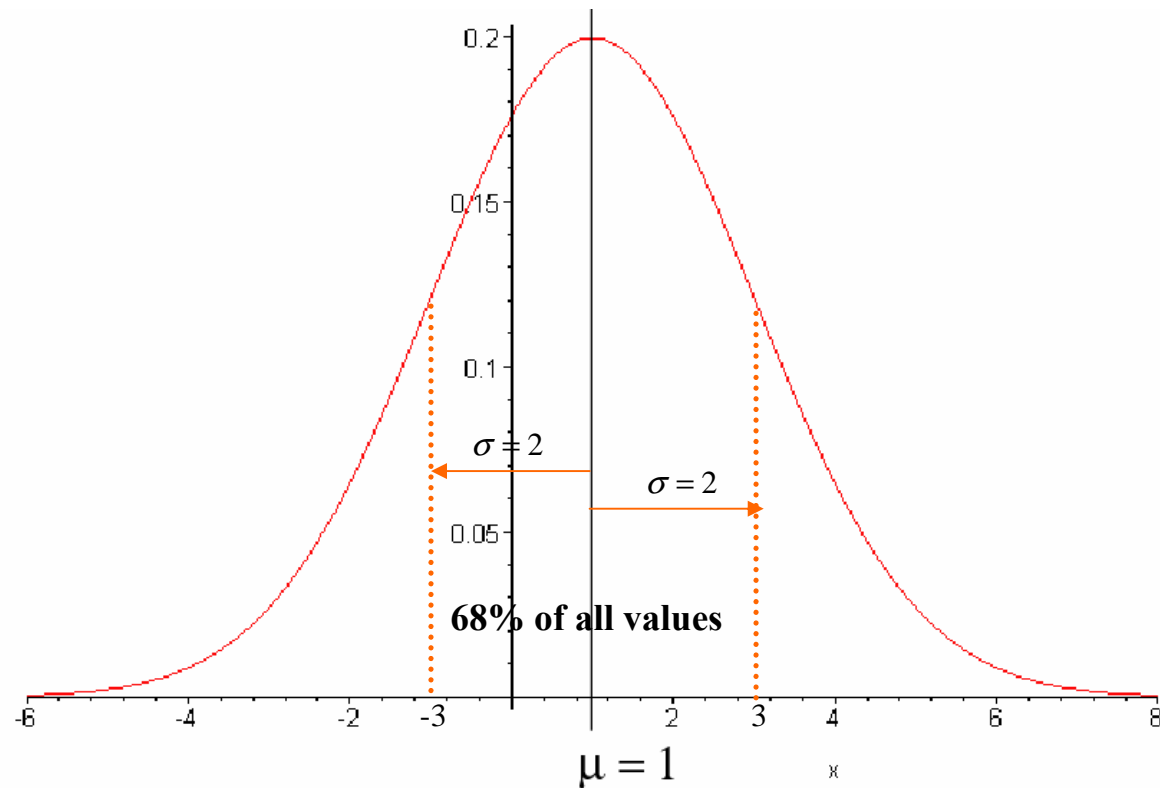
$$S_n = X_1 + X_2 + \dots + X_n \quad \text{and} \quad \bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{S_n}{n}$$

where  $n=50$ .

- Calculate the mean and the standard deviation of  $S_n$  and  $\bar{X}_n$ .
- Calculate  $P(E[X_1] - 1 \leq X_1 \leq E[X_1] + 1)$
- Calculate  $P(E[S_n] - 1 \leq S_n \leq E[S_n] + 1)$
- Calculate  $P(E[\bar{X}_n] - 1 \leq \bar{X}_n \leq E[\bar{X}_n] + 1)$

## Exercise 6.1

Probability density function of Normal distribution with mean value = 1 and standard deviation = 2



a) Calculate the mean and the standard deviation of  $S_n$ .

$$S_n = X_1 + X_2 + \dots + X_n$$

Sum of Normal random variables is also a Normal distributed random variable!

Remember the following formulas:

$$E[X + Y] = E[X] + E[Y]$$

$$Var[X + Y] = Var[X] + Var[Y]$$

$$\mu_{S_n} = E[S_n] =$$

$$\sigma_{S_n}^2 = Var[S_n] =$$

$$N(\mu_{S_n}, \sigma_{S_n}^2) = N(\dots, \dots)$$



## Exercise 6.1

a) Calculate the mean and the standard deviation of  $\bar{X}_n$  .

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{S_n}{n}$$

Remember the following formulas:

$$E[cX] = cE[X]$$

$$Var[cX] = c^2Var[X]$$

$$E[\bar{X}_n] =$$

$$V[\bar{X}_n] =$$

$$N(\mu_{\bar{X}_n}, \sigma_{\bar{X}_n}^2) = N(\dots, \dots)$$

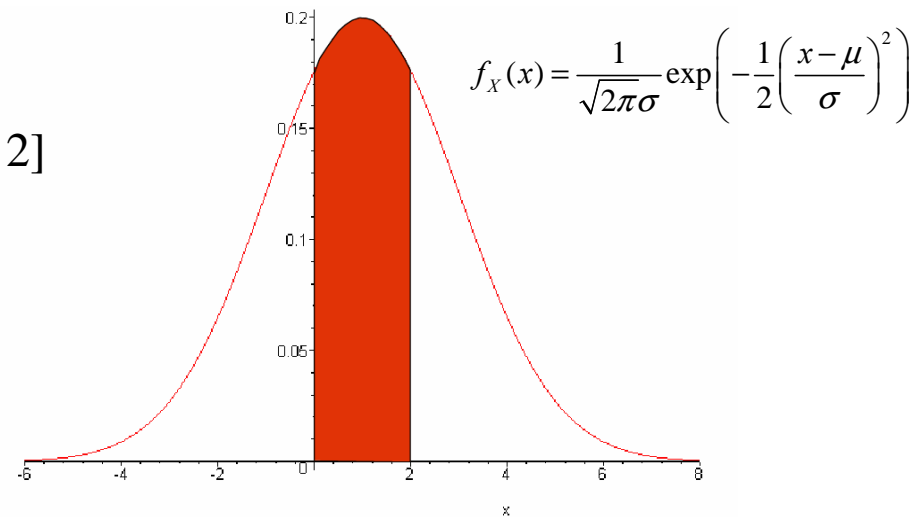
## Exercise 6.1

b) Calculate  $P(E[X_1]-1 \leq X_1 \leq E[X_1]+1)$

$$P(E[X_1]-1 \leq X_1 \leq E[X_1]+1) = P[0 \leq X_1 \leq 2]$$

$$= \int_0^2 \frac{1}{\sqrt{2\pi} \cdot 2} \exp\left(-\frac{(x-1)^2}{2 \cdot 2^2}\right) dx$$

Use table in the script page D-17



What is a probability table?

How to integrate??? -> Utilize the probability table!

In order to utilize the probability table, the variable has to be **standardized!**

**Standardization**

$$P(E[X_1]-1 \leq X_1 \leq E[X_1]+1) = P[0 \leq X_1 \leq 2]$$

$$Z = \frac{X_1 - \mu}{\sigma}$$

$$= P[0-1 \leq X_1-1 \leq 2-1] \quad \leftarrow \text{Mean value of } \mu = 1$$

$$= P\left[\frac{0-1}{2} \leq \frac{X_1-1}{2} \leq \frac{2-1}{2}\right] \quad \leftarrow \text{Standard deviation of } \sigma = 2$$

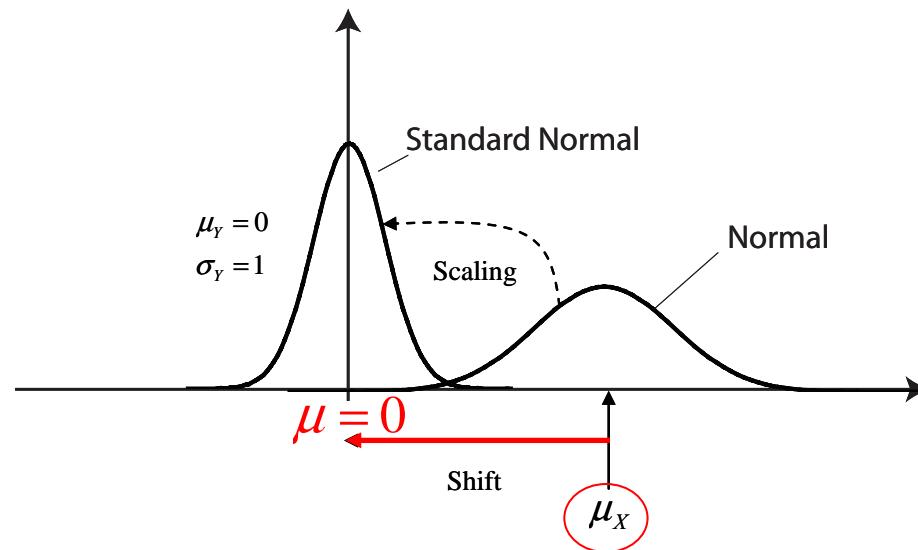
$$= P\left[-\frac{1}{2} \leq \frac{X_1-1}{2} \leq \frac{1}{2}\right]$$

$$= P\left[-\frac{1}{2} \leq Z \leq \frac{1}{2}\right]$$

$$= \Phi\left(\frac{1}{2}\right) - \Phi\left(-\frac{1}{2}\right)$$

$\Phi(z)$  is the cumulative distribution function for the Standard Normal distributed random variable  $N(0,1^2)$

## Standardization



$$Z = \frac{X_1 - \mu}{\sigma}$$

$$E[Z] = E\left[\frac{X_1 - \mu}{\sigma}\right] = \frac{E[X_1] - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = \underline{0}$$

$$\begin{aligned} \text{Var}[Z] &= \text{Var}\left[\frac{X_1 - \mu}{\sigma}\right] = \frac{1}{\sigma^2} \text{Var}[X_1 - \mu] \\ &= \frac{1}{\sigma^2} \text{Var}[X_1] = \frac{\sigma^2}{\sigma^2} = \underline{1} \end{aligned}$$

$\Phi(z)$  is the cumulative distribution function for the Standard Normal distributed random variable  $N(0,1^2)$

Exercise 6.1

Probability table

$$P(E[X_1]-1 \leq X_1 \leq E[X_1]+1)$$

$$= \Phi\left(\frac{1}{2}\right) - \Phi\left(-\frac{1}{2}\right)$$

Where is  $\Phi(-0.5)$ ?? Nowhere in the table!

because...  $\Phi(-z) = 1 - \Phi(z)$

so that

$$\Phi(-0.5) = 1 - \Phi(0.5)$$

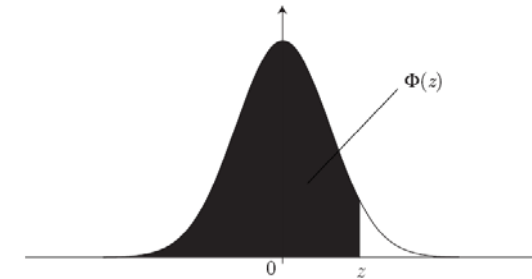
$$= 1 - 0.6915$$

$$= 0.3085$$

Finally

$$P(E[X_1]-1 \leq X_1 \leq E[X_1]+1)$$

$$= 0.6915 - 0.3085 = 0.3830$$



Probability density function of the standard normal random variable.

z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)	z	Φ(z)
0.00	0.5000	0.50	0.6915	1.00	0.8413	1.50	0.9332	2.00	0.9772
0.01	0.5040	0.51	0.6950	1.01	0.8438	1.51	0.9345	2.10	0.9821356
0.02	0.5080	0.52	0.6985	1.02	0.8461	1.52	0.9357	2.20	0.9860966
0.03	0.5120	0.53	0.7019	1.03	0.8485	1.53	0.9370	2.30	0.9892759
0.04	0.5160	0.54	0.7054	1.04	0.8508	1.54	0.9382	2.40	0.9918025
0.05	0.5199	0.55	0.7088	1.05	0.8531	1.55	0.9394	2.50	0.9937903
0.06	0.5239	0.56	0.7123	1.06	0.8554	1.56	0.9406	2.60	0.9953388
0.07	0.5279	0.57	0.7157	1.07	0.8577	1.57	0.9418	2.70	0.9965330
0.08	0.5319	0.58	0.7190	1.08	0.8599	1.58	0.9429	2.80	0.9974449
0.09	0.5359	0.59	0.7224	1.09	0.8621	1.59	0.9441	2.90	0.9981342
0.10	0.5398	0.60	0.7257	1.10	0.8643	1.60	0.9452	3.00	0.9986501
0.11	0.5438	0.61	0.7291	1.11	0.8665	1.61	0.9463	3.10	0.9990324
0.12	0.5478	0.62	0.7324	1.12	0.8686	1.62	0.9474	3.20	0.9993129
0.13	0.5517	0.63	0.7357	1.13	0.8708	1.63	0.9484	3.30	0.9995166
0.14	0.5557	0.64	0.7389	1.14	0.8729	1.64	0.9495	3.40	0.9996631
0.15	0.5596	0.65	0.7422	1.15	0.8749	1.65	0.9505	3.50	0.9997674
0.16	0.5636	0.66	0.7454	1.16	0.8770	1.66	0.9515	3.60	0.9998409
0.17	0.5675	0.67	0.7486	1.17	0.8790	1.67	0.9525	3.70	0.9998922
0.18	0.5714	0.68	0.7517	1.18	0.8810	1.68	0.9535	3.80	0.9999277
0.19	0.5753	0.69	0.7549	1.19	0.8830	1.69	0.9545	3.90	0.9999519
0.20	0.5793	0.70	0.7580	1.20	0.8849	1.70	0.9554	4.00	0.9999683
0.21	0.5832	0.71	0.7611	1.21	0.8869	1.71	0.9564	4.10	0.9999793

Table T.1 in Annex T of the Script

c) Calculate  $P(E[S_n]-1 \leq S_n \leq E[S_n]+1)$

→ same steps:

$$P(E[S_n]-1 \leq S_n \leq E[S_n]+1)$$

$$= P[\dots \leq S_n \leq \dots]$$

$$= P[\dots \leq \dots \leq \dots]$$

$$= \Phi(\dots) - \Phi(\dots)$$

=

Find the standardized form  $Z = \frac{X_1 - \mu}{\sigma}$

Look up the values in the probability table

Subtract  $F_b - F_a$

d) Calculate  $P(E[\bar{X}_n] - 1 \leq \bar{X}_n \leq E[\bar{X}_n] + 1)$

→ same steps:

-Standardization

-Probability table

$$P(E[\bar{X}_n] - 1 \leq \bar{X}_n \leq E[\bar{X}_n] + 1)$$

$$= P[\dots \leq \bar{X}_n \leq \dots]$$

$$= P[\dots \leq \dots \leq \dots]$$

$$= \Phi(\dots) - \Phi(\dots)$$

=

## Exercise 6.2

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

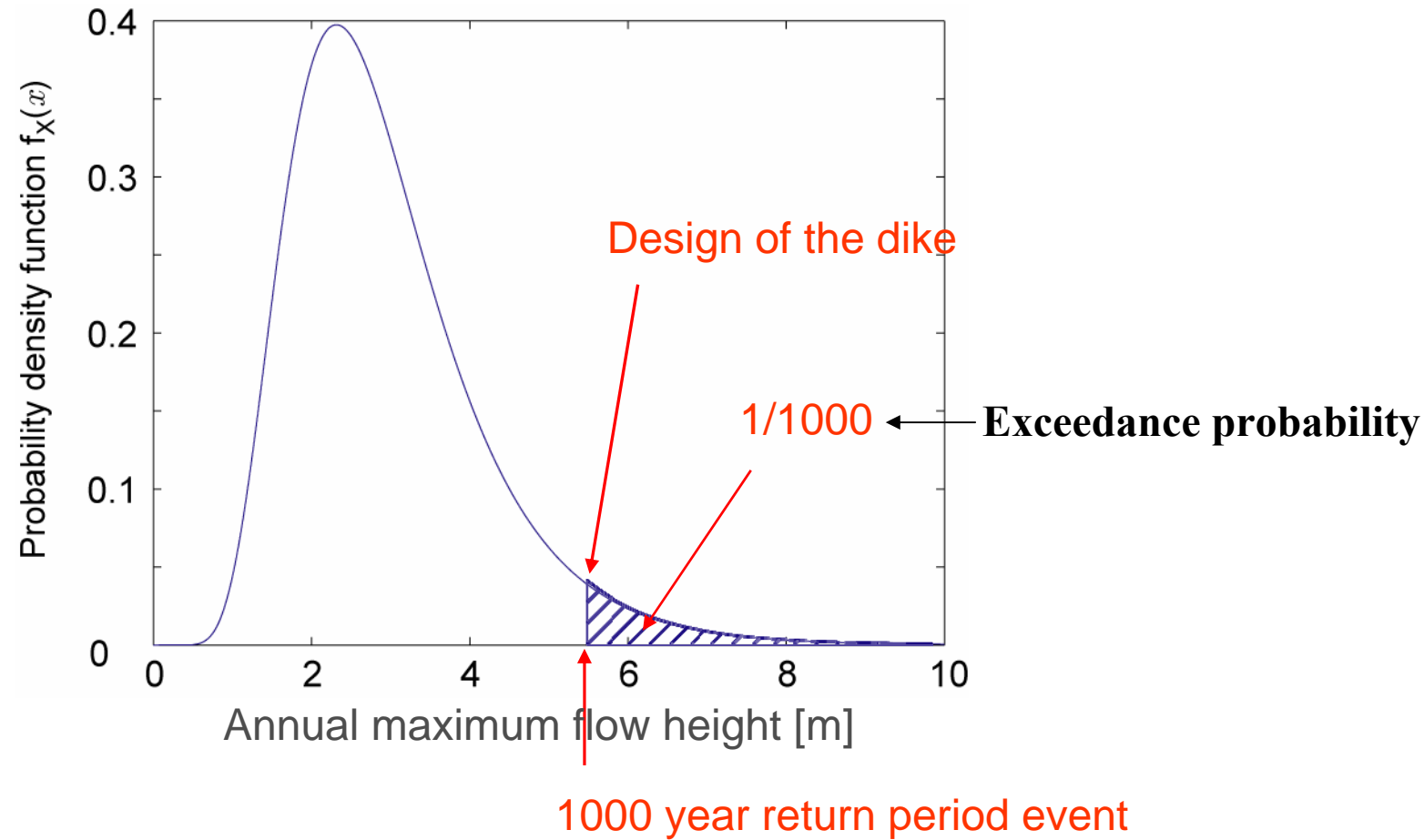
- a) During a 10 year period, for the first time in a given year?
- b) During a 10 year period, twice?
- c) Will not overflow during a 10 year period?
- d) During a 10 year period, at most once?
- e) During a 100 year period, 10 times?
- f) -
- g) During a 1000 year period, once or more often?

(It is assumed that flood occurs once in a year.)





Return period  $T$ :



Return period  $T$ :

Annual exceedance probability is  $p (= 1/T)$ .

Random variable  $N =$  time until a flood occurs for the first time

The probability that a flood occurs in the  $n^{\text{th}}$  year is

$$\begin{aligned} P[N = n] &= \underbrace{(1-p)(1-p)\dots(1-p)}_{n-1} p \leftarrow \text{Geometric distribution} \\ &= (1-p)^{n-1} p \end{aligned}$$

Expected value of  $N$ ,  $E[N]$  is

$$E[N] = \sum_{n=1}^{\infty} nP[N = n] = \sum_{n=1}^{\infty} n(1-p)^{n-1} p = \frac{1}{p} = T$$

Exercise 6.2

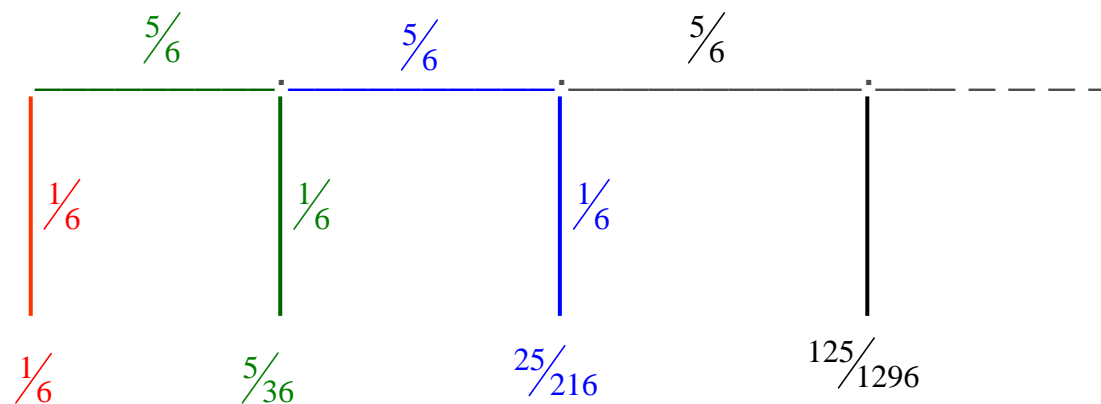
The probability that a flood occurs in the  $n^{\text{th}}$  year is

$$P[N = n] = \underbrace{(1 - p)(1 - p)\dots(1 - p)}_{n-1} p \quad \leftarrow \text{Geometric distribution}$$

$$= (1 - p)^{n-1} p$$

An easy example of a geometric distribution:

Probability of getting a 5 with a dice



$$P[N = 1] = \frac{1}{6}$$

$$P[N = 2] = (1 - \frac{1}{6}) \cdot \frac{1}{6}$$

$$P[N = 3] = \underbrace{(1 - \frac{1}{6})(1 - \frac{1}{6})}_{t-1} \cdot \frac{1}{6}$$

$$= (1 - \frac{1}{6})^{3-1} \cdot \frac{1}{6}$$

## Exercise 6.2

---

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

- a) During a 10 year period, for the first time in a given year?  
(for example, in the 10th year)

The probability that a flood occurs in the  $n^{\text{th}}$  year is

$$P[N = n] = \underbrace{(1 - p)(1 - p)\dots(1 - p)}_{n-1} p \quad \leftarrow \text{Geometric distribution}$$
$$= (1 - p)^{n-1} p$$

- 
- a) The event of overflow in the 10th year during a 10 year period may be described by a geometric distribution:

$$P(H_{\text{overflow},1}) = (p) \cdot (1 - p)^{n-1} = (0.001) \cdot (0.999)^9 = 0.000991$$

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

b) During a 10 year period, twice?

Probability that in 10 years ( $n$  trials) you get 2 overflows ( $y$  successes)

→ Binomial distribution

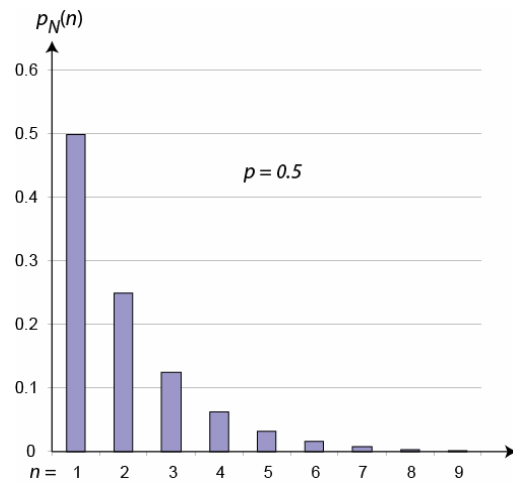
$$P[Y = y] = \binom{n}{y} p^y (1 - p)^{n-y}$$

According to the Binomial distribution it is

$$P(H_{\text{overflow},2}) = \frac{10!}{2! \cdot (10-2)!} (p)^2 \cdot (1-p)^{10-2}$$

Review:

Geometric distribution  
Time till first success

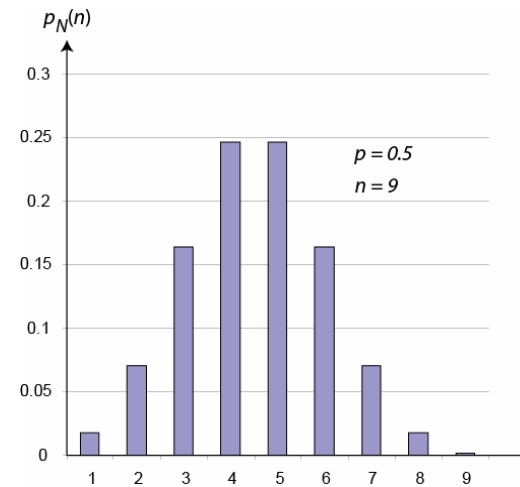


$$P[T = t] = (1 - p)^{t-1} p$$

$$E[T] = \frac{1}{p}$$

$$Var[T] = \frac{1 - p}{p^2}$$

Binomial distribution  
Number of success



$$P[N = y] = \binom{n}{y} (1 - p)^{n-y} p^y$$

$$E[N] = np$$

$$Var[N] = np(1 - p)$$

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

c) Will not overflow in a 10 year period?

Probability that in 10 years ( $n$  trials) you get 0 overflows ( $y$  successes)

→ Binomial distribution

$$P[Y = y] = \binom{n}{y} p^y (1 - p)^{n-y}$$


$$P(H_{\text{overflow},0}) = \frac{10!}{0! (10-0)!} (p)^0 \cdot (p-1)^{10-0} = (0.001)^0 \cdot (0.999)^{10} = \dots$$

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

d) During a 10 year period, at most once?

Probability that in 10 years ( $n$  trials) you get or 0 or 1 overflows ( $y$  successes)  
→ Binomial distribution

$$P(H_{\max,1}) = P(H_{\text{overflow},0}) + P(H_{\text{overflow},1})$$

$$P(H_{\text{overflow},0}) = \frac{10!}{0! \cdot (10-0)!} (p)^0 (p-1)^{10-0}$$
$$P(H_{\text{overflow},1}) = \frac{10!}{1! \cdot (10-1)!} (p)^1 (p-1)^{10-1}$$



## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

e) During a 100 year period, 10 times?

Probability that in 100 years ( $n$  trials) you get 10 overflows ( $y$  successes)  
→ Binomial distribution

$$P(H_{\text{overflow},10}) = \frac{100!}{10!(100-10)!} (p)^{10} \cdot (p-1)^{100-10}$$

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

g) During a 1000 year period, once or more often?

Probability that in 1000 years ( $n$  trials) you get any other result than 0 overflows ( $y$  successes)

→ Binomial distribution

$$P(H_{\text{overflow},0}) = \frac{1000!}{0!(1000-0)!} (p)^0 \cdot (p-1)^{1000-0} = (0.001)^0 (0.999)^{1000} = 0.368$$

And the required probability is the probability of the complementary event:

$$P(H_{\text{overflow},\geq 1}) = 1 - 0.368 = 0.632$$

### Exercise 6.3 (Group exercise/or, during the exercise tutorial)

An environmental planning engineering company obtains a project in return for a project proposal with the success rate of 27%.

Assume that you have taken over this company and you need to make the business plan for the forthcoming years.

- a. How large is the probability that the company will have at least one success after 12 project proposals?
- b. How large is the probability that only the last of 10 project proposals is accepted?
- c. How large is the probability that at most 2 out of 13 project proposals are accepted?

## Exercise 6.3 (Group exercise/or, during the exercise tutorial)

What is given?      Success rate = 27%  $\rightarrow p = 0.27$

- a. How large is the probability that the company will have **at least one success** after **12** project proposals?

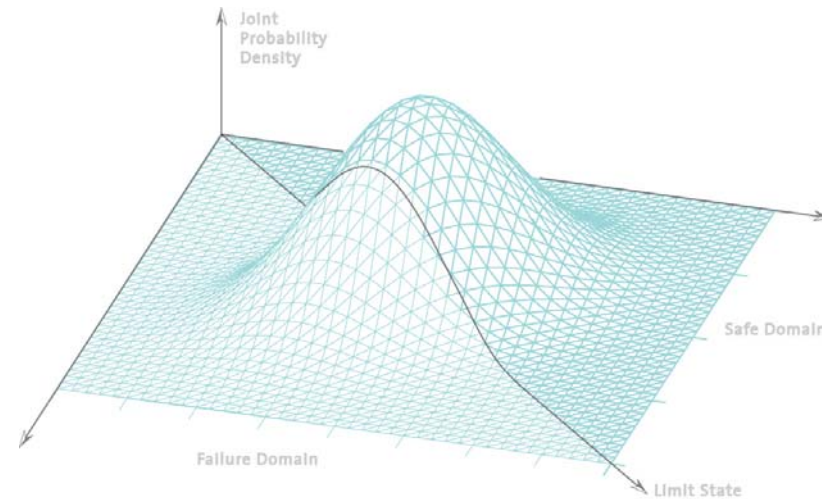
What is required: Time till first success – or number of successes?  
How can you express “at least”?

- b. How large is the probability that only the **last of 10** project proposals is accepted?

What is required: Time till first success – or number of successes?

- b. How large is the probability that **at most 2 out of 13** project proposals are accepted?

What is required: Time till first success – or number of successes?  
How can you express “at most”?



## Exercises Tutorial 7

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

---

## Assessment 1

- Corrections are still ongoing
- Results to be announced on the web page on the 22nd of May
- Multiple choice 1.13 (the one with the pipelines) will be omitted (many found the text “imprecise”)
- Multiple choice 1.23 (the last one) will be omitted – Printing error...there was actually no correct answer.
- So your mark will be based on 21 multiple choice questions (instead of 23)

---

## Assessment 1

- The Assessment itself and the solution will be published in the web page at the END of the semester
- You are welcomed to come during the office hours or after making an appointment and check your own assessment – after May 22  
(identify the areas where you are strong and those where you need to pay attention)
- An overall discussion of the assessment results will follow on the next exercise Tutorial, May 24

## Exercise 6.3 (Group exercise)

What is given?      Success rate = 27%  $\rightarrow p = 0.27$

- a. How large is the probability that the company will have **at least one success** after **12** project proposals?

What is required: Trials till first success – or number of successes?  
How can you express “at least”?

- b. How large is the probability that only the **last of 10** project proposals is accepted?

What is required: Trials till first success – or number of successes?

- c. How large is the probability that **at most 2 out of 13** project proposals are accepted?

What is required: Trials till first success – or number of successes?  
How can you express “at most”?



## Exercise 6.3 (Group exercise)

What is given?      Success rate = 27%  $\rightarrow p = 0.27$

- a. How large is the probability that the company will have **at least one success** after 12 project proposals?

What is required: Trials till first success – or **number of successes?**  
How can you express “at least”?      **Binomial distribution**

At least: One or more successes; anything else than 0 successes out of 12 proposals  
 $\rightarrow$  complementary event



### Exercise 6.3 (Group exercise)

What is given?      Success rate = 27%  $\rightarrow p = 0.27$

b. How large is the probability that only the last of 10 project proposals is accepted?

What is required: Trials till first success – or number of successes?

Geometric distribution



## Exercise 6.3 (Group exercise)

What is given?      Success rate = 27%  $\rightarrow p = 0.27$

c. How large is the probability that at most 2 out of 13 project proposals are accepted?

What is required: Trials till first success – or number of successes?

How can you express “at most”?

**Binomial distribution**

At most: Or zero, or one, or two successes.

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

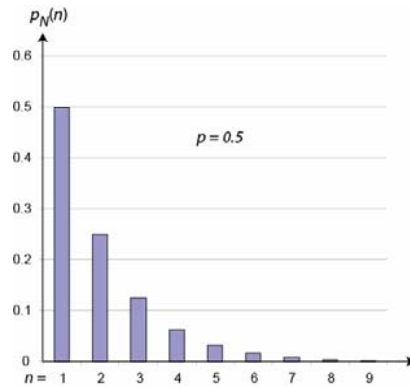
- a) During a 10 year period, for the first time in a given year?
- b) During a 10 year period, twice?
- c) Will not overflow during a 10 year period?
- d) During a 10 year period, at most once?
- e) During a 100 year period, 10 times?
- f) -
- g) During a 1000 year period, once or more often?

(It is assumed that flood occurs at most once in a year.)

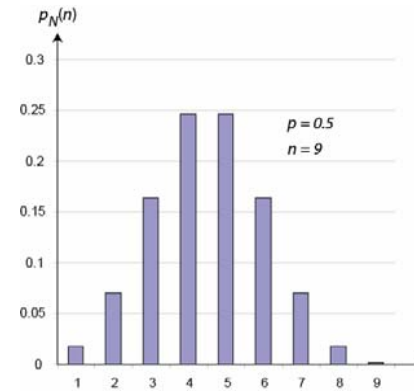


Exercise 6.2

Geometric distribution  
Time till first success



Binomial distribution  
Number of success



*How large is the probability that water will overflow the dike ...*

a) ... during a 10 year period, **for the first time** in a given year (for example, in the 10<sup>th</sup> year) ?

→ Geometric distribution:

$$P(H_{\text{overflow},1}) = (p)(1-p)^{n-1}$$

$$= (0.001)(0.999)^9 = 0.000991$$

b) ... during a 10 year period, **twice**? → Binomial distribution

$$P[Y = y] = \binom{n}{y} p^y (1-p)^{n-y}$$

$$P(H_{\text{overflow},2}) = \frac{10!}{2! \cdot (10-2)!} (p)^2 (1-p)^{10-2}$$

$$= \frac{10 \cdot 9}{2} (0.001)^2 (0.999)^8 = 0.000045$$

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

c) Will not overflow in a 10 year period?

→ Binomial distribution

$$P(H_{\text{overflow},0}) = \frac{10!}{0!(10-0)!} (p)^0 \cdot (p-1)^{10-0} = (0.001)^0 \cdot (0.999)^{10} = 0.99004$$

d) During a 10 year period, at most once?

Probability that in 10 years you get **or 0 or 1** overflows → Binomial distribution

$$P(H_{\text{max},1}) = P(H_{\text{overflow},0}) + P(H_{\text{overflow},1}) = 0.99995$$

$$P(H_{\text{overflow},1}) = \frac{10!}{1!(10-1)!} (p)^1 (p-1)^{10-1} = 10(0.001)^1 (0.999)^9 = 0.00991$$

## Exercise 6.2

A dike is designed to withstand a “1000-year return period flood”. How large is the probability that water will overflow the dike in the following conditions:

e) During a 100 year period, 10 times?

$$\begin{aligned} P(H_{\text{overflow},10}) &= \frac{100!}{10!(100-10)!} (p)^{10} \cdot (p-1)^{100-10} = \frac{100!}{10!(90)!} (0.001)^{10} \cdot (0.999)^{90} \\ &= 1.73 \cdot 10^{13} \cdot 10^{-30} \cdot 0.914 = 1.6 \cdot 10^{-17} \end{aligned}$$

g) During a 1000 year period, once or more often?

Probability that in 1000 years you get **any other result than 0** overflows

→ Binomial distribution

→ the required probability is the probability of the complementary event “0 overflows”

$$P(H_{\text{overflow},0}) = \frac{1000!}{0!(1000-0)!} (p)^0 \cdot (p-1)^{1000-0} = (0.001)^0 (0.999)^{1000} = 0.368$$

$$P(H_{\text{overflow},\geq 1}) = 1 - 0.368 = 0.632$$

## Exercise 6.1

Let  $\{X_i\}_{1 \leq i \leq 50}$  be independent, identically Normal distributed with mean value of  $\mu = 1$  and standard deviation of  $\sigma = 2$ . Define:

$$S_n = X_1 + X_2 + \dots + X_n \quad \text{and} \quad \bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{S_n}{n}$$

where  $n=50$ .

a) Calculate the mean and the standard deviation of  $S_n$  and  $\bar{X}_n$ . ✓ done

b) Calculate  $P(E[X_1]-1 \leq X_1 \leq E[X_1]+1)$  ✓ done

c) Calculate  $P(E[S_n]-1 \leq S_n \leq E[S_n]+1)$

d) Calculate  $P(E[\bar{X}_n]-1 \leq \bar{X}_n \leq E[\bar{X}_n]+1)$



Exercise 6.1

c) Calculate  $P(E[S_n]-1 \leq S_n \leq E[S_n]+1)$

$$E[S_n] = 50$$

$$P(E[S_n]-1 \leq S_n \leq E[S_n]+1)$$

$$= P[49 \leq S_n \leq 51]$$

$$= P\left[\frac{49-50}{\sqrt{200}} \leq \frac{S_n-50}{\sqrt{200}} \leq \frac{51-50}{\sqrt{200}}\right]$$

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

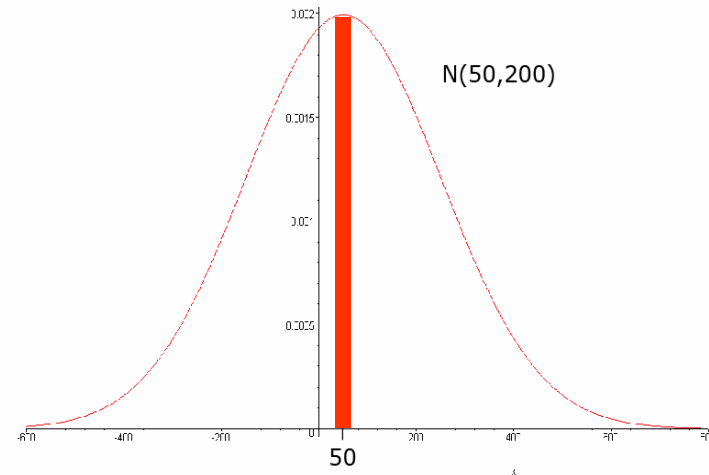
Standardization

$$= \Phi\left(\frac{1}{\sqrt{200}}\right) - \Phi\left(-\frac{1}{\sqrt{200}}\right)$$

$$= \Phi\left(\frac{1}{\sqrt{200}}\right) - \left(1 - \Phi\left(\frac{1}{\sqrt{200}}\right)\right)$$

$$= 2 \cdot \Phi\left(\frac{1}{\sqrt{200}}\right) - 1 = 2 \cdot \Phi(0.07) - 1$$

$$= 2 \cdot 0.5279 - 1 = 0.06$$



$z$	$\Phi(z)$	$z$	$\Phi(z)$
0.00	0.5000	0.50	0.6915
0.01	0.5040	0.51	0.6950
0.02	0.5080	0.52	0.6985
0.03	0.5120	0.53	0.7020
0.04	0.5160	0.54	0.7054
0.05	0.5199	0.55	0.7088
0.06	0.5239	0.56	0.7122
0.07	0.5279	0.57	0.7156
0.08	0.5319	0.58	0.7190
0.09	0.5359	0.59	0.7224
0.10	0.5398	0.60	0.7257
0.11	0.5438	0.61	0.7290
0.12	0.5478	0.62	0.7324
0.13	0.5517	0.63	0.7357

d) Calculate  $P(E[\bar{X}_n] - 1 \leq \bar{X}_n \leq E[\bar{X}_n] + 1)$

steps:

- Standardization
- Probability table

$$E[\bar{X}_n] = 1$$

$$V[\bar{X}_n] = 0.08$$

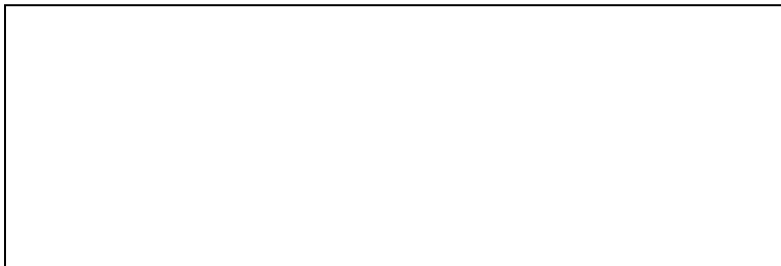
✓ Calculated last time

$$P(E[\bar{X}_n] - 1 \leq \bar{X}_n \leq E[\bar{X}_n] + 1)$$

$$= P(0 \leq \bar{X}_n \leq 2)$$

$$= P\left(\frac{0-1}{\sqrt{0.08}} \leq Z \leq \frac{2-1}{\sqrt{0.08}}\right)$$

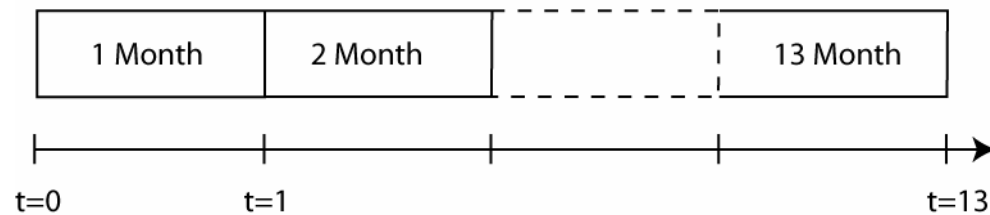
$$= P(-3.5 \leq Z \leq 3.5)$$



$z$	$z$	$\Phi(z)$	$z$	$\Phi(z)$	$z$	$\Phi(z)$
15	1.00	0.8413	1.50	0.9332	2.00	0.9772
50	1.01	0.8438	1.51	0.9345	2.10	0.9821356
85	1.02	0.8461	1.52	0.9357	2.20	0.9860966
19	1.03	0.8485	1.53	0.9370	2.30	0.9892759
54	1.04	0.8508	1.54	0.9382	2.40	0.9918025
88	1.05	0.8531	1.55	0.9394	2.50	0.9937903
23	1.06	0.8554	1.56	0.9406	2.60	0.9953388
57	1.07	0.8577	1.57	0.9418	2.70	0.9965330
90	1.08	0.8599	1.58	0.9429	2.80	0.9974449
24	1.09	0.8621	1.59	0.9441	2.90	0.9981342
57	1.10	0.8643	1.60	0.9452	3.00	0.9986501
91	1.11	0.8665	1.61	0.9463	3.10	0.9990324
24	1.12	0.8686	1.62	0.9474	3.20	0.9993129
57	1.13	0.8708	1.63	0.9484	3.30	0.9995166
89	1.14	0.8729	1.64	0.9495	3.40	0.9996631
22	1.15	0.8749	1.65	0.9505	3.50	0.9997674
54	1.16	0.8770	1.66	0.9515	3.60	0.9998409
86	1.17	0.8790	1.67	0.9525	3.70	0.9998922
17	1.18	0.8810	1.68	0.9535	3.80	0.9999377

### Exercise 7.1

The occurrence of rainfall in an area in a year may be described by a non-homogeneous Poisson process with the intensity, namely, the mean rate of occurrence of rainfall per unit time,  $\lambda(t)$ , where  $t$  is defined in the interval  $[0,13]$  and describes the time in a monthly *unit* (i.e., 4 weeks).



$$\lambda(t) = \begin{cases} \frac{2 \cdot t}{3} & \text{for } 0 \leq t \leq 3 \\ 2 & \text{for } 3 < t \leq 7 \\ \frac{13-t}{3} & \text{for } 7 < t \leq 13 \end{cases}$$

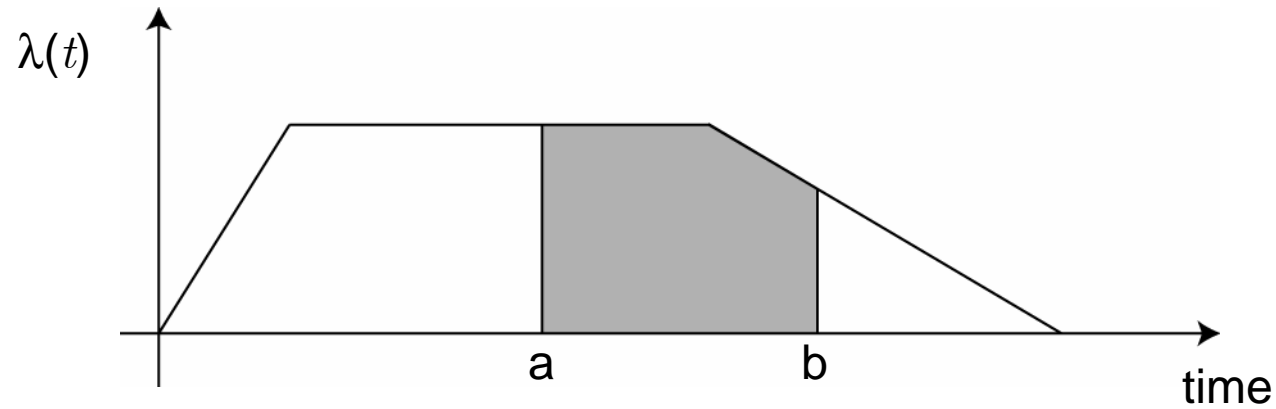
*Please correct it in tutorials book*

Hint: For a non-homogeneous Poisson process, the intensity varies with time.

The mean occurrence rate for any time interval  $(t_1, t_2)$  of the Poisson process can be described by:

$$\nu = \int_{t_1}^{t_2} \lambda(t) dt$$

Mean occurrence rate of an event per unit time  $\lambda(t)$



Mean occurrence rate in time  $[a, b]$ :  $\nu$        $\nu = \int_a^b \lambda(t) dt$

$$P_n(t) = \frac{\left( \int_0^t \lambda(\tau) d\tau \right)^n}{n!} \exp\left( -\int_0^t \lambda(\tau) d\tau \right) \longrightarrow P_n(t) = \frac{\nu^n}{n!} e^{-\nu}$$

$n$  = number of occurrence of the event  
 $(t)$  = period of interest

Exercise 7.1

- a) Calculate the probability that in the first 5 months of a year, three or more rainfalls occur.

$$\lambda(t) = \begin{cases} \frac{2 \cdot t}{3} & \text{for } 0 \leq t \leq 3 \\ 2 & \text{for } 3 < t \leq 7 \\ \frac{13-t}{3} & \text{for } 7 < t \leq 13 \end{cases}$$

Steps:

Random variable  $T$  = number of rainfalls in the first 5 months

Obtain the parameter in the Poisson process:  $\nu = \int_0^5 \lambda(t) dt = \int_0^3 \frac{2 \cdot t}{3} dt + \int_3^5 2 dt = \dots$

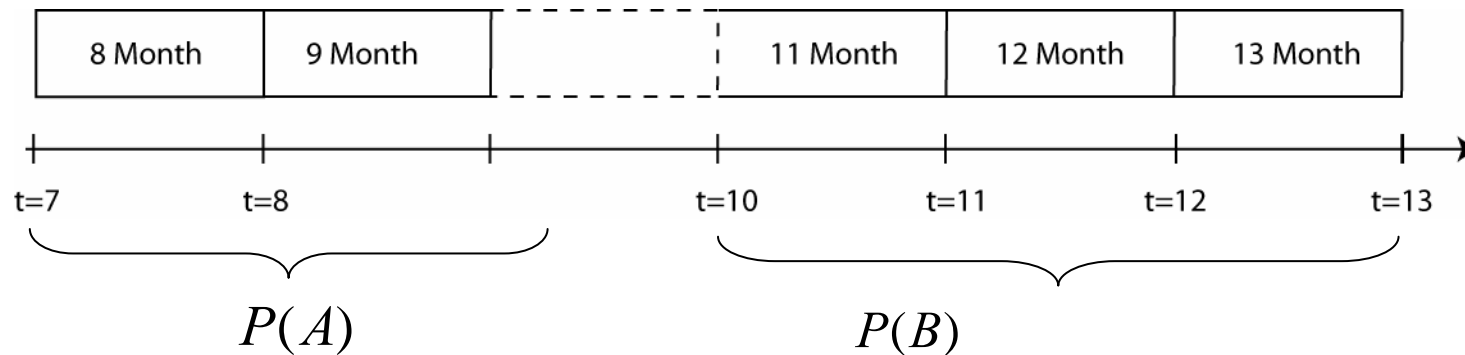
Calculate the probability:

$$\sum_{i=3}^{\infty} P_T(5) = 1 - [P_0(5) + P_1(5) + P_2(5)]$$

$$\left\{ \begin{array}{l} P_0(5) = \frac{\nu^n}{n!} e^{-\nu} = \dots \\ P_1(5) = \dots \\ P_2(5) = \dots \end{array} \right.$$

## Exercise 7.1

- b) Calculate the probability that a rainfall occurs  
at most once during the 8<sup>th</sup>, 9<sup>th</sup> and 10<sup>th</sup> month and  
at most once during the last 3 months of a year.



Let Event  $A$  represent the number of occurrences of a rainfall  
in the 8th, 9th, 10th month ( $\Delta t_a$ )

Let Event  $B$  represent the number of occurrences of a rainfall  
in the 11th, 12th and 13th month ( $\Delta t_b$ )

$$P[A \cap B] = P(A) \cdot P(B) \quad \text{Independency!}$$

## Exercise 7.1

- b) Calculate the probability that a rainfall occurs at most once during the 8<sup>th</sup>, 9<sup>th</sup> and 10<sup>th</sup> month and at most once during the last 3 months of a year.

$$\lambda(t) = \begin{cases} \frac{2 \cdot t}{3} & \text{for } 0 \leq t \leq 3 \\ 2 & \text{for } 3 < t \leq 7 \\ \frac{13-t}{3} & \text{for } 7 < t \leq 13 \end{cases}$$

**Steps**

Calculate the mean occurrence rate for each period:

$$\nu_1 = \int_7^{10} \frac{1}{3}(13-t) dt = \dots$$

$$\nu_2 = \int_{10}^{13} \frac{1}{3}(13-t) dt = \dots$$

Calculate the probability of at most one rainfall (or 0, or 1) :

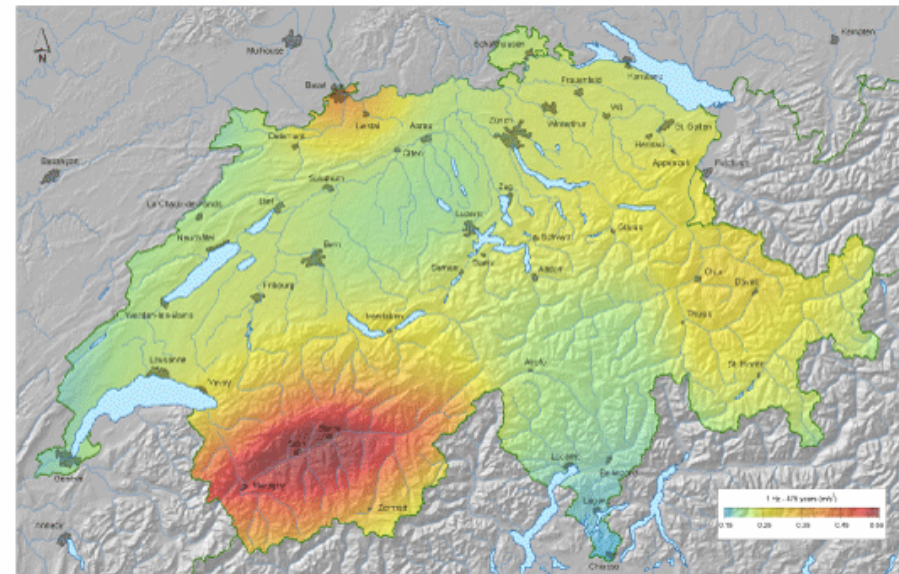
## Exercise 7.2

**Exercise 7.2**

An earthquake hazard map is often represented in terms of peak ground acceleration and a return period of 475 years is adopted in the map in many countries.

a. Show that the event with a return period of 475 years corresponds to the event whose occurrence probability is 10% in 50 years, under the assumption that an event follows a homogeneous Poisson process.

b. What is the probability that an earthquake with a return period of 475 years will occur within the next 475 years?



**Seismic Shaking Hazards in Switzerland**

**Probabilistic Seismic Hazards Assessment**  
**10% probability of being exceeded**  
**in 50 years**

[www.earthquake.ethz.ch](http://www.earthquake.ethz.ch)



## Exercise 7.2

### What is given?

Mapping of earthquakes with return periods of 475 years

Earthquake event

- follows a homogeneous Poisson process
- occurrence probability is 10% in 50 years

The 10% probability of exceeding (a certain ground motion) in 50 years maps depict an annual probability of 1 in 475 of being exceeded each year ( $\rightarrow$  90% chance that these ground motions will NOT be exceeded).

a. Verify this relationship!

Event with a return period of 475 years = Event with probability 10% in 50 years

b. What is the probability that an earthquake with a return period of 475 years will occur within the next 475 years?

**Exercise 7.2**

a. Verify this relationship!

Event with a return period of 475 years = Event with probability 10% in 50 years

Annual probability of an event with a return period of 475 years:

$$P_A(1) = \frac{1}{475}$$

Probability of occurring in the next 50 years is:

(probability of occurring in the 1st year + probability of occurring in the 2nd year + ... + ... + probability of occurring in the 50th year)

$$\sum_{i=1}^n p(1-p)^{i-1} = 1 - (1-p)^n$$

- Geometric distribution, summed up over 50 years
- Cumulative distribution function of the Geometric distribution!

$$P_A(50) = 1 - (1 - P_A(1))^{50} = 1 - \left(1 - \frac{1}{475}\right)^{50} = 0.1$$

---

**Exercise 7.2**

b. What is the probability that an earthquake with a return period of 475 years will occur within the next 475 years?

Probability of occurring in the next 475 years is:

Type of distribution:

$$P_A(475) = \input{type="text"}$$

**Exercise 7.2 (Think in another way...!)**

What is given?

Return period

So...annual probability of occurrence:

(check script after Equation D.60)

Average time till “success”- (here till occurrence)

(check script Equation D.59)

Time till and between events, described by the Poisson process, is Exponential distributed (script Equation D.64)

a.  $P[T \leq 50 \text{ Jahren}] =$

(script Table D.1)

b.  $P[T \leq 475 \text{ Jahren}] =$

## Exercise 7.2

## Exercise 7.2 (Think in another way...!)

What is given?      Return period       $T = 475$  years

So...annual probability of occurrence:  $p = \frac{1}{T} = \frac{1}{475}$  (check script after Equation D.60)

Average time till “success”- (here till occurrence)  $E[N] = \frac{1}{p} = \frac{1}{\frac{1}{475}} = 475$  (check script Equation D.59)

Time till and between events, described by the Poisson process, is Exponential distributed (script Equation D.64)

a.  $P[T \leq 50 \text{ Jahren}] = 1 - e^{-\lambda t} = 1 - e^{-\frac{1}{475} \cdot 50} = 10\%$        $\lambda = \frac{1}{E[N]} = \frac{1}{475}$  (script Table D.1)

b.  $P[T \leq 475 \text{ Jahren}] = 1 - e^{-\lambda t} = 1 - e^{-\frac{1}{475} \cdot 475} = 1 - e^{-1} = 63.2\%$

---

**Exercise 7.4 (Group exercise)**

The operational time of a diesel engine until a breakdown, is assumed to follow an Exponential distribution with mean  $\mu_T = 24$  months.

Normally such an engine is inspected every 6 months and in case that a default is observed this is fully repaired. It is assumed that a default is a serious damage that leads to breakdown if the engine is not repaired.

a. Calculate the probability that such an engine will need repair before the first inspection.

We are looking for:  $P(T \leq 6 \text{ months}) = \dots$

We know that the time until breakdown is exponentially distributed:  $F_T(t) = 1 - e^{-\lambda t}$

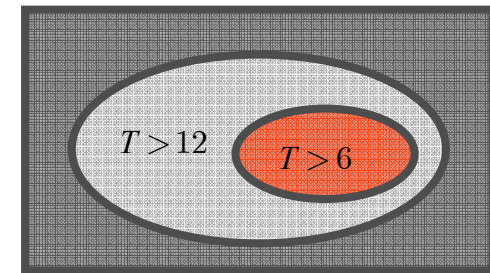
(Script page D-17)

**Exercise 7.4 (Group exercise)**

- b. Assume that the first inspection has been carried out and no repair was required. Calculate the probability that the diesel engine will operate normally until the next scheduled inspection.

$P(\text{no repair up to the second inspection} | \text{no repair at the first inspection})$

*Hint:* 
$$P[T > 12 | T > 6] = \frac{P[T > 12 \cap T > 6]}{P[T > 6]} = \frac{P[T > 12]}{P[T > 6]}$$



*Please correct in exercise tutorials book: The hint is for part b (not c)*

$$0 \leq P(T > 6) \leq P(T > 12)$$

- c. Calculate the probability that the diesel engine will fail between the first and the second inspection.

$$P[6 \leq T \leq 12]$$

---

**Exercise 7.4 (Group exercise)**

- d. A nuclear power plant owns **6 such engines**. The operational lives  $T_1, T_2 \dots T_6$  of the diesel engines are assumed statistically independent. What is the probability that at most 1 engine will need repair at the first scheduled inspection?

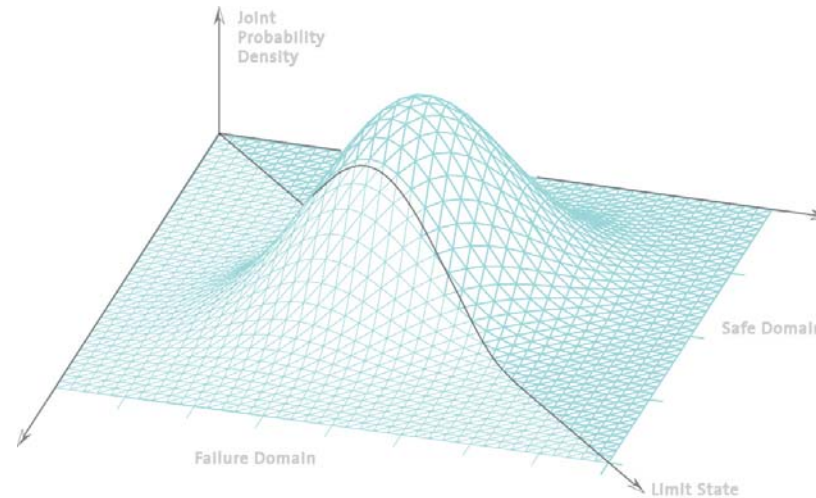
We are looking for: **At most 1** repair out of 6 engines → Binomial distribution

The probability of 1 engine needing repair at the first inspection, has been calculated in a)

- e. It is a requirement that the probability of repair at each scheduled inspection is not more than 60%. The operational lives  $T_1, T_2 \dots T_6$  of the diesel engines are assumed statistically independent. What should be the inspection interval?

$$P(\text{no engine needs repair at the time } t \text{ of the inspection}) = 0.6$$





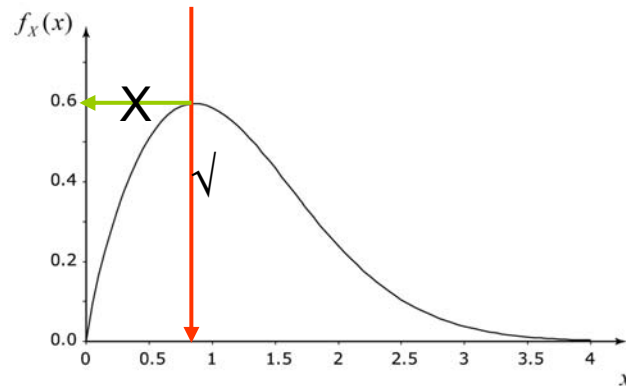
## Exercise Tutorial 8

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

## Mode

The probability density function of a continuous random variable is shown in the following figure.



What is the mode of the data set?

## Properties of estimator

The cost associated with the occurrence of an event  $A$  is a function of a random variable  $X$  with mean value  $\mu_x = 100 \text{ CHF}$  and standard deviation  $\sigma_x = 10 \text{ CHF}$ .

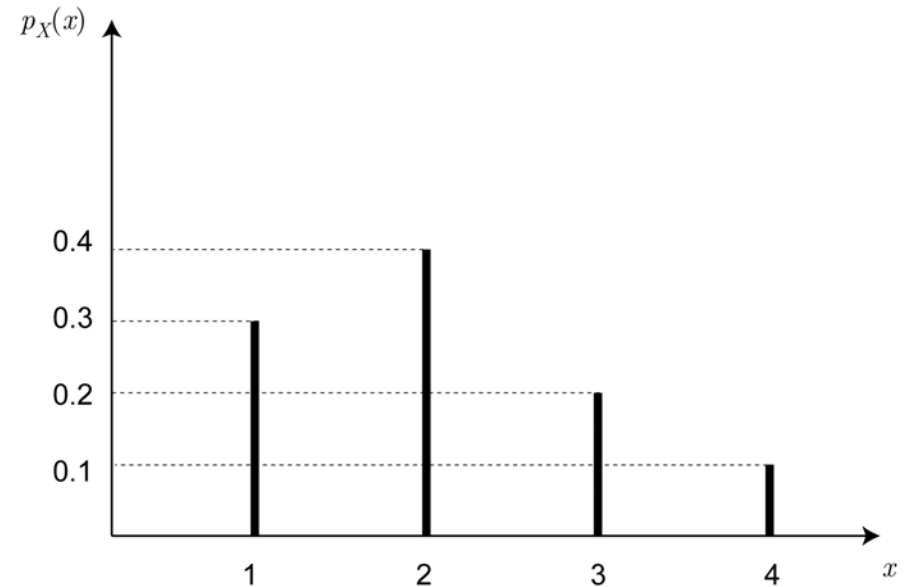
Their relation can be written as:  $C_A = a + bX + cX^2$ , where  $a$ ,  $b$  and  $c$  are constants and equal to 10, 0.5 and 1 respectively.

What is the expected cost of event  $A$ ?

## Moments – discrete distributions

A discrete random variable is represented by a probability density function:

$$p_X(x) = \begin{cases} 0.3 & , x = 1 \\ 0.4 & , x = 2 \\ 0.2 & , x = 3 \\ 0.1 & , x = 4 \\ 0 & \text{otherwise} \end{cases}$$



Calculate the mean and the standard deviation of the random variable  $X$ .

## Exercise 7.3

The annual maximum discharge of a particular river is assumed to follow the Gumbel distribution with mean =10.000 m<sup>3</sup>/s and standard deviation =3.000 m<sup>3</sup>/s.

a. Calculate the probability that the annual maximum discharge will exceed 15.000 m<sup>3</sup>/s.

The Gumbel distribution function is expressed as:

$$-\infty < x < \infty$$

$$F_X(x) = \exp(-\exp(-\alpha(x-u)))$$

$$\mu_X = u + \frac{0.577216}{\alpha}$$

$$\sigma_X = \frac{\pi}{\alpha\sqrt{6}}$$

Script Table D.2

$\mu_X$  – Mean

$\sigma_X$  – Standard deviation

$u$  – Parameter of the distribution

$\alpha$  – Parameter of the distribution

### Exercise 7.3

a. Calculate the probability that the annual maximum discharge will exceed 15.000 m<sup>3</sup>/s.

The mean  $\mu_x$  and standard deviation  $\sigma_x$  of the distribution are given.

Find the parameters  $u$  and  $\alpha$

$$\alpha = \frac{\pi}{\sigma_x \sqrt{6}} = \frac{\pi}{3.000 \sqrt{6}} = 4.2752 \cdot 10^{-4}$$

$$u = \mu_x - \frac{0,57722}{\alpha} = 10.000 - \frac{0,57722}{4.2752 \cdot 10^{-4}} = 8649,809$$

### Exercise 7.3

a. Calculate the probability that the annual maximum discharge will exceed 15.000 m<sup>3</sup>/s.

The mean  $\mu_x$  and standard deviation  $\sigma_x$  of the distribution are given.

Find the parameters  $u$  and  $\alpha$

$$\alpha = \frac{\pi}{\sigma_x \sqrt{6}} = \frac{\pi}{3.000 \sqrt{6}} = 4.2752 \cdot 10^{-4}$$

$$u = \mu_x - \frac{0,57722}{\alpha} = 10.000 - \frac{0,57722}{4.2752 \cdot 10^{-4}} = 8649,809$$

The probability that the annual maximum discharge will exceed 15.000 m<sup>3</sup>/s is:

$$P[\text{annual max} \geq 15.000] = 1 - F_x(x < 15.000) = 1 - e^{-e^{-\alpha(15.000-u)}}$$

$F_x(x) = \exp(-\exp(-\alpha(x-u)))$   
Script Table D.2

$$= 1 - F_x(x < 15.000) = 1 - e^{-e^{-4.2752 \cdot 10^{-4} (15.000 - 8649.81)}} = 1 - e^{-e^{-2.715}} = 1 - 0.9359 = 0.0641$$

## Exercise 7.3

The annual maximum discharge of a particular river is assumed to follow the Gumbel distribution with mean = 10.000 m<sup>3</sup>/s and standard deviation = 3.000 m<sup>3</sup>/s.

b. What is the discharge that corresponds to a return period of 100 years?

Recall from tutorial 6 that if the return period is  $T$ , the exceedance probability is  $p = 1/T$

For  $T = 100$  years, the exceedance probability is  $1/100 = 0.01$

Using the cumulative distribution function, find the value of  $x$  corresponding to the exceedance probability of 0.01

Script Table D.2  $F_X(x) = e^{-e^{-\alpha(x-u)}} \Rightarrow 1 - \frac{1}{100} = e^{-e^{-\alpha(x-u)}} \Rightarrow 0.99 = e^{-e^{-\alpha(x-u)}} \Rightarrow$

$$\ln(-\ln(0.99)) = -\alpha(x-u) \Leftrightarrow \frac{\ln(-\ln(0.99))}{-\alpha} + u = x \Leftrightarrow \frac{\ln(-\ln(0.99))}{-4,2752 \cdot 10^{-4}} + 8649.809 = x \Leftrightarrow$$

$$10760.08 + 8649.809 = x \Leftrightarrow 19409.889 = x$$

The discharge that corresponds to a return period of 100 years is 19410 m<sup>3</sup>/s .



### Exercise 7.3

The annual maximum discharge of a particular river is assumed to follow the Gumbel distribution with mean =10.000 m<sup>3</sup>/s and standard deviation =3.000 m<sup>3</sup>/s.

c. Find an expression for the cumulative distribution function of the river's maximum discharge over the 20 year lifetime of an anticipated flood-control project. Assume that the individual annual maxima are independent random variables.

For independent random variables, the cumulative distribution function of the largest extreme in a period of  $nT$  is:

$$F_{X,nT}^{\max}(x) = \left\{ F_{X,T}^{\max}(x) \right\}^n \quad (\text{Script Equation D.76})$$

For  $n=20$ ,

$$F_Y(y) = P[Y \leq y] = [F_X(x)]^{20} = F_Y(y) = \left( e^{-e^{-\alpha(x-u)}} \right)^{20} = F_Y(y) = e^{-20e^{-\alpha(x-u)}}$$

### Exercise 7.3

The annual maximum discharge of a particular river is assumed to follow the Gumbel distribution with mean =10.000 m<sup>3</sup>/s and standard deviation =3.000 m<sup>3</sup>/s.

d. What is the probability that the 20-year-maximum discharge will exceed 15.000 m<sup>3</sup>/s?

The probability that the 20-year-maximum discharge will exceed 15.000 m<sup>3</sup>/s is:  
(Use the cumulative distribution function derived in part c)

$$1 - F_Y(15000) = 1 - e^{-20e^{-4.2756 \cdot 10^{-4}(15000 - 8649.81)}} = 1 - e^{-1.324} = 1 - 0.266 = 0.734$$

## Exercise 7.4 (Group Exercise)

Diesel engines are used, among others, for electrical power generation. The operational time  $T$  of a diesel engine until a breakdown, is assumed to follow an Exponential distribution with mean  $\mu_T = 24$  months. Normally such an engine is inspected every 6 months and in case that a default is observed this is fully repaired. It is assumed herein that a default is a serious damage that leads to breakdown if the engine is not repaired.

## Exercise 7.4 (Group Exercise)

a. Calculate the probability that such an engine will need repair before the first inspection.

We are looking for:  $P(T \leq 6 \text{ months}) = \dots$

It is given that the time until breakdown is exponentially distributed.

The probability of failure of an event described by a variable that is exponentially distributed is given by: (Script Table D.1)

$$F_T(t) = P(T \leq t) = 1 - e^{-\lambda t} \quad \text{where } \lambda = \frac{1}{\mu_T}$$

Here  $\lambda = \frac{1}{24}$

So the probability that a repair will be required before the first inspection is:

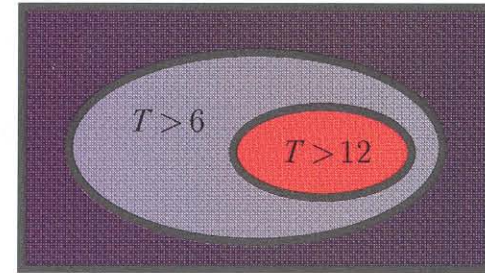
$$P(T \leq 6 \text{ months}) = 1 - e^{-\frac{1}{24} \cdot 6} = 0.221 = 22.1\%$$

### Exercise 7.4 (Group Exercise)

- b. Assume that the first inspection has been carried out and no repair was required. Calculate the probability that the diesel engine will operate normally until the next scheduled inspection.

$P(\text{no repair up to the second inspection} | \text{no repair at the first inspection}) =$

$$P[T > 12 | T > 6] = \frac{P[T > 12 \cap T > 6]}{P[T > 6]} = \frac{P[T > 12]}{P[T > 6]}$$



$$0 \leq P(T > 12) \leq P(T > 6)$$

$$\frac{P[T > 12]}{P[T > 6]} = \frac{1 - P[T \leq 12]}{1 - P[T \leq 6]} = \frac{1 - F_T(12)}{1 - F_T(6)} = \frac{1 - (1 - e^{-\frac{1}{24}12})}{1 - (1 - e^{-\frac{1}{24}6})} = \frac{0.607}{0.7788} = 0.7794 = 77.94\%$$

### Exercise 7.4 (Group Exercise)

- c. Calculate the probability that the diesel engine will fail between the first and the second inspection.

We need to find out  $P[6 \leq T \leq 12]$

$P(6 \text{ months} \leq T \leq 12 \text{ months}) =$

### Exercise 7.4 (Group Exercise)

d. A nuclear power plant owns 6 such diesel engines. The operational lives  $T_1, T_2, \dots, T_6$  of the diesel engines are assumed statistically independent. What is the probability that at most 1 engine will need repair at the first scheduled inspection?

We are looking for: **At most 1 repair out of 6 engines** → Binomial distribution

$$P(\text{max 1 engine needs repair at } t=6 \text{ months}) = \binom{6}{0} p_R^0 (1 - p_R)^6 + \binom{6}{1} p_R^1 (1 - p_R)^5$$

### Exercise 7.4 (Group Exercise)

d. A nuclear power plant owns 6 such diesel engines. The operational lives  $T_1, T_2, \dots, T_6$  of the diesel engines are assumed statistically independent. What is the probability that at most 1 engine will need repair at the first scheduled inspection?

We are looking for: **At most 1 repair out of 6 engines** → Binomial distribution

$$P(\text{max 1 engine needs repair at } t=6 \text{ months}) = \binom{6}{0} p_R^0 (1 - p_R)^6 + \binom{6}{1} p_R^1 (1 - p_R)^5$$

The probability of 1 engine needing repair at the first inspection, has been calculated in part a) as 0.221

$$\text{So } p_R = P_T(6) = 0.221$$

$$\begin{aligned} P(\text{max 1 engine needs repair at } t=6 \text{ months}) &= \binom{6}{0} 0.221^0 (1 - 0.221)^6 + \binom{6}{1} 0.221^1 (1 - 0.221)^5 = \\ &= 0.223 + 0.38 = 0.603 = 60.3\% \end{aligned}$$



### Exercise 7.4 (Group Exercise)

e. It is a requirement that the probability of repair at each scheduled inspection is not more than 60%. The operational lives  $T_1, T_2, \dots, T_6$  of the diesel engines are assumed statistically independent. What should be the inspection interval?

This can be expressed as:

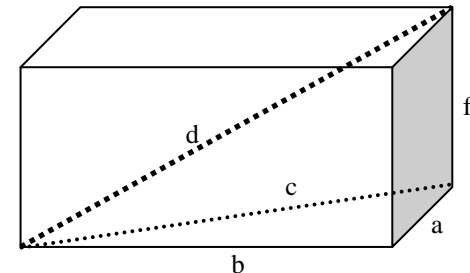
$$P(\text{no engine needs repair at the time of inspection } t) = 0.6 \Rightarrow$$

$$\binom{6}{0} (1 - e^{-\frac{1}{24}t})^0 (1 - (1 - e^{-\frac{1}{24}t}))^6 = 0.6 \Rightarrow e^{-\frac{1}{4}t} = 0.6 \Rightarrow t \approx 2 \text{ months}$$

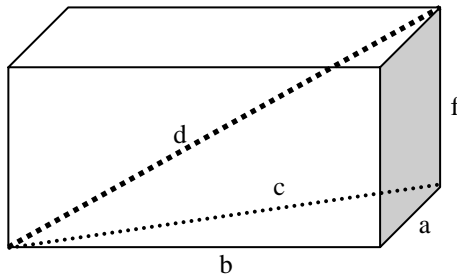
## Exercise 8.1

Consider the following three-dimensional shape. Measurements have been performed on  $a$ ,  $b$  and  $f$ . It is assumed that the measurements are performed with the same absolute error  $\varepsilon$  that is assumed to be **Normal** distributed, unbiased and with standard deviation  $\sigma_\varepsilon$ .

- Obtain the probability density function and cumulative distribution function of the error in  $d$  when this is assessed using the above measurements.
- If the same measurements are used for the assessment of  $c$ , how large is the probability that the error in  $c$  is larger than  $2.4\sigma_\varepsilon$  ?



a)



Geometrical relation:

$$d^2 = f^2 + a^2 + b^2$$

$$\varepsilon_d = \sqrt{\varepsilon_f^2 + \varepsilon_a^2 + \varepsilon_b^2}$$

$$Z \equiv \frac{\varepsilon_d}{\sigma_\varepsilon} = \sqrt{\left(\frac{\varepsilon_f}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_a}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_b}{\sigma_\varepsilon}\right)^2}$$

← Standardized!

$X_1, X_2, \dots, X_n$  follow  $N(0, 1^2)$ .  
then,

$$Y \equiv X^2 \equiv X_1^2 + X_2^2 + \dots + X_n^2$$

$Y$  follows  $\chi^2$ -distribution  
with  $n$  degrees of freedom  
and  
 $X$  follows  $\chi$ -distribution.

$Z$  follows the  $\chi$ -distribution  
with 3 degrees of freedom

$\chi$ -distribution with  $n$  degrees of freedom

Probability density function:

$$f_Z(z) = \frac{z^{(n-1)}}{2^{(n/2)-1} \Gamma(n/2)} e^{(-z^2/2)}$$

Script  
Equation E.4

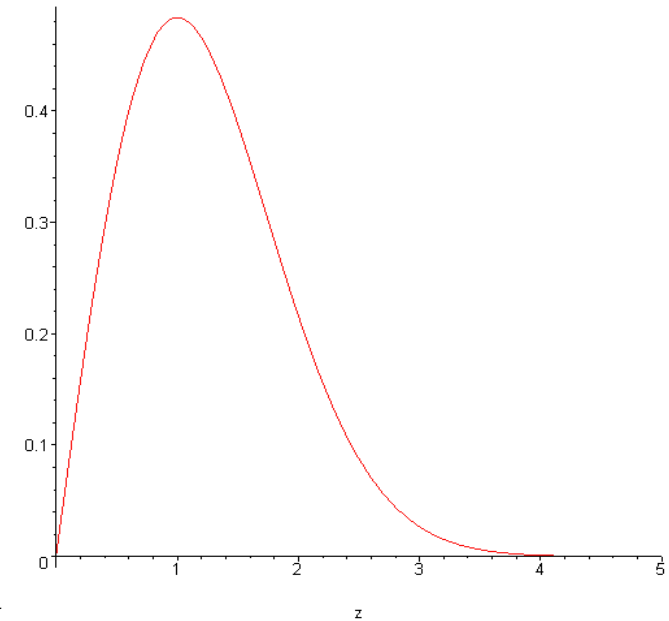
Here the number of degrees of freedom is 3 so,

$$f_Z(z) = \frac{z^{(3-1)}}{2^{(3/2)-1} \Gamma(3/2)} e^{(-z^2/2)}$$

The Gamma function can be defined as  $\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$

$$\left\{ \begin{array}{l} \Gamma(1) = 1 \\ \Gamma(1/2) = \sqrt{\pi} \\ \Gamma(a+1) = a\Gamma(a) \end{array} \right.$$

So  $\Gamma(3/2) = \frac{\sqrt{\pi}}{2}$



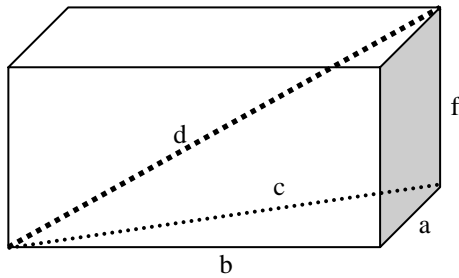
The probability density function is thus obtained as:

$$f_Z(z) = \frac{z^{(n-1)}}{2^{(n/2)-1} \Gamma(n/2)} e^{(-z^2/2)} = \frac{z^{(3-1)}}{2^{(3/2)-1} \Gamma(3/2)} e^{(-z^2/2)} = \sqrt{\frac{2}{\pi}} \cdot z^2 \cdot e^{(-z^2/2)}$$

The cumulative distribution function is:

$$F_Z(z) = \int_{-\infty}^z \sqrt{\frac{2}{\pi}} y^2 e^{(-y^2/2)} dy$$

b)



Geometrical relation:

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

$$\varepsilon_c = \sqrt{\varepsilon_a^2 + \varepsilon_b^2}$$

$$Z \equiv \frac{\varepsilon_c}{\sigma_\varepsilon} = \sqrt{\left(\frac{\varepsilon_a}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_b}{\sigma_\varepsilon}\right)^2}$$

(Z follows the  $\chi$ -distribution with 2 degrees of freedom)

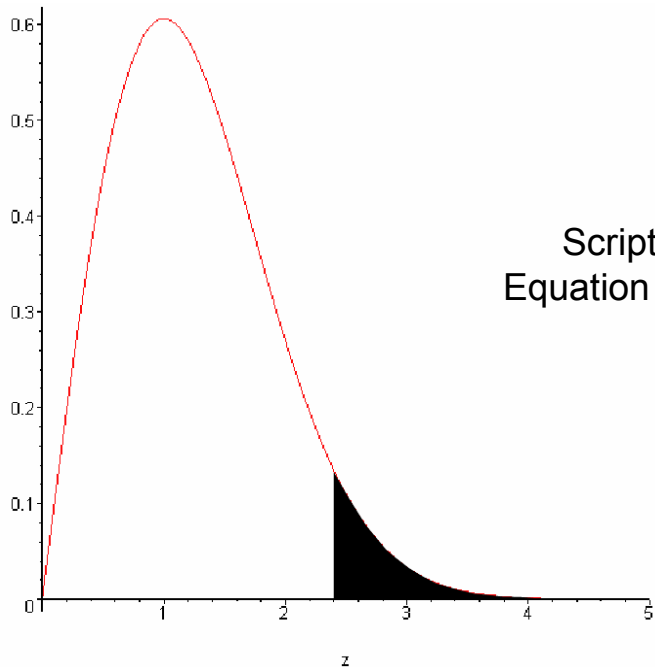
Script  
Equation E.4

$$f_Z(z) = \frac{z^{(n-1)}}{2^{n/2-1}\Gamma(n/2)} e^{(-z^2/2)} = \frac{z^{(2-1)}}{2^{2/2-1}\Gamma(2/2)} e^{(-z^2/2)}$$

$$P(\varepsilon_c \geq 2.4\sigma_\varepsilon) = P(\varepsilon_c / \sigma_\varepsilon \geq 2.4) = P(Z \geq 2.4)$$

$$= \int_{2.4}^{\infty} z e^{-z^2/2} dz$$

$$= \left\{ -e^{-z^2/2} \right\}_{2.4}^{\infty} = 5.6\%$$



## Exercise Extra

It is known from experience that the traveling time by car from Zug to the Zurich airport can be described by a Normal distributed random variable  $X$  with mean value  $\mu_x$  and standard deviation  $\sigma_x = 3$  minutes

A guy works at the airport and lives in Zug, so he travels by car everyday to his work. In the next  $n=13$  days he measured the traveling time from Zug to the airport. He obtained a sample mean:

$$\bar{x} = 22.3 \text{ minutes}$$

1. Estimate the confidence interval in which the sample average will lie in with a probability of 95%, i.e. at the 5% significance level
2. Estimate at the 5% significance level the confidence interval of the mean

## Exercise Extra

Known:

Normal distributed random variable  $X$ :  $N(\mu_X, \sigma_X)$

standard deviation :  $\sigma_X = 3$  minutes

Number of measurements:  $n = 13$  days

confidence level:  $\alpha = 0.05$

1. Estimate the confidence interval in which the sample average will lie in with a probability of 95%, i.e. at the 5% significance level

$$P\left(-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} < k_{\alpha/2}\right) = 1 - \alpha \Rightarrow P\left(-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{3}{\sqrt{13}}} < k_{\alpha/2}\right) = 1 - 0.05$$

Script Equation E.22



## Exercise Extra

Known:

Normal distributed random variable  $X$ :  $N(\mu_X, \sigma_X)$

standard deviation :  $\sigma_X = 3$  minutes

Number of measurements:  $n = 13$  days

confidence level:  $\alpha = 0.05$

1. Estimate the confidence interval in which the sample average will lie in with a probability of 95%, i.e. at the 5% significance level

$$P\left(-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} < k_{\alpha/2}\right) = 1 - \alpha \Rightarrow P\left(-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{3}{\sqrt{13}}} < k_{\alpha/2}\right) = 1 - 0.05$$

Script Equation E.22

How to estimate  $k_{\alpha/2}$

$$k_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \Phi^{-1}\left(1 - \frac{0.05}{2}\right) = \Phi^{-1}(0.975)$$

Script Equation E.24

## Exercise Extra

Known:

Normal distributed random variable  $X$ :  $N(\mu_X, \sigma_X)$

standard deviation :  $\sigma_X = 3$  minutes

Number of measurements:  $n = 13$  days

Confidence level:  $\alpha = 0.05$

1. Estimate the confidence interval in which the sample average will lie in with a probability of 95%, i.e. at the 5% significance level

$$P\left(-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} < k_{\alpha/2}\right) = 1 - \alpha \Rightarrow P\left(-k_{\alpha/2} < \frac{\bar{X} - \mu_X}{\frac{3}{\sqrt{13}}} < k_{\alpha/2}\right) = 1 - 0.05$$

Script Equation E.22

$$\Rightarrow P\left(-1.96 < \frac{\bar{X} - \mu_X}{\frac{3}{\sqrt{13}}} < 1.96\right) = 0.95 \Rightarrow P\left(-1.63 < \bar{X} - \mu_X < 1.63\right) = 0.95$$

## Exercise Extra

Known:

Normal distributed random variable  $X$ :  $N(\mu_X, \sigma_X)$

standard deviation :  $\sigma_X = 3$  minutes

Number of measurements:  $n = 13$  days

confidence level:  $\alpha = 0.05$

2. Estimate at the 5% significance level the confidence interval of the mean

$$P(-1.63 < \bar{X} - \mu_X < 1.63) = 0.95$$

Measurements are made:  $\bar{X} \rightarrow \bar{x} = 22.3$  minutes

## Exercise Extra

Known:

Normal distributed random variable  $X$ :  $N(\mu_X, \sigma_X)$

standard deviation :  $\sigma_X = 3$  minutes

Number of measurements:  $n = 13$  days

confidence level:  $\alpha = 0.05$

2. Estimate at the 5% significance level the confidence interval of the mean

$$P(-1.63 < \bar{X} - \mu_X < 1.63) = 0.95$$

Measurements are made:  $\bar{X} \rightarrow \bar{x} = 22.3$  minutes

$$\begin{aligned} P(-1.63 < \bar{x} - \mu_X < 1.63) = 0.95 &\equiv -1.63 - \bar{x} < -\mu_X < 1.63 - \bar{x} \\ &-1.63 - 22.3 < -\mu_X < 1.63 - 22.3 \\ &-23.93 < -\mu_X < -20.67 \\ &20.67 < \mu_X < 23.93 \end{aligned}$$

With a 95% probability the interval [20.67,23.93] contains the value of the true mean

## Ppt Lecture 9 Slide 22

- If we then observe that the sample mean is equal to e.g. 400 we know that **with a probability of 0.95 the following interval will contain the value of the true mean**

$$P[-9.8 < \bar{X} - \mu_X < 9.8] = 0.95$$

~~$$P[390.2 < \mu_X < 409.8] = 0.95$$~~

$$390.2 < \mu_X < 409.8$$

- Typically confidence intervals are considered for mean values, variances and characteristic values – e.g. lower percentile values.
- Confidence intervals represent/describe the (statistical) uncertainty due to lack of data.

## Exercise 8.4

In a laboratory, 30 measurements are taken to control the water quality every day.

Each measurement result is assumed to follow the Normal distribution with a mean of  $\mu = 23 \text{ ng/ml}$  and a standard deviation of  $\sigma = 4.3 \text{ ng/ml}$

a. How large is the probability that a measurement result is less than  $23 \text{ ng/ml}$  ?

How large is the probability that a measurement result lies in the interval  $[19.5 \text{ ng/ml}; 20.5 \text{ ng/ml}]$  ?

b. How large is the probability of the daily mean being less than  $20 \text{ ng/ml}$  ?

## Exercise 8.4

In a laboratory, 30 measurements are taken to control the water quality every day.

Each measurement result is assumed to follow the Normal distribution with a mean of  $\mu = 23 \text{ ng/ml}$  and a standard deviation of  $\sigma = 4.3 \text{ ng/ml}$

a. How large is the probability that a measurement result is less than  $23 \text{ ng/ml}$  ?

$$P[X < 23] = P\left[\frac{X - \mu_X}{\sigma_X} < \frac{23 - \mu_X}{\sigma_X}\right] = P\left[\frac{X - 23}{4.3} < \frac{23 - 23}{4.3}\right] = \Phi(0) = 0.5$$

## Exercise 8.4

In a laboratory, 30 measurements are taken to control the water quality every day.

Each measurement result is assumed to follow the Normal distribution with a mean of  $\mu = 23 \text{ ng/ml}$  and a standard deviation of  $\sigma = 4.3 \text{ ng/ml}$

a. How large is the probability that a measurement result lies in the interval  $[19.5 \text{ ng/ml}; 20.5 \text{ ng/ml}]$  ?

$$\begin{aligned} P[19.5 < X \leq 20.5] &= P\left[\frac{19.5 - 23.0}{4.3} < \frac{X - 23.0}{4.3} \leq \frac{20.5 - 23.0}{4.3}\right] \\ &= \Phi(-0.58) - \Phi(-0.81) = \dots \end{aligned}$$

Check Table of standard  
Normal distribution....

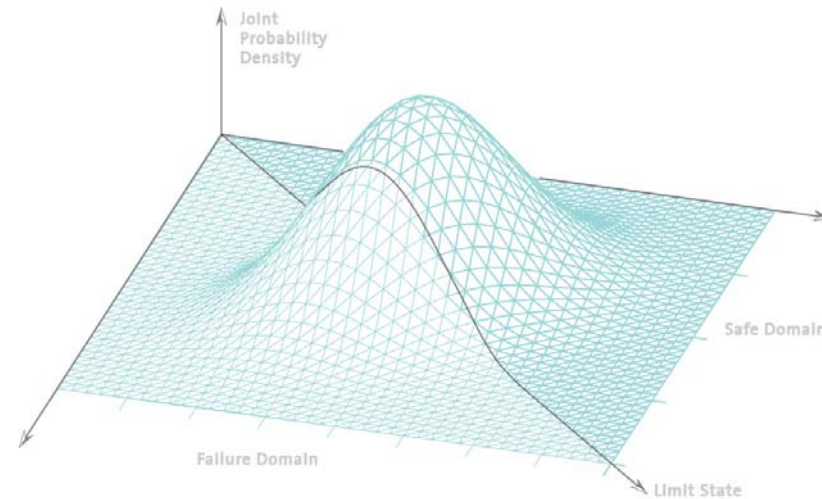


## Exercise 8.4

In a laboratory, 30 measurements are taken to control the water quality every day. Each measurement result is assumed to follow the Normal distribution with a mean of  $\mu = 23 \text{ ng/ml}$  and a standard deviation of  $\sigma = 4.3 \text{ ng/ml}$

b. How large is the probability of the daily sample mean being less than  $20 \text{ ng/ml}$  ?





## Exercise Tutorial 9

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

### Exercise 7.3

(from the last tutorial: please correct the mistake, also in the solution given in the script)

c. Find an expression for the cumulative distribution function of the river's maximum discharge over the 20 year lifetime of an anticipated flood-control project. Assume that the individual annual maxima are independent random variables.

For independent random variables, the cumulative distribution function of the largest extreme in a period of  $nT$  is:

$$F_{X,20T}^{\max}(x) = \{ F_{X,T}^{\max}(x) \}^{20} = \left( e^{-e^{-\alpha(x-u)}} \right)^{20}$$

$$\cancel{F_{X,20T}^{\max}(x) = e^{-e^{-20\alpha(x-u)}}} \longrightarrow F_{X,20T}^{\max}(x) = e^{-20e^{-\alpha(x-u)}}$$

$$1 - F_{X,20T}^{\max}(x) = 1 - F_{X,20T}^{\max}(15000) = 1 - e^{-20e^{-4.2756 \cdot 10^{-4}(15000 - 8649.81)}}$$

$$= 1 - e^{-20e^{-2.7150}} = 1 - e^{-1.324} = 1 - 0.266 = 0.734$$

## What is hypothesis testing?

An example:

1. We want to judge whether a coin is fair or not.
2. Therefore we toss the coin 20 times and count the number of heads/tails.
3. As a result, we obtain 16 heads.
4. We had already decided that we judge the coin to be fair if the number of heads is between 6 and 14 and otherwise we judge the coin to be not fair.
5. In accordance with the rule we had decided, we conclude that the coin is not fair.

What is important in this example are:

1. how many trials are necessary to judge the hypothesis?
2. how should we decide the operating rule?

## Truth, hypothesis and judgment

Assume that the world consists of two sets,  $\Theta_0$  and  $\Theta_1$ .  $\Theta_0$  and  $\Theta_1$  are complementary and one of them is the truth.

We want to judge in which set the truth belongs. Since we can never know the truth, what we can do is to believe one of them to be the truth by some artificial rule which we create by ourselves. **So it is possible that we might make misjudgments!!**

Hypothesis:

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta_1$$

Operating rule:

$$\mathbf{x} \in C \Rightarrow \text{reject } H_0 \text{ (= accept } H_1)$$

$$\mathbf{x} \in \bar{C} \Rightarrow \text{accept } H_0$$

$\mathbf{x}$  is the value of sample statistic

$C$  is the critical region

		Truth	
		$\Theta_0$	$\Theta_1$
Hypothesis	Judgment		
	Accept $H_0$	Very good!	Type 2 error
	Accept $H_1$	Type 1 error	Very good!

Control the type 1 error

The probability that we make type 1 error is  $P[\mathbf{X} \in C \mid \theta \in \Theta_0]$

We use the term “**level of significance**”  $\alpha$  in order to state to which degree we allow the type 1 error as:

$$P[\mathbf{X} \in C \mid \theta \in \Theta_0] = \alpha$$

or equivalently

$$P[\mathbf{X} \in \bar{C} \mid \theta \in \Theta_0] = 1 - \alpha$$

Finally we can create the operating rule in accordance with the above equations.

We can use the knowledge we have gained about probability to create the rule!!!

---

Coin problem again!

We want to judge whether the coin is fair or not.

The world consists of

$$\Theta_0 = \{p \mid p = 0.5\} = \{0.5\}$$

$$\Theta_1 = \{p \mid p \neq 0.5\} = \{\text{the set of all possible values between 0 and 1 excluding 0.5}\}$$

Hypotheses are

$$H_0: p = 0.5$$

$$H_1: p \neq 0.5 \quad (p \text{ is the probability that the coin lands with the head up.})$$

We toss the coin 20 times and count the number of heads.

Statistic utilized for this purpose : **the number of heads =  $N$**

Note that  $N$  is a random variable **before** we begin tossing the coin.

We formulate a suitable operating rule as:

The null-hypothesis cannot be rejected at the  $\alpha$  level of significance if

$$10 - k \leq N \leq 10 + k, \quad k \text{ is an integer } > 0$$

We select a level of significance  $\alpha = 0.05$ , meaning that we accept the occurrence of a type 1 error with a probability of 5%

We thus need to judge

$$P[\mathbf{X} \in \bar{C} \mid p = 0.5] = 1 - \alpha$$

The operating rule hence becomes:

$$P[10 - k \leq N \leq 10 + k \mid p = 0.5] = 1 - \alpha = 1 - 0.05 = 0.95$$

The probability that  $N$  is between  $10 - k$  and  $10 + k$  given  $p = 0.5$  is

$$P[10 - k \leq N \leq 10 + k \mid p = 0.5] = \sum_{i=10-k}^{10+k} \binom{20}{i} 0.5^i (1-0.5)^{20-i} = 0.95$$

$\mathbf{X} \in \bar{C}$

$\theta \in \Theta_0$

By solving this equation we obtain  $k = 4$ .



Therefore the operating rule is obtained as:

“ $H_0$  cannot be rejected if the realization of  $N$  is between 6 and 14, otherwise  $H_0$  can be rejected, i.e.  $H_1$  can be accepted”

## Summary

We want to judge whether the coin is fair,  $H_0: p = 0.5$

We obtained 16 heads out of 20.

In accordance with the operating rule, we reject  $H_0$  (= accept  $H_1$ ).

We state this: “the hypothesis  $H_0$  is rejected at the 5% level of significance.”

As a conclusion, the coin is not fair.

## General procedure for hypothesis testing

1. Specify what you want to judge as the null hypothesis:  $H_0$  (complimentary set is  $H_1$ )
2. Determine the condition of sampling (what kind of and how many data?)
3. Create the operating rule (as a function of sample statistic)
4. Choose the level of significance  $\alpha$  and evaluate the operating rule
5. Execute the sampling and obtain the result.
6. Judge the hypothesis  $H_0$  depending on the result.

### Exercise 8.2

In order to check the quality of the concrete production in a construction site, the compressive strength of the concrete produced were measured. Based on previous experience the compressive strength is assumed to follow the Normal distribution and its variance is assumed known and equal to  $16.36 \text{ (MPa)}^2$ . Acceptance criterion of the concrete production is that the mean of the population of the produced concrete is ~~equal to or more than 30 (MPa)~~ **equal to  $30 \pm \Delta \text{ (MPa)}$** ,

Make this correction in the exercise question; the solution is correct

15 samples have been tested and the results are shown in Table 8.2.1.

Should the quality of the concrete production be accepted?  
 Test the hypothesis at significance levels of 10% and 1%.

Sample Number	Compressive Strength [MPa]	Sample Number	Compressive Strength [MPa]
1	24.4	9	30.3
2	26.5	10	39.7
3	27.8	11	38.4
4	29.2	12	33.3
5	39.2	13	33.5
6	37.8	14	28.1
7	35.1	15	34.6
8	30.8		

1. Specify what you want to judge as the null hypothesis:  $H_0$

$H_0$ : The mean of the compressive strength  $\mu$  is **equal to** 30 [MPa]

$H_1$ : The mean of the compressive strength  $\mu$  is **not equal to** 30 [MPa]

## 2. Determine the condition of sampling (what kind of and how many data?)

15 concrete samples are taken from the construction site and their compressive strengths are measured .

Number of samples:  $N = 15$

---

3. Create the operating rule (as a function of sample statistic)

$$30 - \Delta \leq \bar{X} \leq 30 + \Delta$$

where  $\bar{X}$  is the sample mean value of the compressive strength

#### 4. Choose the level of significance $\alpha$ and evaluate the operating rule

Level of significance  $\alpha = 0.1$  (10%)

The sample statistic used is the sample mean,  $\bar{X}$

The operating rule is evaluated as

$$P(30 - \Delta \leq \bar{X} \leq 30 + \Delta) = 0.9 \Rightarrow P\left(\frac{(30 - \Delta) - \mu}{\sigma / \sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{(30 + \Delta) - \mu}{\sigma / \sqrt{n}}\right) = 0.9$$

$$P\left(\frac{-\Delta}{\sqrt{16.36/15}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{\Delta}{\sqrt{16.36/15}}\right) = 0.9 \Rightarrow \Phi\left(\frac{\Delta}{\sqrt{16.36/15}}\right) = 0.95$$

By solving this equation we obtain  $\Delta = 1.72$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{15} X_i$$

is Normal distributed  
with mean = 30  
variance = 16.36/15



#### 4. Choose the level of significance $\alpha$ and evaluate the operating rule

Level of significance  $\alpha = 0.1$  (10%)

The sample statistic used is the sample mean,  $\bar{X}$

The operating rule is evaluated as

$$P(30 - \Delta \leq \bar{X} \leq 30 + \Delta) = 0.9 \Rightarrow P\left(\frac{(30 - \Delta) - \mu}{\sigma / \sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{(30 + \Delta) - \mu}{\sigma / \sqrt{n}}\right) = 0.9$$

$$P\left(\frac{-\Delta}{\sqrt{16.36/15}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{\Delta}{\sqrt{16.36/15}}\right) = 0.9 \Rightarrow \Phi\left(\frac{\Delta}{\sqrt{16.36/15}}\right) = 0.95$$

By solving this equation we obtain  $\Delta = 1.72$  MPa

The operating rule is hence obtained as:

“ $H_0$  cannot be rejected at the 10% level of significance if the sample mean lies between **28.28** MPa and **31.72** MPa, otherwise  $H_0$  can be rejected (and  $H_1$  can be accepted)”

$\bar{X} = \frac{1}{n} \sum_{i=1}^{15} X_i$   
is Normal distributed  
with mean = 30  
variance = 16.36/15

## 5. Execute the sampling and obtain the result.

The sample mean is

$$\begin{aligned}\bar{x} &= \frac{1}{15} (24.4 + 26.5 + 27.8 + 29.2 + 39.2 + 37.8 + 35.1 + 30.8 + \\ &\quad 30.3 + 39.7 + 38.4 + 33.3 + 33.5 + 28.1 + 34.6) \\ &= 32.58 \text{ MPa}\end{aligned}$$

6. Judge the hypothesis  $H_0$  depending on the result.

The sample mean is 32.58 MPa

Since the sample mean lies **outside** the interval  $[28.28MPa \leq \bar{x} \leq 31.72MPa]$ ,  
the null hypothesis is rejected at the **10% level of significance**.

Following the same procedure, the interval for accepting the null hypothesis at the **1% level of significance** is obtained as  $[27.31MPa \leq \bar{X} \leq 32.69MPa]$ .

In this case , since the sample mean (32.58 MPa) is **within** this interval, the null hypothesis cannot be rejected at the **1% level of significance**.

### Exercise 8.5

The weekly working hours in the building industry were decided to be reduced by 2 hours. On a construction site it is to be tested whether this new rule is applied or not since the labor union insists that the workers work the same hours as before the reduction. 9 workers were selected randomly and their weekly working hours were measured before ( $X$ ) and ( $Y$ ) after the reduction of the working hours. It is assumed that  $X$  and  $Y$  are Normal distributed. The variance of the weekly working hours before and after the reduction is assumed to be  $\sigma_X^2 = \sigma_Y^2 = 9.5 \text{ hours}^2$ .

a) Can it be said, based on the data, that the mean of the working hours before the reduction is 40 hours per week at 5% significance level.

Number of workers	Working time Before reduction	Working time After reduction
1	38	38
2	41	39
3	40	41
4	42	39
5	43	40
6	40	40
7	39	39
8	37	38
9	43	40

1. Specify what you want to judge as the null hypothesis:  $H_0$

$H_0$ : The mean of the working hours per week before the reduction  $\mu_X$  is **equal to** 40 hours per week.

$H_1$ : The mean of the working hours per week before the reduction  $\mu_X$  is **not equal to** 40 hours per week.

## 2. Determine the condition of sampling (what kind of and how many data?)

9 workers are selected randomly and their weekly working hours are measured .

Number of samples:  $N = 9$

---

3. Create the operating rule (as a function of sample statistic)

$$40 - \Delta \leq \bar{X} \leq 40 + \Delta$$

where  $\bar{X}$  is the sample mean value of the weekly working hours of the workers before the reduction

#### 4. Choose the level of significance $\alpha$ and evaluate the operating rule

Level of significance  $\alpha = 0.05$  (5%)

The sample statistic used is the sample mean,  $\bar{X}$

The operating rule is evaluated as

$$P(40 - \Delta \leq \bar{X} \leq 40 + \Delta) = 0.95 \Rightarrow P\left(\frac{(40 - \Delta) - \mu}{\sigma / \sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{(40 + \Delta) - \mu}{\sigma / \sqrt{n}}\right) = 0.95$$

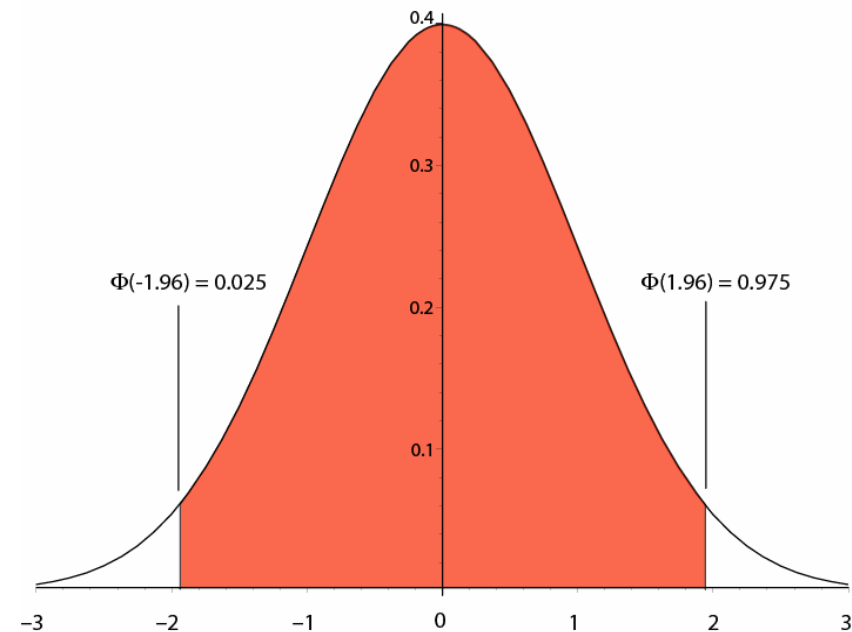
$$P\left(\frac{-\Delta}{\sqrt{9.5/9}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{\Delta}{\sqrt{9.5/9}}\right) = 0.95$$

$$\Rightarrow \Phi\left(\frac{\Delta}{\sqrt{9.5/9}}\right) = 0.975$$

By solving this equation we obtain  
 $\Delta = 2.01$  hours

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{15} X_i$$

is Normal distributed  
 with mean = 40  
 variance = 9.5/9





#### 4. Choose the level of significance $\alpha$ and evaluate the operating rule

Level of significance  $\alpha = 0.05$  (5%)

The sample statistic used is the sample mean,  $\bar{X}$

The operating rule is evaluated as

$$P(40 - \Delta \leq \bar{X} \leq 40 + \Delta) = 0.95 \Rightarrow P\left(\frac{(40 - \Delta) - \mu}{\sigma / \sqrt{n}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{(40 + \Delta) - \mu}{\sigma / \sqrt{n}}\right) = 0.95$$

$$P\left(\frac{-\Delta}{\sqrt{9.5/9}} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq \frac{\Delta}{\sqrt{9.5/9}}\right) = 0.95 \Rightarrow \Phi\left(\frac{\Delta}{\sqrt{9.5/9}}\right) = 0.975$$

$\bar{X} = \frac{1}{n} \sum_{i=1}^{15} X_i$   
is Normal distributed  
with mean = 40  
variance = 9.5/9

By solving this equation we obtain  $\Delta = 2.01$  hours

The operating rule is hence obtained as:

“ $H_0$  cannot be rejected at the 5% level of significance if the sample mean of the weekly working hours of the 9 workers lies between **37.99** hours and **42.01** hours, otherwise  $H_0$  can be rejected (and  $H_1$  can be accepted)”

## 5. Execute the sampling and obtain the result.

The sample mean is

$$\begin{aligned}\bar{x} &= \frac{1}{9}(39 + 41 + 40 + 42 + 43 + 40 + 39 + 37 + 43) \\ &= 40.33 \text{ hours / week}\end{aligned}$$

6. Judge the hypothesis  $H_0$  depending on the result.

The sample mean is 40.33 hours/week

Since the sample mean lies **within** the interval  $[37.99\text{hours} \leq \bar{x} \leq 42.01\text{hours}]$ ,  
the null hypothesis cannot be rejected at the **5% level of significance**.

### Exercise 8.5

The weekly working hours in the building industry were decided to be reduced by 2 hours. On a construction site it is to be tested whether this new rule is applied or not since the labor union insists that the workers work the same hours as before the reduction. 9 workers were selected randomly and their weekly working hours were measured before ( $X$ ) and ( $Y$ ) after the reduction of the working hours. It is assumed that  $X$  and  $Y$  are Normal distributed. The variance of the weekly working hours before and after the reduction is assumed to be  $\sigma_X^2 = \sigma_Y^2 = 9.5 \text{ hours}^2$ .

b) Test the claim of the labor union at the 5% significance level.

Number of workers	Working time Before reduction	Working time After reduction
1	38	38
2	41	39
3	40	41
4	42	39
5	43	40
6	40	40
7	39	39
8	37	38
9	43	40

## 1. Specify what you want to judge as the null hypothesis: $H_0$

The variables  $X$  and  $Y$  refer to the weekly working hours **before** and **after** the reduction of the working hours respectively.

The claim of the labor union is taken as the null hypothesis

$$H_0 : \mu_X = \mu_Y \Leftrightarrow \mu_X - \mu_Y = 0$$

$$H_1 : \mu_X \neq \mu_Y$$

## 2. Determine the condition of sampling (what kind of and how many data?)

9 workers are selected randomly and their weekly working hours **before** and **after** the reduction are measured .

Number of samples:  $N = 9$

### 3. Create the operating rule (as a function of sample statistic)

Test two datasets in order to compare two populations.

We are interested in the difference  $\mu_X$  and  $\mu_Y$ .

So we adopt  $Z = \bar{X} - \bar{Y}$  as the sampling statistic for judgment.

Then Z has the following properties:

$$\mu_Z = \mu_{\bar{X}} - \mu_{\bar{Y}} = \mu_X - \mu_Y; \quad \sigma_Z^2 = \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 = \frac{\sigma_X^2}{k} + \frac{\sigma_Y^2}{l}$$

According to the null hypothesis,  $\mu_X = \mu_Y \Rightarrow \mu_Z = \mu_{\bar{X}} - \mu_{\bar{Y}} = \mu_X - \mu_Y = 0$

$$\text{Also } \sigma_Z^2 = \sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2 = \frac{\sigma_X^2}{k} + \frac{\sigma_Y^2}{l} = \frac{9.5}{9} + \frac{9.5}{9} = 2.11$$

What is more, Z is normally distributed.

$$\bar{X} = \frac{1}{k} \sum_{i=1}^k X_i$$
$$\bar{Y} = \frac{1}{l} \sum_{i=1}^l Y_i$$

### 3. Create the operating rule (as a function of sample statistic)

$$\bar{X} - \bar{Y} \leq \Delta$$

$$Z = \bar{X} - \bar{Y} \Rightarrow Z \leq \Delta$$

where  $\bar{X}$  and  $\bar{Y}$  are the sample mean values of the weekly working hours of the workers **before** and **after** the reduction respectively and  $Z$  is the difference between  $\bar{X}$  and  $\bar{Y}$



#### 4. Choose the level of significance $\alpha$ and evaluate the operating rule

Level of significance  $\alpha = 0.05$  (5%)

The sample statistic is the difference between the sample means,  $Z = \bar{X} - \bar{Y}$

The operating rule is evaluated as

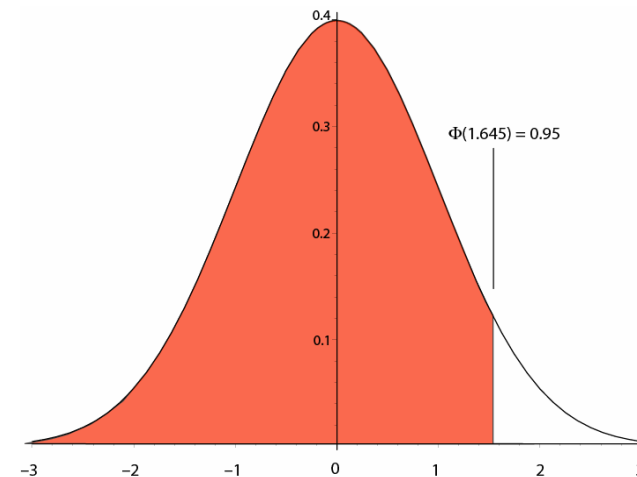
$$P(Z \leq \Delta) = 1 - 0.05 = 0.95$$

$$P\left(\frac{Z - \mu_Z}{\sigma_Z} \leq \frac{\Delta - \mu_Z}{\sigma_Z}\right) = 0.95 \Rightarrow P\left(\frac{Z - 0}{\sqrt{2.11}} \leq \frac{\Delta - 0}{\sqrt{2.11}}\right) = 0.95$$

$$\Rightarrow \Phi\left(\frac{\Delta}{\sqrt{2.11}}\right) = 0.95$$

**Z is Normal distributed  
with mean = 0  
variance = 2.11**

By solving this equation we obtain  
 $\Delta = 2.39$  hours



#### 4. Choose the level of significance $\alpha$ and evaluate the operating rule

Level of significance  $\alpha = 0.05$  (5%)

The sample statistic is the difference between the sample means,  $Z = \bar{X} - \bar{Y}$

The operating rule is evaluated as

$$P(Z \leq \Delta) = 1 - 0.05 = 0.95$$

$$P\left(\frac{Z - \mu_Z}{\sigma_Z} \leq \frac{\Delta - \mu_Z}{\sigma_Z}\right) = 0.95 \Rightarrow P\left(\frac{Z - 0}{\sqrt{2.11}} \leq \frac{\Delta - 0}{\sqrt{2.11}}\right) = 0.95$$

$$\Rightarrow \Phi\left(\frac{\Delta}{\sqrt{2.11}}\right) = 0.95$$

By solving this equation we obtain  $\Delta = 2.39$  hours

The operating rule is hence obtained as:

“ $H_0$  cannot be rejected at the 5% level of significance if the difference in the sample mean of the weekly working hours before and after the reduction is smaller than or equal to **2.39** hours, otherwise  $H_0$  can be rejected (and  $H_1$  can be accepted)”

*Z is Normal distributed  
with mean = 0  
variance = 2.11*

## 5. Execute the sampling and obtain the result.

The difference in the sample mean values is

$$\bar{x} = 40.33 \text{ hours}$$

$$\bar{y} = 39.33 \text{ hours}$$

$$z = \bar{x} - \bar{y} = 1 \text{ hour}$$

6. Judge the hypothesis  $H_0$  depending on the result.

The difference in the sample mean is  $z = 1$  hour

Since the sample mean lies **within** the interval  $[z \leq 2.39 \text{ hours}]$ ,

the null hypothesis cannot be rejected at the **5% level of significance**.

## Exercise 8.6

Table 8.3 provides a number of data on the daily traffic flow in Rosengartenstrasse in Zürich.

- a) Produce the probability paper for the triangular distribution given by

$$f_X(x) = \begin{cases} \frac{2}{10000^2} x & 0 \leq x \leq 10000 \\ 0 & \textit{otherwise} \end{cases}$$

- b) Check if the daily traffic flow is triangularly distributed with the help of the probability paper.

Day ( $i$ )	Number of cars
1	3600
2	4500
3	5400
4	6500
5	7000
6	7500
7	8700
8	9000
9	9500

What is a probability paper?

We are interested in whether the sample originates from the given distribution.

We then create the probability paper. If the sample comes from the distribution, the plotted points on the paper are on the straight line.

Probability paper, P-P plot Q-Q plot

All are useful to check the suitability of the modeling (assumed distribution)

But they differ. They are applied...

Before parameter estimation → Probability paper

- ✓ You do not need to estimate the parameters in advance.
- ✓ The suitability is checked whether or not the data on a straight line (not necessarily tangent of 45 degree)

After parameter estimation → P-P plot, Q-Q plot

- ✓ You need to estimate the parameters in advance.
- ✓ The suitability is checked whether or not the data on a straight line whose tangent is 45 degree.

The probability density function and the cumulative distribution functions are

$$f_X(x) = \begin{cases} \frac{2}{10000^2} x & 0 \leq x \leq 10000 \\ 0 & \textit{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & 0 \leq x \\ \left(\frac{x}{10000}\right)^2 & 0 < x \leq 10000 \\ 1 & x > 10000 \end{cases}$$

Taking the square root of both sides of the first equation equation, a linear relationship between  $F_X(x)$  and the square root of  $x$  is obtained:

$$F_X(x) = \left(\frac{x}{10000}\right)^2 \Leftrightarrow \sqrt{F_X(x)} = \frac{x}{10000}$$



For values of the cumulative distribution function in the interval  $[0;1]$ , the following table for the values of  $\sqrt{F_X(x)}$  and  $F_X(x)$  is obtained

$\sqrt{F_X(x)}$	$F_X(x)$
0	0
0.31	0.1
0.45	0.2
0.55	0.3
0.63	0.4
0.71	0.5
0.77	0.6
0.84	0.7
0.89	0.8
0.94	0.9
1.0	1.0

## Empirical distribution function

We give the probability for  $i^{\text{th}}$  large sample

$$F_X(x_i) = \frac{i}{N+1}$$

where  $N$  is the total number of samples.

<b><math>i</math></b>	<b>No. of cars</b>	$F_X(x_i) = \frac{i}{N+1}$
1	3600	0.1
2	4500	0.2
3	5400	0.3
4	6500	0.4
5	7000	0.5
6	7500	0.6
7	8700	0.7
8	9000	0.8
9	9500	0.9

The probability paper is now created by rescaling the y-axis.

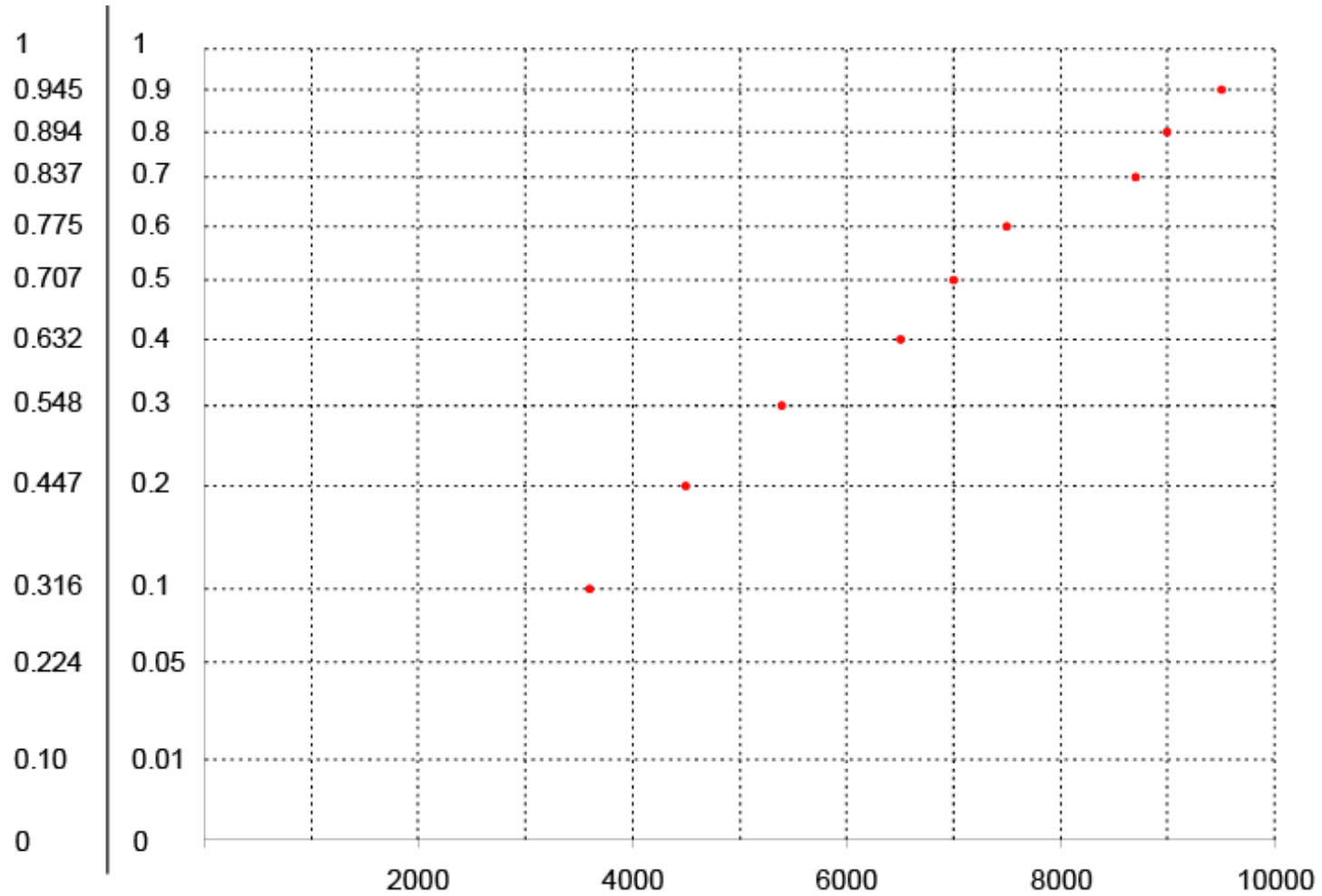
The data are plotted on the probability paper.

If the data fits on a straight line, it follows the triangular distribution.

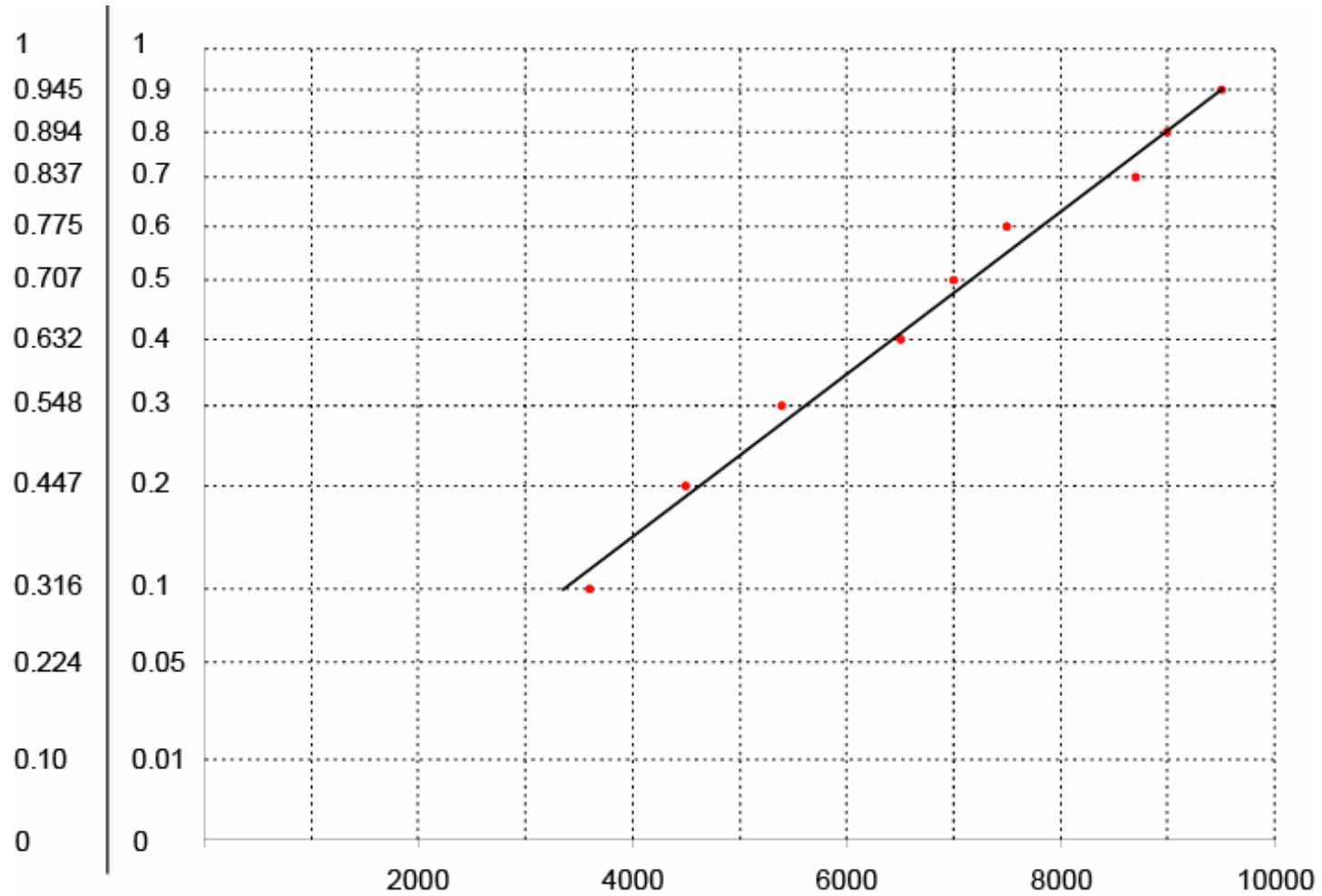
The cumulative distribution function used to plot the data is obtained from the following table:

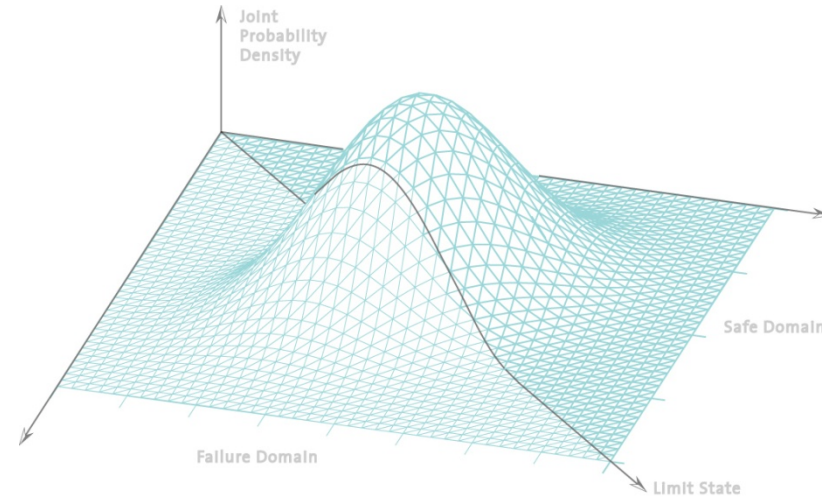
<b>i</b>	<b>No. of cars</b>	$F_X(x_i^o) = \frac{i}{N+1}$
1	3600	0.1
2	4500	0.2
3	5400	0.3
4	6500	0.4
5	7000	0.5
6	7500	0.6
7	8700	0.7
8	9000	0.8
9	9500	0.9

### Plotting the probability paper



### Plotting the probability paper





## Exercise Tutorial 10

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ

## Correction!!

### EXERCISE TUTORIAL 8 SOLUTION

#### Exercise 8.1 Solution

a. → From the Pythagorean Theorem it follows that:

$$\rightarrow f^2 + a^2 + b^2 = d^2$$

Therefore, the error in  $d$  propagates according to  $\varepsilon_d = \sqrt{\varepsilon_f^2 + \varepsilon_a^2 + \varepsilon_b^2}$ .

→ Then,  $\frac{\varepsilon_d}{\sigma_\varepsilon} = \sqrt{\left(\frac{\varepsilon_f}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_a}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_b}{\sigma_\varepsilon}\right)^2}$  is Chi-distributed with three degrees of freedom.

→ The probability density function of  $Z = \frac{\varepsilon_d}{\sigma_\varepsilon}$  is:

$$\rightarrow f_z(z) = \frac{z^{(3-1)}}{2^{3/2-1} \Gamma(3/2)} e^{(-z^2/2)} \quad f_{\varepsilon_d}(\varepsilon_d) = \frac{1}{\sqrt{2}} \left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2 \frac{1}{\sqrt{\pi}/2} e^{\left(-\left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2/2\right)} \left| \frac{dz}{d\varepsilon_d} \right| = \sqrt{\frac{2}{\pi}} \left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2 e^{\left(-\left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2/2\right)} \frac{1}{\sigma_\varepsilon}$$

→ Therefore, the probability density function of  $\varepsilon_d$  is obtained as:

$$\rightarrow f_{\varepsilon_d}(\varepsilon_d) = \frac{1}{2\sqrt{2}} \left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2 \frac{1}{\sqrt{\pi}/2} e^{\left(-\left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2/2\right)} \left| \frac{dz}{d\varepsilon_d} \right| = \frac{1}{\sqrt{2\pi}} \left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2 e^{\left(-\left(\frac{\varepsilon_d}{\sigma_\varepsilon}\right)^2/2\right)} \frac{1}{\sigma_\varepsilon}$$

**Exercise 8.7 (Group Exercise)**

To rebuild a car park, the arrival times of cars were measured. The time interval between arriving cars are shown in the table.

- Check graphically, if the time interval of car arrivals can be represented by an Exponential distribution.
- Calculate the mean value of the time interval of car arrivals. Under the assumption that the time interval is Exponential distributed, determine the mean value of the time interval from the graph produced in part (a.).

$i$	Time interval (seconds)	$i$	Time interval (seconds)
1	1.52	7	30.4
2	6.84	8	30.4
3	9.12	9	34.2
4	10.64	10	60.8
5	15.2	11	78.28
6	21.28	12	95.76



- a. Check graphically, if the time interval of car arrivals can be represented by an exponential distribution.

$i$	$i/(n+1)$	Time interval (seconds)
1	1/13	1.52
2	2/13	6.84
3	3/13	9.12
4	4/13	10.64
5	5/13	15.2
6	6/13	21.28
7	7/13	30.4
8	8/13	30.4
9	9/13	34.2
10	10/13	60.8
11	11/13	78.28
12	12/13	95.76







## Probability paper

### ➤ Exponential distribution

$$F_X(x) = 1 - e^{-\lambda x} \longrightarrow \frac{\ln(1 - F_X(x)) = -\lambda x}{= y}$$

### ➤ Normal distribution

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \longrightarrow \frac{\Phi^{-1}(F_X(x)) = \frac{x - \mu}{\sigma} = \frac{1}{\sigma}x - \frac{\mu}{\sigma}}{= y}$$

### ➤ Gumbel distribution

$$F_X(x) = \exp(-\exp(-a(x - b))) \longrightarrow \frac{-\ln(-\ln(F_X(x))) = a(x - b) = ax - ab}{= y}$$

## Estimation of parameters

There are several ways to estimate the parameters of distributions from observed data. For instance,

1. Method of moments

2. Maximum likelihood method

Let's start with the method of moments.

## The method of moments

Moment in “probability” sense

Moments are defined as:

$$m_k = \int x^k f_X(x) dx \quad \text{for continuous case}$$

$$m_k = \sum_j x_j^k p(x_j) \quad \text{for discrete case}$$

The parameters of the distribution are estimated so that the moments match.

Calculate  
directly from  
assumed distribution

$$m_k = \int x^k f_X(x) dx = E[X^k] = \frac{1}{n} \sum_{i=1}^n \hat{x}_i^k$$

$$m_k = \sum_j x_j^k p(x_j) = E[X^k] = \frac{1}{n} \sum_{i=1}^n \hat{x}_i^k$$

Calculate from  
obtained sample

## The method of moments

---

### A simple example

We want to estimate the parameter of the Exponential distribution.

The probability density function is given as  $f_x(x|\lambda) = \lambda \exp(-\lambda x)$  .

The first moment  $m_1$  is

$$\left. \begin{aligned} m_1 &= \int_0^{\infty} x \lambda e^{-\lambda x} dx = \frac{1}{\lambda} \\ \frac{1}{n} \sum_{i=1}^n \hat{x}_i &= 0.48 \end{aligned} \right\} \lambda = 1/0.48 = 2.08$$

### Realization

1	0.02
2	0.08
3	0.14
4	0.21
5	0.30
6	0.40
7	0.52
8	0.69
9	0.94
10	1.50



## The maximum likelihood method

---

Another method is the maximum likelihood method  
Let's start with an example.

There is a coin. We are interested in the probability that the tail comes out,  $p$ .  
We tossed the coin 20 times and observed 6 tails and 14 heads. How can we estimate the parameter  $p$ ?

Note that  $p$  is the parameter of Bernoulli trials, and is the probability of success.

### The maximum likelihood method

---

We can never know the truth, but...

If  $p=0.1$ , what is the probability that 6 tails and 14 heads come out?

$$L(0.1) = \binom{20}{6} \times 0.1^6 \times 0.9^{14} = 0.0089 \longleftarrow P[6 \text{ tails and } 14 \text{ heads} | p = 0.1]$$

If  $p=0.3$ , what is the probability that 6 tails and 14 heads come out?

$$L(0.3) = \binom{20}{6} \times 0.3^6 \times 0.7^{14} = 0.192 \longleftarrow P[6 \text{ tails and } 14 \text{ heads} | p = 0.3]$$

How should we interpret these numbers?

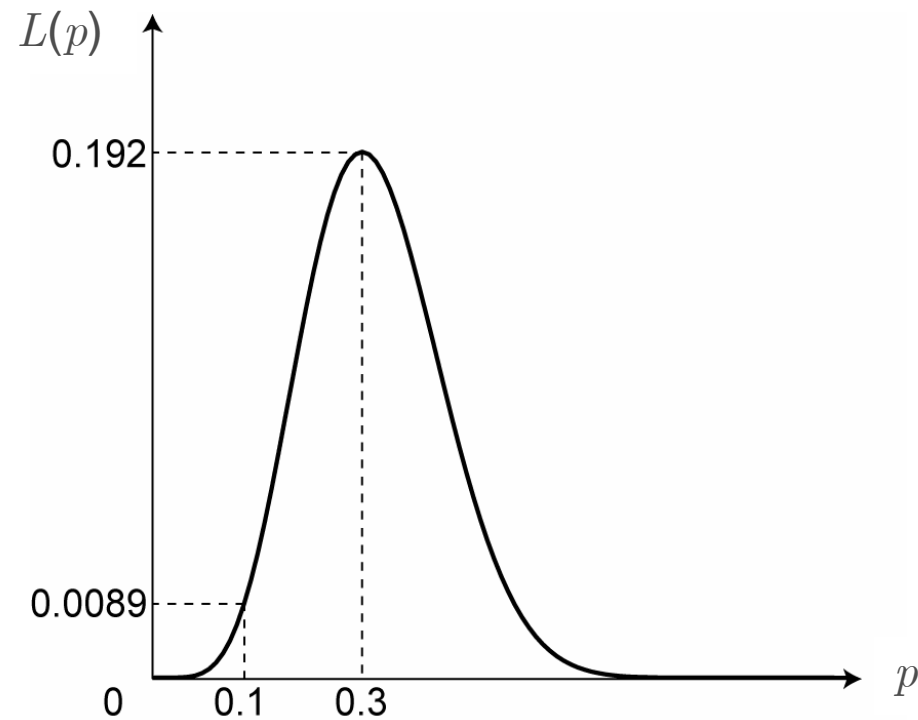
Since the probability that 6 tails and 14 heads come out given  $p=0.1$  is much smaller than that given  $p=0.3$ , we may believe that the outcome comes from “ $p=0.3$ ” rather than “ $p=0.1$ ”.

These are not a probability of “ $p=0.3$ ” nor “ $p=0.1$ ”.

### The maximum likelihood method

---

We can calculate  $L(p)$  for all  $p$  ( $0 \leq p \leq 1$ ) and obtain the following figure.



“ $L(p)$  is called “likelihood function.”

## The maximum likelihood method

---

The likelihood function

Let  $X$  be a random variable whose probability density function is  $f_X(x | \theta)$ .

When you make  $n$  trials, the joint probability density of outcomes is

$$f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n | \theta) = \prod_{i=1}^n f_X(x_i | \theta) \quad \leftarrow \text{before observation}$$

This is the probability density for any given  $\theta$ .

However, we can regard this formulation as the function of  $\theta$  after the observation of outcomes  $\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n$ .

In this interpretation of the above formulation, we call it a “likelihood function”

$$L(\theta | \hat{x}_1, \hat{x}_2, \dots, \hat{x}_n) = \prod_{i=1}^n f_X(\hat{x}_i | \theta) \quad \leftarrow \text{after observation} \\ (\theta \text{ is not known})$$

## The maximum likelihood method

---

The maximum likelihood method (MLM)

The maximum likelihood estimator  $\theta$  is obtained as the value of  $\theta$  that maximizes the likelihood function  $L(\theta)$ .

This is equivalent to maximize the “**log-likelihood function**” defined as:

$$l(\theta | \hat{\mathbf{x}}) = \ln L(\theta | \hat{\mathbf{x}}) = \ln \prod_{i=1}^n f_X(\hat{x}_i | \theta) = \sum_{i=1}^n \ln f_X(\hat{x}_i | \theta)$$

.....

Coin example again!

To be maximized is

$$L(p) = \binom{20}{6} p^6 (1-p)^{14}$$

The log-likelihood function is

$$l(p) = \ln L(p) = \ln \binom{20}{6} + 6 \ln p + 14 \ln(1-p)$$

The MLM estimate is obtained as:

$$\frac{\partial l}{\partial p} = \frac{6}{p} - \frac{14}{1-p} = 0 \Rightarrow p = 0.3$$

## Exercise 9.1

In order to model the concrete compressive strength of a certain concrete production, 20 samples were measured and the result of measurements is shown in the table. It is assumed that the population of the samples follows the Normal distribution  $N(\mu, \sigma^2)$ .

- 1) Describe the likelihood function.
- 2) Estimate the unknown parameters  $(\mu, \sigma)$  with the maximum likelihood method.
- 3) Estimate the unknown parameters with the method of moments.

No. of sample	Compressive strength [MPa]	No. of sample	Compressive strength [MPa]
1	24.4	11	33.3
2	27.6	12	33.5
3	27.8	13	34.1
4	27.9	14	34.6
5	28.5	15	35.8
6	30.1	16	35.9
7	30.3	17	36.8
8	31.7	18	37.1
9	32.2	19	39.2
10	32.8	20	39.7

1) The likelihood function is

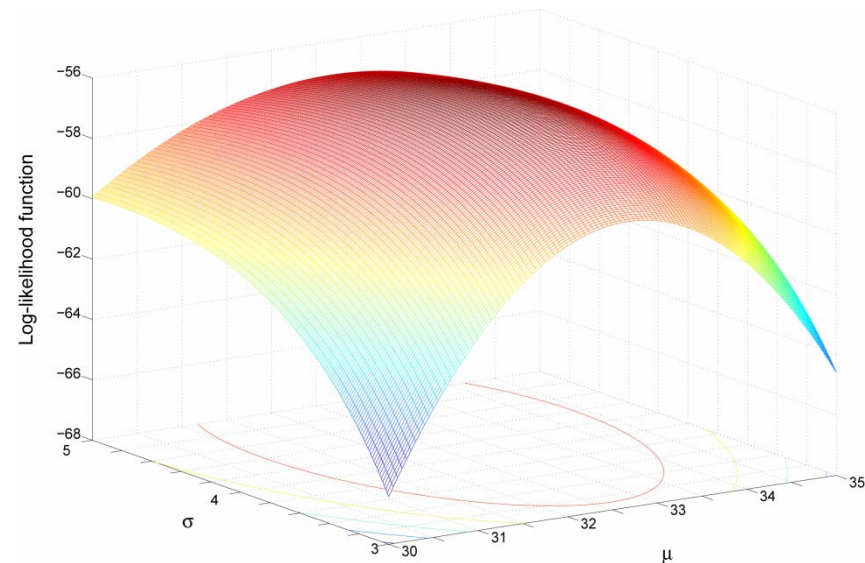
$$L(\mu, \sigma | \hat{\mathbf{x}}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\hat{x}_i - \mu)^2}{2\sigma^2}\right]$$

to be maximized. Instead of the likelihood function we adopt the log-likelihood function.

$$\begin{aligned} l = \ln L &= \sum_{i=1}^n \ln \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\hat{x}_i - \mu)^2}{2\sigma^2}\right] \right] \\ &= -n \ln(\sqrt{2\pi}) - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (\hat{x}_i - \mu)^2 \end{aligned}$$

2) The maximum likelihood estimates are obtained as the parameters at which the likelihood function (accordingly the log-likelihood function) is maximized.

$$\begin{aligned}
 l = \ln L &= \sum_{i=1}^n \ln \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(\hat{x}_i - \mu)^2}{2\sigma^2} \right] \right] \\
 &= -n \ln(\sqrt{2\pi}) - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (\hat{x}_i - \mu)^2
 \end{aligned}$$



$$\left. \begin{aligned}
 \frac{\partial l}{\partial \mu} &= \frac{1}{2\sigma^2} \sum_{i=1}^n 2(\hat{x}_i - \mu) = 0 \\
 \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (\hat{x}_i - \mu)^2 = 0
 \end{aligned} \right\}$$

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

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



$$\mu = 32.67$$

$$\sigma = 4.04$$



3) With the method of moments, the parameters are estimated as:

$$m_1 = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = \mu$$

$$m_2 = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = \sigma^2 + \mu^2$$

$$\left. \begin{aligned} \mu &= m_1 = \frac{1}{n} \sum_{i=1}^n \hat{x}_i \\ \sigma^2 &= m_2 - m_1^2 = \frac{1}{n} \sum_{i=1}^n \hat{x}_i^2 - \left(\frac{1}{n} \sum_{i=1}^n \hat{x}_i\right)^2 \end{aligned} \right\} \begin{aligned} \mu &= 32.67 \\ \sigma &= 4.04 \end{aligned}$$

## Exercise 9.2

What happens if the Exponential distribution is assumed instead of the Normal distribution in exercise 9.1?

- a. Estimate the unknown parameters of the Exponential distribution with the maximum likelihood method.
- b. Draw the cumulative distribution functions together with observed cumulative distribution.

The likelihood function is

$$L = \prod_{i=1}^n \lambda e^{-\lambda \hat{x}_i}$$

The corresponding log-likelihood function is

$$\begin{aligned} l = \ln L &= \sum_{i=1}^n \ln(\lambda e^{-\lambda \hat{x}_i}) \\ &= n \ln \lambda - \lambda \sum_{i=1}^n \hat{x}_i \end{aligned}$$

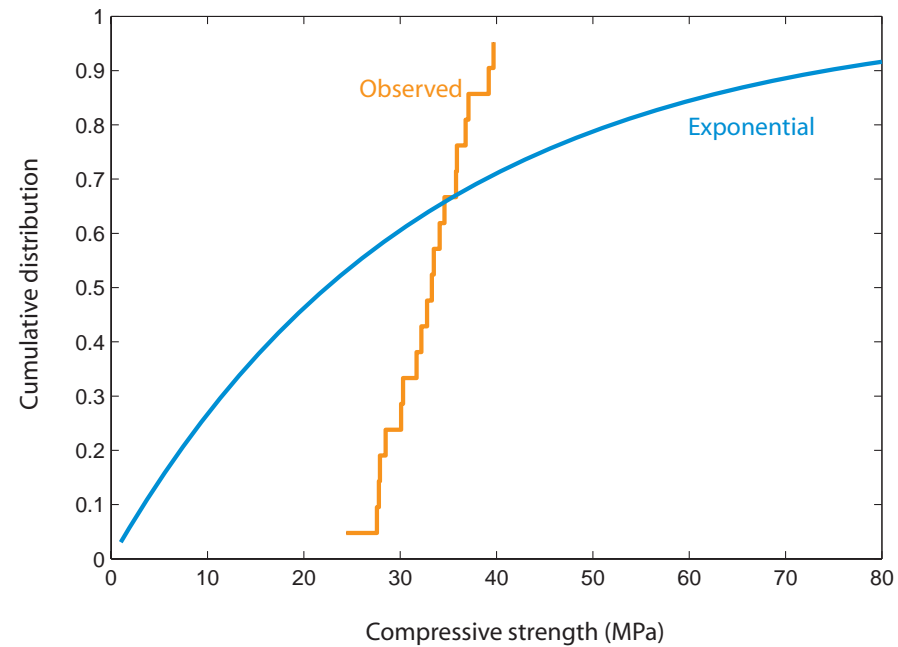
The maximum likelihood method estimate is obtained from

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n \hat{x}_i = 0$$

It is

$$\lambda = \frac{n}{\sum_{i=1}^n \hat{x}_i} \longrightarrow \lambda = 0.031$$

Do you think that the Exponential distribution with the maximum likelihood estimate models the population of the sample nicely???



How to judge that? We will see this the next week!

### Exercise 9.3 (Group exercise)

It is known that the data shown in the table are the realizations of a random variable  $X$  characterized by the cumulative distribution function,

$$F_X(x) = x^\alpha, 0 \leq x \leq 1 \quad \text{with unknown parameter } \alpha.$$

Estimate the parameter  $\alpha$  in the following methods.

- a. Estimate the unknown parameter  $\alpha$  with the *method of moments*.
- b. Estimate the unknown parameter  $\alpha$  with the *maximum likelihood method*.
- c. Draw the cumulative distribution function with the estimated parameter in (b.) and the observed cumulative distribution.

### Exercise 9.3 (Group exercise)

#### a) Method of moments

“sample first moment = analytical first moment”

The sample first moment is calculated by

$$\frac{1}{n} \sum_{i=1}^n \hat{x}_i$$

The analytical first moment is calculated by

$$\int xf(x|\alpha)dx$$

$$\leftarrow f(x|\alpha) = \frac{dF_x(x|\alpha)}{dx} = \frac{dx^\alpha}{dx}$$

Equate and solve

$$\frac{1}{n} \sum_{i=1}^n \hat{x}_i = \int xf(x|\alpha)dx$$

### Exercise 9.3 (Group exercise)

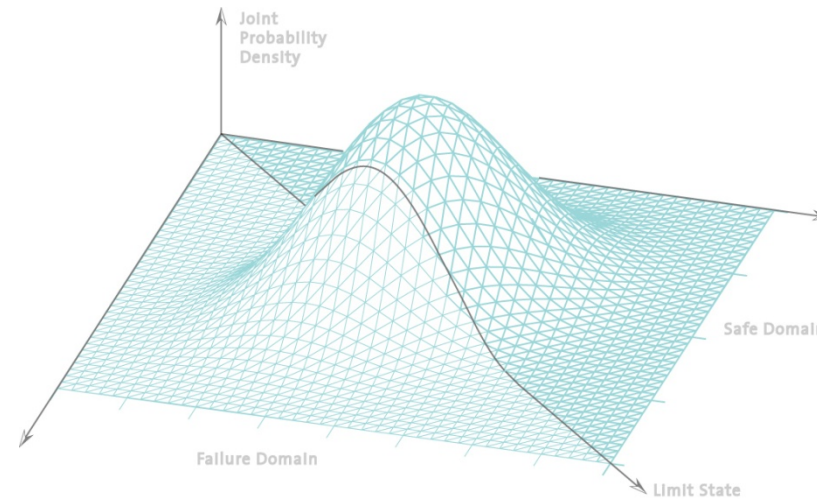
#### b) Maximum likelihood method

The likelihood function is described as:

$$L(\alpha) = \prod_{i=1}^n f(\hat{x}_i | \alpha) \quad \longleftarrow \quad f(x | \alpha) = \frac{dF_X(x | \alpha)}{dx} = \frac{dx^\alpha}{dx}$$

Maximize the log likelihood function with respect to  $\alpha$ .

$$l(\alpha) = \sum_{i=1}^n \ln(f(\hat{x}_i | \alpha))$$



## Exercise Tutorial 10

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber  
Swiss Federal Institute of Technology Zurich  
ETHZ



### Exercise 9.3 (Group exercise)

It is known that the data shown in the table are the realizations of a random variable  $X$  characterized by the cumulative distribution function,

$$F_X(x) = x^\alpha, 0 \leq x \leq 1 \quad \text{with unknown parameter } \alpha.$$

Estimate the parameter  $\alpha$  in the following methods.

- a. Estimate the unknown parameter  $\alpha$  with the *method of moments*.
- b. Estimate the unknown parameter  $\alpha$  with the *maximum likelihood method*.
- c. Draw the cumulative distribution function with the estimated parameter in (b.) and the observed cumulative distribution.

### Exercise 9.3 (Group exercise)

#### a) Method of moments

“sample first moment = analytical first moment”

The sample first moment is calculated by

$$\frac{1}{n} \sum_{i=1}^n \hat{x}_i = 0.656$$

The analytical first moment is calculated by

$$\int_0^1 x f(x | \alpha) dx = \int_0^1 x \alpha x^{\alpha-1} dx = \left[ \frac{\alpha}{\alpha+1} x^{\alpha+1} \right]_0^1 = \frac{\alpha}{\alpha+1}$$

Equate and solve

$$\frac{\alpha}{\alpha+1} = 0.656 \quad \Leftrightarrow \alpha = 1.91$$

No. of data	Realization
1	0.22
2	0.97
3	0.92
4	0.59
5	0.39
6	0.74
7	0.81
8	0.86
9	0.39
10	0.67

### Exercise 9.3 (Group exercise)

#### b) Maximum likelihood method

The likelihood function is described as:

$$L(\alpha) = \prod_{i=1}^n f(\hat{x}_i | \alpha)$$

Maximize the log likelihood function with respect to  $\alpha$ .

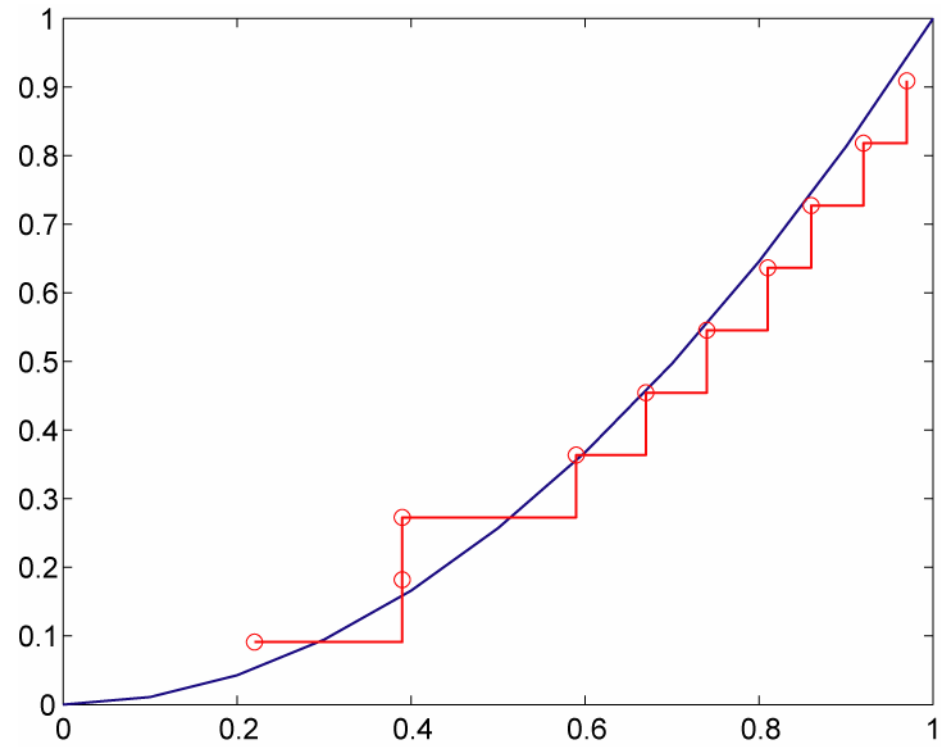
$$l(\alpha) = \sum_{i=1}^n \ln(f(\hat{x}_i | \alpha)) = \sum_{i=1}^n \ln(\alpha x_i^{\alpha-1}) = n \ln \alpha + (\alpha - 1) \sum_{i=1}^n \ln x_i$$

$$\frac{\partial l(\alpha)}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^n \ln \hat{x}_i = 0 \Leftrightarrow \alpha = -\frac{n}{\sum_{i=1}^n \ln \hat{x}_i} = 1.96$$

No. of data	Realization
1	0.22
2	0.97
3	0.92
4	0.59
5	0.39
6	0.74
7	0.81
8	0.86
9	0.39
10	0.67

## Exercise 9.3 (Group exercise)

c)



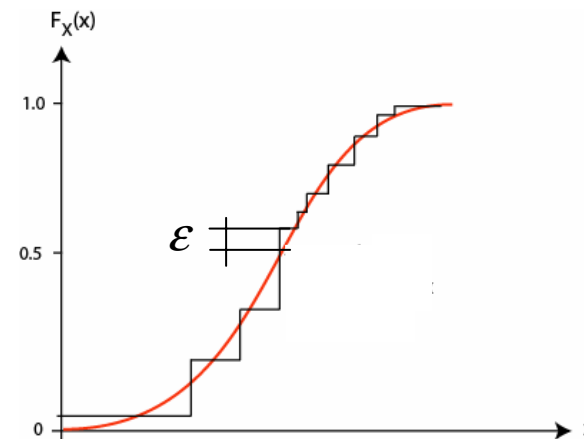
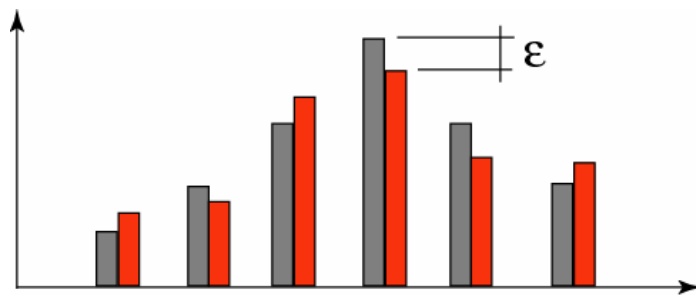
The essence of the goodness-of-fit test is to

measure the deviation between assumed distribution and observations.

There are several ways to measure the deviation.

Let  $K$  represent the deviation, then we judge with  $K$  whether the assumed distribution model the observations well or not.

The operating rule is then simply,  
Accept  $H_0$  if  $K < c$   
Reject  $H_0$  if  $K > c$ .



## Goodness-of-fit test

What can be a good measure of the deviation  $K$ ?

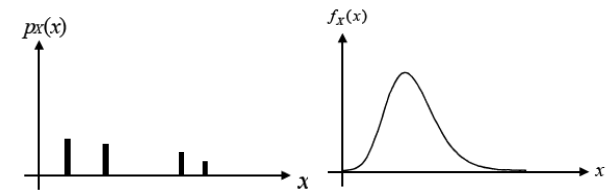
$K$  can be almost arbitrary as long as it is related to the deviation, but it must be the one for which we know the distribution, otherwise we can not determine  $c$  in accordance with the level of significance  $\alpha$ .

$$P[K > c | \theta \in \Theta_0] = \alpha \Leftrightarrow P[K \leq c | \theta \in \Theta_0] = 1 - \alpha$$

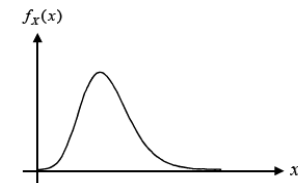
We consider two alternatives:

- 1) Chi-square test (for any cases\*)

\*We discretize in continuous cases.



- 2) Kolmogorov-Smirnov test (only for continuous case)



Let's begin with a simple exercise.



### Exercise 10.1

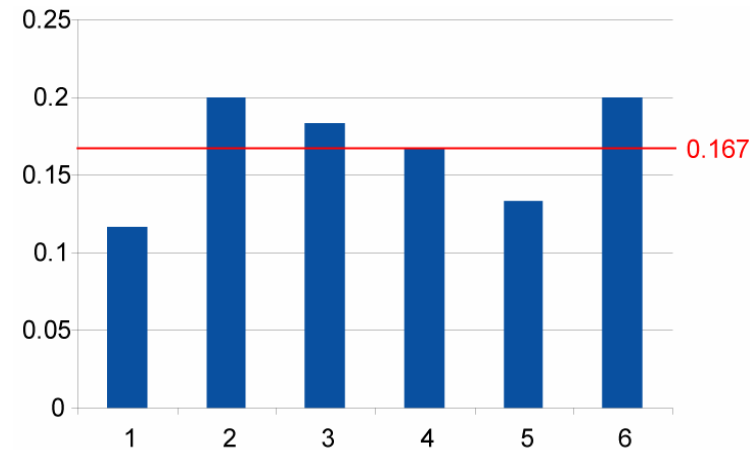
A dice is suspected to be asymmetric, resulting in the inhomogeneity of probability that each side of a dice comes out. In order to judge this suspicion statistically, 60 trials were made and the result is shown in the table.

- 1) Draw the relative frequency histogram, and compare with the uniform probability mass function under the assumption that the dice is symmetric.
- 2) What is the probability that each spot of a dice comes out 10 times respectively in 60 trials when the dice is symmetric?
- 3) Set the symmetry property of the dice as the null hypothesis, and test the hypothesis with the  $\chi^2$  test at the 5% level of significance.

Spot of the dice	No. of realizations
1	7
2	12
3	11
4	10
5	8
6	12
Sum	60

Comparison between the observation and the assumed distribution.

Spot of dice	No. of realization	Relative frequency
1	7	0.1667
2	12	0.2
3	11	0.1833
4	10	0.1667
5	8	0.1333
6	12	0.2
<b>Sum</b>	<b>60</b>	<b>1</b>



There is deviation...but is this too large to say that the dice is asymmetric?



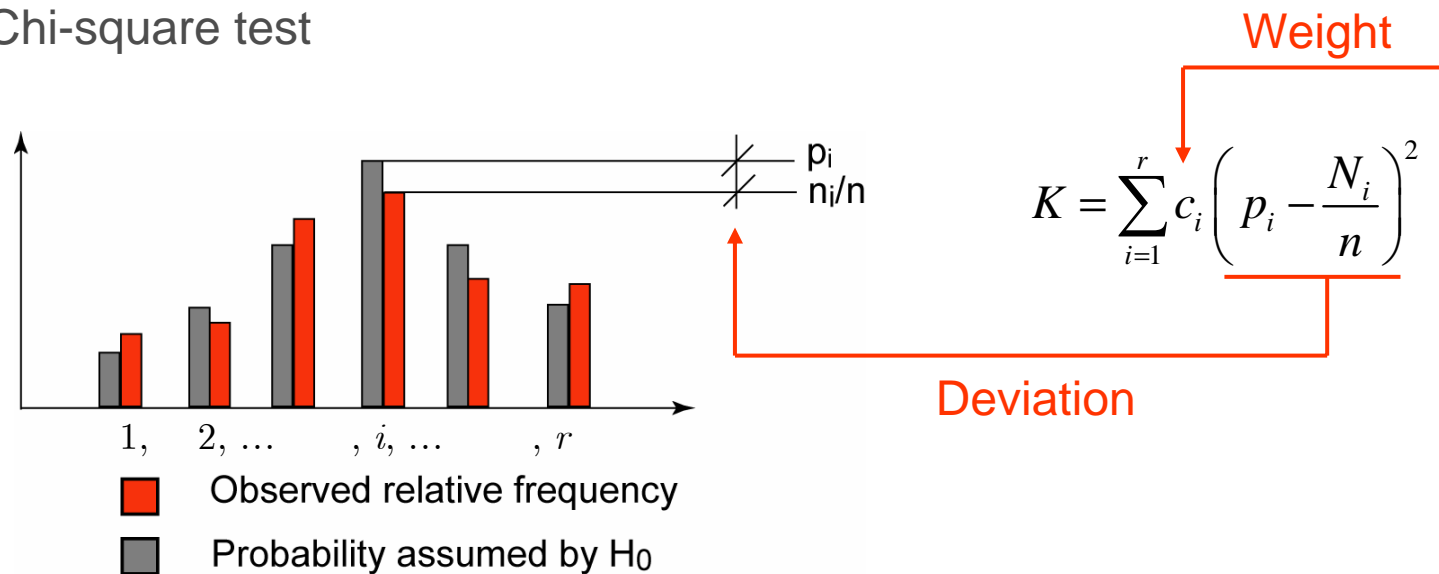
First of all, calculate the probability that the numbers of outcomes completely coincide the expected numbers under the assumption that the dice is symmetric.

The numbers of outcomes of each spots follow the polynomial distribution:

$$\begin{aligned} P[N_i = 10, i = 1, 2, 3, 4, 5, 6] &= \frac{60!}{10!10!10!10!10!10!} \left(\frac{1}{6}\right)^{10} \left(\frac{1}{6}\right)^{10} \left(\frac{1}{6}\right)^{10} \left(\frac{1}{6}\right)^{10} \left(\frac{1}{6}\right)^{10} \left(\frac{1}{6}\right)^{10} \\ &= \frac{60!}{(10!)^6} \left(\frac{1}{6}\right)^{60} \\ &= 0.0000745 \end{aligned}$$

The probability is very very small!!! We should allow the deviation to some extent.

### Chi-square test



Fundamentally, the weights  $c_i$  are arbitrary but if you choose  $c_i = \frac{n}{p_i}$

then it is known that  $K$  follows approximately  $\chi^2$  distribution with  $r - 1$  degrees of freedom.

When you know the distribution of  $K$ , you can select  $c$  in accordance with the level of significance. ( $P[K > c | \theta \in \Theta_0] = \alpha \Leftrightarrow P[K \leq c | \theta \in \Theta_0] = 1 - \alpha$ )

We adopt  $K$  as the sample statistic.

$$K = \sum_{i=1}^r \frac{n}{p_i} \left( p_i - \frac{N_i}{n} \right)^2 = \sum_{i=1}^r \frac{(N_i - np_i)^2}{np_i}$$

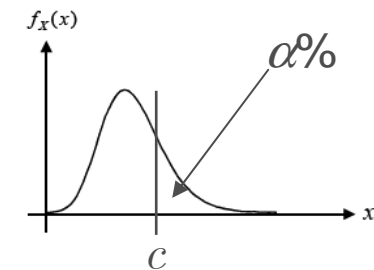
Again  $K$  follows  $\chi^2$  distribution with  $r - 1$  degrees of freedom.

Therefore it is possible to determine  $c$  in accordance with the level of significance.

$$P[K > c | \theta \in \Theta_1] = \alpha$$

In the case of this example,  $r = 6$  and  $\alpha = 5\%$ .

$f$	$\chi^2_{F=0.01}$	$\chi^2_{F=0.05}$	$\chi^2_{F=0.10}$	$\chi^2_{F=0.25}$	$\chi^2_{F=0.50}$	$\chi^2_{F=0.75}$	$\chi^2_{F=0.90}$ $\alpha=10\%$	$\chi^2_{F=0.95}$ $\alpha=5\%$
1	0.0002	0.0039	0.0158	0.1015	0.4549	1.3233	2.7055	3.8415
2	0.0201	0.1026	0.2107	0.5754	1.3863	2.7726	4.6052	5.9915
3	0.1148	0.3518	0.5844	1.2125	2.3660	4.1083	6.2514	7.8147
4	0.2971	0.7107	1.0636	1.9226	3.3567	5.3853	7.7794	9.4877
5	0.5543	1.1455	1.6103	2.6746	4.3515	6.6257	9.2363	11.0705
6	0.8721	1.6354	2.2041	3.4546	5.3481	7.8408	10.6446	12.5916



$H_0: p_i = 1/6$  ( $i=1,2,3,4,5,6$ )

$H_1: p_i \neq 1/6$  at least one of  $i = 1,2,3,4,5,6$ .

Calculate the value of the sample statistic  $K$  based on observations.

Spot of dice	No. of Realization	$(N_i - np_i)^2$	$np_i$	$(N_i - np_i)^2/np_i$
1	7	9	10	9/10
2	12	4	10	4/10
3	11	1	10	1/10
4	10	0	10	0/10
5	8	4	10	4/10
6	12	4	10	4/10
sum	60		$\Sigma(N_i - np_i)^2/np_i$	= 2.20

$$k = 2.2 < 11.07 = c$$



$H_0$  is not rejected, meaning that the dice may be symmetric.

### Exercise 10.2

For the estimation of the concrete compressive strength of a certain concrete production, 20 samples were measured and the result is shown in the table. It is assumed that the concrete compressive strength follows the Normal distribution.

- Estimate the unknown parameters of the distribution with the method of moments.
- Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

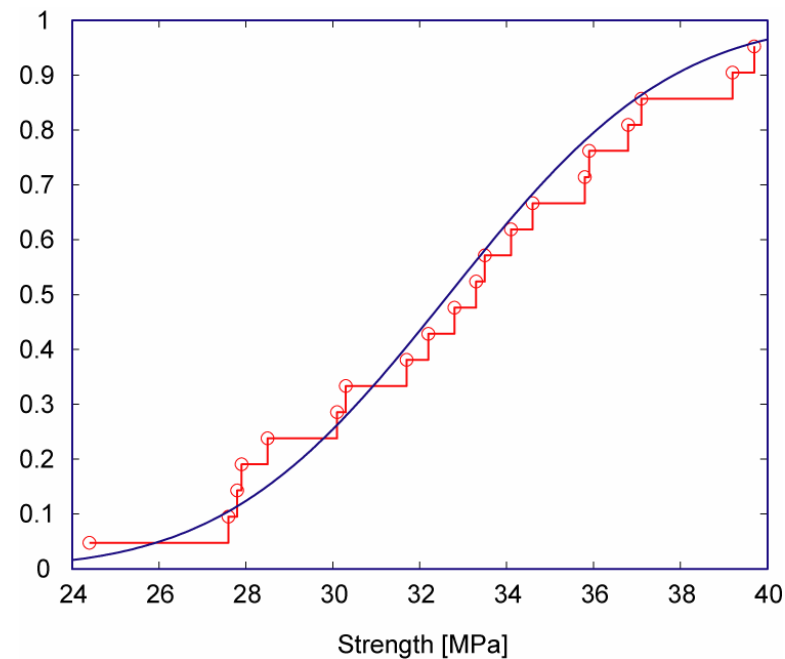
No. of sample ( $i$ )	Compressive strength (MPa)	No. of sample ( $i$ )	Compressive strength (MPa)
1	24.4	11	33.3
2	27.6	12	33.5
3	27.8	13	34.1
4	27.9	14	34.6
5	28.5	15	35.8
6	30.1	16	35.9
7	30.3	17	36.8
8	31.7	18	37.1
9	32.2	19	39.2
10	32.8	20	39.7

## Exercise 10.2

- a. Estimate the unknown parameters of the distribution with the method of moments.

$$\mu = m_1 = 32.67$$

$$\sigma = \sqrt{m_2 - m_1^2} = \sqrt{1083.4 - 32.67^2} = 4.04$$



### Exercise 10.2

b. Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

$$\epsilon_m^2 = \sum_{j=1}^k \frac{(N_{o,j} - N_{p,j})^2}{N_{p,j}}$$



Interval	$N_{o,j}$	$p(x_j)$	$N_{p,j} = np(x_j)$	$\epsilon_m^2$
-30				
30-33				
33-36				
36-				
Sum				

## Exercise 10.2

b. Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

Interval	$N_{o,j}$	$p(x_j)$	$N_{p,j} = np(x_j)$	$\epsilon_m^2$
-30	5	$\Phi\left(\frac{30-32.67}{4.04}\right) = 0.254$	5.08	0.001
30-33	5	$\Phi\left(\frac{33-32.67}{4.04}\right) - \Phi\left(\frac{30-32.67}{4.04}\right) = 0.278$	5.56	0.06
33-36	6	$\Phi\left(\frac{36-32.67}{4.04}\right) - \Phi\left(\frac{33-32.67}{4.04}\right) = 0.263$	5.26	0.10
36-	4	$1 - \Phi\left(\frac{36-32.67}{4.04}\right) = 0.205$	4.10	0.002
Sum	20			= 0.163



### Exercise 10.2

- b. Test the goodness of fit for the distribution with estimated parameters with the test at the 5% significance level.

The sample statistic follows the Chi-square distribution with  $4-1-2=1$  degree of freedom. At the 5% significant level, the null hypothesis should be rejected if the sample statistic is larger than 3.84, see the probability table for the Chi-square distribution (Annex T, Table T.3).

Since the sample statistic is obtained as 0.163 from the observations the null hypothesis cannot be rejected at the 5% significance level, i.e., the assumption that the concrete compressive strength may follow the normal distribution with the mean 32.67MPa and the standard deviation 4.04MPa cannot be rejected.

### Exercise 10.4

The strength of 30 wood samples has been measured and the results are shown in the table. The strength is assumed to follow an Exponential distribution.

- a. Estimate the parameter of the Exponential distribution using the method of moments.
- b. Draw the cumulative distribution function with the estimated parameters, together with the observed cumulative distribution.
- c. Test the goodness of fit for the Exponential distribution with the  $\chi^2$  test at the 10% significance level.
- d. Test the goodness of fit for the Exponential distribution with the Kolmogorov-Smirnov test at the 10% significance level. The parameter of the Exponential distribution is assumed to be equal to  $\lambda=0.04$ .

### Exercise 10.4

The strength of 30 wood samples has been measured and the results are shown in the table. The strength is assumed to follow an Exponential distribution.

No.	Strength (MPa)	No.	Strength (MPa)	No.	Strength (MPa)
1	12.8	11	23.4	21	29.3
2	16.3	12	26.8	22	29.5
3	16.6	13	26.9	23	30.3
4	16.9	14	27	24	32.1
5	17.2	15	27.1	25	32.3
6	17.9	16	27.2	26	33.5
7	19.5	17	27.2	27	33.9
8	21.9	18	27.5	28	35.6
9	22.3	19	27.9	29	39.2
10	22.5	20	28.3	30	43.5

### Exercise 10.4

- a. Estimate the parameters of the exponential distribution using the method of moments.

The cumulative distribution function of the Exponential distribution is written as:

$$F_X(x) = 1 - \exp(-\lambda x)$$

The first moment is obtained as:

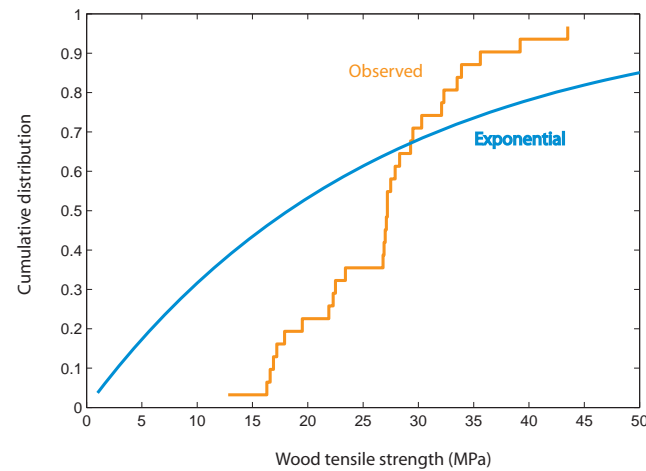
$$m_1 = \frac{1}{\lambda}$$

The parameter is estimated as:

$$\lambda = \frac{1}{m_1} = \frac{1}{1/30 \sum_{i=1}^n \hat{x}_i} = 0.038$$


### Exercise 10.4

- b. Draw the cumulative distribution function with the estimated parameter, together with the observed cumulative distribution.



## Exercise 10.4

c. Test the goodness of fit for the Exponential distribution with the  $\chi^2$  test at the 10% significance level.

$$\epsilon_m^2 = \sum_{j=1}^k \frac{(N_{o,j} - N_{p,j})^2}{N_{p,j}}$$


Interval	$N_{o,j}$	$p(x_j)$	$N_{p,j} = np(x_j)$	$\epsilon_m^2$
-20	7	0.49	14.7	4.03
20-25	4	0.08	2.4	1.07
25-30	11	0.07	2.1	37.72
30-	8	0.36	10.8	0.73
Sum	30			43.55

### Exercise 10.4

- c. Test the goodness of fit for the Exponential distribution with the  $\chi^2$  test at the 10% significance level.

The sample statistic follows the Chi-square distribution with  $4-1-1=2$  degrees of freedom. At the 10% significant level, the null hypothesis shall be rejected if the sample statistic is larger than 4.6, see the probability table for the Chi-square distribution (Annex T, Table T.3).

Since the sample statistic is obtained as 43.55 from the observations, the null hypothesis should be rejected at the 10% significance level.

### Exercise 10.4

- d. Test the goodness of fit for the exponential distribution with the Kolmogorov-Smirnov test at the 10% significance level. The parameter of the Exponential distribution is assumed to be equal to  $\lambda=0.04$ .





### Exercise 10.4

- d. Test the goodness of fit for the exponential distribution with the Kolmogorov-Smirnov test at the 10% significance level. The parameter of the Exponential distribution is assumed to be equal to  $\lambda=0.04$ .

At the 10% significance level and  $n=30$ , the null hypothesis should be rejected if the sample statistic is larger than 0.22, (Annex T, Table T.4).

Since the sample statistic is obtained as 0.412 from the observations the null hypothesis should be rejected at the 10% significance level.