Basic Statistics and Probability Theory in Civil, Surveying and Environmental Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Contents of Today's Lecture

- The organisation of the lecture – practical stuff
- Why statistics and probability in engineering?
- Decision Problems in Engineering
- Examples
- The lecture program

What do we offer to you ?

- It is our intention to provide you to the best of our abilities
  - Motivation and overview of context
  - Targeted presentation of required knowledge
  - Guidance on self study
  - Help on training your abilities
  - Help on your self evaluation
- We are here for you and we take this statement seriously

Structure and organization of the course

- 13 weekly lectures of each two sessions of 45 minutes
- 11 weekly exercise tutorials of each two sessions of 45 minutes
- 2 assessments of each 90 minutes
- Self study estimated to 4 times by 45 minutes per week

The course's web page

http://www.ibk.ethz.ch/fa/education/ss_statistics

What can you find there?

- Course's program and timetable
- Tutorial's timetable
- Script (downloadable/printable)
- Exercises/Solutions for the exercise tutorials (downloadable/printable)
- Presentations of the lecture and of the exercise tutorial (uploaded a day before the respective day)
- Videos of the lecture (uploaded the day after the lecture)
- Glossary (German-English terms)
- Links to helpful web pages
- Past examination papers
- Your exercise tutorial class and group!

Organization of the Lecture

When??

Normally…Tuesdays 8-10

Where??

HIL E1

Exceptions:

Thursday 22.03.07 8-10 HPH G 3 (lecture instead of exercise tutorial)

Other exceptions: Check the course's program!

- Script (English)
  Download from the course's web page

Organization of the Exercise Tutorials

Eva Sabiote
HIL E 22.2

Harikrishna (Hari)
Narasimhan
HIL E 13.1

Kazuyoshi (Kazu)
Nishijima
HIL E 22.3

Vasiliki (Vicky)
Malioka
HIL E 23.1

Organization of the Exercise tutorial

- **When??:**
  Normally….Thursday 8-10
- **Where??**
  HPH G 3 HCI H 2.1 HCI D 8 HCI D 2
- **Where do I go???**
  find out in the "Group lists" link on the course's web page
- **Exceptions….☺**
  First tutorial: Tuesday 27.03.07
  Where???: HIL E1 HIL B 21 HIL D 10.2 HIL F 10.3

Organization of the Exercise Tutorials

2 or more exercises will be presented in steps (based on the content of the latest lecture)

1 or more solution(s) of exercises shown in steps in the last tutorial

Group exercise
1 exercise - steps will be shown

Group Presentation
25 min

Organization of the Exercise Tutorials

Office hours:

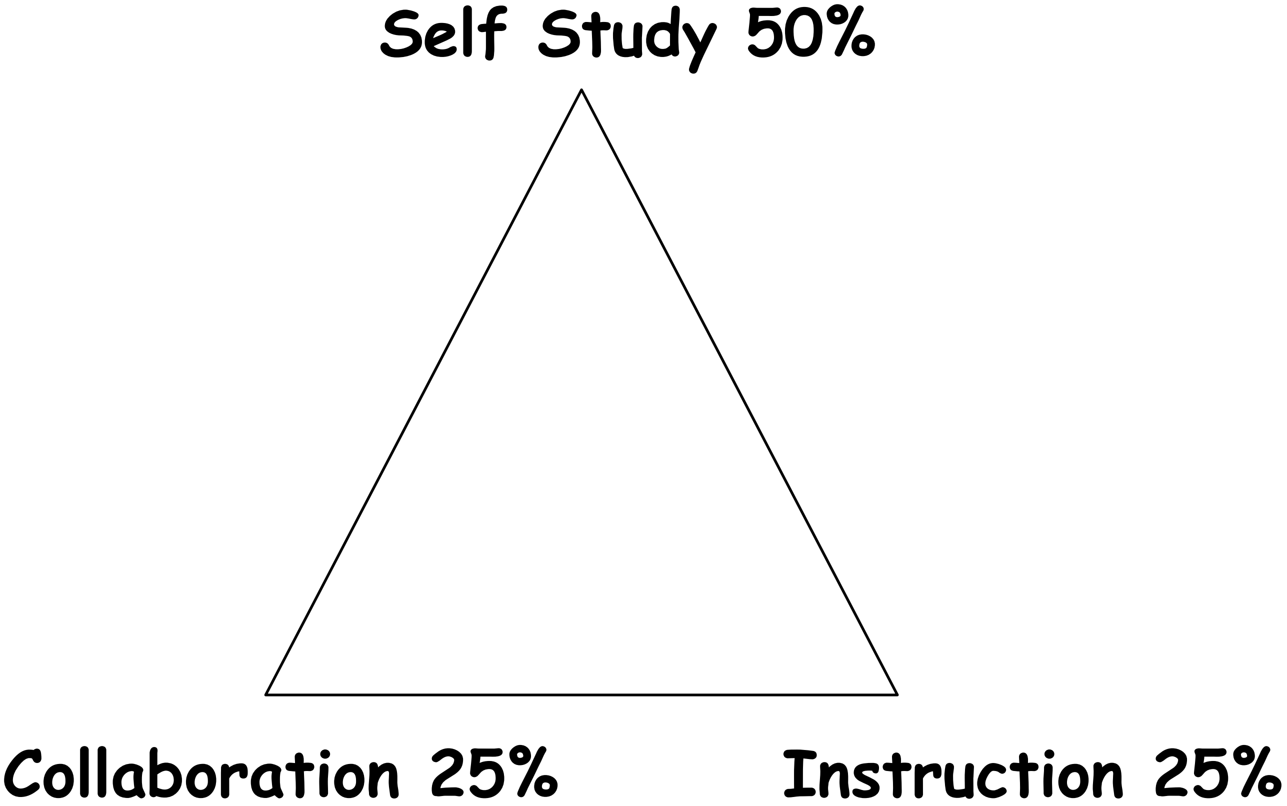
Mondays 11.30 - 12.30

Thursdays 13.30 - 14.30

What do we expect from you?

- Take advantage of the help we provide to you
  - benefit from the lectures
  - benefit from the exercise tutorials
  - benefit from the help of the assistants and professor (office hours)
- Tips and tricks
  - prepare yourself for the lectures
  - ask questions
  - try to understand the topics rather than prepare for examination
  - be curious, interested, open minded but critical to what we tell you

What do we expect from you?



Mode of assessment

- Two assessments during the semester
  one midterm (03.05.07)
  the other one towards the end of the course (14.06.07)

- Final Exam
  October/March….

Final mark= $\frac{1}{3}(two\ assessments)+\frac{2}{3}(final\ exam)$

Programmable calculators are strictly not allowed!
Open book assessments and final exam☺

Read carefully all the information in the "Preamble" of the script!!
If you have any questions ask!

Why Statistics and Probability in Engineering?

- What do engineers do ?
  - Plan, design, build, maintain and decommission

    Infrastructure
    Roads, water supply systems, tunnels, sewage systems, waste deposits, power supply systems, channels

    Structures
    houses, hospitals, schools, industry buildings, dams, powerplants, wind turbines, offshore platforms

  - Safeguard
    - people
    - environment
    - assets

    SUSTAINABLE DEVELOPMENT !

    from natural and man made hazards

Why Statistics and Probability in Engineering?

- What are engineers working with ?

  Real problems - the real world - nature

Concrete

Gravity

Ice

Dynamics

Waves

New materials

Snow

Soil

Temperature

Water

Rocks

Waste

Air

Chemicals

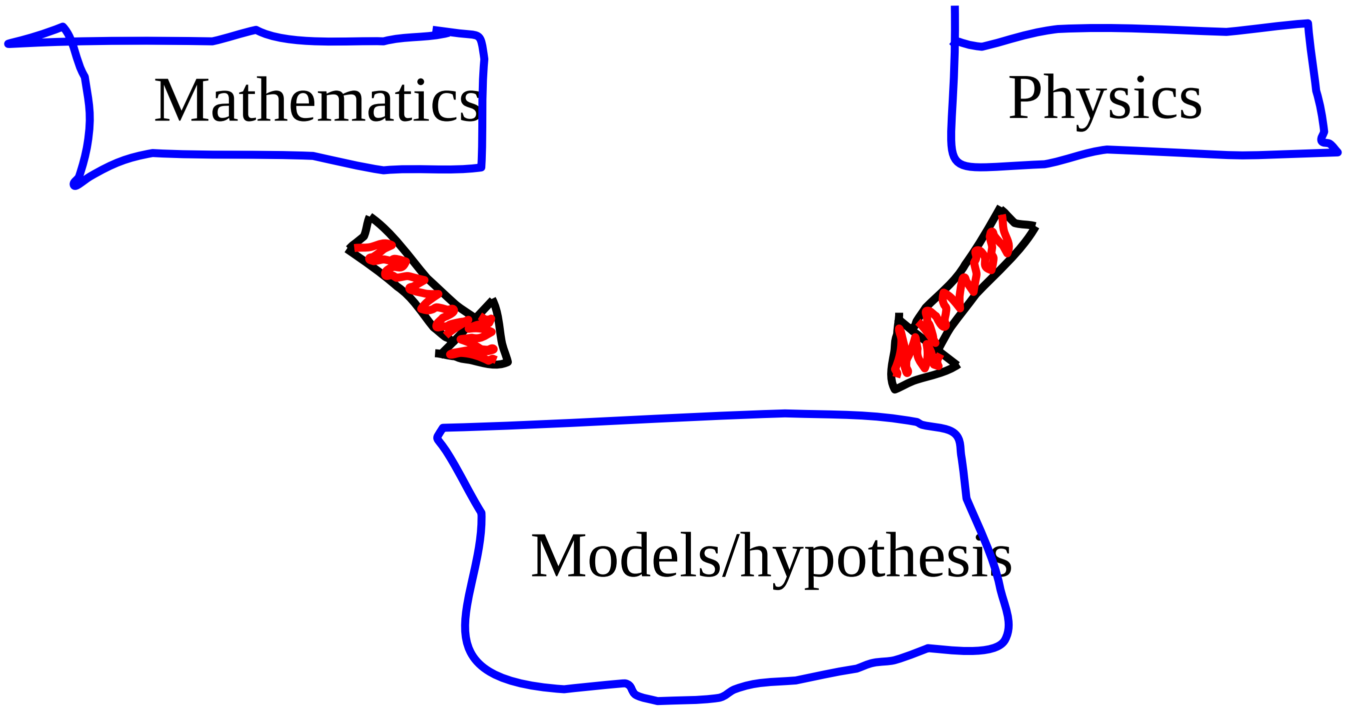
Electricity

Wind

Steel

Why Statistics and Probability in Engineering?

- How do engineers work with the real world ?

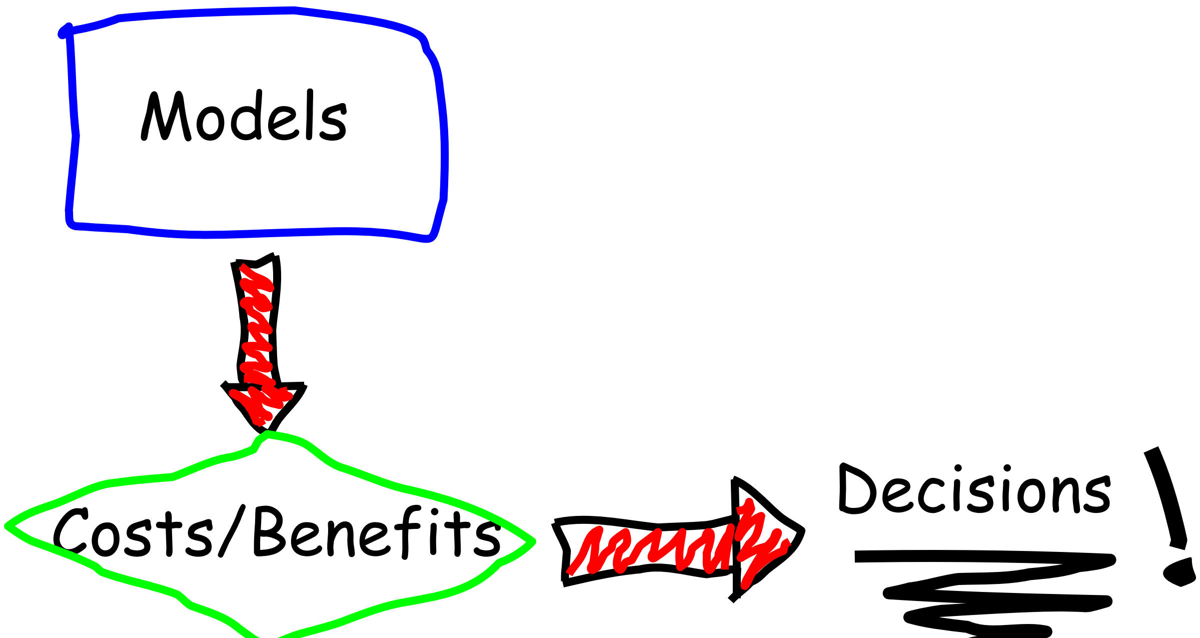


We model the real world to the „best" of our knowledge

Why Statistics and Probability in Engineering?

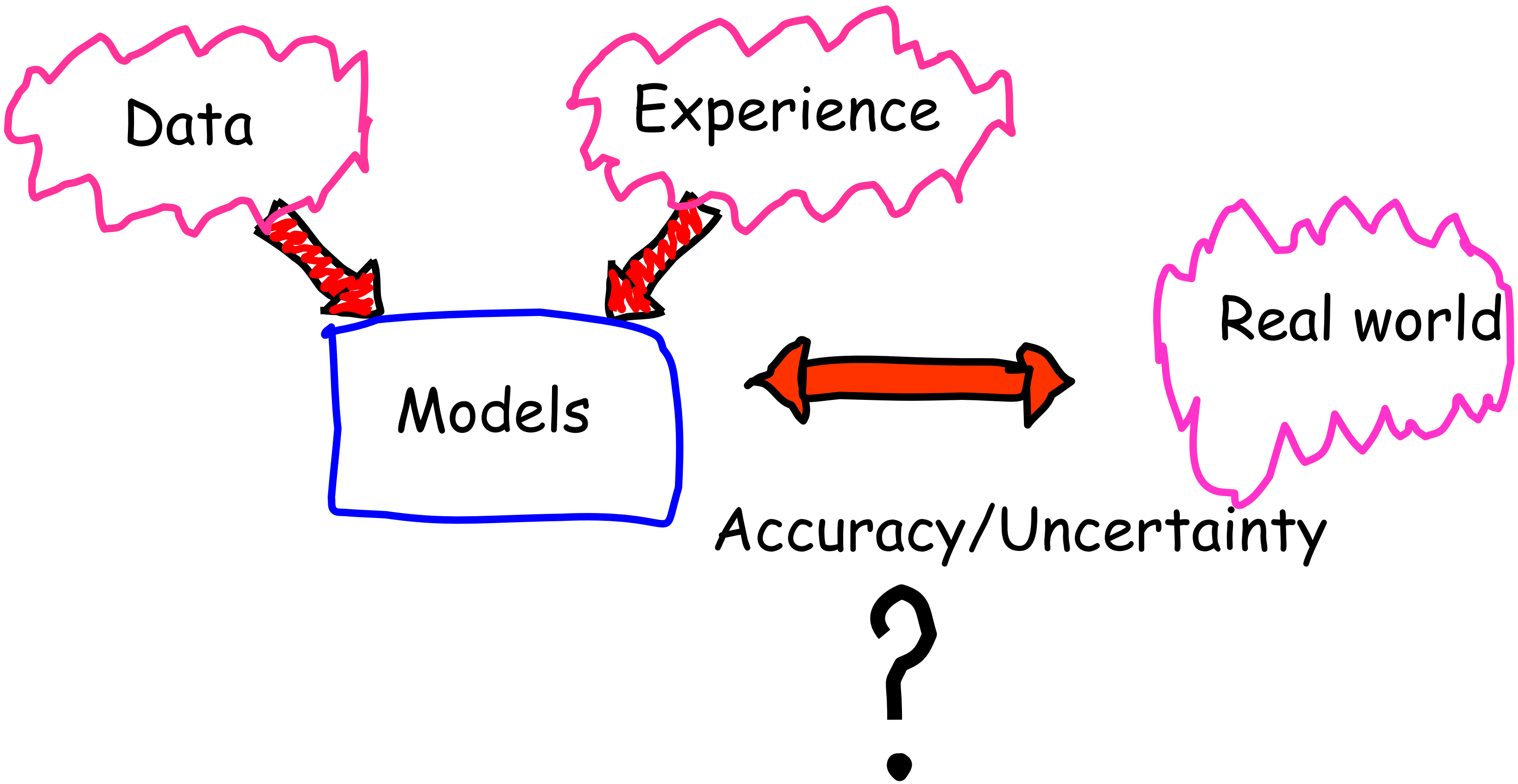
- How do engineers use knowledge

In a perfectly known world



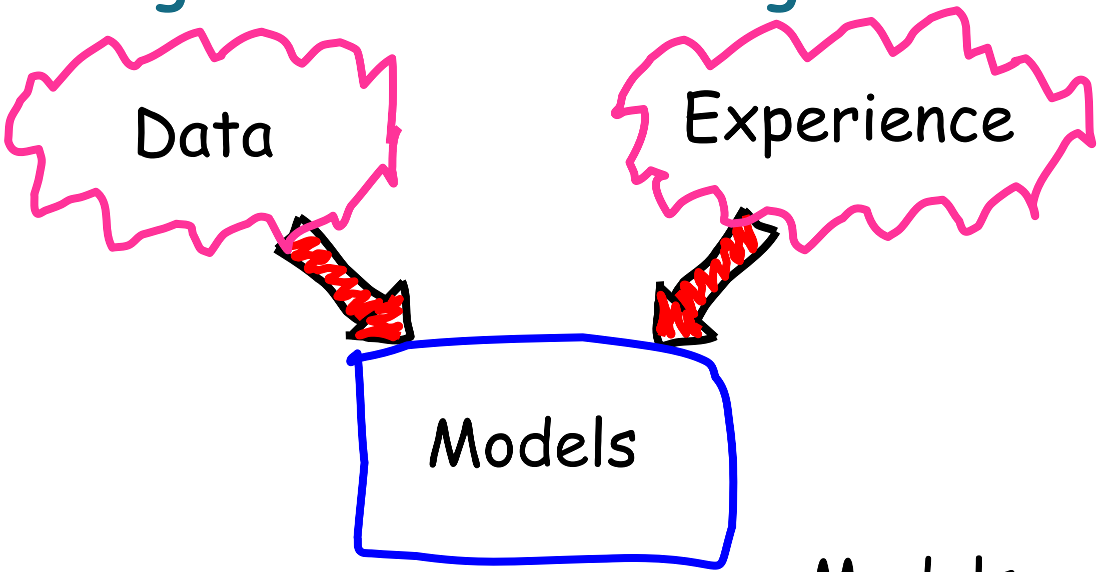
Why Statistics and Probability in Engineering?

- How do engineers establish knowledge



Why Statistics and Probability in Engineering?

- How do engineers use knowledge



Models are not precise

Uncertainty

Data are not sufficient

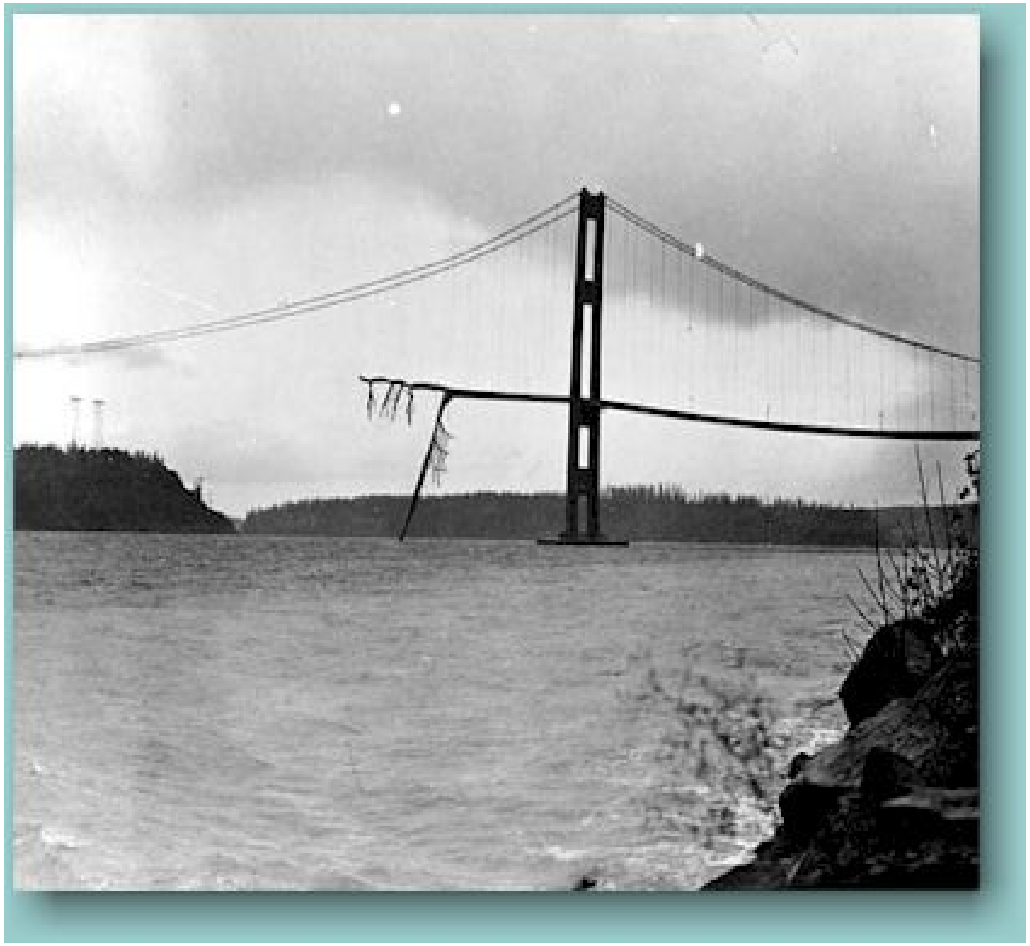
WHY ?

Natural variability

Experience is subjective

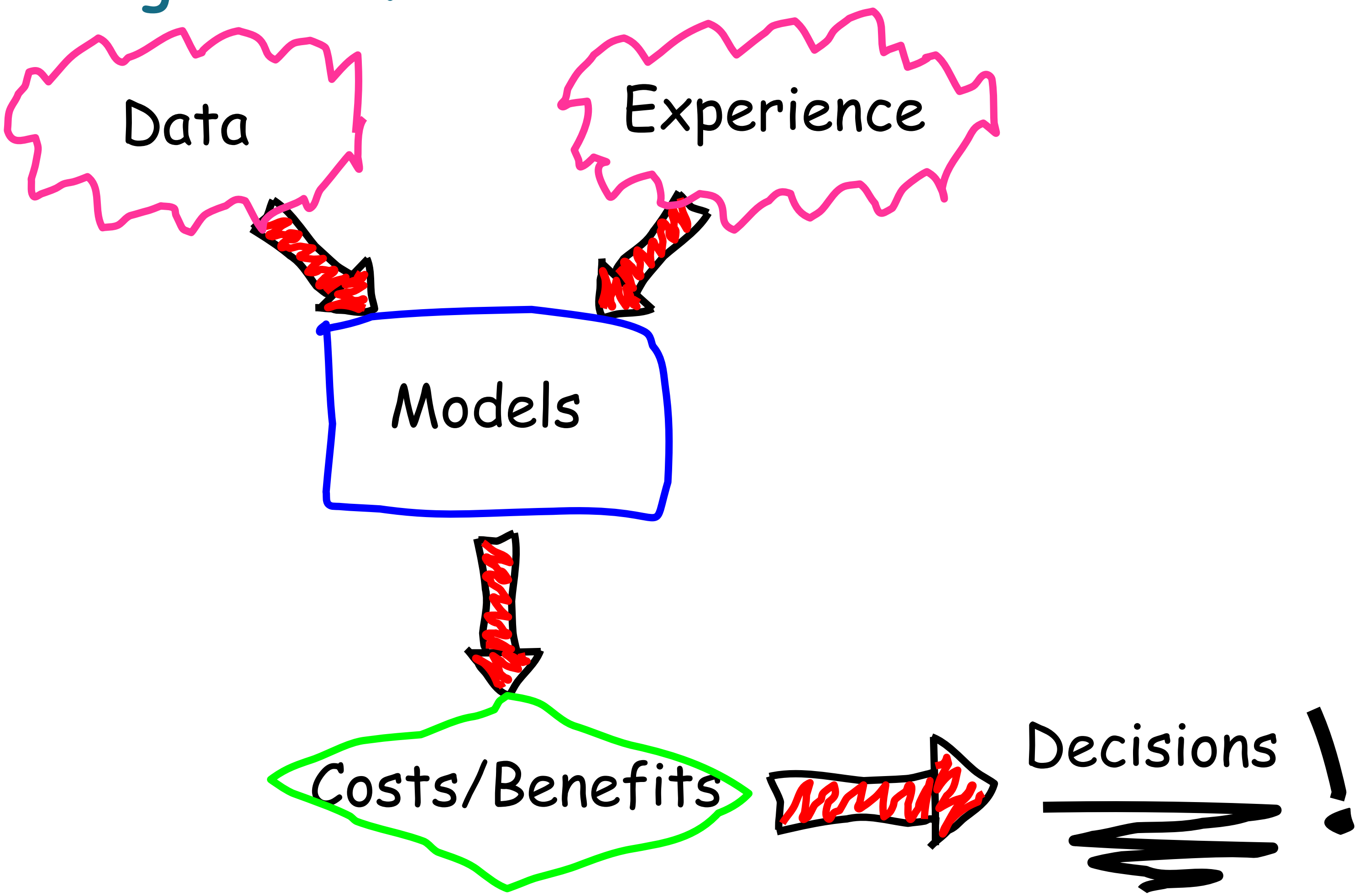
Why Statistics and Probability in Engineering?

- **An example where models were not too representative**



Why Statistics and Probability in Engineering?

- How do engineers make decisions



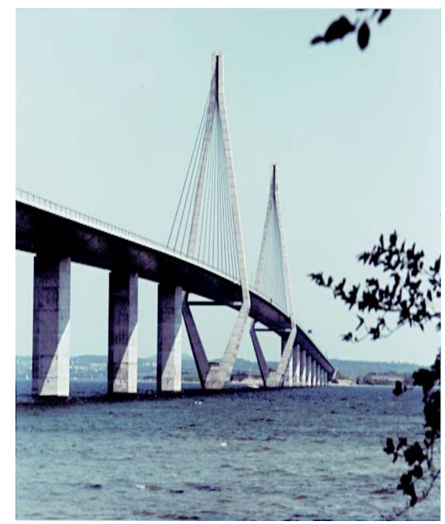
Why Statistics and Probability in Engineering?

All activities are associated with uncertainties

Activities are e.g.

- Transport
- Work
- Sport

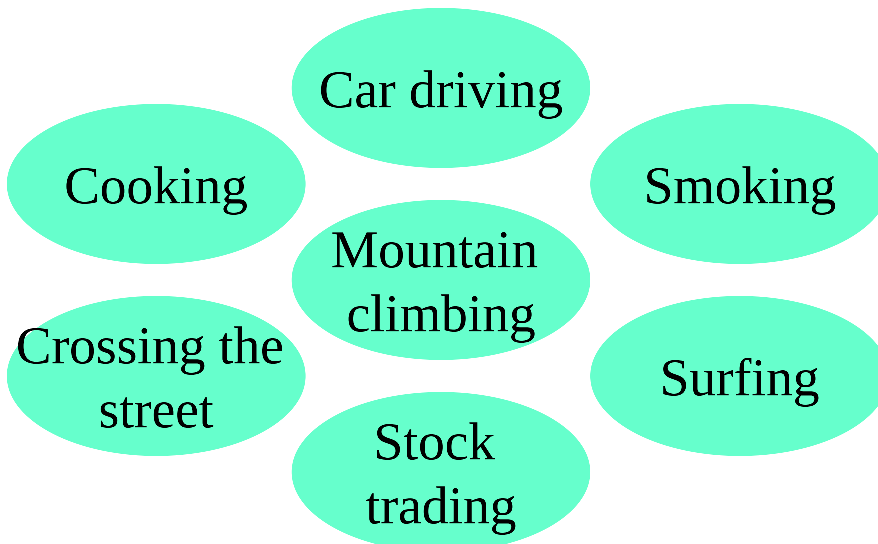
but also



- Production of energy
- Exploitation of resources
- Construction and operation of production and infrastructure projects
- Research and development

Why Statistics and Probability in Engineering?

Every day we must make decisions in regard to activities associated with uncertainties



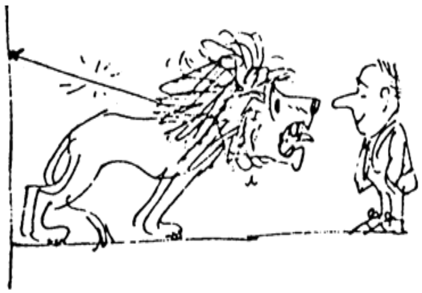
Every one of these activities is associated with uncertainties

We all have an opinion regarding the associated risks

We have gut Feelings !

Why Statistics and Probability in Engineering?

How far can we get with gut feelings ?



An example

After all - maybe it is not so
„straight forward" to comprehend uncertainties ?

What can we learn from the past ?

Why Statistics and Probability in Engineering?

Disasters and accidents have always occurred

Some examples

Tacoma Narrows, Washington, 1940

Fort Mayer, Virginia, 1908

Why Statistics and Probability in Engineering?

Disasters and accidents have always occurred

Some examples

Concord, North Carolina, 2000

Concorde, Paris, 2000

Why Statistics and Probability in Engineering?

Disasters and accidents have always occurred

Some examples

Kobe, 1995

Why Statistics and Probability in Engineering?

Disasters and accidents have always occurred

Some examples

Canada, 1993

Open questions

- did we realise the risks ?
- are the consequences acceptable ?

Why Statistics and Probability in Engineering?

Risk assessment, within the framework of decision analysis, provides a basis for rational decision making subject to uncertain and / or incomplete information

Thereby we can take into account, in a consistent manner, the prevailing uncertainties and quantify their effect on risks

Thus we may find answers to the following questions

- How large is the risk associated with a given activity ?
- How may we reduce and / or mitigate risks ?
- How much does it cost to reduce and / or mitigate risks ?
- What risks must we accept – what can we afford ?

Why Statistics and Probability in Engineering?

Risk is a characteristic of an activity relating to all possible events n_E which may follow as a result of the activity

The risk contribution R_{E_i} from the event E_i is defined through the product between

the Event probability P_{E_i}

and

the Consequences of the event C_{E_i}

The Risk associated with a given activity R_A may then be written as

$$R_A = \sum_{i=1}^{n_E} R_{E_i} = \sum_{i=1}^{n_E} P_{E_i} \cdot C_{E_i}$$

Decision Problems in Engineering

Uncertainties must be considered in the decision making throughout all phases of the life of an engineering facility

Planning and feasibility study ⟷ Investigations and tests ⟷ Design ⟷ Manufacturing ⟷ Execution ⟷ Operation & maintenance ⟷ Decommissioning

Idea & Concept ⟷ Planning and feasibility study

Uncertainties

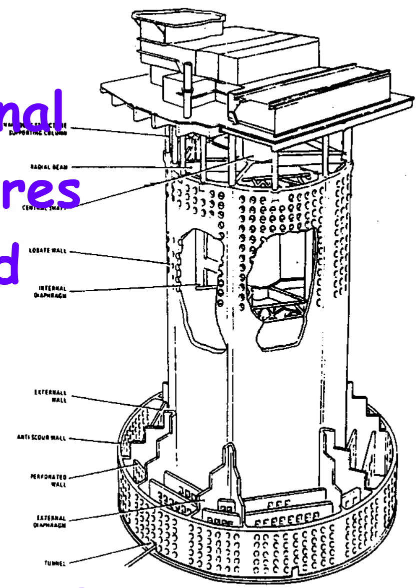
- **Traffic volume**
- **Loads**
- **Resistances (material, soil,..)**
- **Degradation processes**
- **Service life**
- **Manufacturing costs**
- **Execution costs**
- **Decommissioning costs**

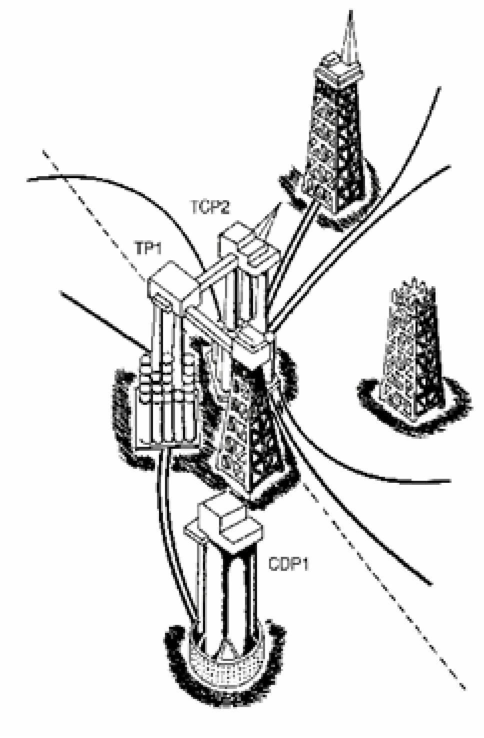
Example – Decommissioning of the Frigg Field

- The Frigg Field – built 1972-1978
  - TCP2
  - TP1
  - CDP1

According to international conventions the structures must be decommissioned

Each structure :





Weight : 250000 t

Costs : 200 – 600 Mio. SFr

- None of the platforms were designed for decommissioning !

Example – Decommissioning of the Frigg Field

- The decision problem

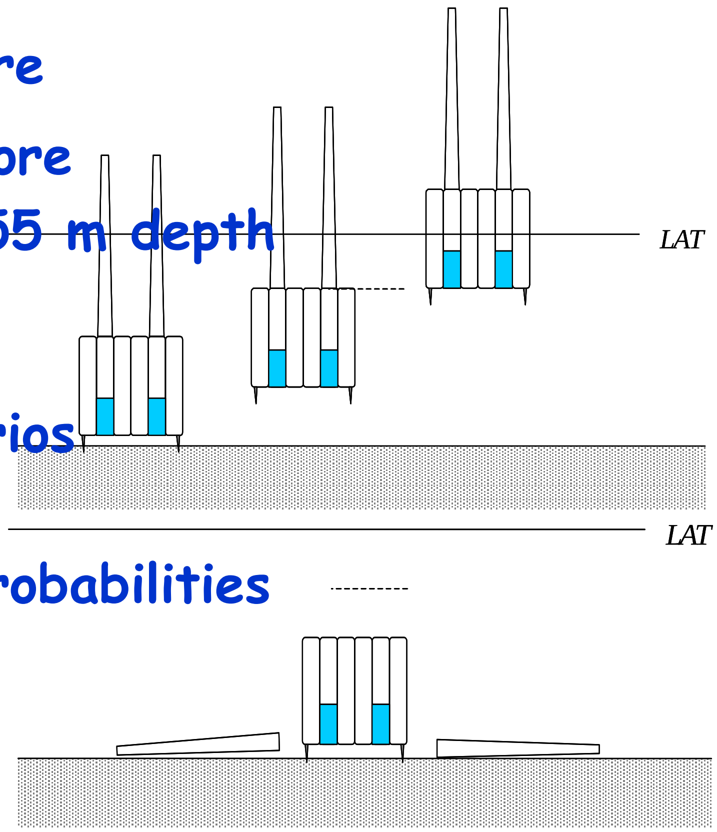
  Decommissioning/removal taking into account

  - Safety of personnel
  - Safety of the environment
  - Costs
  - Interest groups

    Greenpeace
    Fishers
    IMO
    ....

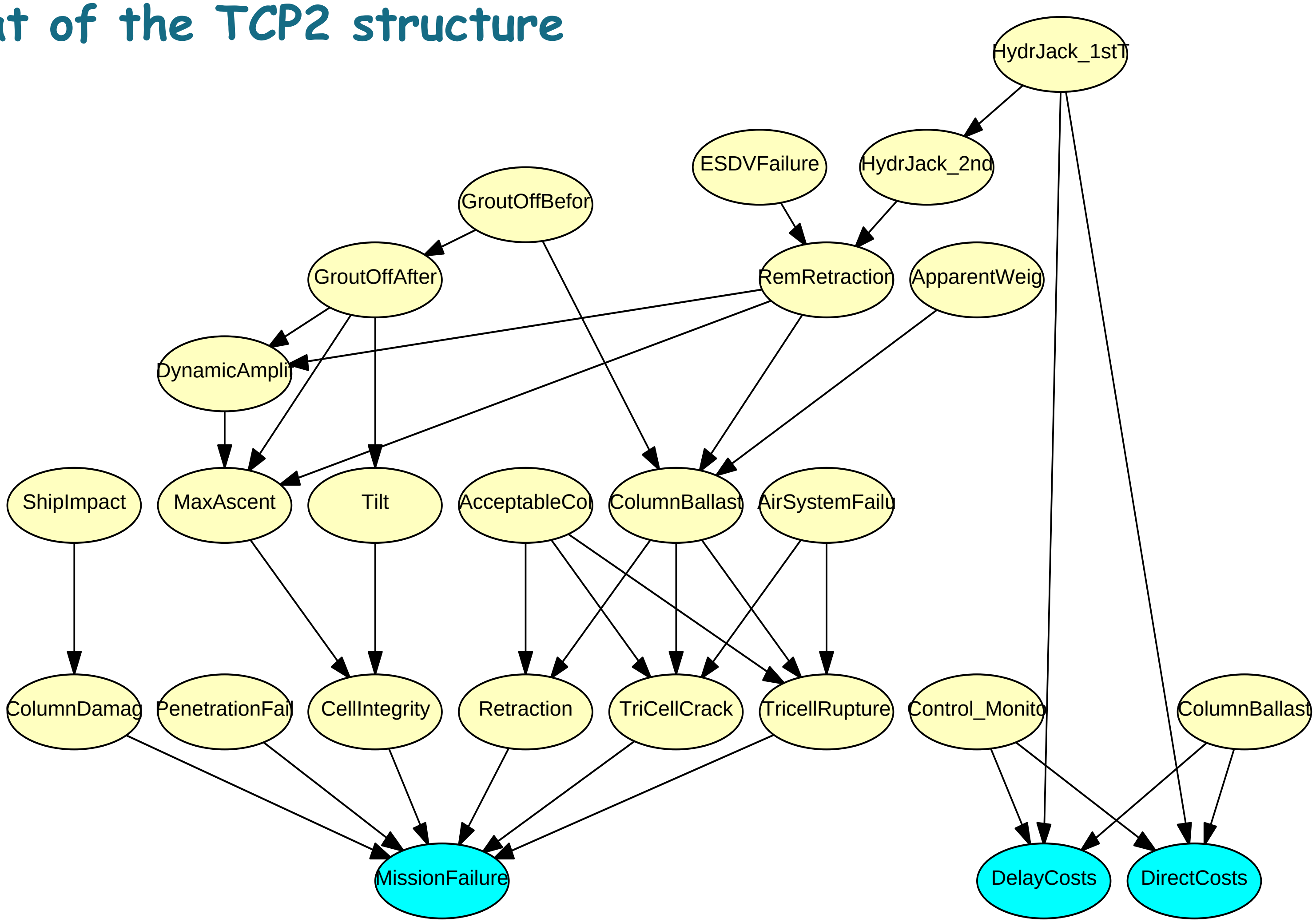
Example – Decommissioning of the Frigg Field

- Three options are considered
  - „Refloat“ and demolition Onshore
  - „Refloat“ and demolition Offshore
  - Removal to a free passage of 55 m depth
- The approach
  - Identification of hazard scenarios chronologically
  - Quantification of occurrence probabilities
  - Quantification of consequences
- Applied approach – Bayesian Nets



Example – Decommissioning of the Frigg Field

- Re-float of the TCP2 structure



Example – Decommissioning of the Frigg Field

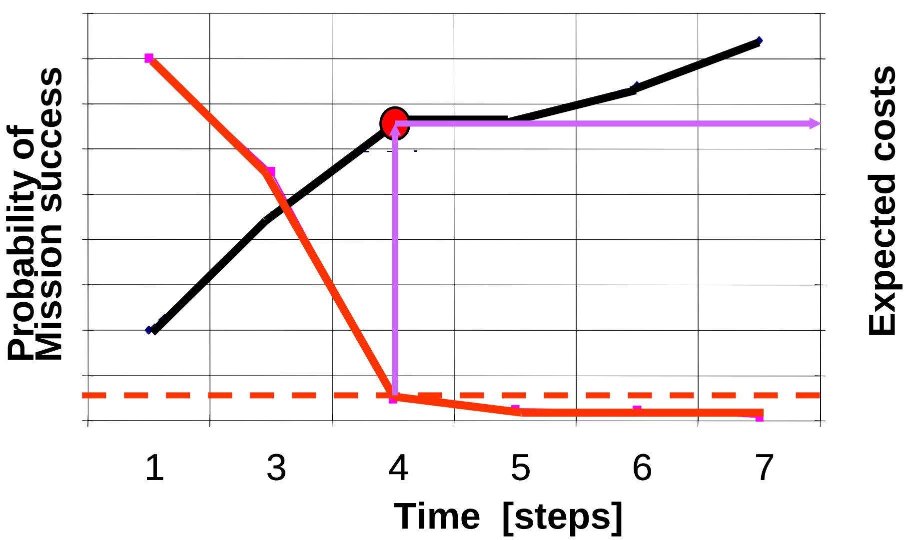
- Results of the decision analysis

Time variation of

- Expected costs
- Probability of mission success

Decision support

- How much to invest before a satisfactorily level of probability of mission success has been reached



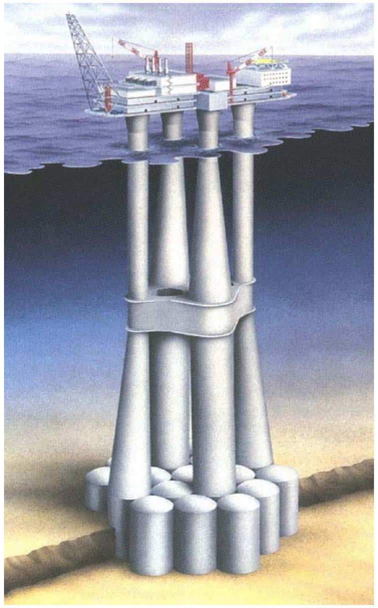
Decision Problems in Engineering

- Structural Design

Exceptional structures are often associated with structures of

„Extreme Dimensions"

Great Belt Bridge
under Construction

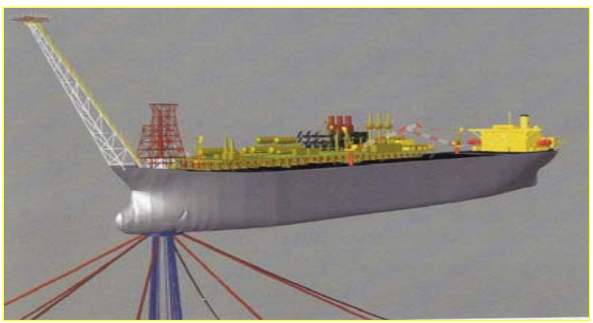
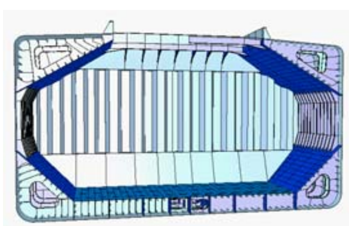


Concept drawing
of the Troll platform

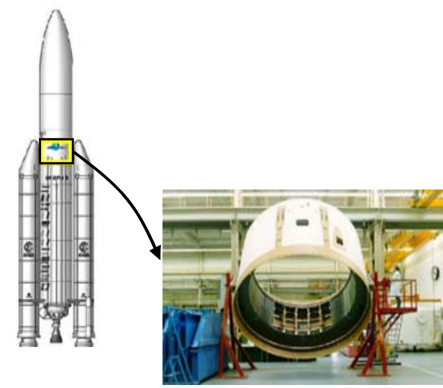
Decision Problems in Engineering

- Structural Design

  or associated with structures fulfilling
  „New and Innovative Purposes“

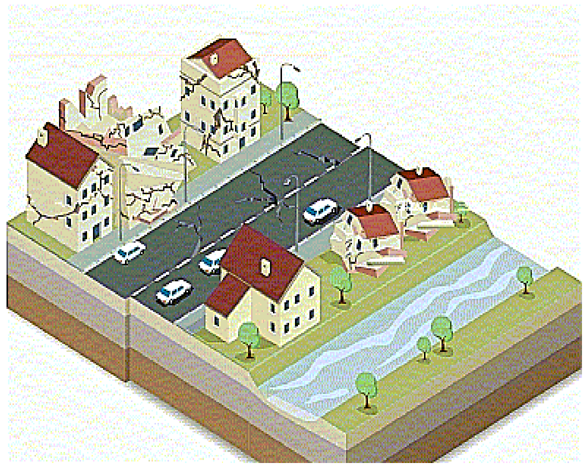
Concept drawing of
Floating Production, Storage and Offloading unit

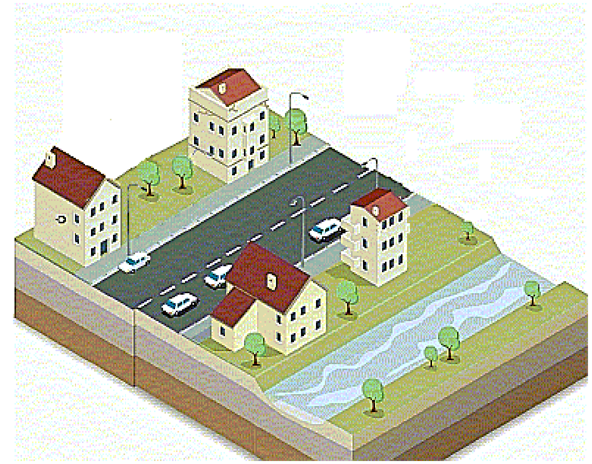
Illustrations of the ARIANE 5 rocket

Decision Problems in Engineering

Before



During



After

Before

Optimal allocation of available resources for risk reduction

- strengthening
- rebuilding

in regard to possible earth-quakes

During

Damage reduction/Control

Emergency help and rescue

After quake hazards

After

Rehabilitation of infrastructure functionality

Condition assessment and updating of reliability and risks

Optimal allocation of resources for rebuilding and strengthening

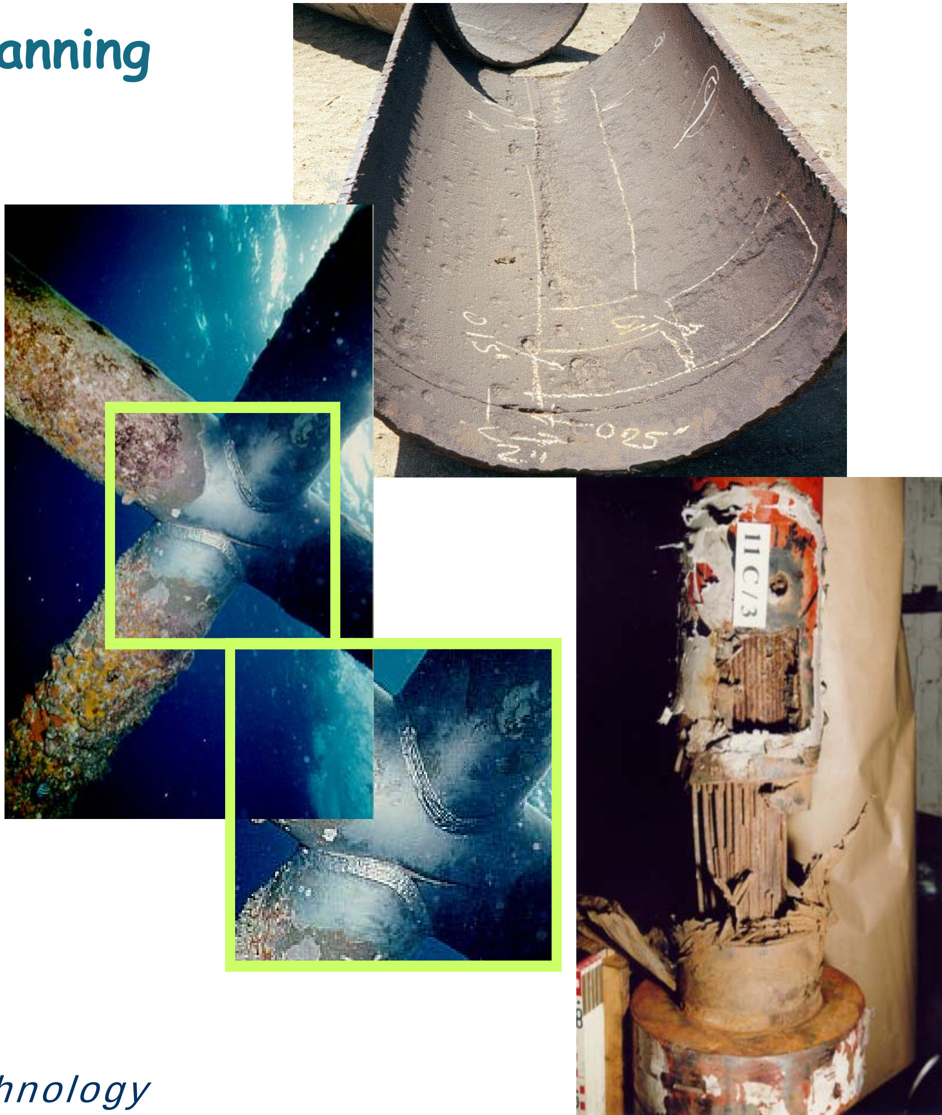
Decision Problems in Engineering

- Inspection and Maintenance Planning

  Due to
  - operational loading
  - environmental exposure

  structures will always to some degree be exposed to degradation processes such as

  - fatigue
  - corrosion
  - scour
  - wear



Why Statistics and Probability in Engineering?

- In summary

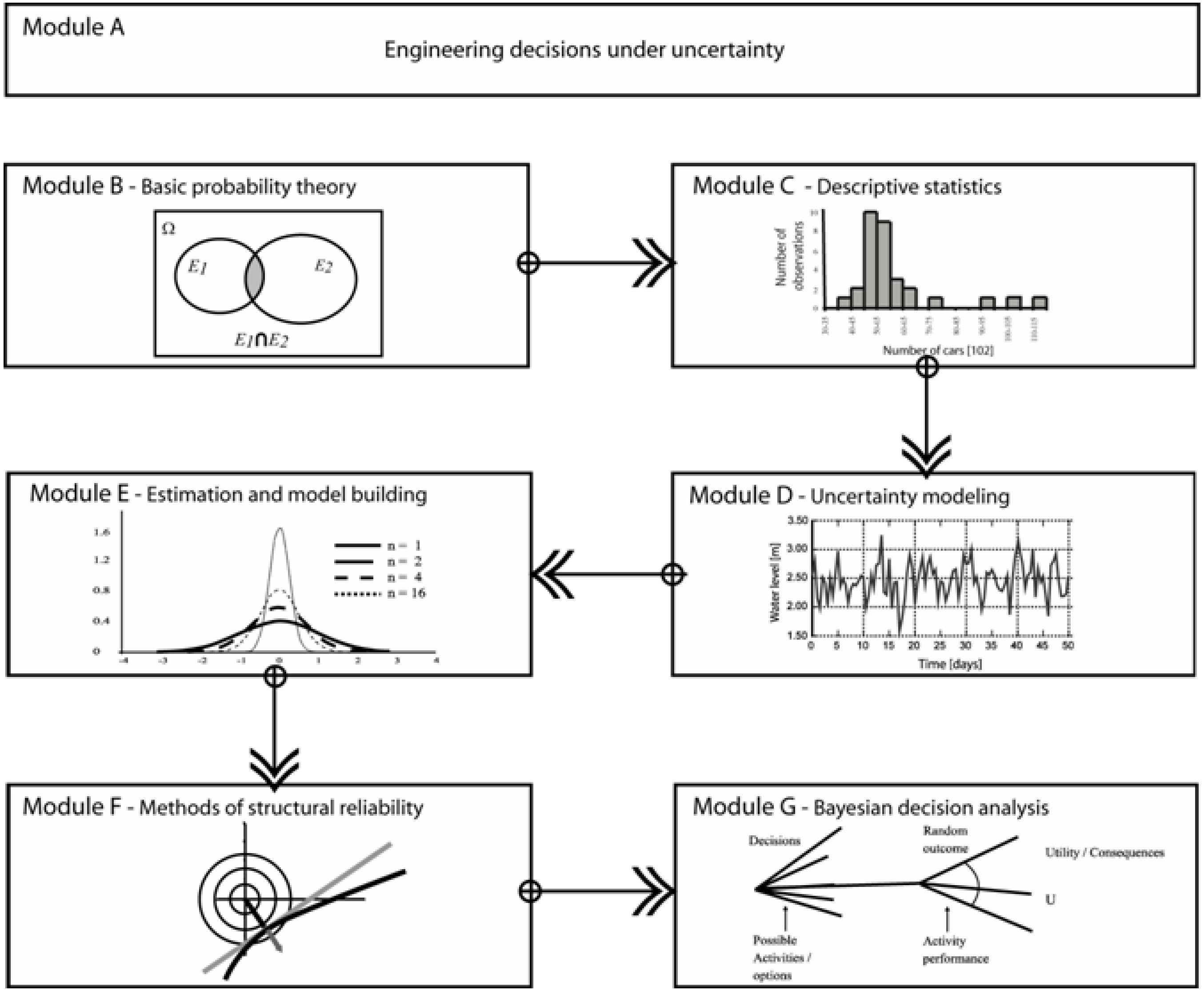
  statistics and probability theory is needed in engineering to

  - quantify the uncertainty associated with engineering models
  - evaluate the results of experiments
  - assess importance of measurement uncertainties
  - safe guard

    safety for persons
    qualities of environment
    assets

ENHANCE DECISION MAKING

Organisation of the Lecture



Basic Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Rooms information

Before…

Group	**Tutorial 1**	**Tutorials 2-9 and 11**	**Tutorial 10**
E	HIL D 10.2	HCI D 2	To be announced
H	HIL B 21	HCI H 2.1	
K	HIL F 10.3	HCI D 8	
V	HIL E 1	HPH G 3	

Now…

Group	**Tutorial 1**	**Tutorials 2-9 and11**	**Tutorial 10**
E	HCI D 2	HCI D 2	To be announced
H	HPT C103	HCI H 2.1	
K	HIL F 10.3	HCI D 8	
V	HIL E 1	HPH G 3	

Time starting (Lecture/Tutorials):

HIL: 8 **Physics/Chemistry Buildings: 7.45**

Contents of Todays Lecture

- Risk and Motivation for Risk Assessment
- Overview of Probability Theory
- Interpretation of Probability
- Sample Space and Events
- The three Axioms of Probability Theory
- Conditional Probability and Bayes's Rule

Why Statistics and Probability in Engineering?

Risk is a characteristic of an activity relating to all possible events n_E which may follow as a result of the activity

The risk contribution R_{E_i} from the event E_i is defined through the product between

the Event probability P_{E_i}

and

the Consequences of the event C_{E_i}

The Risk associated with a given activity R_A may then be written as

$$R_A = \sum_{i=1}^{n_E} R_{E_i} = \sum_{i=1}^{n_E} P_{E_i} \cdot C_{E_i}$$

Decision Problems in Engineering

Uncertainties must be considered in the decision making throughout all phases of the life of an engineering facility

Uncertainties

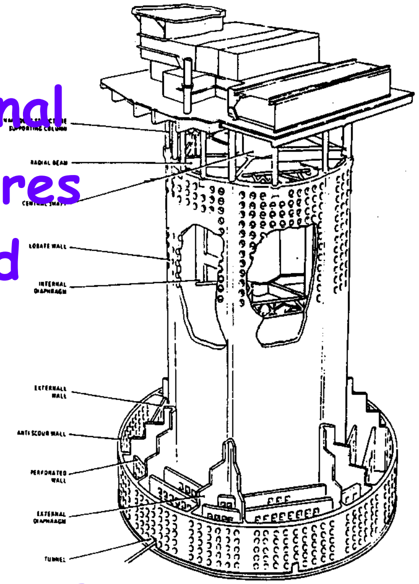
- **Traffic volume**
- **Loads**
- **Resistances (material, soil,..)**
- **Degradation processes**
- **Service life**
- **Manufacturing costs**
- **Execution costs**
- **Decommissioning costs**

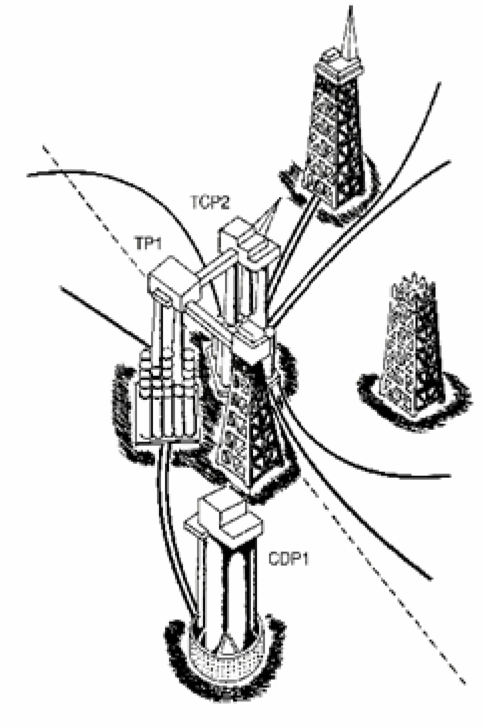
Idea & Concept ↔ Planning and feasibility study ↔ Investigations and tests ↔ Design ↔ Manufacturing ↔ Execution ↔ Operation & maintenance ↔ Decommissioning

Example – Decommissioning of the Frigg Field

- The Frigg Field – built 1972-1978
  - TCP2
  - TP1
  - CDP1

According to international conventions the structures must be decommissioned





Each structure :

Weight : 250000 t
Costs : 200 – 600 Mio. SFr

- None of the platforms were designed for decommissioning !

Example – Decommissioning of the Frigg Field

- The decision problem

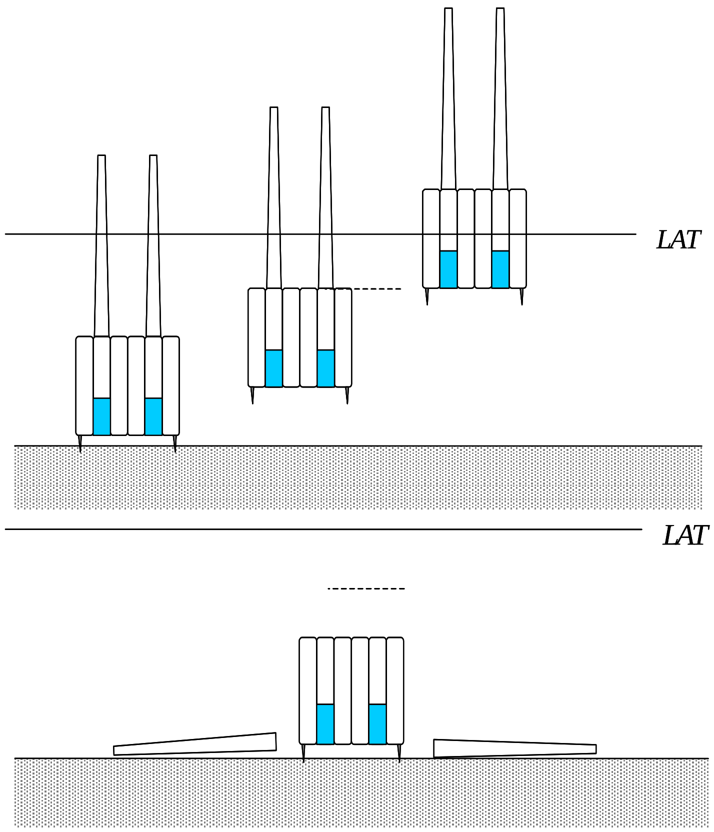
  Decommissioning/removal taking into account

  - Safety of personnel
  - Safety of the environment
  - Costs
  - Interest groups

    Greenpeace
    Fishers
    IMO
    ....

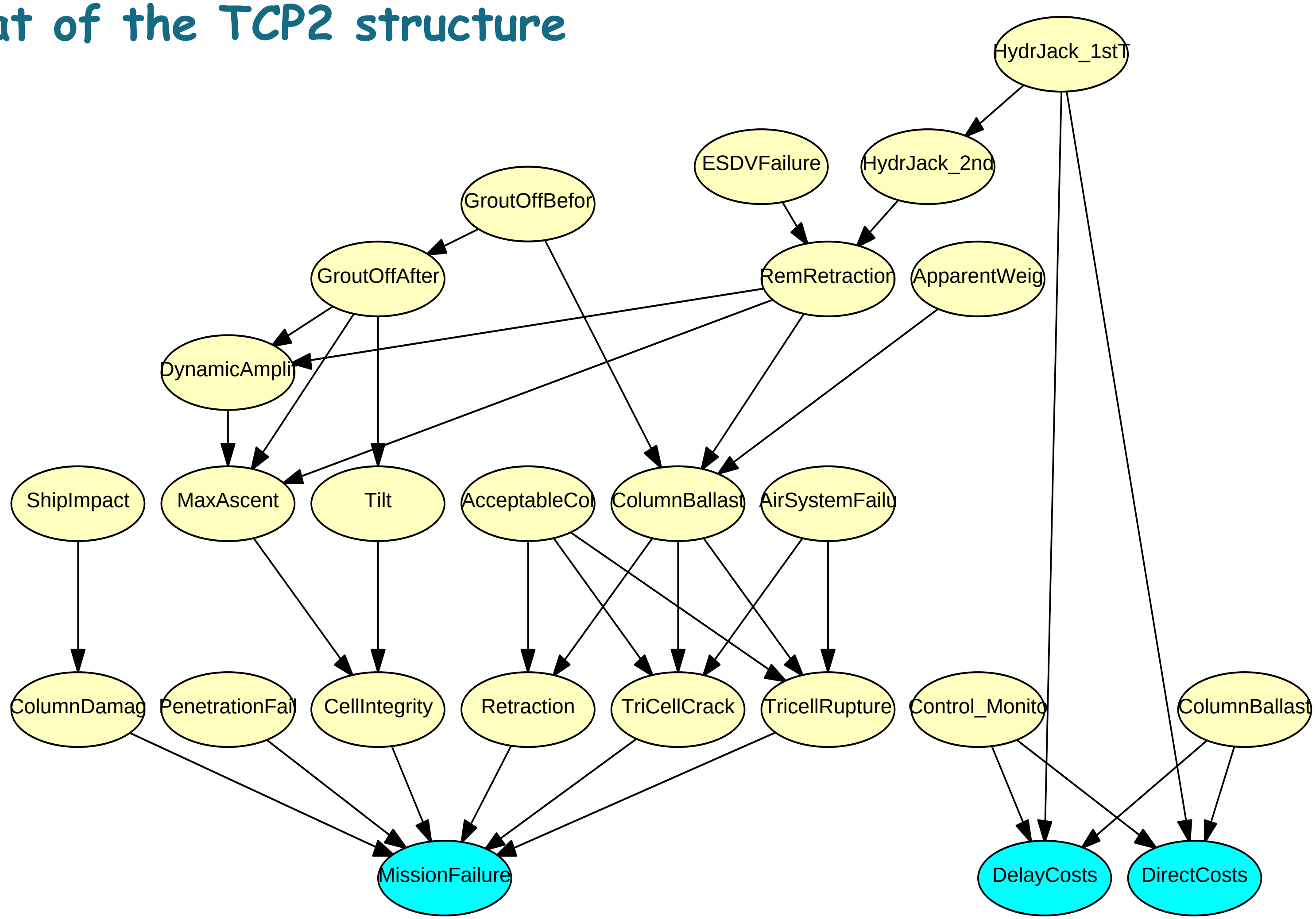
Example – Decommissioning of the Frigg Field

- Three options are considered
  - „Refloat" and demolition Onshore
  - „Refloat" and demolition Offshore
  - Removal to a free passage of 55 m depth
- The approach
  - Identification of hazard scenarios chronologically
  - Quantification of occurrence probabilities
  - Quantification of consequences
- Applied approach – Bayesian Nets



Example – Decommissioning of the Frigg Field

- Re-float of the TCP2 structure

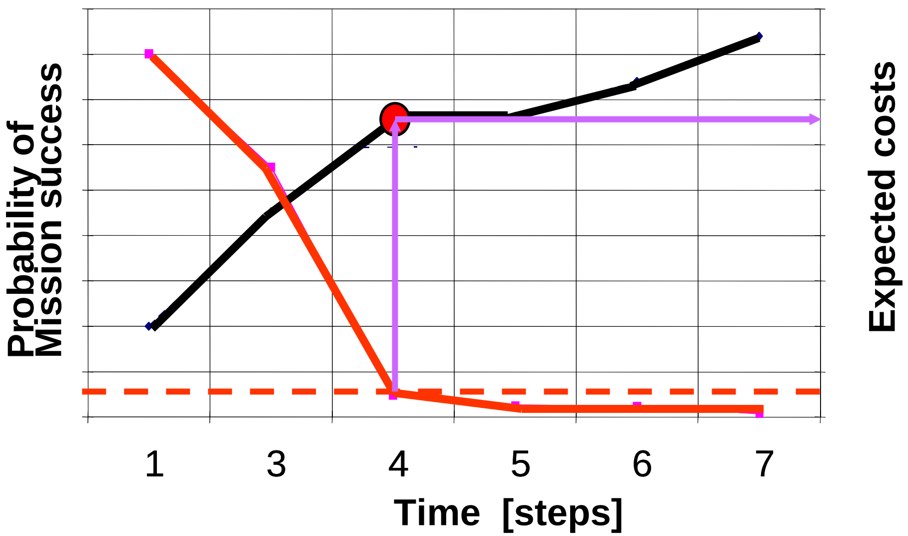


Example – Decommissioning of the Frigg Field

- Results of the decision analysis

Time variation of

- Expected costs
- Probability of mission success



Decision support

- How much to invest before a satisfactorily level of probability of mission success has been reached

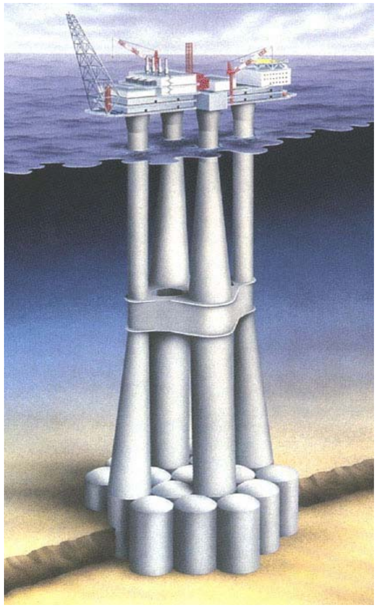
Decision Problems in Engineering

- Structural Design

Exceptional structures are often associated with structures of

„Extreme Dimensions“

Great Belt Bridge
under Construction



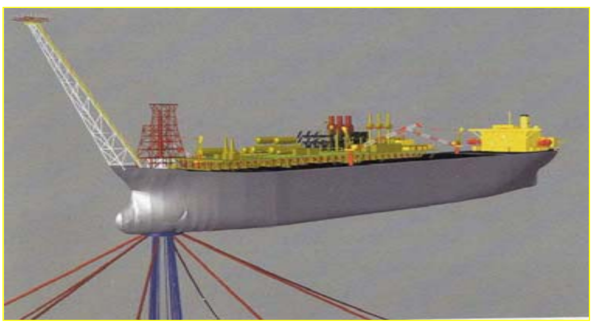
Concept drawing
of the Troll platform

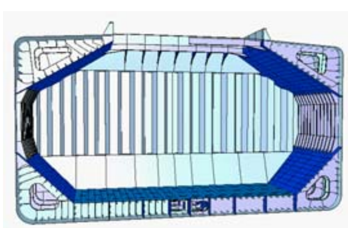
Decision Problems in Engineering

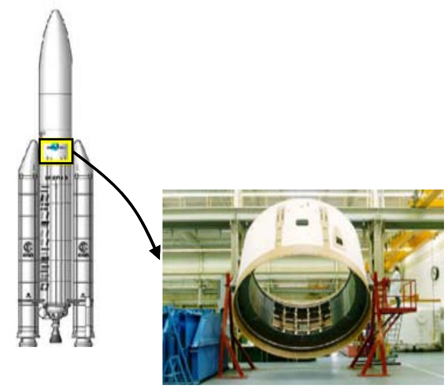
- **Structural Design**

  or associated with structures fulfilling

  „New and Innovative Purposes“





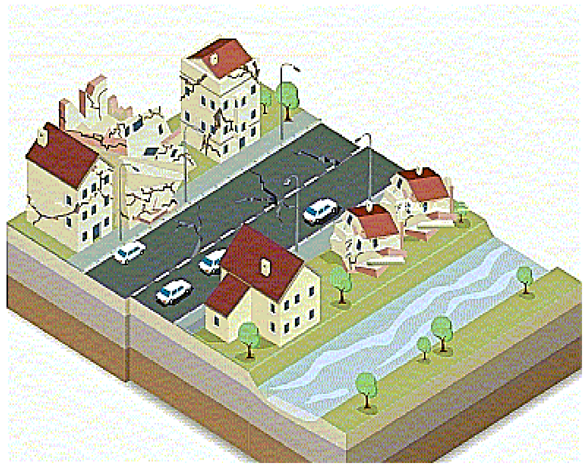


Illustrations of the ARIANE 5 rocket

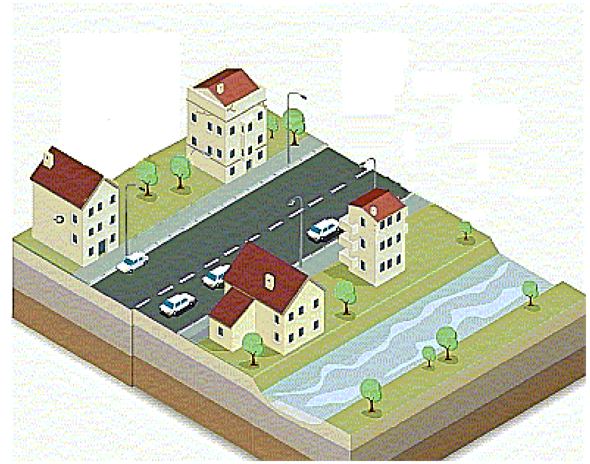
Concept drawing of
Floating Production, Storage and Offloading unit

Decision Problems in Engineering

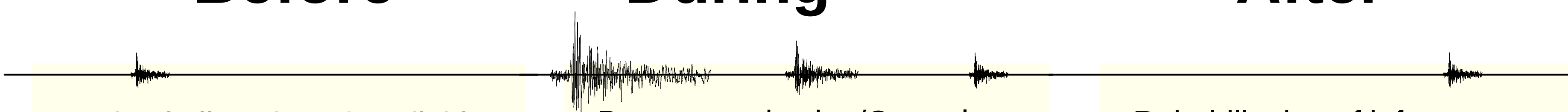
Before



During



After



Before

Optimal allocation of available resources for risk reduction

- strengthening
- rebuilding

in regard to possible earth-quakes

During

Damage reduction/Control

Emergency help and rescue

After quake hazards

After

Rehabilitation of infrastructure functionality

Condition assessment and updating of reliability and risks

Optimal allocation of resources for rebuilding and strengthening

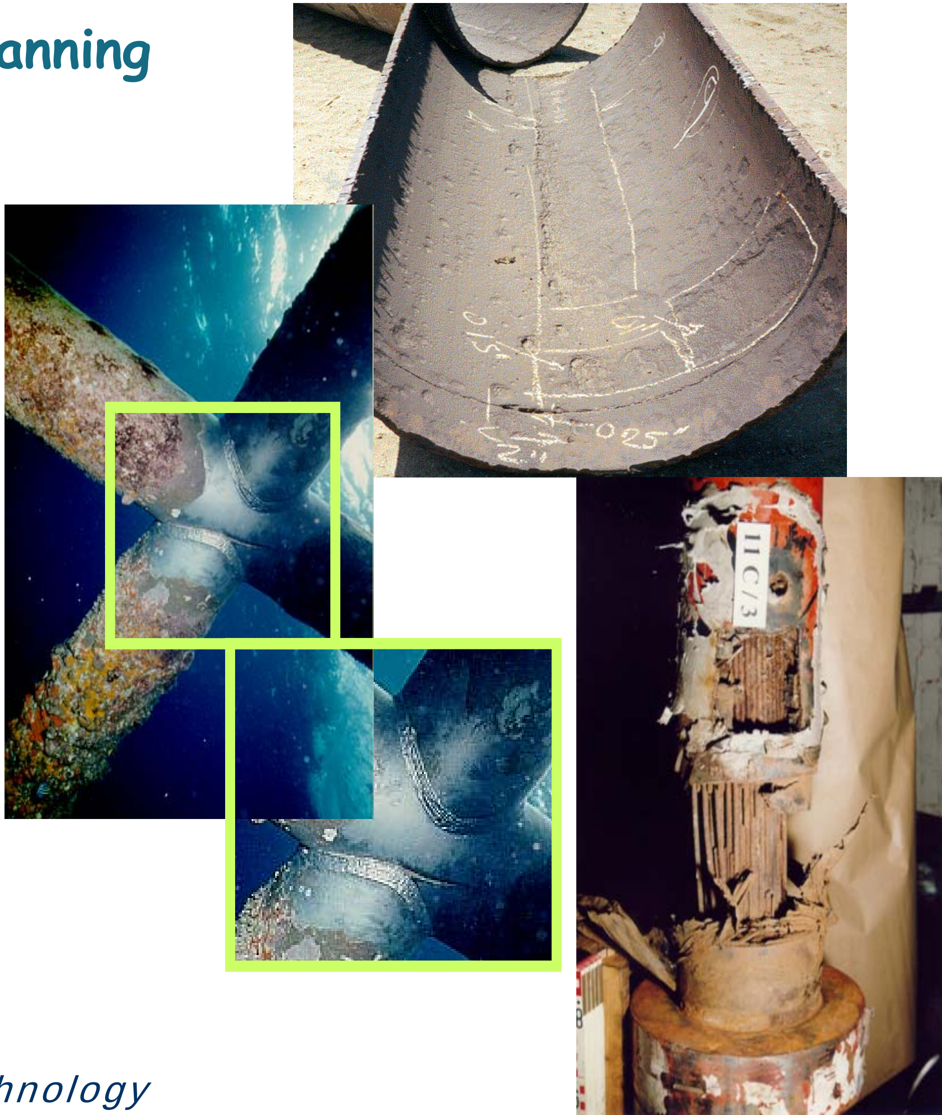
Decision Problems in Engineering

- Inspection and Maintenance Planning

  Due to
  - operational loading
  - environmental exposure

  structures will always to some degree be exposed to degradation processes such as

  - fatigue
  - corrosion
  - scour
  - wear



Why Statistics and Probability in Engineering?

In summary

statistics and probability theory is needed in engineering to

- quantify the uncertainty associated with engineering models
- evaluate the results of experiments
- assess importance of measurement uncertainties
- safe guard

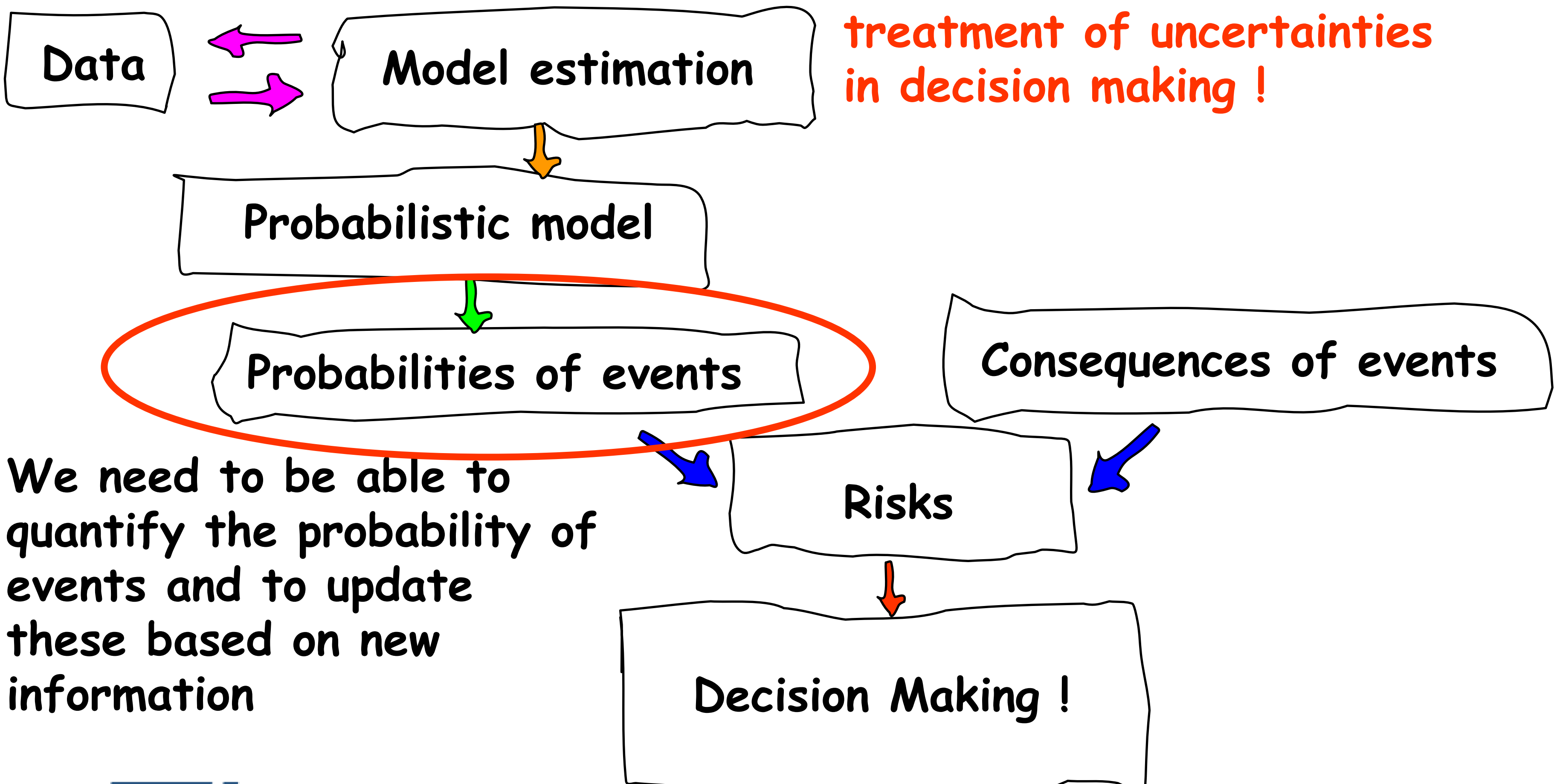
  safety for persons
  qualities of environment
  assets

ENHANCE DECISION MAKING

Overview of Probability Theory

- What are we aiming for ?

The probability theory provides the basis for the consistent treatment of uncertainties in decision making !



We need to be able to quantify the probability of events and to update these based on new information

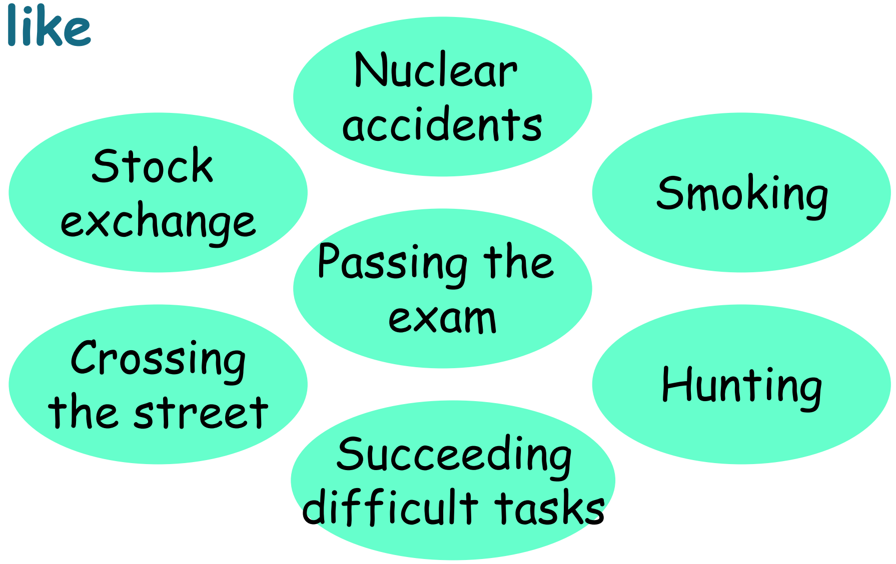
Interpretation of Probability

- What is Probability ?

  We all have some notion of probability !

  and frequently use words like

  - Chance
  - Likelihood
  - Frequency
  - Probability



Interpretation of Probability

States of nature of which we have interest such as:

- a bridge failing due to excessive traffic loads
- a water reservoir being over-filled
- an electricity distribution system „falling out“
- a project being delayed

are in the following denoted „events“

we are generally interested in quantifying the probability that such events take place within a given „time frame“

Interpretation of Probability

- There are in principle three different interpretations of probability

  - Frequentistic $P(A) = \lim \frac{N_A}{n_{\exp}}$ **for** $n_{\exp} \to \infty$
  - Classical $P(A) = \frac{n_A}{n_{tot}}$
  - Bayesian $P(A) = \text{ degree of belief that } A \text{ will occur}$

Interpretation of Probability

Consider the probability of getting a „head“ when flipping a coin

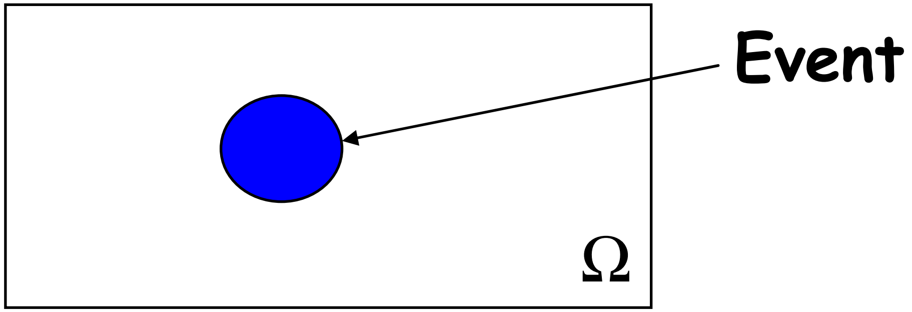
- **Frequentistic** $P(A)=\frac{510}{1000}=0.51$
- **Classical** $P(A)=\frac{1}{2}$
- **Bayesian** $P(A)=0.5$

Sample Space and Events

The set of all possible outcomes of the state of nature e.g. concrete compressive strength test results is called the sample space Ω. For concrete compressive strength test results the sample space can be written as $\Omega =]0;\infty[$

A sample space can be continuous or discrete.

Typically we illustrate the sample space and events using Venn diagrams

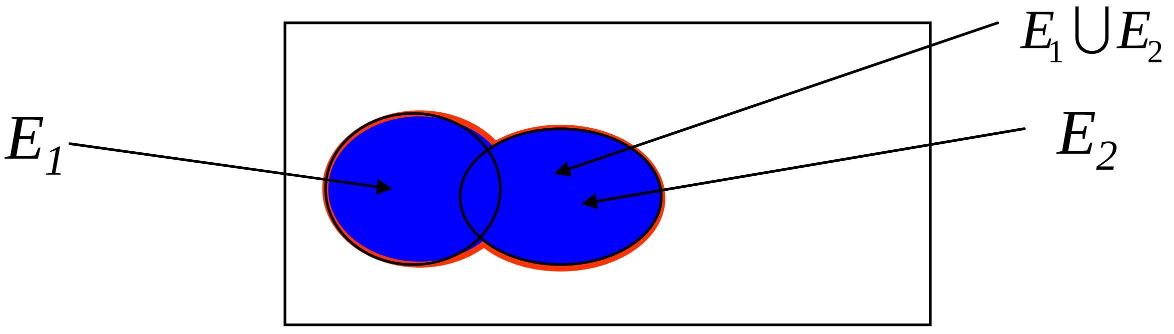


Sample Space and Events

An event is a sub-set of the sample space

- if the sub-set is empty the event is impossible
- if the sub-set contains all of the sample space the event is certain

Consider the two events E_1 and E_2:
The sub-set of sample points belonging to the event E_1 and/or the event E_2 is called the **union** of E_1 and E_2 and is written as : $E_1 \cup E_2$

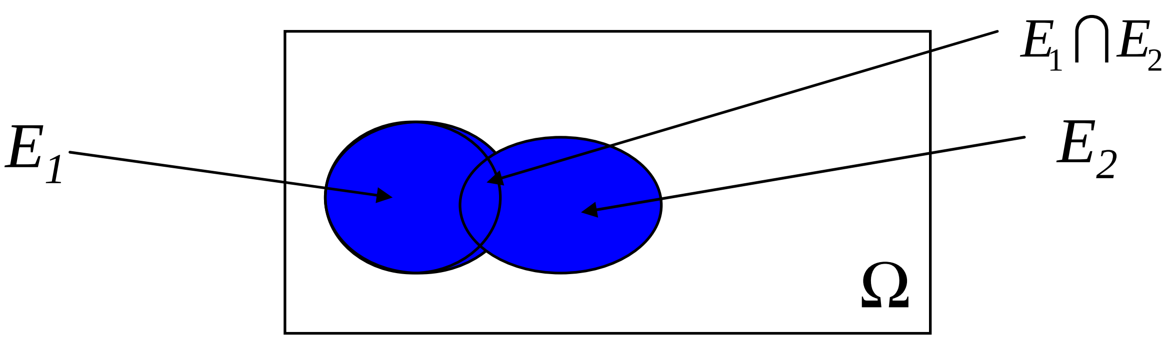


Sample Space and Events

An event is a sub-set of the sample space

- if the sub-set is empty the event is impossible
- if the sub-set contains all of the sample space the event is certain

Consider the two events E_1 and E_2 :
The sub-set of sample points belonging to the event E_1 and the event E_2 is called the **intersection** of E_1 and E_2 and is written as: $E_1 \cap E_2$

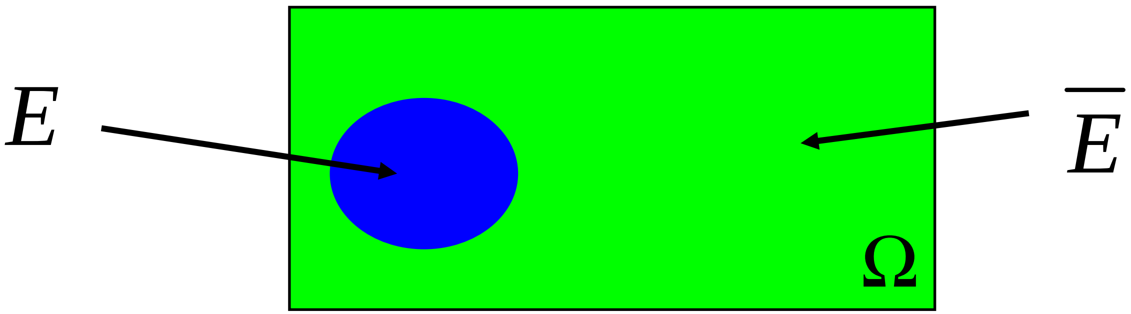


Sample Space and Events

The event containing all sample points in Ω not included in the event E is called the complementary event to E and written as : $\overline{E}$

It follows that $E \cup \overline{E} = \Omega$

and $E \cap \overline{E} = \varnothing$



Sample Space and Events

It can be show that the intersection and union operations obey the following commutative, associative and distributive laws:

$$E_1 \cap E_2 = E_2 \cap E_1$$

Commutative law

$$E_1 \cap (E_2 \cap E_3) = (E_1 \cap E_2) \cap E_3$$

$$E_1 \cup (E_2 \cup E_3) = (E_1 \cup E_2) \cup E_3$$

Associative law

$$E_1 \cap (E_2 \cup E_3) = (E_1 \cap E_2) \cup (E_1 \cap E_3)$$

$$E_1 \cup (E_2 \cap E_3) = (E_1 \cup E_2) \cap (E_1 \cup E_3)$$

Distributive law

Sample Space and Events

From the commutative, associative and distributive laws the so-called De Morgan's laws may be derived:

$$E_1 \cap E_2 = E_2 \cap E_1$$

$$E_1 \cap (E_2 \cap E_3) = (E_1 \cap E_2) \cap E_3$$

$$E_1 \cup (E_2 \cup E_3) = (E_1 \cup E_2) \cup E_3$$

$$E_1 \cap (E_2 \cup E_3) = (E_1 \cap E_2) \cup (E_1 \cap E_3)$$

$$E_1 \cup (E_2 \cap E_3) = (E_1 \cup E_2) \cap (E_1 \cup E_3)$$

$$\Rightarrow$$

$$E_1 \cap E_2 = \overline{\bar{E}_1 \cup \bar{E}_2}$$

$$E_1 \cup E_2 = \overline{\bar{E}_1 \cap \bar{E}_2}$$

The Three Axioms of Probability Theory

The probability theory is built up on – only – three axioms due to Kolmogorov:

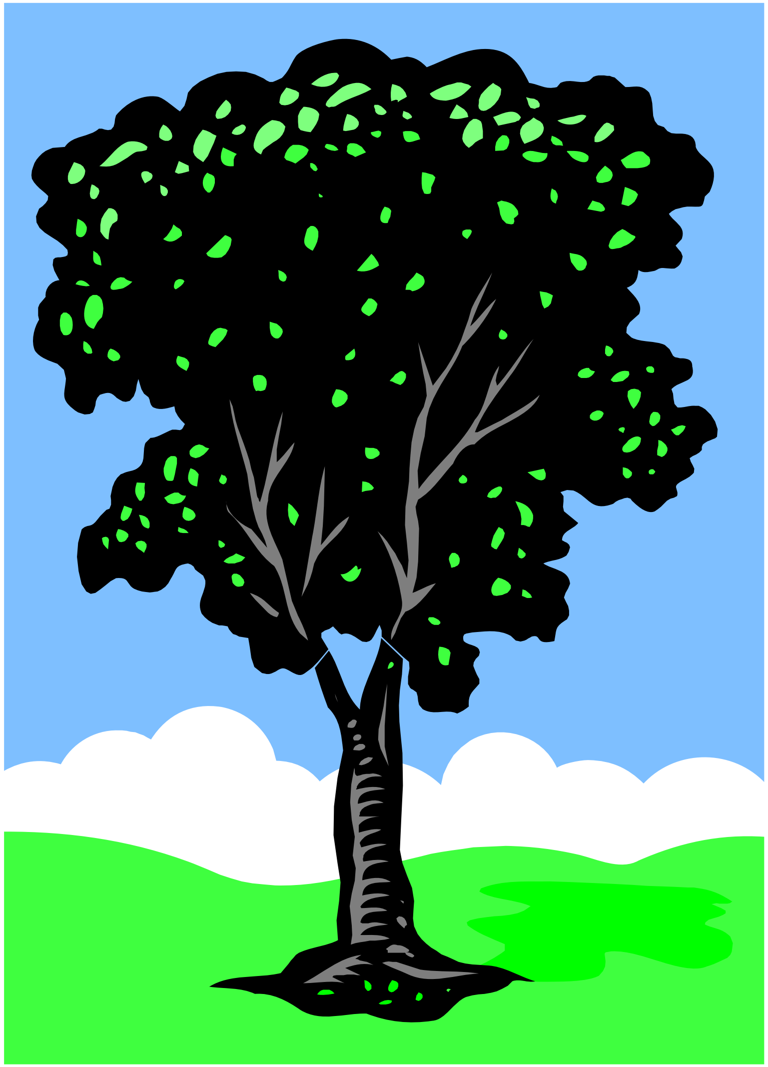
Axiom 1: $0 \leq P(E) \leq 1$

Axiom 2: $P(\Omega) = 1$

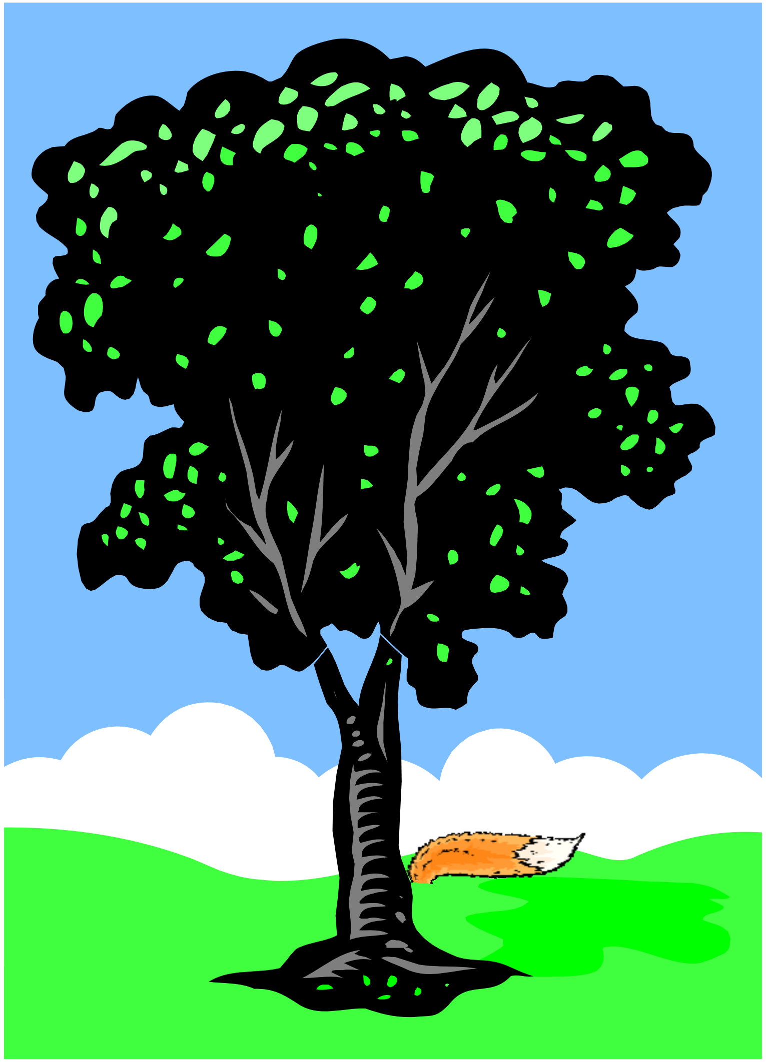
Axiom 3: $P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$

When E_1, E_2,... are mutually exclusive

Conditional Probability and Bayes's Rule



Conditional Probability and Bayes's Rule



Conditional Probability and Bayes's Rule

Conditional Probability and Bayes's Rule

Formulate hypothesis about the world

Utilize existing knowledge

Combine with data

Learn how to develop knowledge !

Conditional Probability and Bayes's Rule

Conditional probabilities are of special interest as they provide the basis for utilizing new information in decision making.

The conditional probability of an event E_1 given that event E_2 has occured is written as:

$$P(E_1 | E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

Not defined if $P(E_2) = 0$

The events E_1 and E_2 are said to be statistically independent if:

$$P(E_1 | E_2) = P(E_1)$$

Conditional Probability and Bayes's Rule

From $$P(E_1 | E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

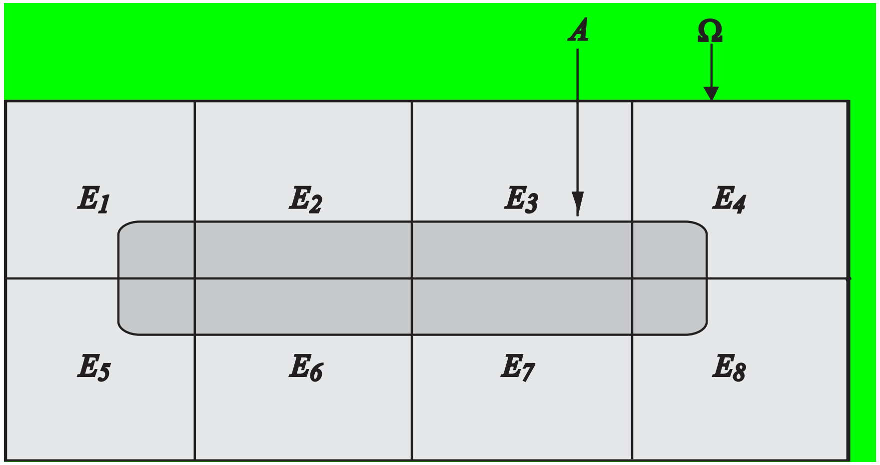
it follows that $P(E_1 \cap E_2) = P(E_2)P(E_1 | E_2)$

and when E_1 and E_2 are statistically independent there is

$$P(E_1 \cap E_2) = P(E_2)P(E_1)$$

Conditional Probability and Bayes's Rule

Consider the sample space Ω divided up into n mutually exclusive events $E_1,\ E_2,\ \ldots,\ E_n$



$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \ldots + P(A \cap E_n)$$

$$P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \ldots + P(A|E_n)P(E_n) =$$

$$\sum_{i=1}^{n} P(A|E_i)P(E_i)$$

Conditional Probability and Bayes's Rule

as there is $P(A \cap E_i) = P(A|E_i)P(E_i) = P(E_i|A)P(A)$

we have

Likelihood

Prior

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)} = \frac{P(A|E_i)P(E_i)}{\sum_{i=1}^{n} P(A|E_i)P(E_i)}$$

Posterior

Bayes Rule

Reverend Thomas Bayes (1702-1764)

Basic Statistics and Probability Theory

in

Civil, Surveying and Environmental Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zürich, Switzerland

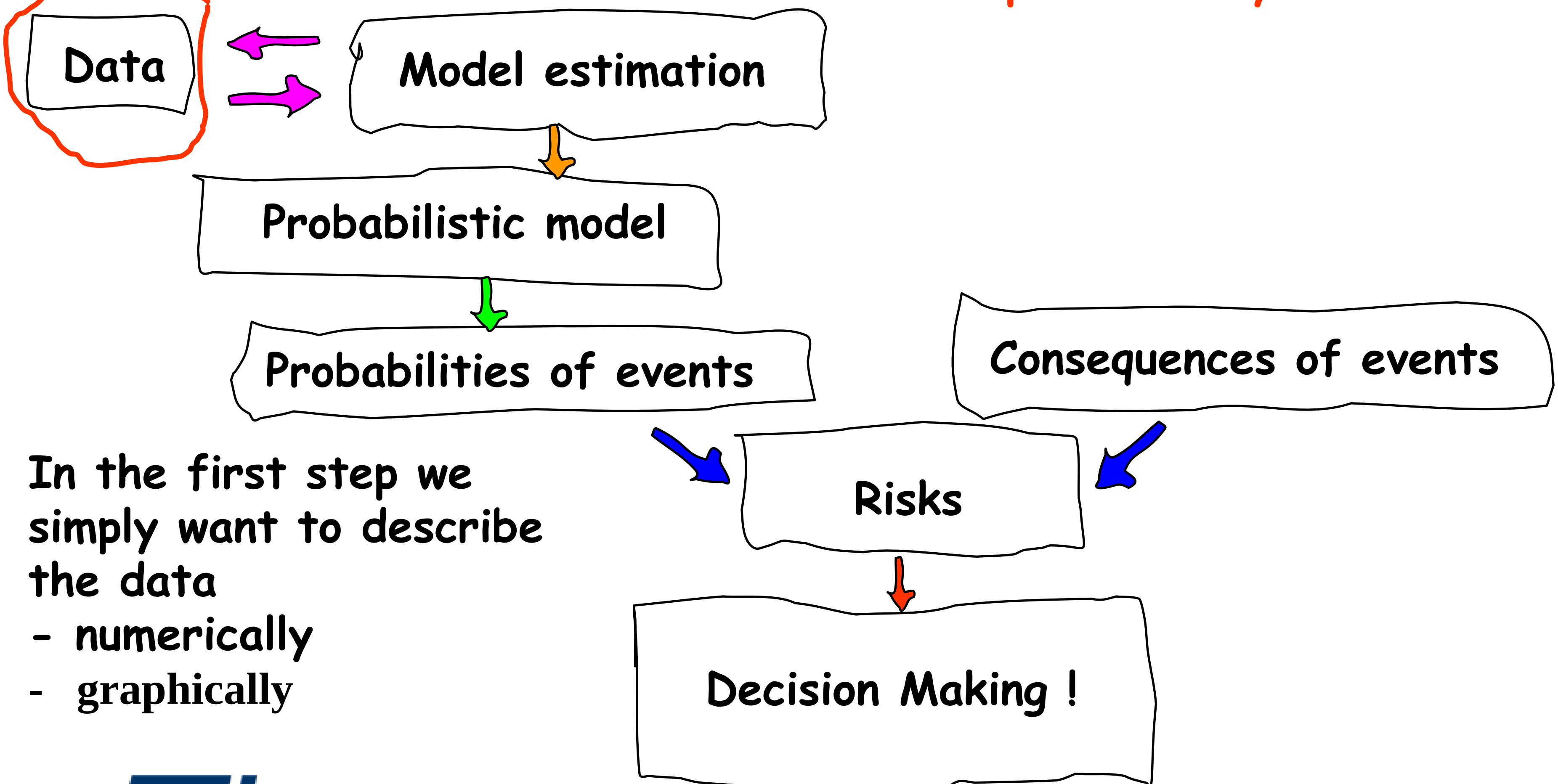
Contents of Todays Lecture

- Overview of descriptive statistics
- Numerical summaries
  - Central measures
  - Dispersion measures
  - Other measures
  - Measures of correlation
- Graphical representations
  - One-dimensional scatter plots
  - Histograms
  - Quantile plots
  - Tukey Box plots
  - Q-Q plots and Tukey mean-difference plot

Overview of Descriptive Statistics

- What we are aiming for ?

Descriptive statistics make no assumptions – only describe !



In the first step we simply want to describe the data
- numerically
- graphically

Numerical Summaries

- Central measures:

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

If one number should be given to represent a data set typically the sample mean would be chosen

Median : The 0.5 quantile (obtained from ordered data sets, see quantile
plots)

Mode : Most frequent value – obtained from histograms

Numerical Summaries

- **Dispersion measures:**

Sample variance: $$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$ s: standard deviation

Indicator of variability around the sample mean

Sample coefficient of variation (CoV): $$\nu = \frac{s}{\bar{x}}$$

Indicator of variability relative to the sample mean

Numerical Summaries

- Other measures:

Sample skewness:

$$\eta = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

Measure of symmetry

Sample kurtosis

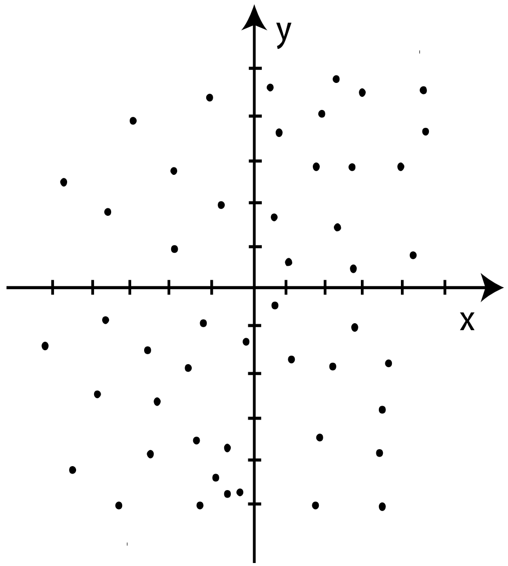
$$\kappa = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4}$$

Measure of peakedness

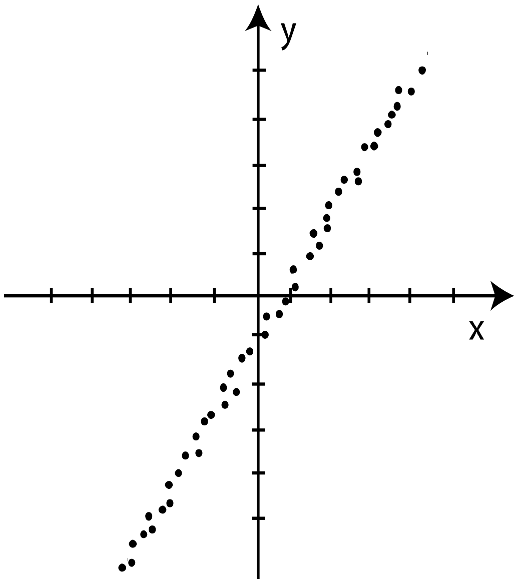
Numerical Summaries

- Measures of correlation (linear dependency between data pairs):

2-dimensional scatter plots



Almost no dependency



Almost full dependency

Numerical Summaries

- Measures of correlation (linear dependency between data pairs):

Sample covariance:

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\cdot(y_i - \overline{y})$$

The sum will get positive contributions in case of low-low or high-high data pairs

Sample coefficient of correlation:

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

r_{XY} is limited in the interval -1 to +1

Numerical Summaries

- Summary:

Central measures:

- sample mean value:	The center of gravity of a data set
- sample median:	The mid value of a data set
- sample mode:	The most frequent value/range of a data set

Dispersion measures:

- sample variance:	The distribution around the sample mean
- sample CoV:	The variability relative to the sample mean

Other measures:

- sample skewness:	The skewness relative to the sample mean
- sample kurtosis:	The peakedness around the sample mean

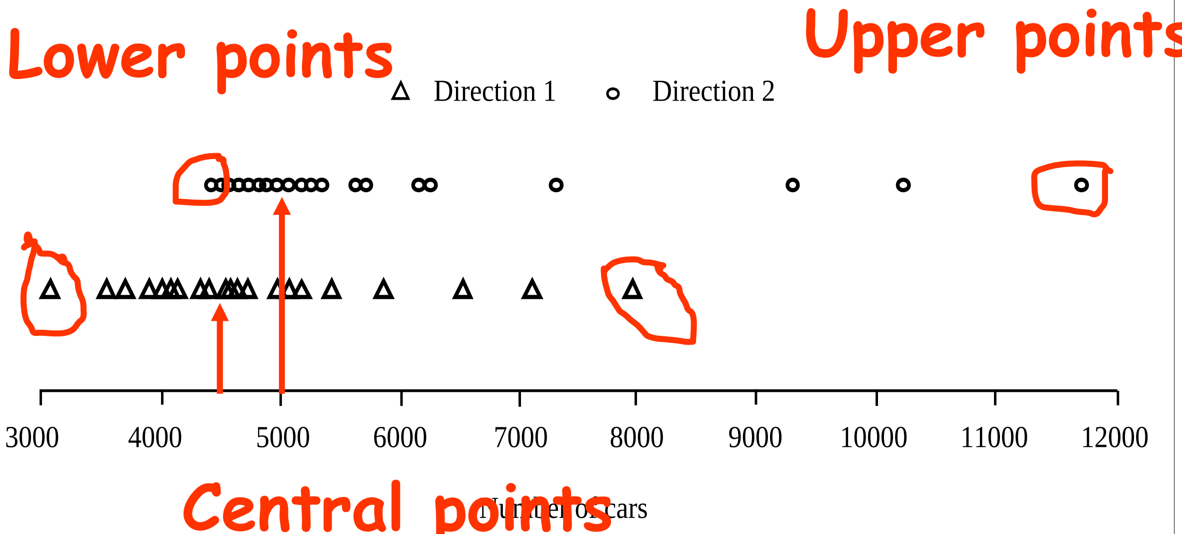
Measures of correlation:

- sample covariance:	Tendency for high-high, low-low and high-low pairs in two data sets
- sample coefficient of correlation :	Normalized coefficient between -1 and +1

Graphical Representations

- Assume that we have a set of data (observations of road way traffic)

The simplest representation of the data is the one-dimensional scatter plot



Date	Direction 1		Direction 2	
	Unordered	Ordered	Unordered	Ordered
01.01	3087	3087	3677	3677
02.01	4664	3578	7357	4453
03.01	4164	3710	9323	4480
04.01	3710	3737	11748	4560
05.01	4029	3906	10256	4635
06.01	4323	4029	4453	4648
07.01	4041	4041	4815	4672
08.01	3737	4085	4757	4757
09.01	4103	4103	4672	4791
10.01	5457	4164	5401	4815
11.01	4563	4323	5688	4880
12.01	3906	4359	6308	4928
13.01	4419	4366	4946	4946
14.01	4359	4368	4635	5005
15.01	4667	4371	5100	5013
16.01	5098	4419	4791	5100
17.01	6551	4563	5235	5220
18.01	4371	4588	4560	5235
19.01	3578	4664	5729	5281
20.01	4366	4667	5005	5318
21.01	4368	4727	4480	5398
22.01	4588	4739	4880	5401
23.01	5001	4741	4928	5679
24.01	7118	5001	5398	5688
25.01	4727	5098	4648	5729
26.01	4085	5193	6183	6183
27.01	4741	5457	5220	6308
28.01	4739	5892	5013	7357
29.01	5193	6551	5281	9323
30.01	5892	7118	5318	10256
31.01	7974	7974	5679	11748

Graphical Representations

- **Histograms**

The data are grouped into intervals

Date	Direction 1		Direction 2	
	Unordered	Ordered	Unordered	Ordered
01.01	3087	3087	3677	3677
02.01	4664	3578	7357	4453
03.01	4164	3710	9323	4480
04.01	3710	3737	11748	4560
05.01	4029	3906	10256	4635
06.01	4323	4029	4453	4648
07.01	4041	4041	4815	4672
08.01	3737	4085	4757	4757
09.01	4103	4103	4672	4791
10.01	5457	4164	5401	4815
11.01	4563	4323	5688	4880
12.01	3906	4359	6308	4928
13.01	4419	4366	4946	4946
14.01	4359	4368	4635	5005
15.01	4667	4371	5100	5013
16.01	5098	4419	4791	5100
17.01	6551	4563	5235	5220
18.01	4371	4588	4560	5235
19.01	3578	4664	5729	5281
20.01	4366	4667	5005	5318
21.01	4368	4727	4480	5398
22.01	4588	4739	4880	5401
23.01	5001	4741	4928	5679
24.01	7118	5001	5398	5688
25.01	4727	5098	4648	5729
26.01	4085	5193	6183	6183
27.01	4741	5457	5220	6308
28.01	4739	5892	5013	7357
29.01	5193	6551	5281	9323
30.01	5892	7118	5318	10256
31.01	7974	7974	5679	11748

Interval (Number of cars x 10^2 /day)	Interval Midpoint (Number of cars x 10^2 /day)	Number of observations	Frequency (%)	Cumulative frequency
35-40	37.5	1	3.2258	0.0323
40-45	42.5	2	6.4516	0.0968
45-50	47.5	10	32.2581	0.4194
50-55	52.5	9	29.0323	0.7097
55-60	57.5	3	9.6774	0.8065
60-65	62.5	2	6.4516	0.8710
65-70	67.5	0	0.0000	0.8710
70-75	72.5	1	3.2258	0.9032
75-80	77.5	0	0.0000	0.9032
80-85	82.5	0	0.0000	0.9032
85-90	87.5	0	0.0000	0.9032
90-95	92.5	1	3.2258	0.9355
95-100	97.5	0	0.0000	0.9355
100-105	102.5	1	3.2258	0.9677
105-110	107.5	0	0.0000	0.9677
110-115	112.5	0	0.0000	0.9677
115-120	117.5	1	3.2258	1.0000

$= \frac{1}{31} \cdot 100$

$\sum = 31$

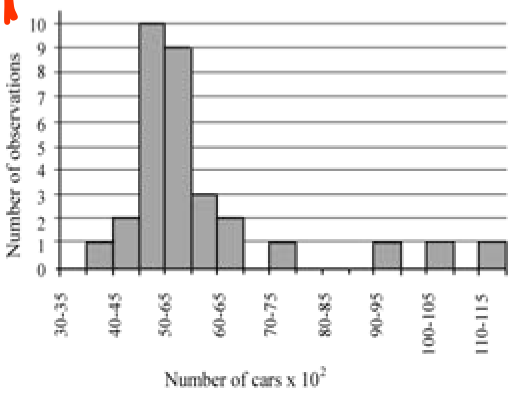
Graphical Representations

- Histograms

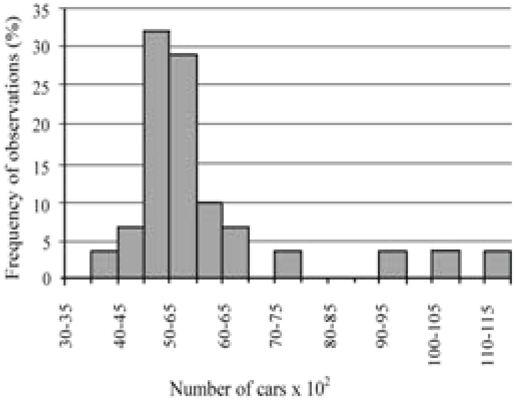
  The grouped data are plotted

mode

Interval (Number of cars $*10^2$)	Interval Midpoint (Number of cars $*10^2$)	Number of observations	Frequency [%]	Cumulative frequency
30-35	32.5	0	0.0000	0.0000
35-40	37.5	1	3.2258	0.0323
40-45	42.5	2	6.4516	0.0968
45-50	47.5	10	32.2581	0.4194
50-55	52.5	9	29.0323	0.7097
55-60	57.5	3	9.6774	0.8065
60-65	62.5	2	6.4516	0.8710
65-70	67.5	0	0.0000	0.8710
70-75	72.5	1	3.2258	0.9032
75-80	77.5	0	0.0000	0.9032
80-85	82.5	0	0.0000	0.9032
85-90	87.5	0	0.0000	0.9032
90-95	92.5	1	3.2258	0.9355
95-100	97.5	0	0.0000	0.9355
100-105	102.5	1	3.2258	0.9677
105-110	107.5	0	0.0000	0.9677
110-115	112.5	1	3.2258	1.0000



Simple histogram



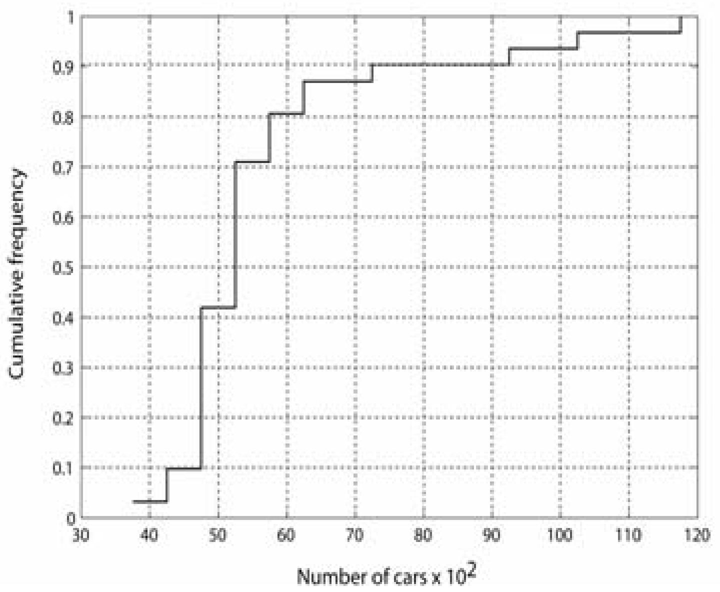
Frequency distribution

Graphical Representations

- Histograms

  The grouped data are plotted

Interval (Number of cars $*10^2$)	Interval Midpoint (Number of cars $*10^2$)	Number of observations	Frequency [%]	Cumulative frequency
30-35	32.5	0	0.0000	0.0000
35-40	37.5	1	3.2258	0.0323
40-45	42.5	2	6.4516	0.0968
45-50	47.5	10	32.2581	0.4194
50-55	52.5	9	29.0323	0.7097
55-60	57.5	3	9.6774	0.8065
60-65	62.5	2	6.4516	0.8710
65-70	67.5	0	0.0000	0.8710
70-75	72.5	1	3.2258	0.9032
75-80	77.5	0	0.0000	0.9032
80-85	82.5	0	0.0000	0.9032
85-90	87.5	0	0.0000	0.9032
90-95	92.5	1	3.2258	0.9355
95-100	97.5	0	0.0000	0.9355
100-105	102.5	1	3.2258	0.9677
105-110	107.5	0	0.0000	0.9677
110-115	112.5	1	3.2258	1.0000



Cumulative frequency distribution

Graphical Representations

- Histograms

  The number of intervals selected will influence the information maintained

No general rule can be given but some suggest the following

$$k = 1 + 3.3 \log n$$

k: **number of intervals**
n: **number of data**

For the traffic flow data set:
$k = 1 + 3.3 \log 31 = 5.92 = 6$

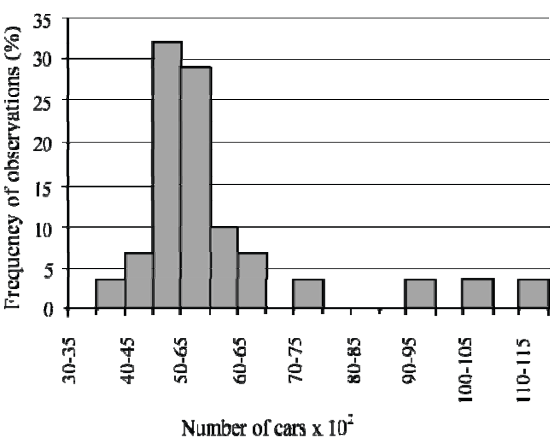
Graphical Representations

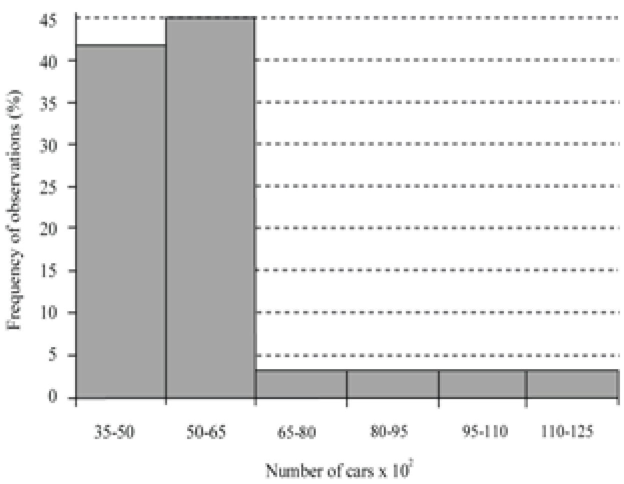
- Histograms

  The number of intervals selected will influence the information maintained

$k=17$

$k=6$





Graphical Representations

- Quantile plots

Definition : the Q-quantile corresponds to the value in a data set which is exceeded by 100% - Q x 100% of the data

e.g. the 0.75 quantile is exceeded by 100% - 0.75 x 100% = 25% of the data

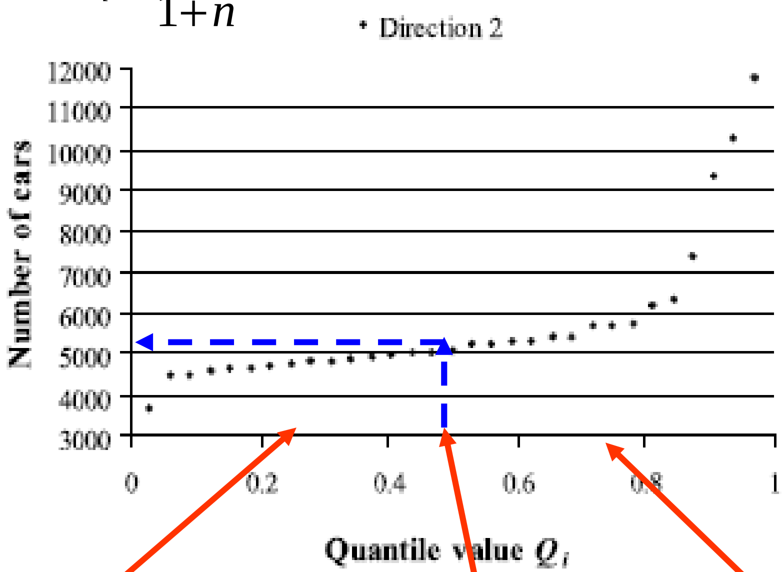
Quantile plots are generated by plotting the data against their quantile values

Graphical Representations

- Quantile plots

The quantiles are calculated from the ordered data set as:

$$Q_i = \frac{i}{1+n}$$



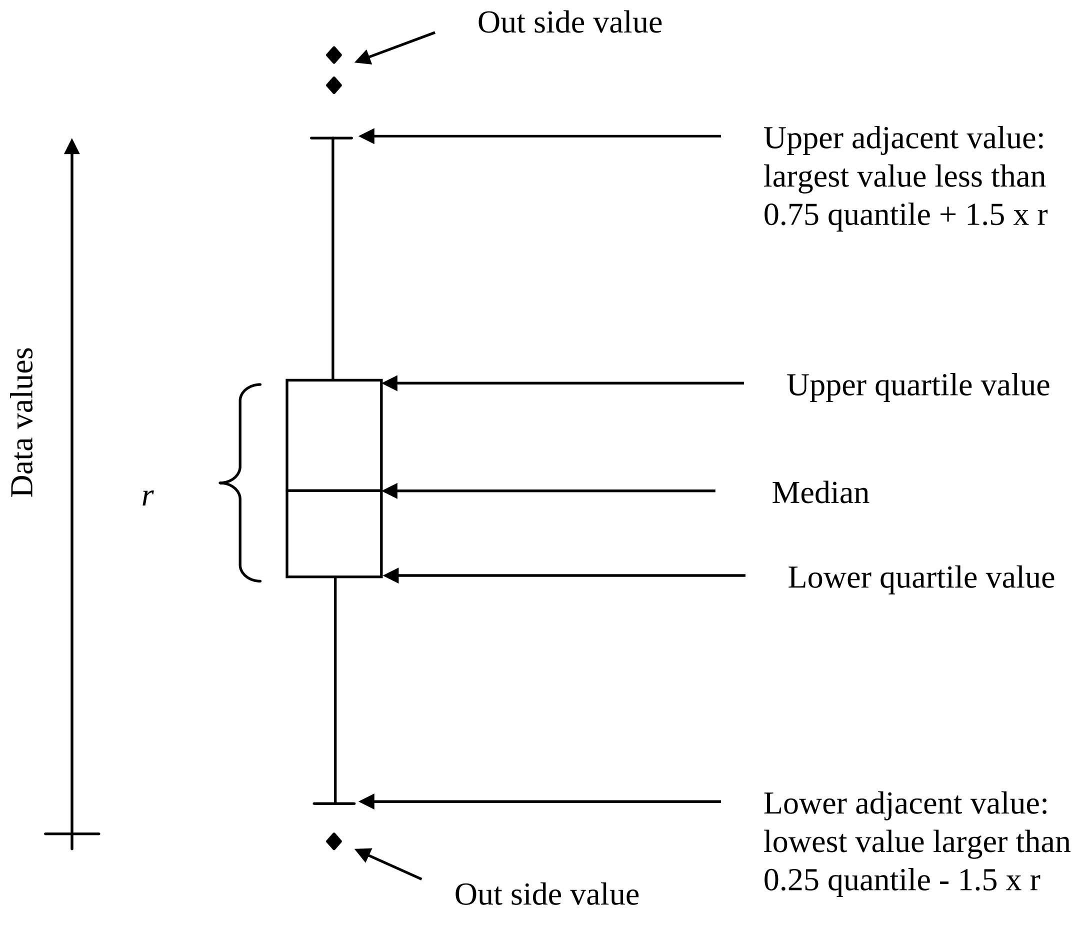
Lower quartile = 0.25 quantile value

Median = 0.5 quantile value

Upper quartile = 0.75 quantile value

x_i	Direction 1	Direction 2	Q_i quantile
1	3087	3677	0.0313
2	3578	4453	0.0625
3	3710	4480	0.0938
4	3737	4560	0.1250
5	3906	4635	0.1563
6	4029	4648	0.1875
7	4041	4672	0.2188
8	4085	4757	0.2500
9	4103	4791	0.2813
10	4164	4815	0.3125
11	4323	4880	0.3438
12	4359	4928	0.3750
13	4366	4946	0.4063
14	4368	5005	0.4375
15	4371	5013	0.4688
16	4419	5100	0.5000
17	4563	5220	0.5313
18	4588	5235	0.5625
19	4664	5281	0.5938
20	4667	5318	0.6250
21	4727	5398	0.6563
22	4739	5401	0.6875
23	4741	5679	0.7188
24	5001	5688	0.7500
25	5098	5729	0.7813
26	5193	6183	0.8125
27	5457	6308	0.8438
28	5892	7357	0.8750
29	6551	9323	0.9063
30	7118	10256	0.9375
31	7974	11748	0.9688

Graphical Representations

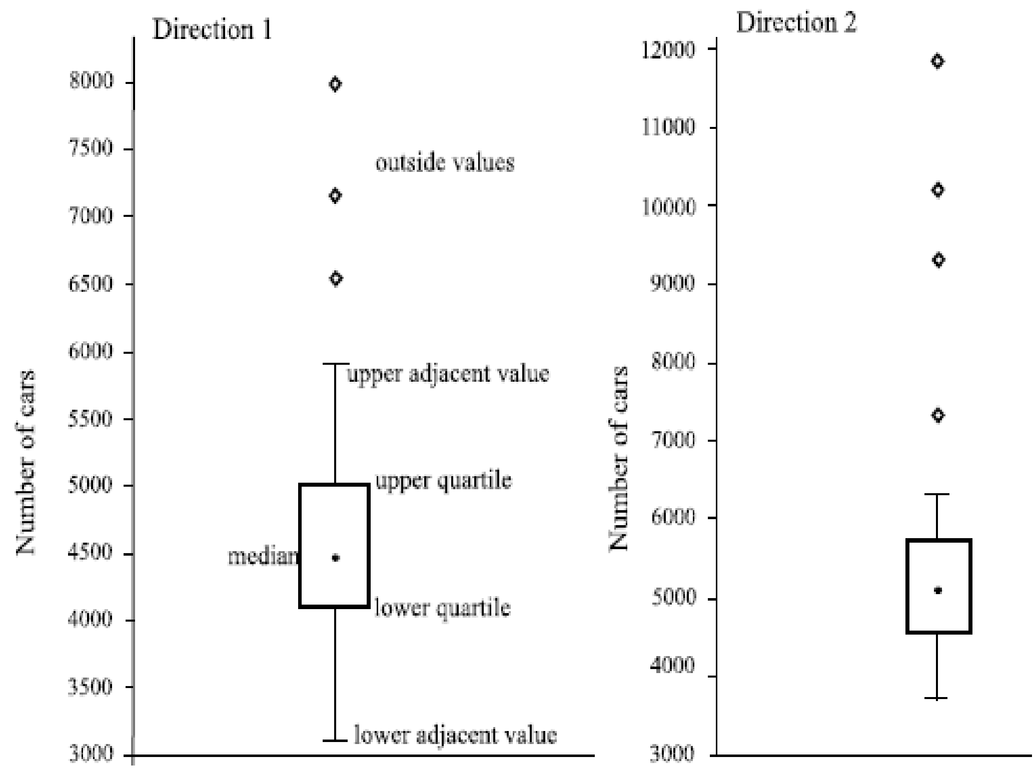


r : Inter-quartile range (50% of data)

Graphical Representations

- Tukey Box plots (traffic data)

Statistic	Direction 1	Direction 2
Lower adjacent value	3087	3677
Lower quartile	4085	4757
Median	4419	5100
Upper quartile	5001	5688
Upper adjacent value	5892	6308
Outside values	6551	7357
	7118	9323
	7974	10256
		11748

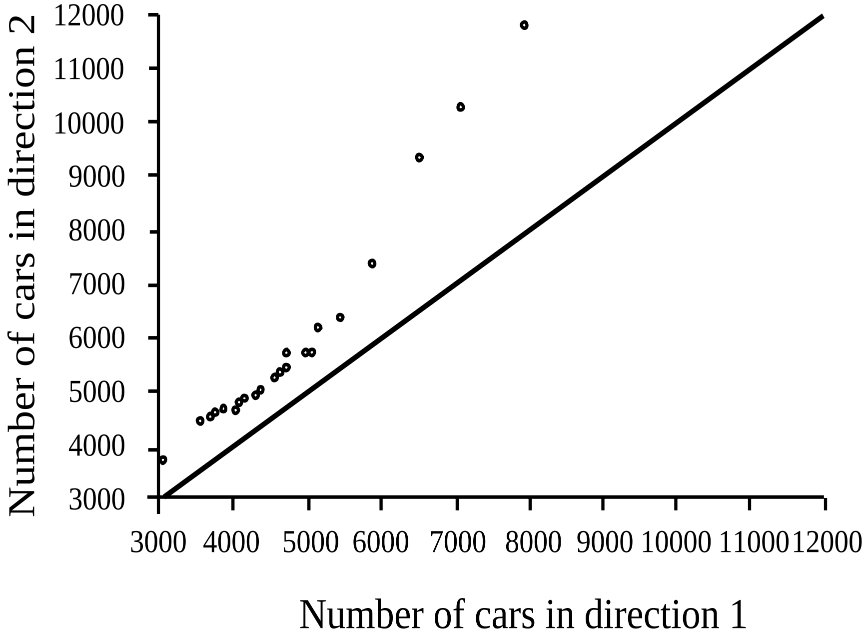


Graphical Representations

- Q-Q plots

Q-Q plots are produced to represent and compare 2 data sets

Data points of the two data sets with the same quantile values are plotted against each other



Graphical Representations

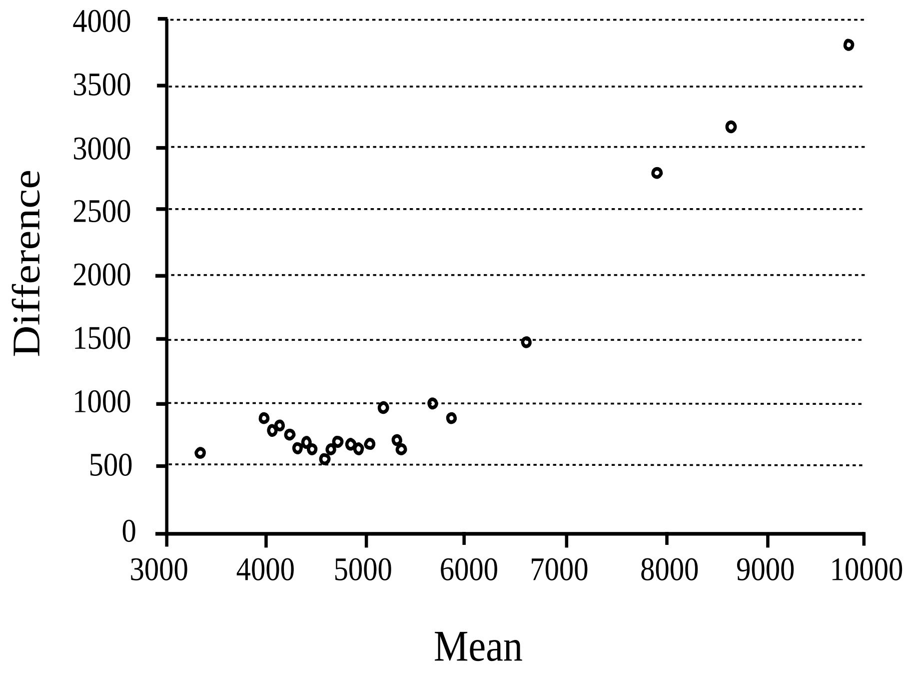
- **Mean vs. difference plots**

Mean vs. difference plots are produced to represent and compare 2 data sets

$$(y_i + x_i)/2$$

is plotted against

$$y_i - x_i$$



Graphical Representations

- Summary

One-dimensional scatter plots :	illustrate the range and distribution of a data sets along one axis, indicate symmetry.
Histograms:	illustrate how the data are distributed over the range of data, indicate mode and symmetry.
Quantile plots: symmetry	Illustrate median, distribution and
Tukey – Box plots:	Illustrate median, upper/lower quartiles, symmetry and distribution
Q-Q plots:	Compare two data set, relative shapes
Mean vs. difference plots:	Compare two data sets, relative shapes

Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zürich, Switzerland

Contents of Today's Lecture

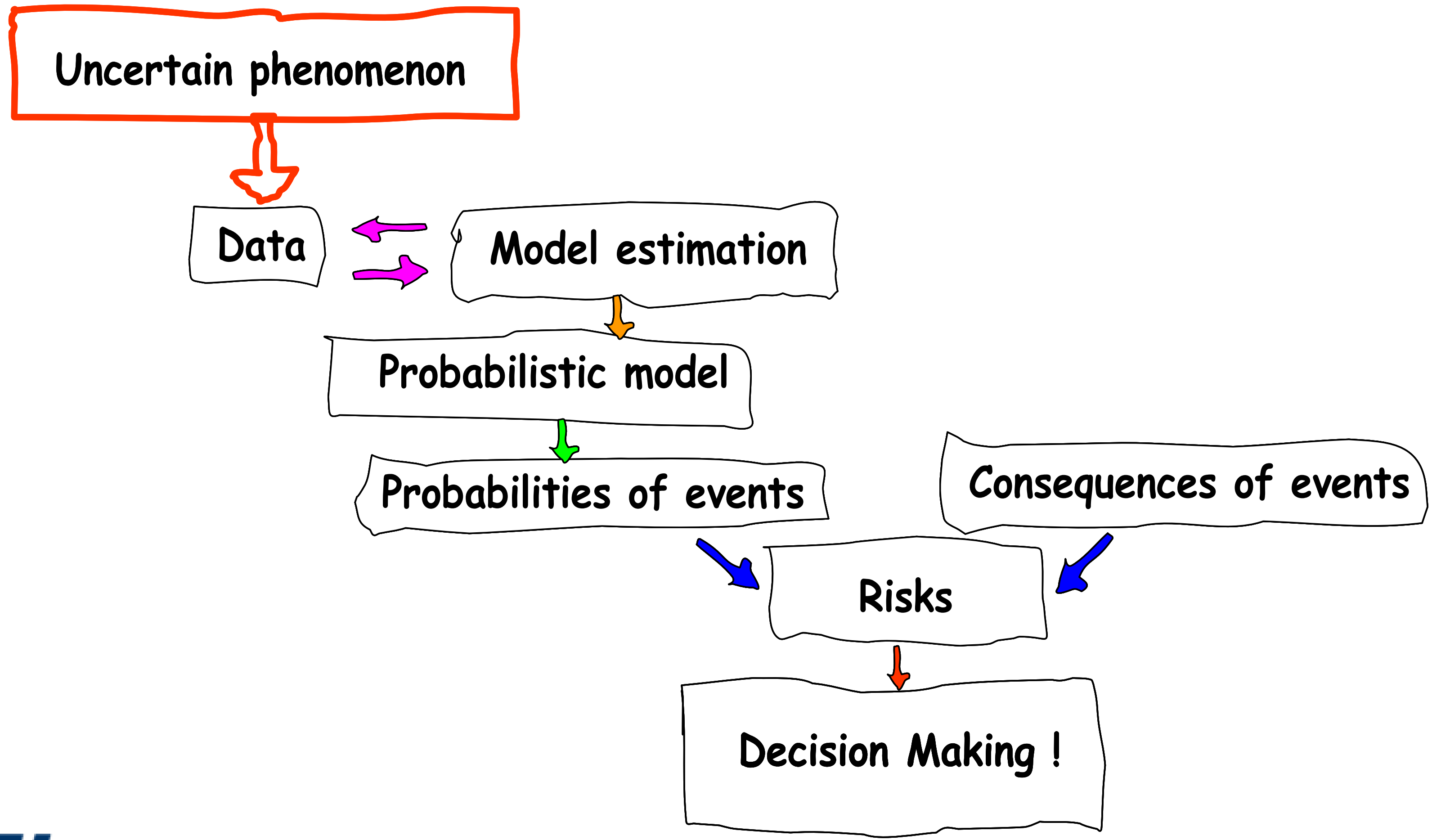
- Overview of Uncertainty Modelling

- Uncertainties in Engineering Problems

- Random Variables
  - discrete cumulative distribution and probability density functions
  - continuous cumulative distribution and probability density functions
  - characterization of random variables
  - moments of random variables
  - the expectation and the variance operator

Overview of Uncertainty Modelling

- Why uncertainty modelling



Uncertainties in Engineering Problems

Different types of uncertainties influence decision making

- Inherent natural variability – aleatory uncertainty
  - result of throwing dices
  - variations in material properties
  - variations of wind loads
  - variations in rain fall

- Model uncertainty – epistemic uncertainty
  - lack of knowledge (future developments)
  - inadequate/imprecise models (simplistic physical modelling)

- Statistical uncertainties – epistemic uncertainty
  - sparse information/small number of data

Uncertainties in Engineering Problems

- Consider as an example a dike structure

  - the design (height) of the dike will be determining the frequency of floods

  - if exact models are available for the prediction of future water levels and our knowledge about the input parameters is perfect then we can calculate the frequency of floods (per year) - a deterministic world !

  - even if the world would be deterministic – we would not have perfect information about it – so we might as well consider the world as random

Uncertainties in Engineering Problems

In principle the so-called

inherent physical uncertainty (aleatory – Type I)

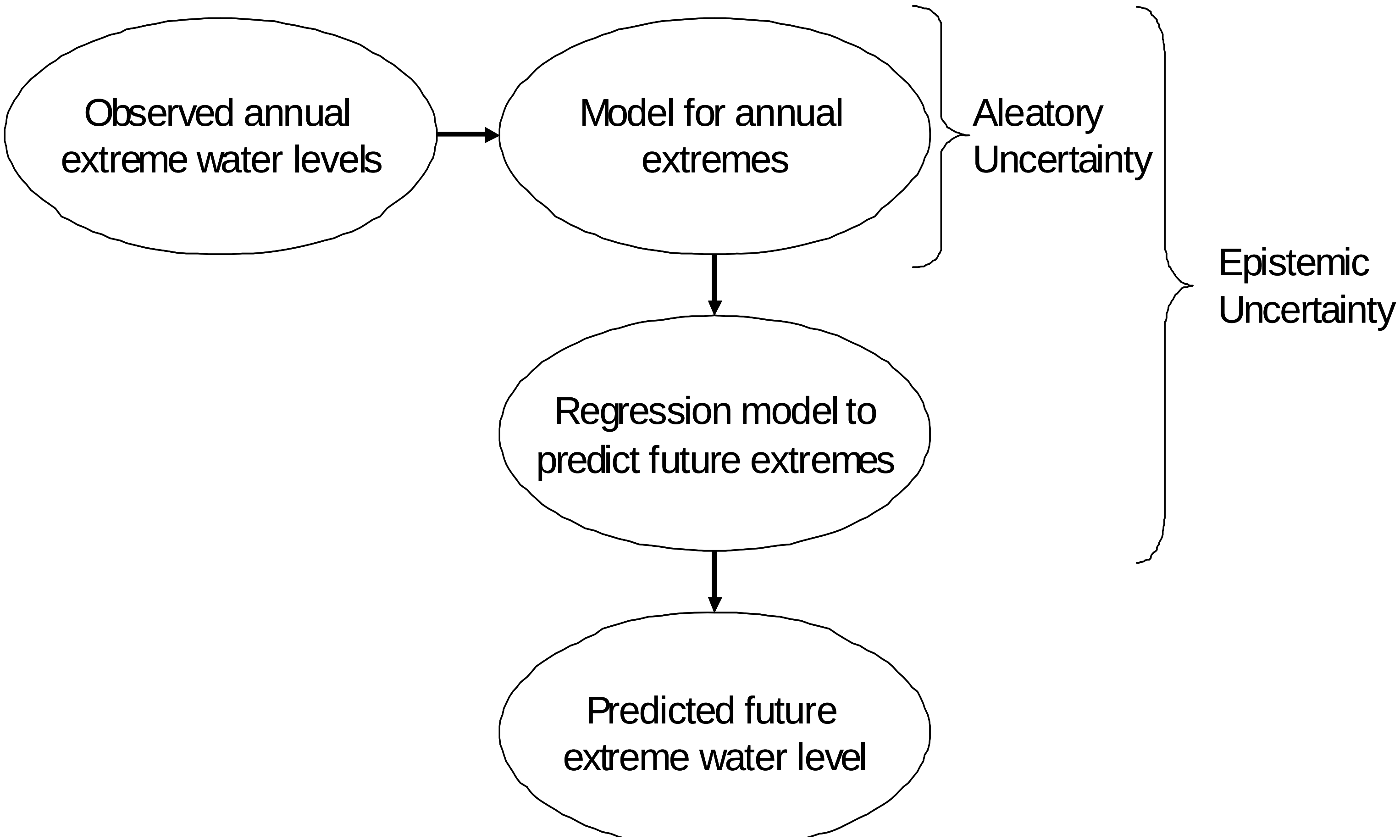
is the uncertainty caused by the fact that the world is random, however, another pragmatic viewpoint is to define this type of uncertainty as

any uncertainty which cannot be reduced by means of collection of additional information

the uncertainty which can be reduced is then the

model and statistical uncertainties (epistemic – Type II)

Uncertainties in Engineering Problems



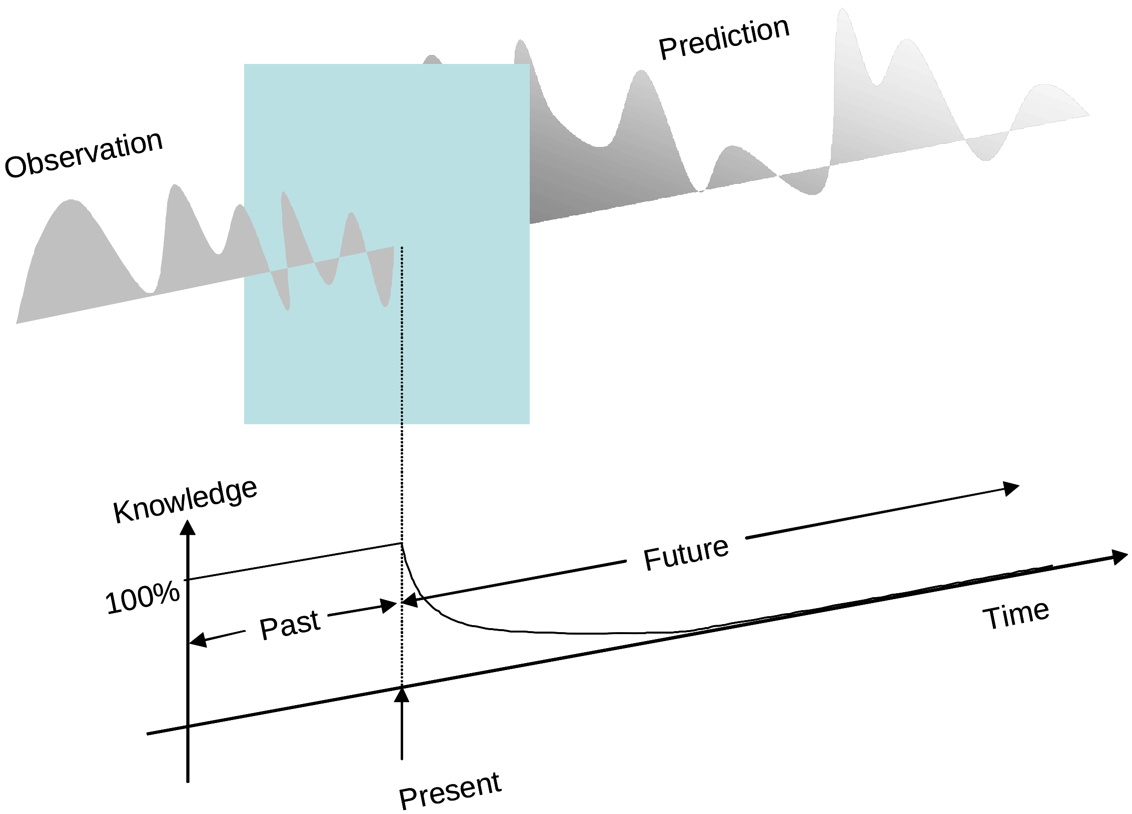
Uncertainties in Engineering Problems

The relative contribution of aleatory and epistemic uncertainty to the prediction of future water levels is thus influenced directly by the applied models

refining a model might reduce the epistemic uncertainty – but in general also changes the contribution of aleatory uncertainty

the uncertainty structure of a problem can thus be said to be scale dependent !

Uncertainties in Engineering Problems



The uncertainty structure changes also as function of time – is thus time dependent !

Random Variables

- Probability density and cumulative distribution functions

  A random variable is denoted with capital letters : X

  A realization of a random variable is denoted with small letters : x

  We distinguish between

  - *continuous random variables* : can take any value in a given range

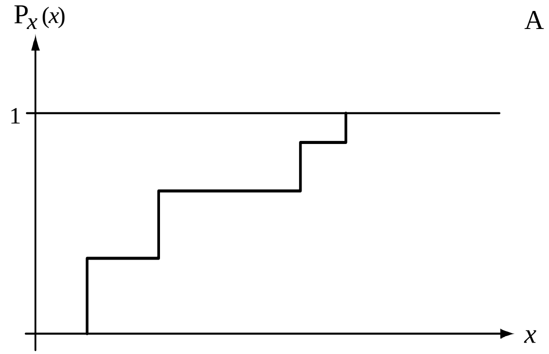
  - *discrete random variables* : can take only discrete values

Random Variables

- Probability density and cumulative distribution functions

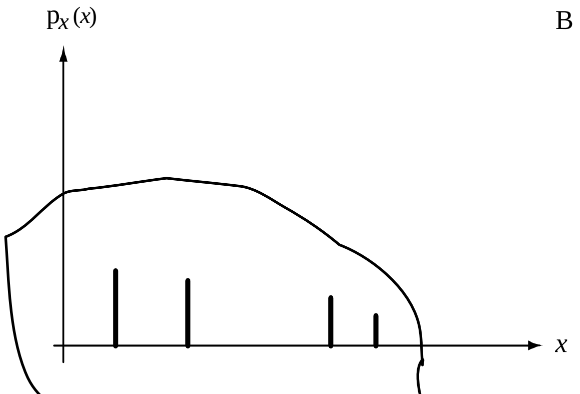
The probability that the outcome of a discrete random variable *X* is smaller than *x* is denoted the *cumulative distribution function*

$$P_X(x) = \sum_{x_i < x} p_X(x_i)$$



The *probability density function* for a discrete random variable is defined by

$$p_X(x_i) = P(X = x_i)$$



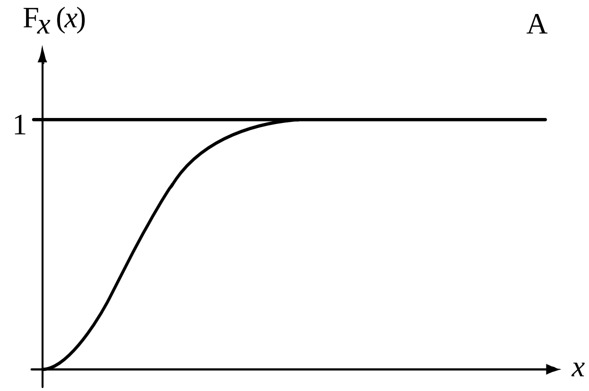
Sum of these must equal 1

Random Variables

- Probability density and cumulative distribution functions

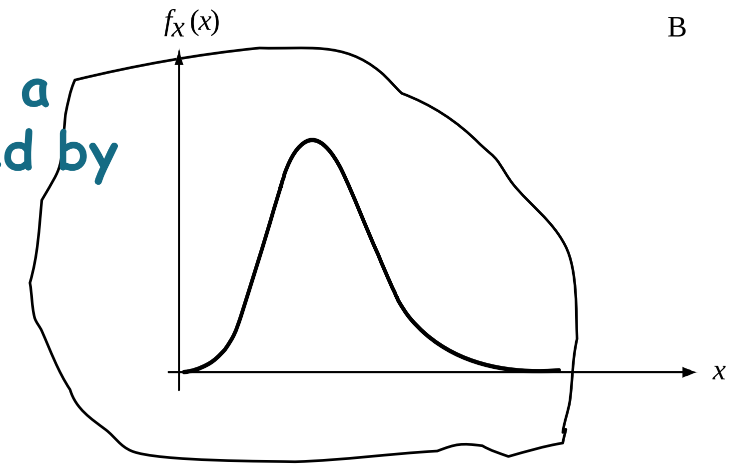
The probability that the outcome of a continuous random variable X is smaller than x is denoted the *cumulative distribution function*

$$F_X(x) = P(X < x)$$



The *probability density function* for a continuous random variable is defined by

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$



Integral of this must equal 1

Random Variables

- Moments of random variables and the expectation operator

  Probability distributions (cumulative distribution function and probability density function) can be described in terms of their paramaters $\mathbf{p}$ or their moments

  Often we write

$$F_X(x, \mathbf{p}) \qquad f_X(x, \mathbf{p})$$

Parameters

  The parameters can be related to the moments and visa versa

Random Variables

- **Moments of random variables and the expectation operator**

  The i'th moment m_i for a continuous random variable X is defined through

$$m_i = \int_{-\infty}^{\infty} x^i f_X(x)dx$$

  The *expected value* $\mathrm{E}[X]$ of a continuous random variable X is defined accordingly as the first moment

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

Random Variables

- **Moments of random variables and the expectation operator**

  The i'th moment m_i for a discrete random variable X is defined through

$$m_i = \sum_{j=1}^{n} x_j^i p_X(x_j)$$

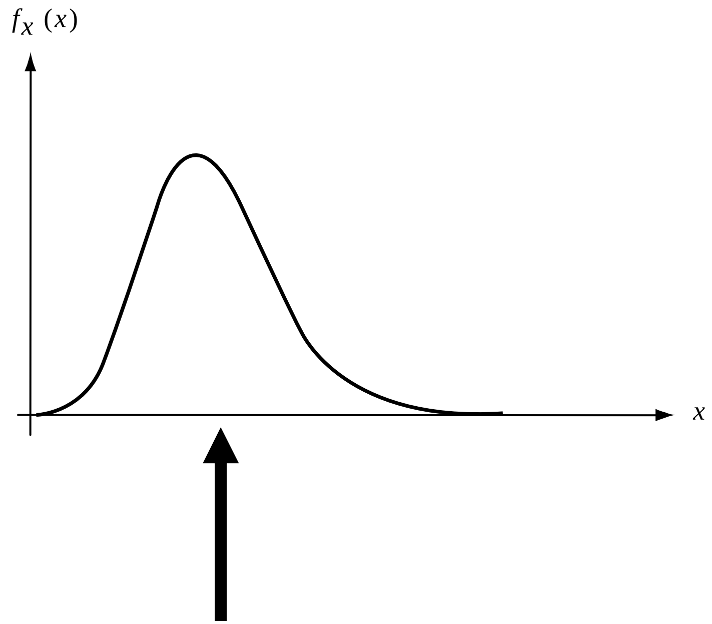
  The *expected value* $\mathrm{E}[X]$ of a discrete random variable X is defined accordingly as the first moment

$$\mu_X = E[X] = \sum_{j=1}^{n} x_j p_X(x_j)$$

Random Variables

- **Moments of random variables and the expectation operator**

The expected value (or mean value) of a random variable can be understood as the *center of gravity* of the probability density function of the random variable !



Random Variables

- **Moments of random variables and the expectation operator**

The *variance* σ_X^2 of a **continuous** random variable is defined as the second central moment i.e. for a continuous random variable *X* we have

$$\sigma_X^2 = Var[X] = E\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx$$

Variance

Mean value

for a **discrete** random variable we have correspondingly

$$\sigma_X^2 = Var[X] = \sum_{j=1}^{n} (x_j-\mu_X)^2 p_X(x_j)$$

Random Variables

- **Moments of random variables and the expectation operator**

The ratio between the standard deviation and the expected value of a random variable is called the *Coefficient of Variation CoV* and is defined as

$$CoV[X] = \frac{\sigma_X}{\mu_X}$$

Dimensionless

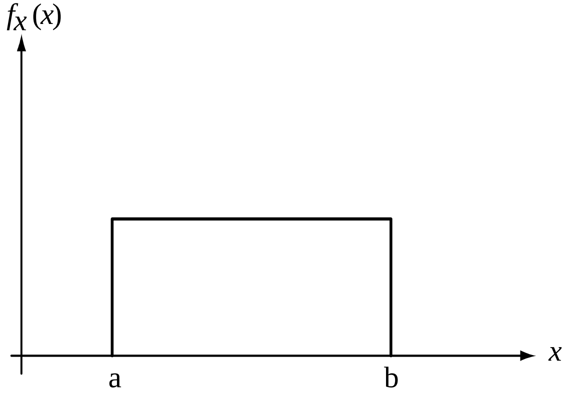
a useful characteristic to indicate the variability of the random variable around its expected value

Random Variables

- Example – uniformly distributed random variable

  probability density and cumulative distribution functions

$$f_X(x) = \begin{cases} 0, & x < a \\ \dfrac{1}{b-a}, & a \le x \le b \\ 0, & b < x \end{cases}$$

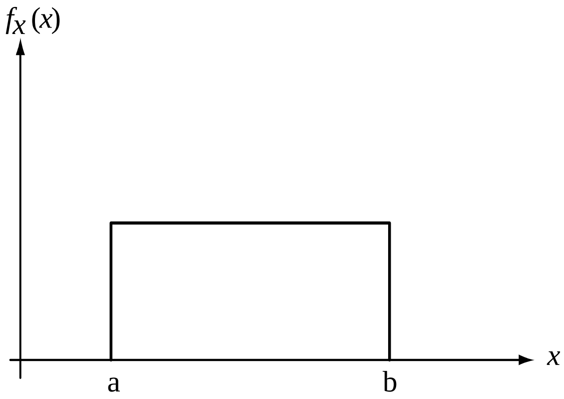


$$F_X(x) = \begin{cases} 0, & x < a \\ \int_a^x f_X(y)dy = \int_a^x \dfrac{1}{b-a}dy = \dfrac{(x-a)}{(b-a)}, & a \le x \le b \\ 1, & b < x \end{cases}$$

Random Variables

- Example – uniformly distributed random variable

  expected value and variance



$$\mu_X = E[X] = \int_a^b x\ f_X(x)dx = \int_a^b \frac{x}{b-a}dx = \left.\frac{x^2}{2(b-a)}\right|_a^b$$

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

$$\sigma_X^2 = E\left[(X-\mu_X)^2\right] = \int_a^b (x-\mu_X)^2 f_X(x)dx = \int_a^b \frac{(x-\mu_X)^2}{(b-a)}dx = \left.\frac{\frac{1}{3}x^3 - x^2\mu_X + x\mu_X{}^2}{(b-a)}\right|_a^b$$

$$= \frac{1}{12}(b-a)^2$$

Statistics and Probability Theory

in

Civil, Surveying and Environmental Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Contents of Todays Lecture

- Overview of Uncertainty Modeling

- Random Variables
  - properties of the expectation operator
  - random vectors and joint moments
  - conditional distributions and conditional moments
  - the probability distribution for the sum of two random variables
  - the probability distribution for functions of random variables

Overview of Uncertainty Modeling

- Random variables and their characteristics

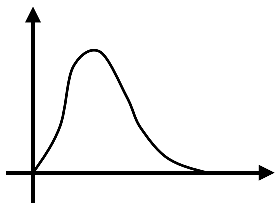
Real world

Uncertain phenomenon

↓

Data/observations

Model



Random variables

↓

x,y	p(x,y)
1,10	0.033
1,20	0.067
1,30	0.033
1,40	0.033
2,10	0.067
2,20	0.100
2,30	0.067
2,40	0.033
3,10	0.067
3,20	0.133
3,30	0.100
3,40	0.067
4,10	0.033
4,20	0.067
4,30	0.067
4,40	0.033

Random Variables

- **Properties of the expectation operator**

  The expectation operator facilitates that we can assess the expected value and the variance of a random variable

  By understanding how the expectation operator works we will be able to assess the expected value and the variance of functions of random variables

  This is useful if we want to analyze engineering models involving one or more random variables in regard to their expected values and their variances

  E.g.: Duration of a construction process as a function of the duration of its individual processes

Random Variables

- Properties of the expectation operator

The expectation operator possesses the following properties:

$$E[c] = c$$

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

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

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

Random Variables

- Properties of the expectation operator

The variance can thus be written as:

$$
\begin{aligned}
Var[X] &= E\left[(X-\mu_X)^2\right] \\
&= E\left[X^2+\mu_X^2-2\mu_X X\right] \\
&= \mu_X^2 + E\left[X^2\right] - 2\mu_X E[X] \\
&= \mu_X^2 + E\left[X^2\right] - 2\mu_X^2 = E\left[X^2\right] - \mu_X^2
\end{aligned}
$$

Random Variables

- **Properties of the expectation operator**

$$E[c] = c$$
$$E[cX] = cE[X]$$
$$E[a + bX] = a + bE[X]$$
$$E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$$

Furthermore there is

$$Var[c] = 0$$
$$Var[cX] = c^2 Var[X]$$
$$Var[a + bX] = b^2 Var[X]$$

Random Variables

- Properties of the expectation operator

From the result

$$Var[X] = E\left[(X-\mu_X)^2\right] = E\left[X^2 + \mu_X^2 - 2\mu_X X\right] = E\left[X^2\right] - \mu_X^2$$

it is seen that there in general is $E[g(X)] \neq g(E[X])$

$E[g(X)] \geq g(E[X])$ for convex functions - Jensen's inequality !

↑

Equality only for linear functions

Random Variables

- Random vectors and joint moments

Often we are dealing with models involving not only one random variable but several random variables

These random variables can be collected in a vector

In general the components of the vector are dependent

E.g. Rainfall and water level

It is thus necessary that we establish probabilistic models which include this dependency – we can do this through the joint cumulative distributions and the joint moments.

Random Variables

- Random vectors and joint moments

  Now we consider not just one continuous random variable but a vector of continuous random variables

$$\mathbf{X} = \left(X_1, X_2, \ldots, X_n\right)^T$$

The *joint cumulative distribution function* is given by

$$F_{\mathbf{X}}\left(\mathbf{x}\right) = P\left(X_1 \le x_1 \cap X_2 \le x_2 \cap \ldots \cap X_n \le x_n\right)$$

and the *joint probability density function* is given by

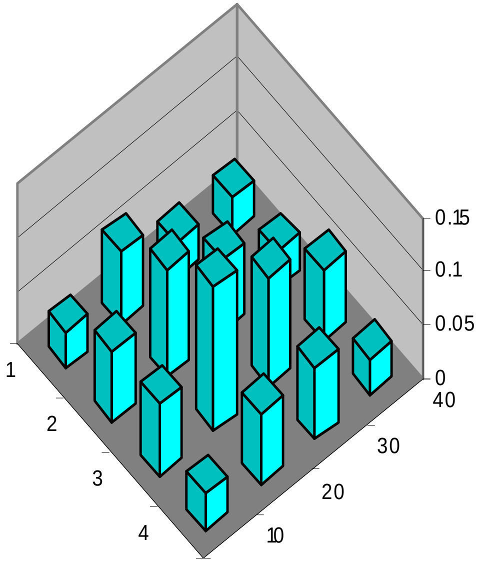
$$f_{\mathbf{X}}\left(\mathbf{x}\right) = \frac{\partial^n}{\partial z_1 \partial z_2 \ldots \partial z_n} F_{\mathbf{X}}\left(\mathbf{x}\right)$$

Random Variables

- Random vectors and joint moments

Consider the two dimensional discrete probability density function:

x,y	p(x,y)
1,10	0.033
1,20	0.067
1,30	0.033
1,40	0.033
2,10	0.067
2,20	0.100
2,30	0.067
2,40	0.033
3,10	0.067
3,20	0.133
3,30	0.100
3,40	0.067
4,10	0.033
4,20	0.067
4,30	0.067
4,40	0.033



Random Variables

- **Random vectors and joint moments**

The *marginal probability density function* of a random variable *X_i* is defined by

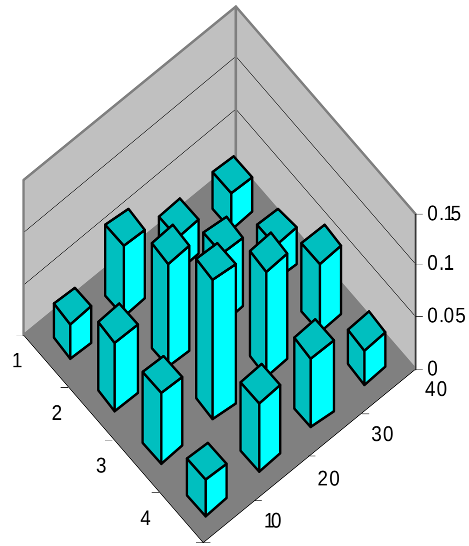
$$f_{X_i}(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (n-1 \ \text{ fold})\, f_{\mathbf{X}}(\mathbf{x})\, dx_1 .. dx_{i-1} dx_{i+1} .. dx_n$$

Random Variables

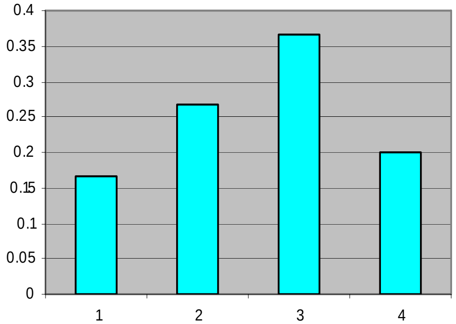
- Random vectors and joint moments

Consider the two dimensional discrete probability density function:

x,y	p(x,y)
1,10	0.033
1,20	0.067
1,30	0.033
1,40	0.033
2,10	0.067
2,20	0.100
2,30	0.067
2,40	0.033
3,10	0.067
3,20	0.133
3,30	0.100
3,40	0.067
4,10	0.033
4,20	0.067
4,30	0.067
4,40	0.033



Discrete joint density



Marginal density for x

Random Variables

- Random vectors and joint moments

The *covariance* between the i'th and the j'th component of the random vector of continuous random variables is defined as the *joint central moment* i.e. by

$$C_{X_iX_j} = E\left[(X_i - \mu_{X_i})(X_j - \mu_{X_j})\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(x_i - \mu_{X_i}\right)\left(x_j - \mu_{X_j}\right) f_{X_iX_j}\left(x_i, x_j\right) dx_i dx_j$$

$$C_{X_iX_i} = Var\left[X_i\right]$$

From where we see that for i = j we get the variance for X_i

Correlation coefficient $\rho_{X_iX_j} = \dfrac{C_{X_iX_j}}{\sigma_{X_i}\sigma_{X_j}}$ $\qquad \rho_{X_iX_i} = 1$

Random Variables

- **Random vectors and joint moments**

The expected value and the variance of a linear function

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

are given by

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

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

Random Variables

- Conditional distributions and conditional moments

Some times it is useful to be able to assess the probability of an event given that we know something about one of the random variables which are used to define the event

E.g. assume we want to calculate the probability that a project will be delayed under the condition that one of the processes will exceed its planned duration by 50%.

Random Variables

- **Conditional distributions and conditional moments**

  The *conditional probability density function* for the random variable X_1 given the outcome of the random variable X_2 is given by

$$f_{X_1|X_2}(x_1|x_2)=\frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}$$

  where if X_1 and X_2 are independent

$$f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1)$$

  The *conditional cumulative distribution function* is obtained by integration as

$$F_{X_1|X_2}(x_1|x_2)=\frac{\int_{-\infty}^{x_1} f_{X_1,X_2}(z,x_2)dz}{f_{X_2}(x_2)}$$

Random Variables

- **Conditional distributions and conditional moments**

  The *un-conditional cumulative distribution function* for the random variable X_1 can be derived from the conditional comulative distribution function by use of the *total probability theorem*

$$F_{X_1}(x_1) = \int_{-\infty}^{\infty} F_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2) dx_2$$

  The *conditional expected value* is defined by

$$\mu_{X_1|X_2} = E\left[X_1 | X_2 = x_2\right] = \int_{-\infty}^{\infty} x_1 \, f_{X_1|X_2}\left(x|x_2\right) dx_1$$

Random Variables

- In many cases we are interested in assessing the probabilites of functions of random variables

  The functions are useful for describing the events we are interested in – they are our engineering models.

  A simple case is the sum of two random variables – it is useful to derive the cumulative distribution function for such a sum.

  A more general case concerns monotonic functions of random variables – we will also derive the cumulative distribution for this case.

Random Variables

- The cumulative distribution function for the sum of two random variables

  Consider the sum $Y = X_1 + X_2$

  and assume that we have $f_{X_1,X_2}(x_1, x_2)$

  First we derive the density function for $Y = x_1 + X_2$

  assuming that X_1 is given i.e. $f_{X_2|X_1}(x_2|x_1) = \dfrac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)}$

$$f_{Y|X_1}(y|x_1) = f_{X_2|X_1}(y - x_1|x_1)$$

and we get $f_{Y,X_1}(y,x_1) = f_{X_2|X_1}(y-x_1|x_1)\,f_{X_1}(x_1) = f_{X_2,X_1}(y-x_1,x_1)$

Random Variables

- The cumulative distribution function for the sum of two random variables

  The marginal probability density function for *Y* is now achieved by integrating out over X_1, i.e.

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2,X_1}(y - x_1, x_1)\,dx_1$$

  For the case where X_1 and X_2 are independent we get the so-called *convolution integral*

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1)\,dx_1$$

Random Variables

- The cumulative distribution function for functions of random variables

  Consider the more general problem of deriving the cumulative distribution function for a function of a random variables i.e. $Y = g(X)$ where the probability distribution function of *X* is given as $F_X(x)$

  If $g(x)$ is monotonically increasing and represents a one-to-one mapping, a realization of *Y* is only smaller than y_0 if the realization of *X* is smaller than x_0 where $x_0 = g^{-1}(y_0)$

$$F_Y(y) = P(Y \leq y) = P(X \leq g^{-1}(y))$$

  The cumulative distribution function for Y is then given by

$$F_Y(y) = F_X(g^{-1}(y))$$

Random Variables

- The cumulative distribution function for functions of random variables

starting now with $F_Y(y) = F_X(g^{-1}(y))$

we have $$f_Y(y) = \frac{\partial F_X(g^{-1}(y))}{\partial y}$$

$$f_Y(y) = \frac{\partial}{\partial y} g^{-1}(y) \, f_X(g^{-1}(y)) \quad \Longrightarrow \quad f_Y(y) = \frac{\partial x}{\partial y} f_X(x)$$

Random Variables

- The cumulative distribution function for functions of random variables

  In case the function $g(x)$ is monotonically decreasing, a realization of Y is only smaller than y_0 if the realization of X is larger than x_0 , and in this case we have to change the sign i.e.

$$F_Y(y) = -F_X(g^{-1}(y))$$

  yielding $$f_Y(y) = -\frac{\partial x}{\partial y} f_X(x)$$

  In the general case – for monotonically increasing or decreasing functions there is thus

$$f_Y(y) = \left|\frac{\partial x}{\partial y}\right| f_X(x)$$

Random Variables

- The cumulative distribution function for functions of random variables

For the case where the components of a random vector $\mathbf{Y}=(Y_1,Y_2,..Y_n)^T$ can be given as one-to-one mappings of monotonically increasing or decreasing functions $g_i, i=1,2,..n$ of the components of a random vector $\mathbf{X}=(X_1,X_2,..X_n)^T$

in the form: $Y_i = g_i(\mathbf{X})$

there is $f_{\mathbf{Y}}(\mathbf{y}) = |\mathbf{J}|\, f_{\mathbf{X}}(\mathbf{x})$

with $|\mathbf{J}|$ being the absolute value of the determinant of

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & ..... & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & ..... & \frac{\partial x_n}{\partial_n} \end{bmatrix}$$

Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

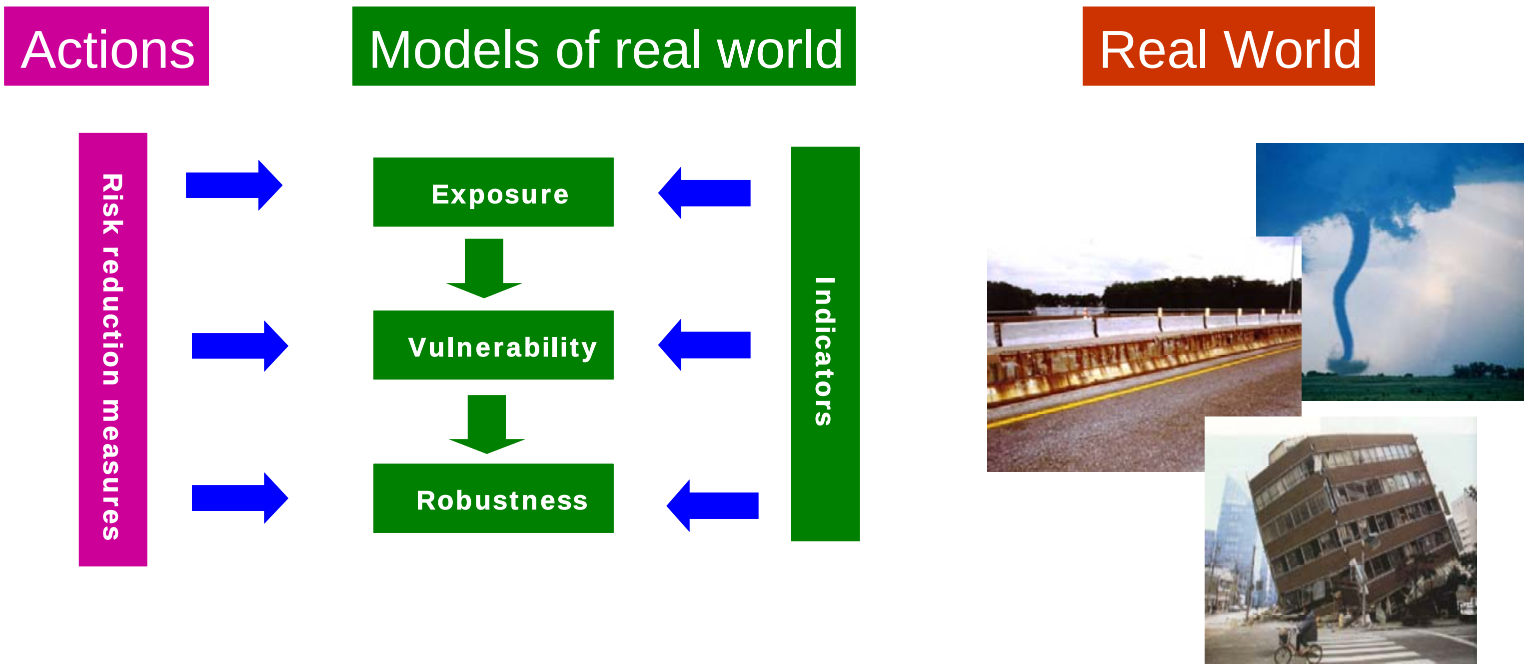
Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Overview of Uncertainty Modeling

- Random variables and their characteristics



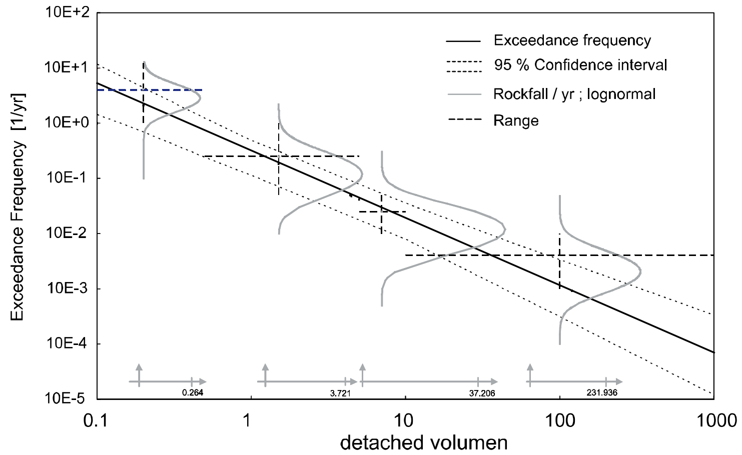
Overview of Uncertainty Modeling

- **Random variables and their characteristics**

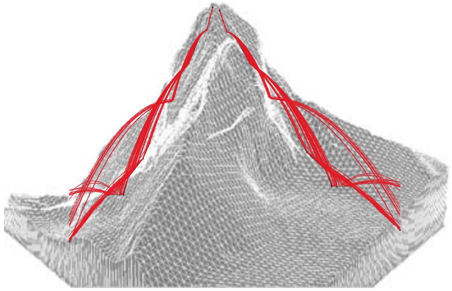
Design of rock-fall galleries

Overview of Uncertainty Modeling

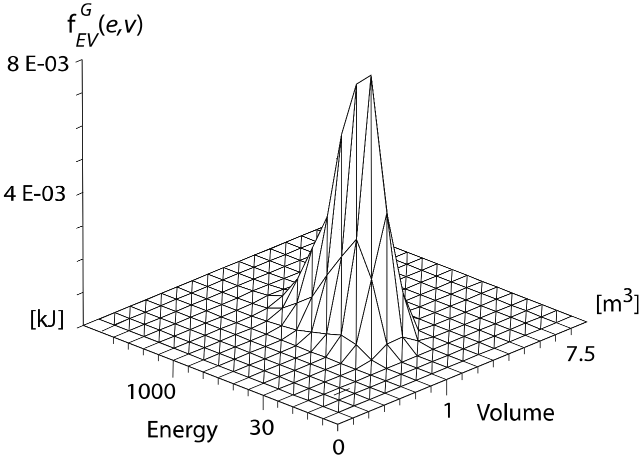
- Random variables and their characteristics



Detachment modeling



Fall modeling



Tools in Uncertainty Modeling

- Engineering problems – also those involving uncertainty are very often specific – unique !

  Being able to solve such problems requires

  - basic tools (physical, mathematical, natural sciences, human sciences, engineering,…)

  - innovation (being able to identify ways of solving problems)

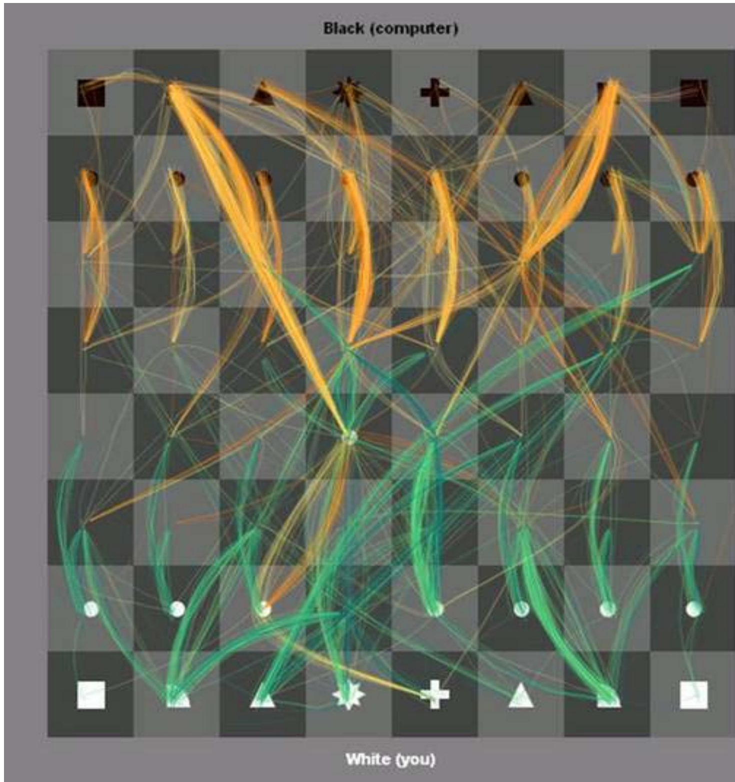
  - training !

  Training is important because it provides experience.

  By training we start to recognize patterns !

Tools in Uncertainty Modeling

- Pattern recognition helps to identify:



the usefulness of solution strategies from previous problems

the potential of the available tools in a given context

Tools in Uncertainty Modeling

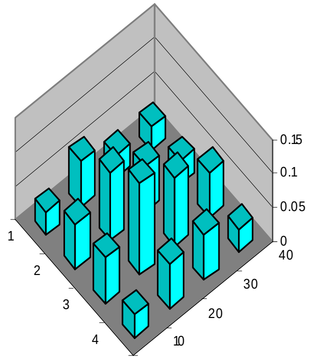
- Random variables and their characteristics

The expectation operator

$$E[c] = c$$
$$E[cX] = cE[X]$$
$$E[a + bX] = a + bE[X]$$
$$E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$$

The variance operator

$$Var[c] = 0$$
$$Var[cX] = c^2 Var[X]$$
$$Var[a + bX] = b^2 Var[X]$$



Jointly distributed random variables

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1 \cap X_2 \leq x_2 \cap \ldots \cap X_n \leq x_n)$$

Tools in Uncertainty Modeling

- Random variables and their characteristics

  Functions of random variables

  - sum of two random variables

$$Y = X_1 + X_2$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1$$

  - non-linear function of random variables

$$Y = g(X)$$

$$f_Y(y) = \left| \frac{\partial x}{\partial y} \right| f_X(x)$$

Tools in Uncertainty Modeling

- Random variables and their characteristics

Functions of random variables

$$\mathbf{Y} = (Y_1, Y_2, ..Y_n)^T$$

$$Y_i = g_i(\mathbf{X}), \qquad X_i = f_i(\mathbf{Y})$$

$$f_{\mathbf{Y}}(\mathbf{y}) = |\mathbf{J}|\, f_{\mathbf{X}}(\mathbf{x})$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots\cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots\cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

Contents of Todays Lecture

- Random variables
  - The Central Limit Theorem
  - The Normal distribution
  - The Log-Normal distribution

- Stochastic Processes and Extremes
  - Random sequences (Bernoulli trials)
  - Binomial distribution
  - Geometric distribution

Random Variables

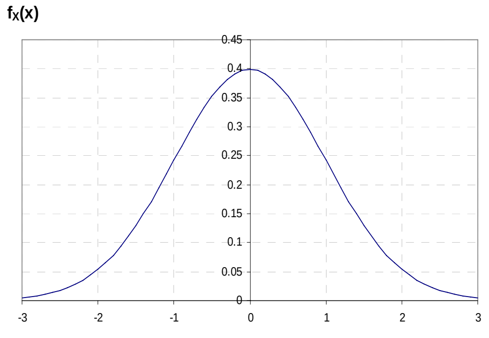
- The Central Limit Theorem states:

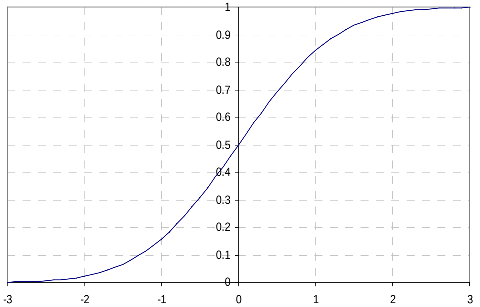
  The probability distribution function of a sum of a number of random variables approaches the Normal (Gaussian) distribution as the number becomes large

$$Y = X_1 + X_2 + \ldots + X_n$$

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

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx$$





Random Variables

- The Central Limit Theorem

  Conditions for the validity of the theorem:

$$Y = X_1 + X_2 + \ldots + X_n$$

  The sum should not be dominated by one or a few components

  The statistical dependency between components should not be strong

  No requirements to the type of distribution of the components

  If the components have skew distributions the number increases

Random Variables

- Illustration:

  A structural member is measured using a ruler.
  - The ruler has limited length (2 m).
  - The smallest unit on the ruler is 1 mm.

  All measurements are rounded to the closest unit on the ruler.

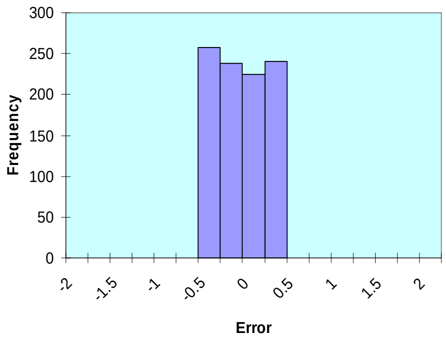
  Each measurement is subject to a measurement uncertainty uniformly distributed in the range of +/- 0.5 mm.

  We now consider the accumulated error associated with measurements over lengths
  - up to 2 m (one measurement)
  - between 2 and 4 m (two measurements)
  - between 6 and 8 m (four measurements)
  - between 14 and 16 m (eight measurements)

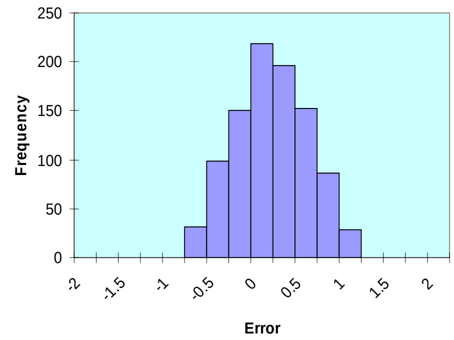
Random Variables

- Illustration:

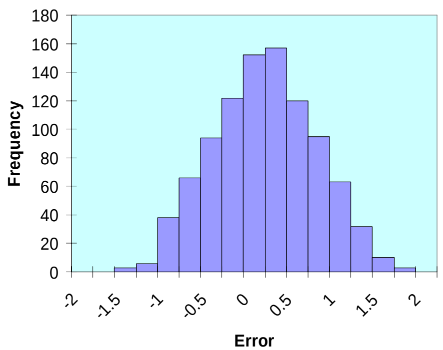
N=1



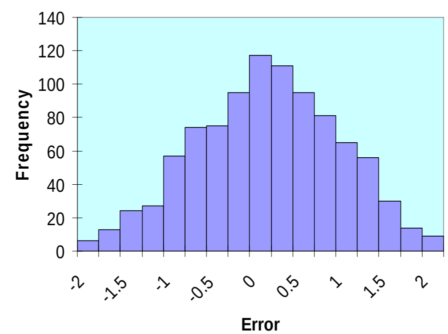
N=2



N=4



N=8



Random Variables

- The Normal distribution

  The analytical form of the Normal distribution may be derived by repeated use of the result regarding the probability density function for the sum of two random variables

  The Normal distribution is very frequently applied in engineering modeling when a random quantity can be assumed to be composed as a sum of a number of individual contributions: $X_i, i=1,2,..,n$

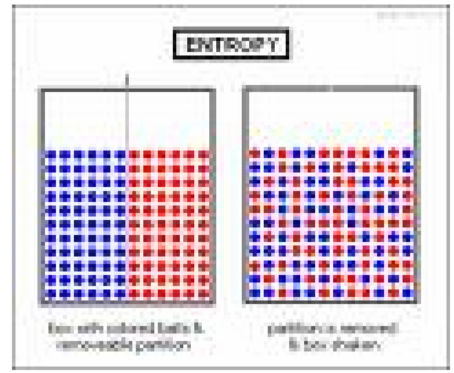
  A linear combination *S* of *n* Normal distributed random variables $S = a_0 + \sum_{i=1}^{n} a_i X_i$ is thus also a Normal distributed random variable

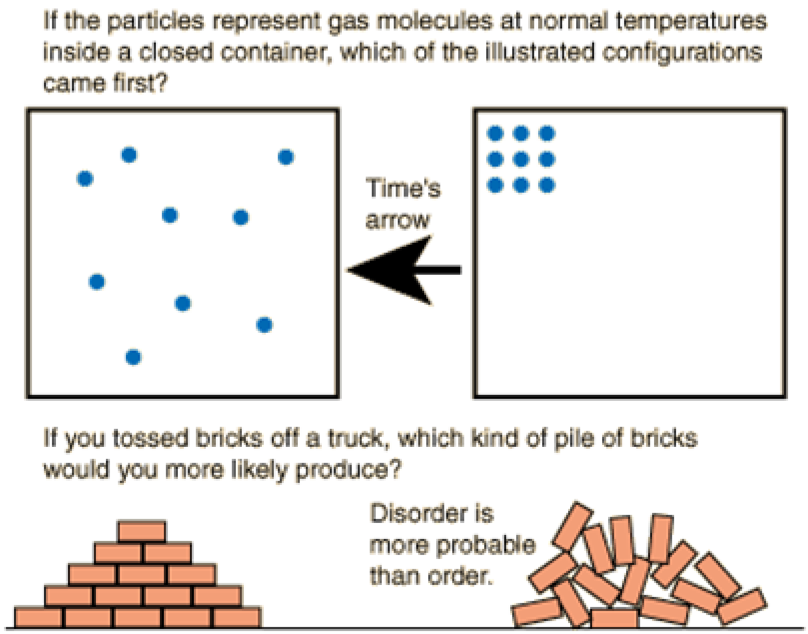
Random Variables

- The Normal distribution

  The Normal distribution also results from other considerations

  The distribution of energy in an isolated system

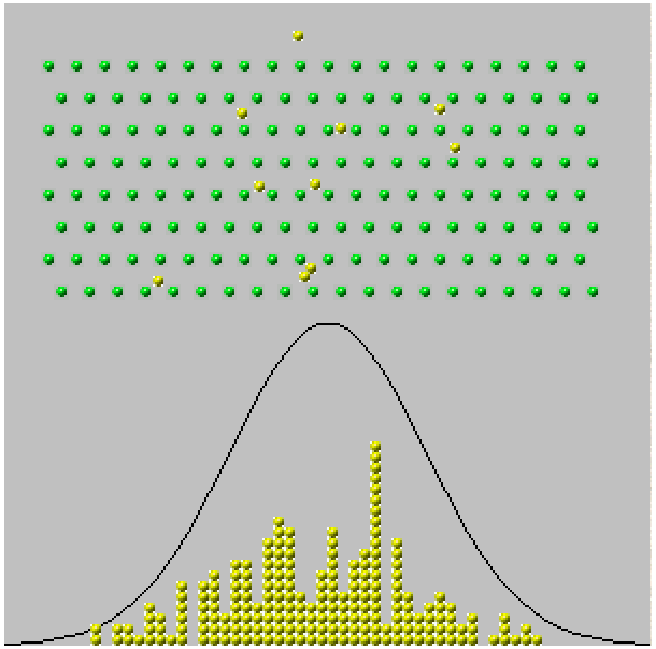




Random Variables

- The Normal distribution

  The accumulation of random movements

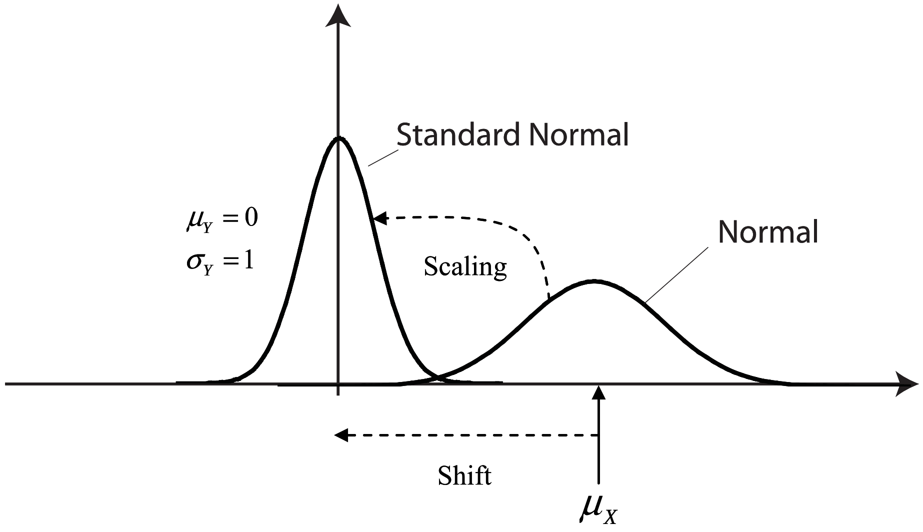


Random Variables

- The Normal distribution:

  In the case where the mean value is equal to zero and the standard deviation is equal to 1 the random variable is said to be *standardized*.

$$Y = \frac{X - \mu_X}{\sigma_X}$$ **Standardized random variable**



Random Variables

- The Normal distribution:

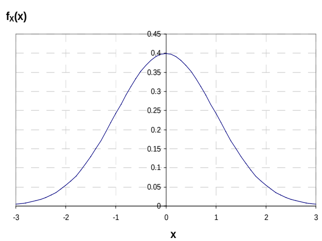
  In the case where the mean value is equal to zero and the standard deviation is equal to 1 the random variable is said to be *standardized*.

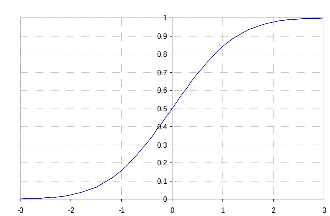
$$Y = \frac{X - \mu_X}{\sigma_X}$$ **Standardized random variable**

$$f_Y(y) = \varphi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y^2\right)$$

$$F_Y(y) = \Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{1}{2} x^2\right) dx$$

Standard normal





Random Variables

When the logarithm of a random variable *X* i.e.

$Y = \ln(X), \qquad Y : N(\mu_Y, \sigma_Y)$

is Normal distributed the random variable *X* is said to be Log-Normal distributed

$$X : LN(\lambda, \zeta)$$

$$f_X(x) = \frac{1}{x\zeta\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\ln(x) - \lambda}{\zeta} \right)^2 \right) \qquad \mu_X = \exp\left(\lambda + \frac{\zeta^2}{2} \right)$$

$$F_X(x) = \Phi\left(\frac{\ln(x) - \lambda}{\zeta} \right) \qquad \sigma_X = \exp\left(\lambda + \frac{\zeta^2}{2} \right) \sqrt{\exp(\zeta^2) - 1}$$

Random Variables

Where the Normal distribution follows from the sum of random variables – **Central Limit Theorem**

the Log-Normal distribution follows from the product of random variables

$$\ln(X_1 \cdot X_2 \cdots X_n) = \ln(\prod_{i=1}^{n} X_i) = \sum_{i=1}^{n} \ln(X_i)$$

Random Variables

The Log-Normal distribution has the useful property that if

$$P = \prod_{i=1}^{n} Y_i^{a_i}$$

and all Y_i are independent Log-Normal distributed random variables with parameters ζ_i , λ_i and $\varepsilon_i = 0$ then P is also Log-Normal with parameters

$$\lambda_P = \sum_{i=1}^{n} a_i \lambda_i \qquad \zeta_P^2 = \sum_{i=1}^{n} a_i^2 \zeta_i^2 \qquad f_P(p) = \frac{1}{p\zeta_P\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(p)-\lambda_P}{\zeta_P}\right)^2\right)$$

Random Variables

The Log-Normal distribution is often used to model

- uncertain parameters which cannot have negative realizations

- fatigue lives

- steel and concrete resistance

- daily river flows

- whenever a random variable results as a product of several random variables

Random Variables

Concrete compression strength

$$\mu_X = \exp\left(\lambda + \frac{\zeta^2}{2}\right)$$

$$\sigma_X = \exp\left(\lambda + \frac{\zeta^2}{2}\right)\sqrt{\exp(\zeta^2) - 1}$$

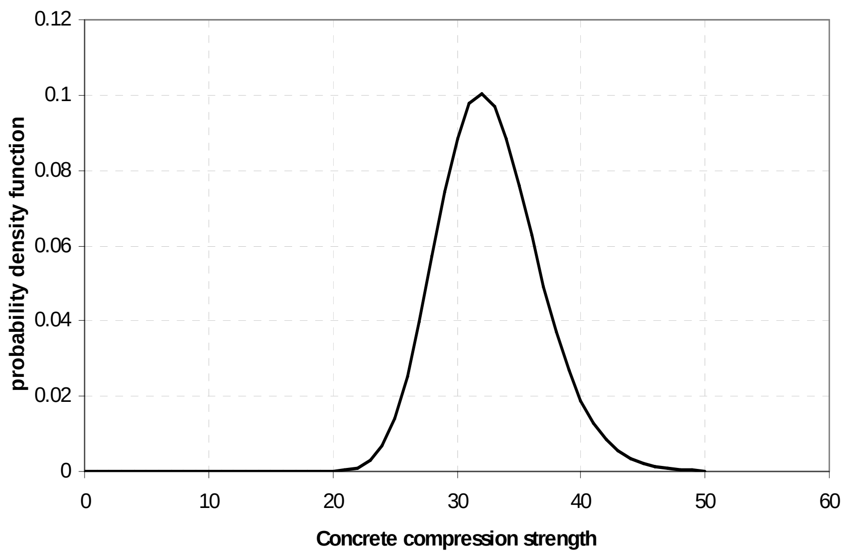
$$\Downarrow$$

$$V_X = \frac{\sigma_X}{\mu_X} = \sqrt{\exp(\zeta^2) - 1} = \frac{4.05}{32.67} = 0.12 \Rightarrow \zeta = 0.12, \lambda = 3.48$$

Probability of value lower than 25 MPa

$$F_X(25) = \Phi\left(\frac{\ln(25) - 3.48}{0.12}\right) = 0.018$$

i	x_i
1	24.4
2	27.6
3	27.8
4	27.9
5	28.5
6	30.1
7	30.3
8	31.7
9	32.2
10	32.8
11	33.3
12	33.5
13	34.1
14	34.6
15	35.8
16	35.9
17	36.8
18	37.1
19	39.2
20	39.7



Random Variables

There exist a large number of different probability density and cumulative distribution functions:

Uniform
Normal
Log-normal
Exponential
Beta
Gamma
...
...

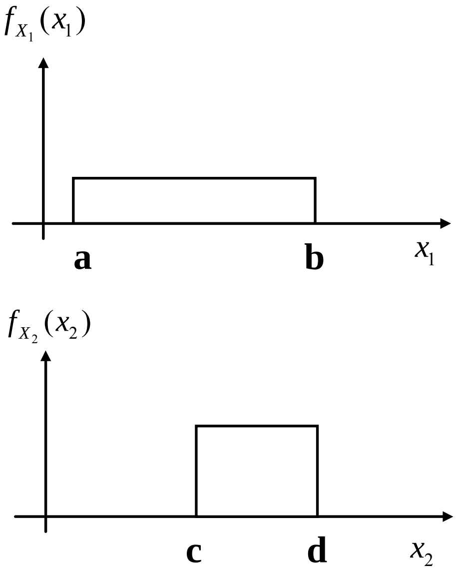
Distribution type	Parameters	Moments
Uniform, $a \le x \le b$ $f_X(x) = \frac{1}{b-a}$ $F_X(x) = \frac{x-a}{b-a}$	a b	$\mu = \frac{a+b}{2}$ $\sigma = \frac{b-a}{\sqrt{12}}$
Normal $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ $F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt$	μ $\sigma > 0$	μ σ
Shifted Lognormal, $x > \varepsilon$ $f_X(x) = \frac{1}{(x-\varepsilon)\zeta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right)^2\right)$ $F_X(x) = \Phi\left(\frac{\ln(x-\varepsilon)-\lambda}{\zeta}\right)$	λ $\zeta > 0$ ε	$\mu = \varepsilon + \exp\left(\lambda + \frac{\zeta^2}{2}\right)$ $\sigma = \exp\left(\lambda + \frac{\zeta^2}{2}\right)\sqrt{\exp(\zeta^2)-1}$
Shifted Exponential, $x \ge \varepsilon$ $f_X(x) = \lambda \exp(-\lambda(x-\varepsilon))$ $F_X(x) = 1 - \exp(-\lambda(x-\varepsilon))$	ε $\lambda > 0$	$\mu = \varepsilon + \frac{1}{\lambda}$ $\sigma = \frac{1}{\lambda}$
Gamma, $x \ge 0$ $f_X(x) = \frac{b^p}{\Gamma(p)} \exp(-bx) x^{p-1}$ $F_X(x) = \frac{\Gamma(bx, p)}{\Gamma(p)}$	$p > 0$ $b > 0$	$\mu = \frac{p}{b}$ $\sigma = \frac{\sqrt{p}}{b}$

Small Example 1

We remember the convolution integral which we used for establishing the probability density function for the sum of two random variables:

$$Y = X_1 + X_2$$

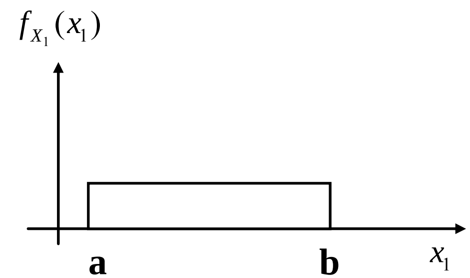
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1$$

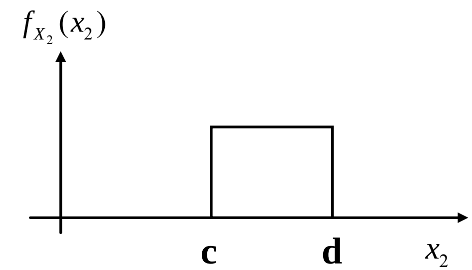


Let us see how easily this works for two uniformly distributed random variables:

Small Example 1

Assuming that the two random variables are independent we can write the convolution integral as:

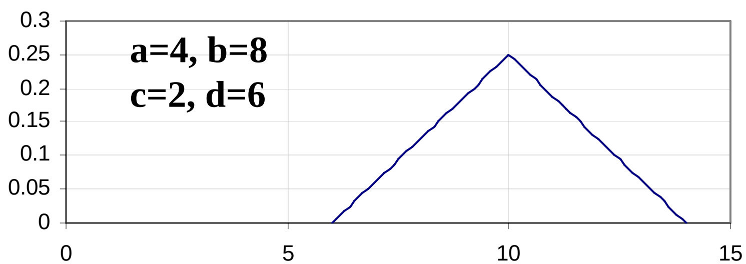


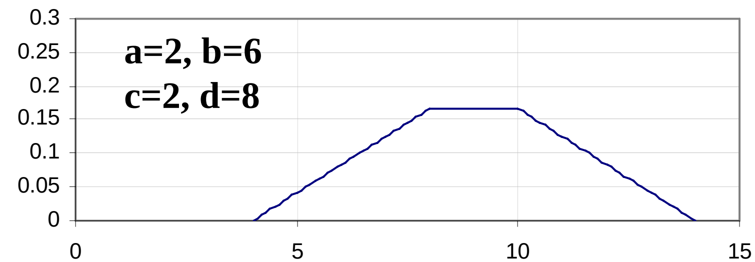


$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1$$

$$= \frac{1}{(b-a)(d-c)} \int_a^b \Uparrow (y - x_1 \in [c;d]) dx_1$$

$$= \frac{1}{(b-a)(d-c)} [x_1]_{\max(c, y-a)}^{\min(d, y-b)}, \quad a + c \le y \le b + d$$





Stochastic Processes and Extremes

- Random quantities may be "time variant" in the sense that they take new values at different times or at new trials.

  - If the new realizations occur at discrete times and have discrete values the random quantity is called a **random sequence**

    failure events, traffic congestions,...

  - If the new realizations occur continuously in time and take continues values the random quantity is called a **random process** or **stochastic process**

    wind velocity, wave heights,...

Stochastic Processes and Extremes

- Random sequences

  - A sequence of experiments with only two possible and mutually exclusive outcomes is called a **Bernoulli trial**
  - Typically the outcomes of Bernoulli trials are denoted **successes** or **failures**

  If the probability of success in one trial is constant and equal to *p* the probability density of *Y* successes in *n* trials, i.e. $p_Y(y)$ is given by:

$$p_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1,2 \ldots n$$

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

Binomial probability density function

Binomial operator

Stochastic Processes and Extremes

- Random sequences

  - A sequence of experiments with only two possible and mutually exclusive outcomes is called a **Bernoulli trial**

  The **Binomial cumulative distribution function** then follows as:

$$P_Y(y) = \sum_{i=0}^{y} \binom{y}{i} p^i (1-p)^{n-i}, \qquad y = 0,1,2,\ldots n$$

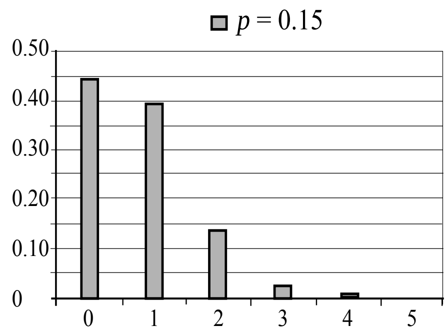
Stochastic Processes and Extremes

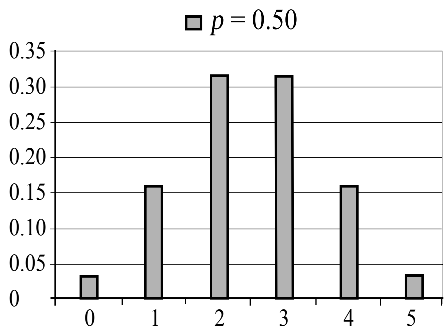
- Random sequences

  - A sequence of experiments with only two possible and mutually exclusive outcomes is called a **Bernoulli trial**

  Illustration:

  Binomial probability density function for n=5 and p=0.15 and p=0.5





Small Example 2

We remember that we can establish the probability density function of a function of a random variable through:

$$Y = g(X)$$

$$f_Y(y) = \left|\frac{\partial x}{\partial y}\right| f_X(x)$$

Small Example 2

Let us see how easily this works:

$$Y = X^2$$

$$\Downarrow$$

$$X = \sqrt{Y}$$

$$f_Y(y) = \left| \frac{\partial x}{\partial y} \right| f_X(x)$$

$$\frac{\partial x}{\partial y} = \frac{\partial \sqrt{y}}{\partial y} = \frac{1}{2} y^{-\frac{1}{2}}$$

$$f_Y(y) = \left| \frac{1}{2} y^{-\frac{1}{2}} \right| f_X(\sqrt{y})$$

Stochastic Processes and Extremes

- Random sequences

  The expected value and the variance of a **binomially distributed** random variable *Y* is given by:

$$E[Y] = np$$

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

Stochastic Processes and Extremes

- Random sequences

  The probability density function for the number of (independent) trials before the first success can be given as:

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

← **Geometric probability density**

and the corresponding cumulative distribution function is thus

$$P_N(n) = \sum_{i=1}^{n} p(1-p)^{i-1} = 1-(1-p)^n$$

↖ **Geometric cumulative distribution**

Small Example 3

We remember that we could establish the probability density function of a vector of random variables *Y* which were given as functions of a vector of random variables *X*

$$\mathbf{Y} = (Y_1, Y_2, ..Y_n)^T$$

$$\mathbf{X} = (X_1, X_2, ..X_n)^T \qquad Y_i = g_i(\mathbf{X}) \qquad X_i = f_i(\mathbf{Y})$$

$$f_{\mathbf{Y}}(\mathbf{y}) = |\mathbf{J}| \, f_{\mathbf{X}}(\mathbf{x})$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots\cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots\cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

Small Example 3

Let us see how easily this approach can be applied for the following problem:

$$Y_1 = X_1 + X_2 \qquad X_1 = Y_1 - Y_2$$

$$Y_2 = X_2 \qquad X_2 = Y_2$$

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \cdots\cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots\cdots & \frac{\partial x_n}{\partial y_n} \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \qquad \det(\mathbf{J}) = 1 \times 1 - 0 \times 1 = 1 \Rightarrow |\mathbf{J}| = 1$$

$$\boxed{f_{\mathbf{Y}}(y_1, y_2) = f_{\mathbf{X}}(y_1 - y_2, y_2)}$$

Stochastic Processes and Extremes

The median of the geometric distribution provides information in regard to how "long" we need to play a game with probability *p* of winning per time unit.

Time units might be
- tosses (dices)
- years (earthquakes)

The median is defined through

$$P_N(n) = 0.5 = 1 - (1 - p)^n$$

All we need to determine is *n* as a function of *p*

Stochastic Processes and Extremes

The median of the geometric distribution provides information in regard to how "long" we need to play a game with probability *p* of winning per time unit.

$$P_N(n) = 0.5 = 1 - (1-p)^n$$

We take the natural logarithm on both sides and get:

$$\ln(0.5) = n \ln(1-p)$$

$$\Downarrow$$

$$0.7 \approx -n \ln(1-p)$$

Now we use that the natural logarithm of

$$\ln(1-p) = -p + \frac{1}{2}p^2 - \frac{1}{3}p^3 + \ldots = \sum_{k=1}^{\infty}(-1)^k \frac{p^k}{k}$$

$$\Downarrow$$

$$\ln(1-p) \approx -p \quad \text{for small } p$$

$$0.7 \approx np \Rightarrow n = \frac{0.7}{p}$$

Stochastic Processes and Extremes

We can now apply this result:

50% chance of getting a 6 requires (*n* tosses):
***n* = 0.7 x 6 = 4 tosses**

50% chance of getting two 6 (with 2 dices) requires:
***n* = 0.7 x 36 = 25 tosses**

50% chance experiencing an earthquake with an annual probability of 0.001 requires (*n* years):
***n* = 0.7 x 1000 = 700 years**

Stochastic Processes and Extremes

- Random sequences

  The expected value and the variance of a random variable with a *Geometrically* distributed random variable are given by:

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

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

If *p* is the annual probability of e.g. an extreme earthquake *E[N]* is the return period of such earthquakes

Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

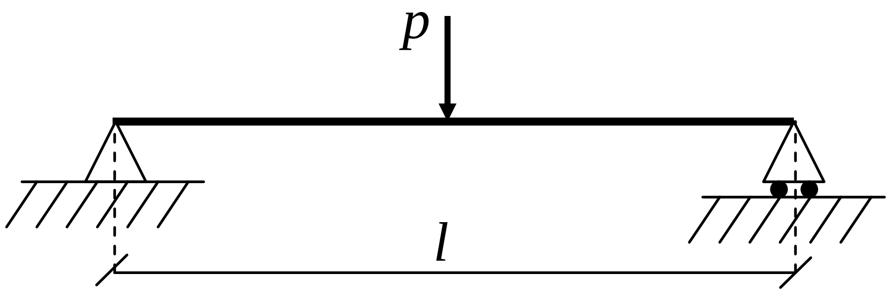
Contents of Todays Lecture

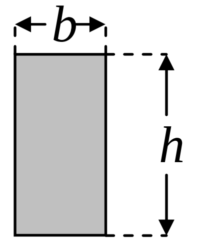
- Presentation on the result of the classroom assessment
- What is a random variable?
- The decision context!
- What are we doing today?
- Details will follow ☺

What is a random variable?

- Let us consider a very simple structural engineering problem!

- We want to design a steel beam – and assume – based on experience that the design controlling load effect is the midspan bending moment *M*
  - the design variable being the moment of resistance *W* of the cross section
  - the load *p* and the yield stress s_y of the beam are associated with uncertainty





Mid span cross-section

$$W = \frac{1}{6} b h^2$$

What is a random variable?

- **The moment capacity of the cross-section R_M and the mid span moment M are calculated as:**

$$R_M = W\sigma_y$$

R_M moment capacity of cross section

W moment of resistance

σ_y yield stress of the steel

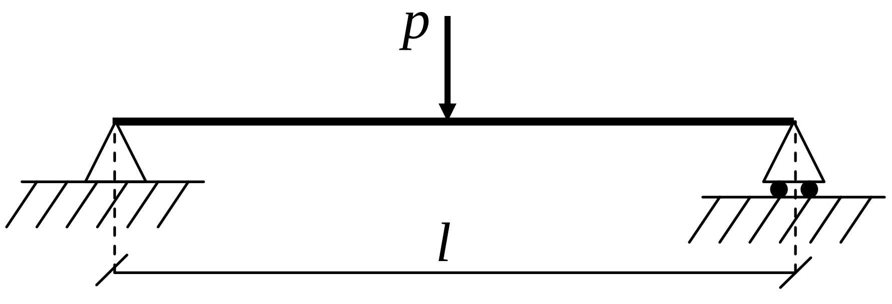
$$W = \frac{1}{6}bh^2$$

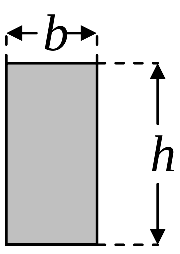
$$M = \frac{1}{4}pl$$

M mid span moment

p load

l length of beam





Mid span
cross-section

What is a random variable?

- **We can now establish a design equation as:**

$$R_M(b,h) - M \geq 0$$

$$\Downarrow$$

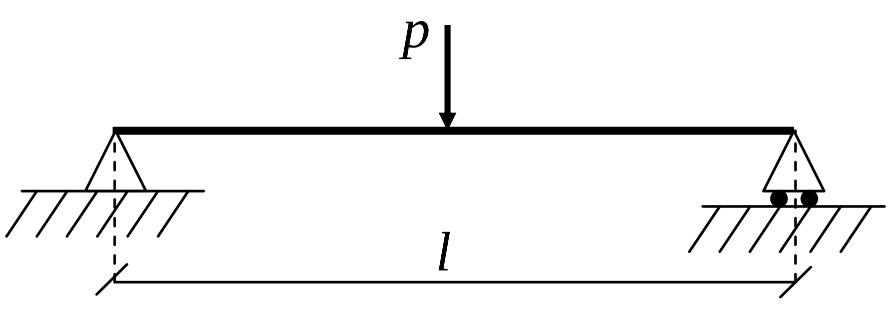
$$W(b,h)\sigma_y - \frac{1}{4}Pl \geq 0$$

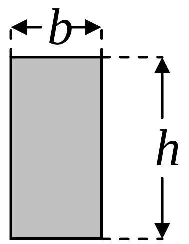
$$\Downarrow$$

$$\frac{1}{6}bh^2\sigma_y - \frac{1}{4}Pl \geq 0$$

The engineer must now select *W*, or rather *b* and *h* such that the design equation is fulfilled

But as *p* and σ_y are associated with uncertainty – she/he must take this uncertainty into account !

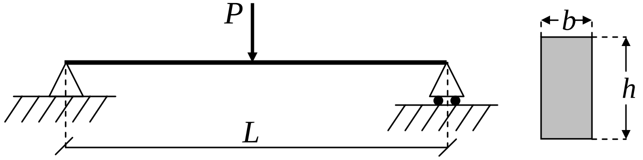




Mid span cross-section

What is a random variable?

- The uncertainty is accounted for by representing *p* and s_y in the design equation as two random variables.



$$P: \quad \text{Normal distributed: } N(\mu_P, \sigma_P)$$

$$\Sigma_y: \quad \text{Normal distributed: } N(\mu_{\Sigma_y}, \sigma_{\Sigma_y})$$

The random variable *P* represents the random variability of the load *p* during a period of one year

The random variable S_y represents the random variability of the steel yield stress s_y - produced by an unknown steel producer.

What is a random variable?

- As the load and yield stress are uncertain the design equation cannot be fulfilled with certainty – independent on the choice of *b* and *h*.

- However, it can be fulfilled with probability !

- The beam can be designed such that the probability of failure is less or equal to a given number – the requirement to safety.

What is a random variable?

- **Let us assume that the load and yield stress are given as:**

$$P: \qquad N(\mu_P, \sigma_P) = \quad N(100\text{kN}, 20\text{kN})$$

$$\Sigma_y: \qquad N(\mu_{\Sigma_y}, \sigma_{\Sigma_y}) = \quad N(370\text{mPa}, 15\text{mPa})$$

we can now write the event of failure as:

$$\left.\begin{aligned} &\frac{1}{6}bh^2\Sigma_y - \frac{1}{4}Pl \le 0 \\ &\Downarrow \\ &\Sigma_y - \frac{3}{2bh^2}Pl \le 0 \end{aligned}\right\} \quad \boxed{S = \Sigma_y - \frac{3}{2bh^2}Pl \le 0}$$

This is called a safety margin!

let us further assume that *l*=5000mm and *b*=50mm

- **Let us now determine *h* such that the annual probability of failure is equal to 10^{-3}**

What is a random variable?

- We have already learned that a linear combination of Normal distributed random variables is also Normal distributed
  The **expected value** of *S* is equal to:

$$\mu_S = \mu_{\Sigma_y} - \frac{3}{2\cdot 0.05\cdot h^2}\mu_P\cdot 5$$
$$= 370 - \frac{3}{2\cdot 0.05\cdot h^2}\mu_P\cdot 5 = 370 - \frac{150000}{h^2}$$

The **variance** of *S* is equal to:

$$\sigma_S^2 = \sigma_{\Sigma_y}^2 + \left(\frac{3}{2\cdot 0.05\cdot h^2}\right)^2\sigma_P^2$$
$$= 15^2 + \left(\frac{30}{h^2}\right)^2\cdot 20000^2 = 225 + \frac{3.6\cdot 10^{11}}{h^4}$$

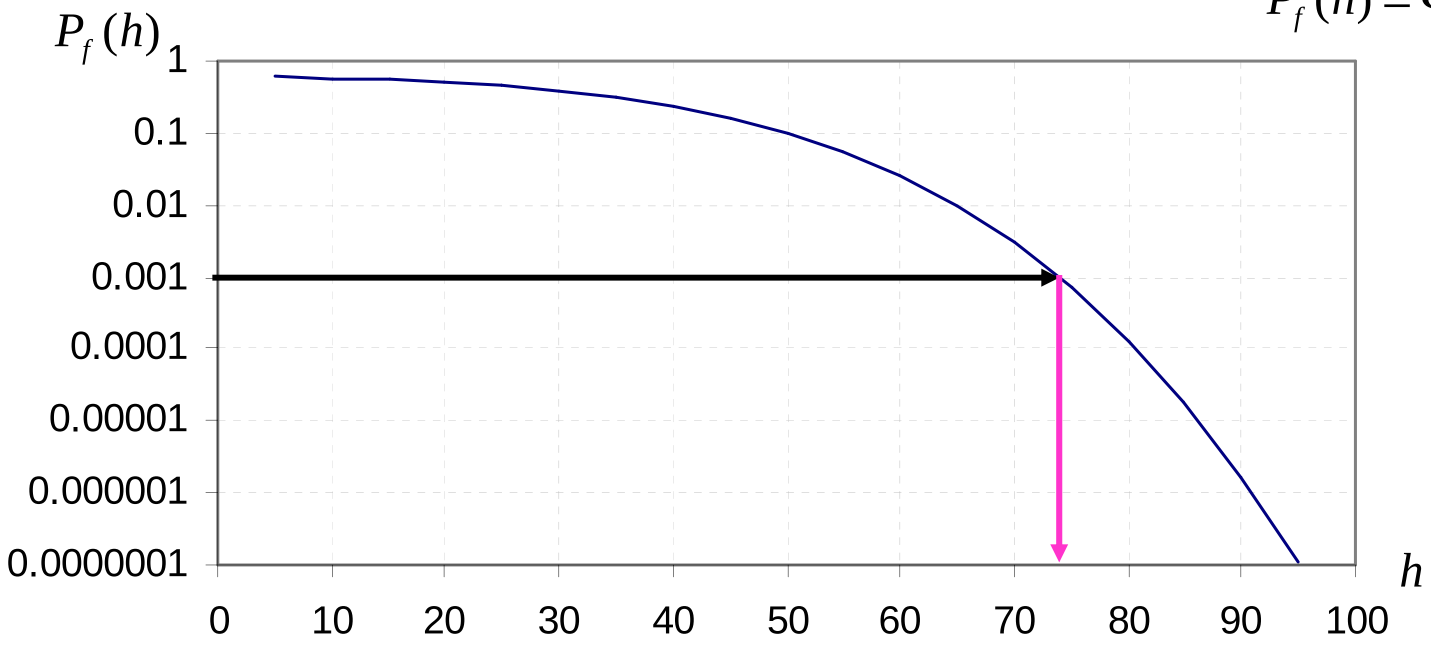
The probability of failure is now easily determined from the standard Normal cumulative distribution function

$$P_f(h) = \Phi\left(\frac{0-\mu_S(h)}{\sigma_S(h)}\right)$$

What is a random variable?

- **Calculating the probability of failure as a function of *h* we get:**

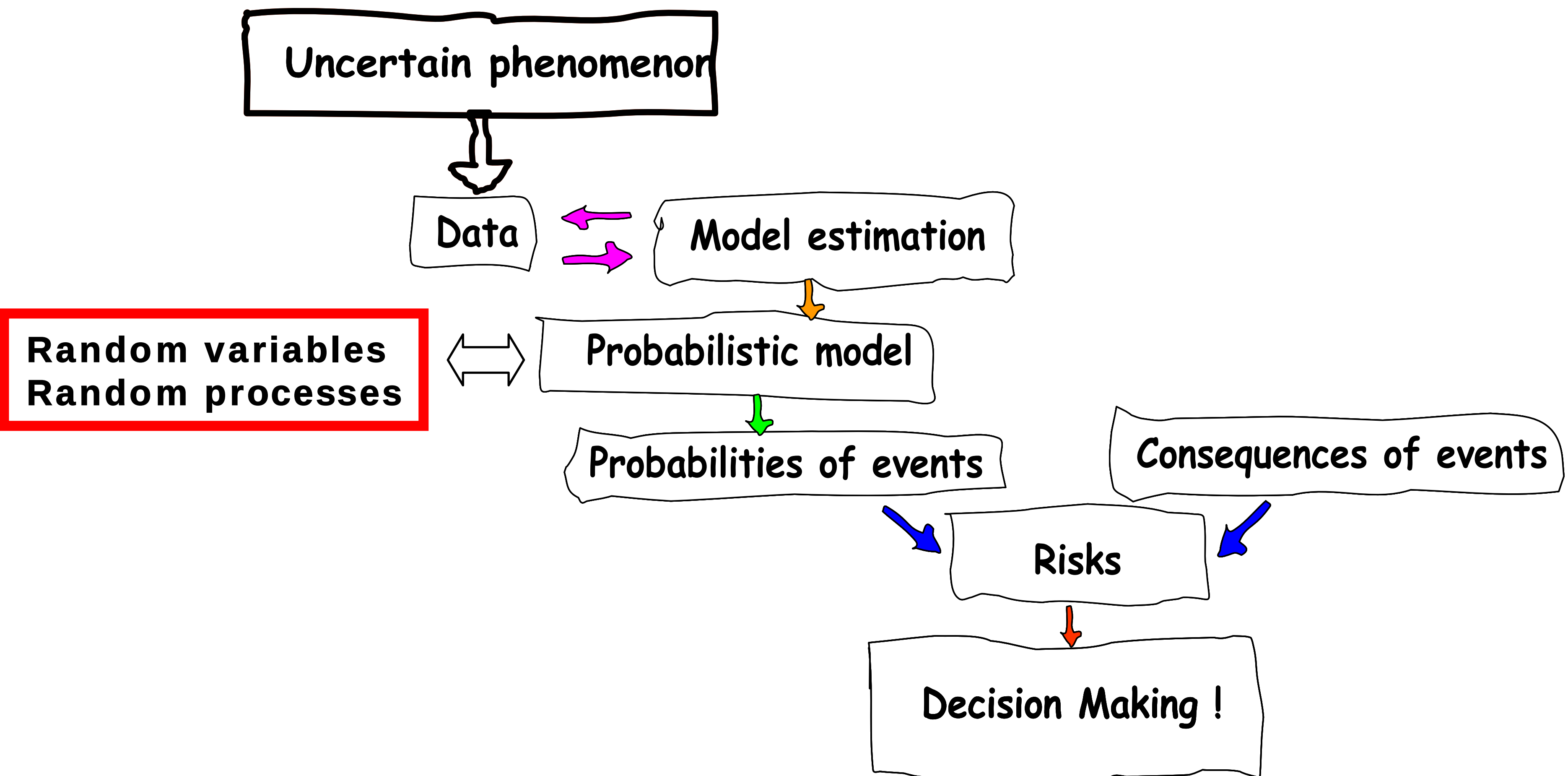
$$P_f(h) = \Phi\left(\frac{0-\mu_S(h)}{\sigma_S(h)}\right)$$



The height of the beam must thus be equal to 73mm!

The decision context!

- **Why uncertainty modeling?**



What are we doing today?

- We have already introduced random variables as a means of representing uncertainties which we may quantify based on observations – related to time frames from which we have experience and observations!

- In many real problems of decision making we need to take into account what might happen in the far future – exceeding the time frames for which we have experience!

  - 475 year design earthquake!

  - 100 year storm / flood

  - 100 year maximum truck load

  - etc..

Thus we need to develop models which can support us in the modeling of extremes of uncertain / random phenomena !

What are we doing today?

- We have already introduced random variables as a means of representing uncertainties which we may quantify based on observations.

- Often we use random variables to represent uncertainties which do not vary in time:

  - Model uncertainties (lack of knowledge)
  - Statistical uncertainties (lack of data).

- Or we use such random variables to represent the random variations which can be observed within a given reference period.

What are we doing today?

- In many engineering problems we need to be able to describe the random variations in time more specifically:

  The occurrences of events at random discrete points in time (rock-fall, earthquakes, accidents, queues, failures, etc.)
  - Poisson process, exponential and Gamma distribution

  The random values of events occurring continuously in time (wind pressures, wave loads, temperatures, etc.)
  - Continuous random processes (Normal process)

Discrete event of flood

Continuous stress variations due to waves

What are we doing today?

- However, we are also interested in modeling extreme events such as:

  the maximum value of an uncertain quantity within a given reference period
  - extreme value distributions

  the expected value of the time till the occurrence of an event exceeding a certain severity
  - return period

Extreme water level

Maximum wave load

What are we doing today?

- In summary we will look at:

  - Random sequences (**Poisson process**)

  - Waiting time between events (**Exponential and Gamma distributions**)

  - **Continuous random processes** (the Normal process)

  - Criteria for extrapolation of extremes (**stationarity and ergodicity**)

  - The maximum value within a reference period (**extreme value distributions**)

  - Expected value of the time till the occurrence of an event exceeding a certain severity (**return period**)

Random Sequences

- **The Poisson counting process is one of the most commonly applied families of probability distributions applied in reliability theory**

  The process $N(t)$ denoting the number of events in a (time) interval $(t,t+\Delta t[$ is called a Poisson process if the following conditions are fulfilled:

  1) the probability of one event in the interval $(t,t+\Delta t[$ is asymptotically proportional to Δt.
  2) the probability of more than one event in the interval $(t,t+\Delta t[$ is a function of higher order of Δt for $\Delta t \to 0$.
  3) events in disjoint intervals are mutually independent.

Random Sequences

- **The Poisson process can be described completely by its intensity *n(t)***

$$\nu(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(\text{one event in } [t, t+\Delta t[)$$

if *n*(*t*) = constant, the Poisson process is said to be homogeneous, otherwise it is inhomogeneous.

The probability of *n* events in the time interval (0,*t*[is:

$$P_n(t) = \frac{\left(\int_0^t \nu(\tau)d\tau\right)^n}{n!} \exp\left(-\int_0^t \nu(\tau)d\tau\right)$$

$$P_n(t) = \frac{\nu t^n}{n!} \exp(-\nu t)$$

Homogeneous case !

Random Sequences

- The mean value and variance of the random variable describing the number of events in a given time interval (0,t[are given as:

$$E\left[N(t)\right] = Var\left[N(t)\right] = \int_0^t \nu(\tau)d\tau$$

Inhomogeneous case !

$$E\left[N(t)\right] = Var\left[N(t)\right] = \nu t$$

Homogeneous case !

Random Sequences

- **The Exponential distribution**

  The probability of no events in a given time interval (0,t[is often of special interest in engineering problems

  - no severe storms in 10 years
  - no failure of a structure in 100 years
  - no earthquake next year
  - …….

  This probability is directly achieved as:

$$P_0(t) = \frac{\left(\int_0^t \nu(\tau)d\tau\right)^0}{0!}\exp\left(-\int_0^t \nu(\tau)d\tau\right)$$

$$= \exp\left(-\int_0^t \nu(\tau)d\tau\right)$$

$$P_0(t) = \exp(-\nu t)$$

Homogeneous case !

Random Sequences

- The probability distribution function of the (waiting) time till the first event T_1 is now easily derived recognizing that the probability of $T_1 > t$ is equal to $P_0(t)$ we get:

$$F_{T_1}(t_1) = 1 - P_0(t_1)$$

$$= 1 - \exp\left(-\int_0^t \nu(\tau)\, d\tau\right)$$

Homogeneous case !

$$F_{T_1}(t_1) = 1 - \exp(-\nu t)$$

↑

Exponential cumulative distribution

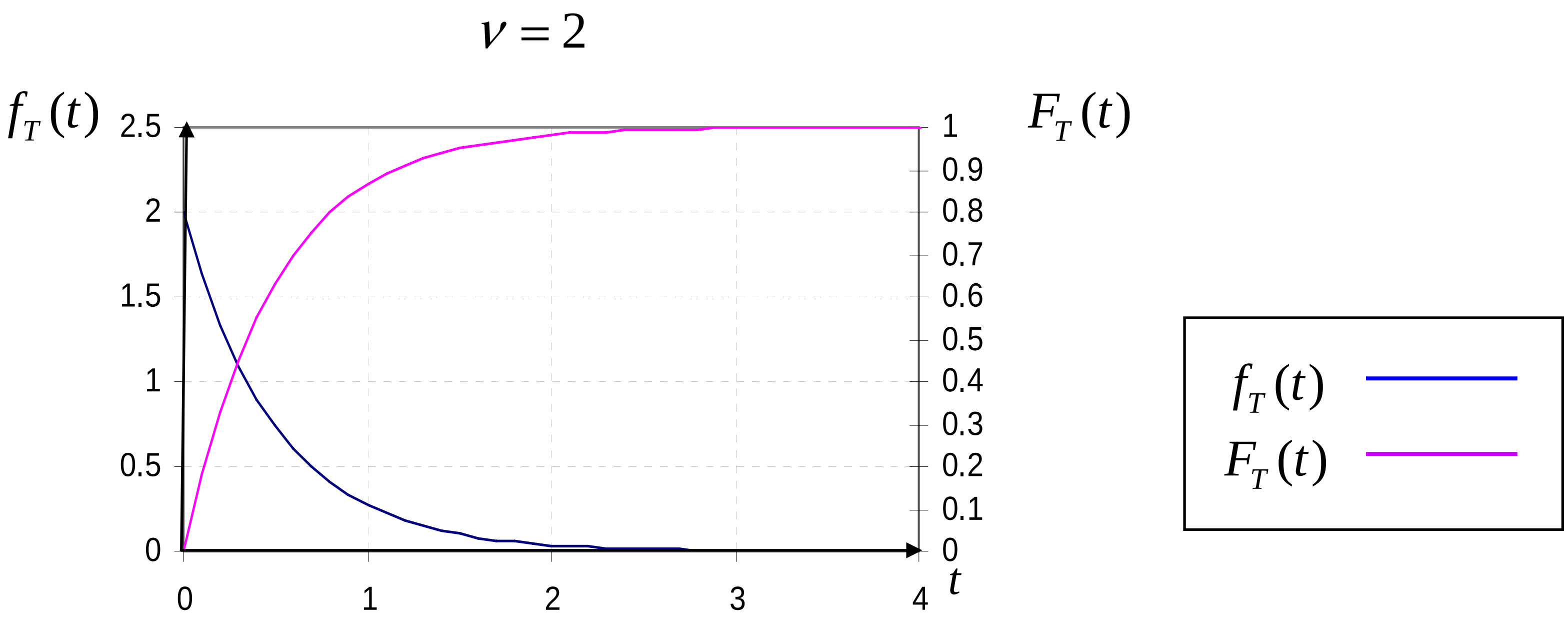
Exponential probability density

↓

$$f_{T_1}(t_1) = \nu \exp(-\nu t)$$

Random Sequences

The Exponential probability density and cumulative distribution functions



Random Sequences

- **The exponential distribution is frequently applied in the modeling of waiting times**

  - time till failure
  - time till next earthquake
  - time till traffic accident
  - ....

$$f_{T_1}(t_1) = \nu \exp(-\nu t)$$

The expected value and variance of an exponentially distributed random variable T_1 are:

$$E[T_1] = \sqrt{Var[T_1]} = 1/\nu$$

Random Sequences

- Sometimes also the time *T* till the *n*'th event is of interest in engineering modeling:

  - repair events
  - flood events
  - arrival of cars at a roadway crossing

  If T_i, i=1,2,..n are independent exponentially distributed waiting times, then the sum *T* i.e.:

$$T = T_1 + T_2 + \ldots + T_{n-1} + T_n$$

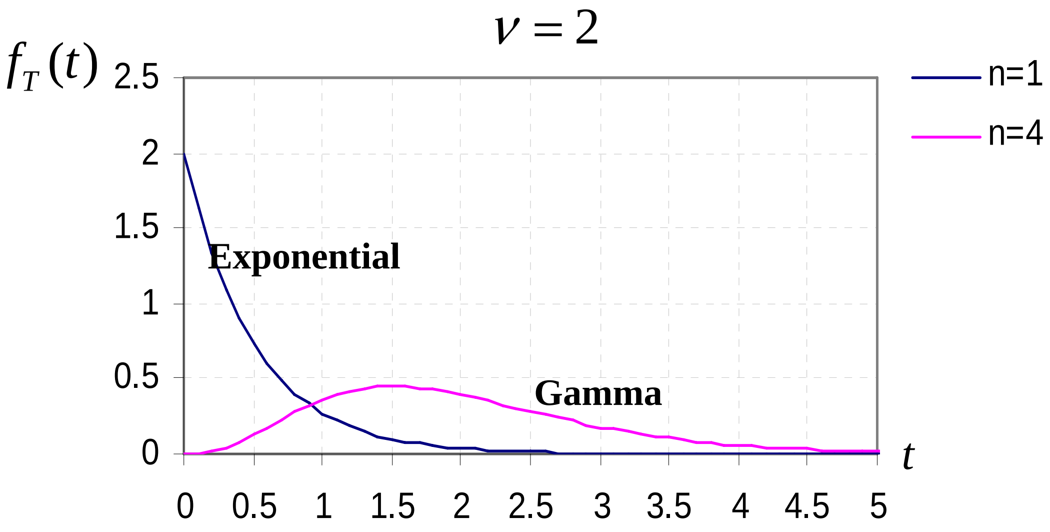
follows a **Gamma distribution**:

$$f_T(t) = \frac{\nu(\nu t)^{(n\text{-}1)} \exp(-\nu t)}{(n-1)!}$$

This follows from repeated use of the result of the distribution of the sum of two random variables

Random Sequences

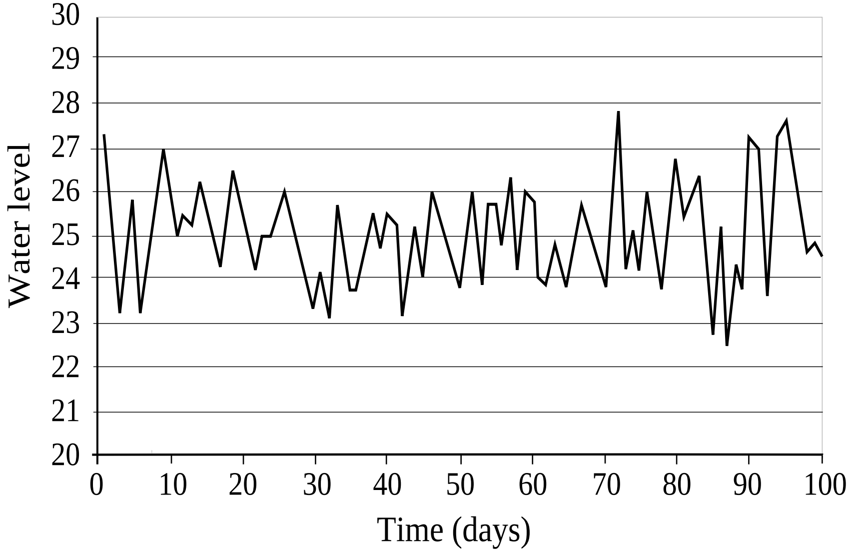
The Gamma probability density function



Random Processes

- **Continuous random processes**

  A continuous random process is a random process which has realizations continuously over time and for which the realizations belong to a continuous sample space.



Variations of;
water levels
wind speed
rain fall
.
.
.

Realization of continuous scalar valued random process

Random Processes

- **Continuous random processes**

  The mean value of the possible realizations of a random process is given as:

$$\mu_X(t) = E\left[X(t)\right] = \int_{-\infty}^{\infty} x \; f_X(x\,;t)dx$$

Function of time !

The correlation between realizations at any two points in time is given as:

$$R_{XX}(t_1,t_2) = E\left[X(t_1)X(t_2)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 \; x_2 \;\; f_{XX}(x_1,x_2;t_1,t_2)dx_1dx_2$$

Auto-correlation function – refers to a scalar valued random process

Random Processes

- **Continuous random processes**

The auto-covariance function is defined as:

$$C_{XX}(t_1,t_2) = E\left[(X(t_1)-\mu_X(t_1))(X(t_2)-\mu_X(t_2))\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_X(t_1))\,(x_2-\mu_X(t_2))\; f_{XX}(x_1,x_2;t_1,t_2)dx_1dx_2$$

for $t_1 = t_2 = t$ the auto-covariance function becomes the covariance function:

$$\sigma_X^2(t) = C_{XX}(t,t) = R_{XX}(t,t) - \mu_X^2(t)$$

$\sigma_X(t)$ **Standard deviation function**

Random Processes

- **Continuous random processes**

A vector valued random process is a random process with two or more components:

$$\mathbf{X}(t) = (X_1(t), X_2(t),.., X_n(t))^T$$

with covariance functions:

$$C_{X_iX_j}(t_1,t_2) = E\left[(X_i(t_1)-\mu_{X_i}(t_1))(X_j(t_2)-\mu_{X_j}(t_2))\right]$$

$i = j$ **auto-covariance functions**

$i \neq j$ **cross-covariance functions**

The correlation coefficient function is defined as:

$$\rho\left[X_i(t_1), X_j(t_2)\right] = \frac{C_{X_iX_j}(t_1,t_2)}{\sigma_{X_i}(t_1)\cdot\sigma_{X_j}(t_2)}$$

Random Processes

- **Normal or Gauss process**

A random process $X(t)$ is said to be Normal if:

For any set; $X(t_1), X(t_2),\ldots,X(t_j)$

the joint probability distributions of $X(t_1), X(t_2),\ldots,X(t_j)$

is the Normal distribution.

Random Processes

- **Stationarity and ergodicity**

A random process is said to be *strictly stationary* if all its moments are invariant to a shift in time.

A random process is said to be *weakly stationary* if the first two moments i.e. the mean value function and the auto-correlation function are invariant to a shift in time

$$\left.\begin{aligned} &\mu_X(t) = cst \\ &R_{XX}(t_1, t_2) = f(t_2 - t_1) \end{aligned}\right\} \quad \textbf{Weakly stationary}$$

Random Processes

- **Stationarity and ergodicity**

  - A random process is said to be *strictly ergodic* if it is strictly stationary and in addition all its moments may be determined on the basis of one realization of the process.
  - A random process is said to be *weakly ergodic* if it is weakly stationary *and in addition* its first two moments may be determined on the basis of one realization of the process.

- **The assumptions in regard to stationarity and ergodicity are often very useful in engineering applications.**

  - If a random process is ergodic we can extrapolate probabilistic models of extreme events within short reference periods to any longer reference period.

Extreme Value Distributions

In engineering we are often interested in extreme values i.e. the smallest or the largest value of a certain quantity within a **certain time interval** e.g.:

The largest earthquake in 1 year

The highest wave in a winter season

The largest rainfall in 100 years

Extreme Value Distributions

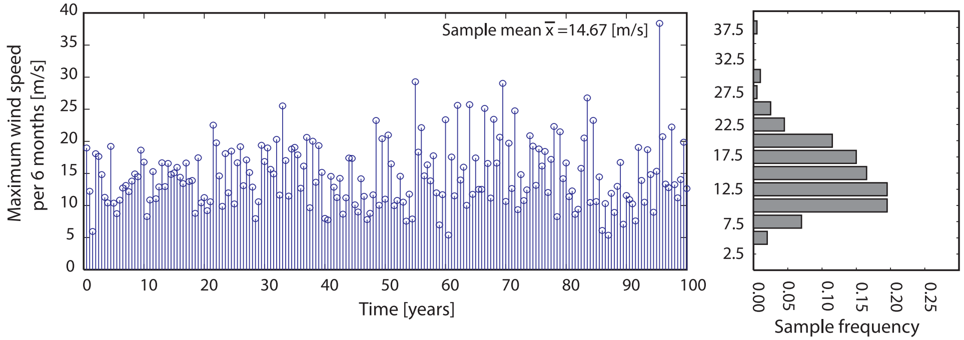
We could also be interested in the smallest or the largest value of a certain quantity within a **certain volume or area** unit e.g.:

The largest concentration of pesticides in a volume of soil

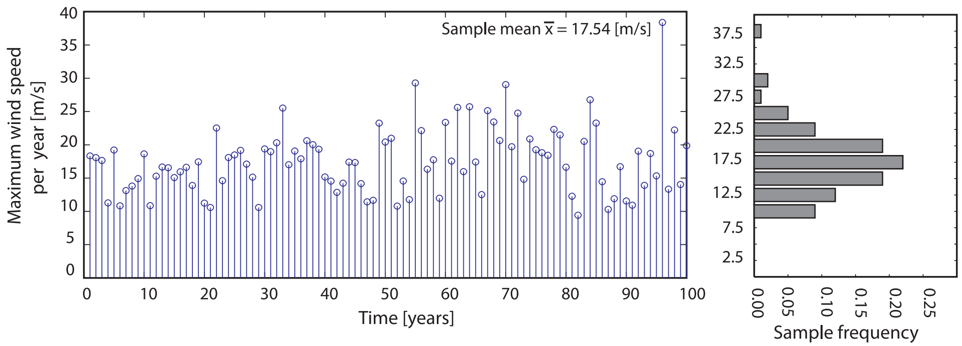
The weakest link in a chain

The smallest thickness of concrete cover

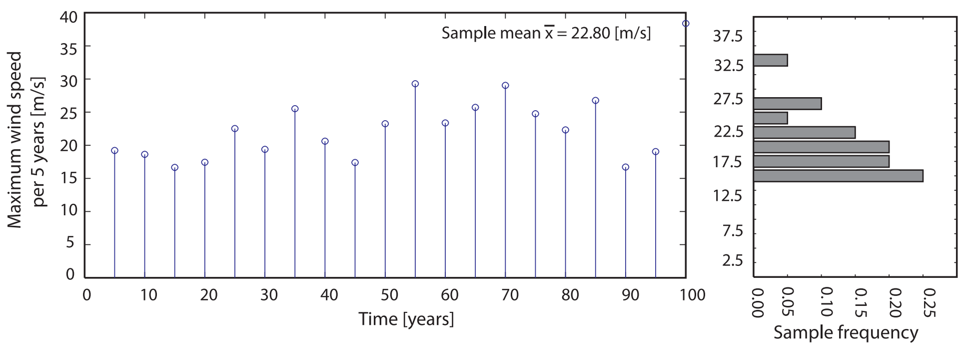
Extreme Value Distributions



Observed monthly extremes



Observed annual extremes



Observed 5-year extremes

Extreme Value Distributions

If the extremes within the period *T* of an ergodic random process *X*(*t*) are independent and follow the distribution:

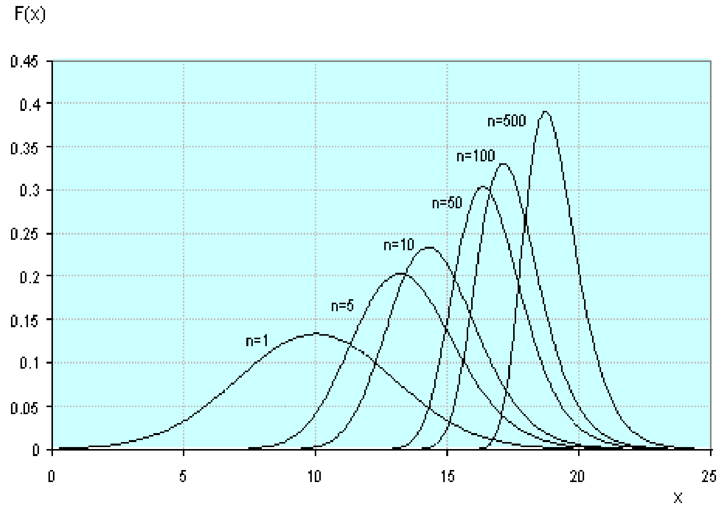
$$F_{X,T}^{\max}(x)$$

Then the extremes of the same process within the period:

$$n \cdot T$$

will follow the distribution:

$$F_{X,nT}^{\max}(x) = \left(F_{X,T}^{\max}(x)\right)^n$$



Extreme Value Distributions

Extreme Type I – Gumbel Max

When the upper tail of the probability density function falls off exponentially (exponential, Normal and Gamma distribution) then the maximum in the time interval *T* is said to be Type I extreme distributed

$$f_{X,T}^{\max}(x) = \alpha \exp(-\alpha(x-u) - \exp(-\alpha(x-u)))$$

$$F_{X,T}^{\max}(x) = \exp(-\exp(-\alpha(x-u)))$$

$$\mu_{X_T^{\max}} = u + \frac{\gamma}{\alpha} = u + \frac{0.577216}{\alpha}$$

$$\sigma_{X_T^{\max}} = \frac{\pi}{\alpha\sqrt{6}}$$

For increasing time intervals the variance is constant but mean value increases as:

$$\mu_{X_{nT}^{\max}} = \mu_{X_T^{\max}} + \frac{\sqrt{6}}{\pi}\sigma_{X_T^{\max}} \ln(n)$$

Extreme Value Distributions

Extreme Type II – Frechet Max

When a probability density function is downwards limited at zero and upwards falls off in the form

$$F_X(x) = 1 - \beta(\frac{1}{x})^k$$

then the maximum in the time interval *T* is said to be Type II extreme distributed

$$F_{X,T}^{\max}(x) = \exp(-\left(\frac{u}{x}\right)^k)$$

$$f_{X,T}^{\max}(x) = \frac{k}{u}\left(\frac{u}{x}\right)^{k+1}\exp(-\left(\frac{u}{x}\right)^k)$$

Mean value and standard deviation

$$\mu_{X_T^{\max}} = u\Gamma(1-\frac{1}{k})$$

$$\sigma^2_{X_T^{\max}} = u^2\left[\Gamma(1-\frac{2}{k}) - \Gamma^2(1-\frac{1}{k})\right]$$

Extreme Value Distributions

Extreme Type III – Weibull Min

When a probability density function is downwards limited at ε and the lower tail falls off towards ε in the form

$$F(x) = c(x-\varepsilon)^k$$

then the maximum in the time interval *T* is said to be Type III extreme distributed

$$F_{X,T}^{\min}(x) = 1 - \exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right)$$

$$f_{X,T}^{\min}(x) = \frac{k}{u-\varepsilon}\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{k-1}\exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right)$$

Mean value and standard deviation

$$\mu_{X_T^{\min}} = \varepsilon + (u-\varepsilon)\Gamma(1+\frac{1}{k})$$

$$\sigma_{X_T^{\min}}^2 = (u-\varepsilon)^2\left[\Gamma(1+\frac{2}{k}) - \Gamma^2(1+\frac{1}{k})\right]$$

Return Period

The **return period** for extreme events T_R may be defined as:

$$T_R = n \cdot T = \frac{1}{(1 - F_{X,T}^{\max}(x))} T$$

Example:

Let us assume that - according to the cumulative probability distribution of the annual maximum traffic load - the annual probability that a truck load is larger than 100 ton is equal to 0.02 – then the return period of such heavy truck events is:

$$T_R = n \cdot T = \frac{1}{0.02} 1 = 50 \text{ years}$$

T=1 since we speak for annual probability of the extreme load event

Statistics and Probability Theory in Civil, Surveying and Environmental Engineering

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

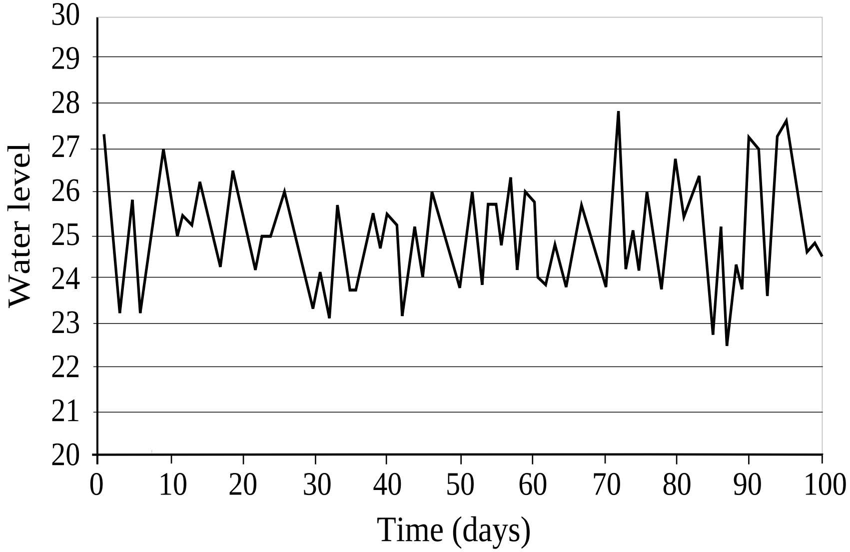
Contents of Today‘s Lecture

- Presentation on the result of the classroom assessment
- Catching up with the lecture from last time
  - Continuous random processes
  - Extremes of random processes
- Overview of Estimation and Model Building
- Probability Distribution Functions in Statistics
- Estimators for Sample Descriptors – Sample Statistics
  - statistical characteristics of the sample average
  - statistical characteristics of the sample variance
  - confidence intervals on estimators

Random Processes

- **Continuous random processes**

  A continuous random process is a random process which has realizations continuously over time and for which the realizations belong to a continuous sample space.



Variations of;
water levels
wind speed
rain fall
.
.
.

Realization of continuous scalar valued random process

Random Processes

- **Continuous random processes**

  The **mean value** of the possible realizations of a random process is given as:

$$\mu_X(t) = E\left[X(t)\right] = \int_{-\infty}^{\infty} x\ f_X(x\,;t)dx$$

↑

Function of time !

The **correlation** between realizations at any two points in time is given as:

$$R_{XX}(t_1,t_2) = E\left[X(t_1)X(t_2)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1\ x_2\ \ f_{XX}(x_1,x_2;t_1,t_2)dx_1dx_2$$

Auto-correlation function – refers to a scalar valued random process

Random Processes

- **Continuous random processes**

The auto-covariance function is defined as:

$$C_{XX}(t_1,t_2)=E\left[(X(t_1)-\mu_X(t_1))(X(t_2)-\mu_X(t_2))\right]$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_X(t_1))\,(x_2-\mu_X(t_2))\,f_{XX}(x_1,x_2;t_1,t_2)dx_1dx_2$$

for $t_1=t_2=t$ the auto-covariance function becomes the covariance function:

$$\sigma_X^2(t)=C_{XX}(t,t)=R_{XX}(t,t)-\mu_X^2(t)$$

$\sigma_X(t)$ **Standard deviation function**

Random Processes

- **Continuous random processes**

A vector valued random process is a random process with two or more components:

$$\mathbf{X}(t) = (X_1(t), X_2(t), .., X_n(t))^T$$

with covariance functions:

$$C_{X_i X_j}(t_1, t_2) = E\left[(X_i(t_1) - \mu_{X_i}(t_1))(X_j(t_2) - \mu_{X_j}(t_2))\right]$$

$i = j$ **auto-covariance functions**

$i \neq j$ **cross-covariance functions**

The correlation coefficient function is defined as:

$$\rho\left[X_i(t_1), X_j(t_2)\right] = \frac{C_{X_i X_j}(t_1, t_2)}{\sigma_{X_i}(t_1) \cdot \sigma_{X_j}(t_2)}$$

Random Processes

- **Normal or Gauss process**

A random process $X(t)$ is said to be Normal if:

for any set; $X(t_1), X(t_2),\ldots,X(t_j)$

the joint probability distribution of $X(t_1), X(t_2),\ldots,X(t_j)$

is the Normal distribution.

Random Processes

- **Stationarity and ergodicity**

A random process is said to be *strictly stationary* if all its moments are invariant to a shift in time.

A random process is said to be *weakly stationary* if the first two moments i.e. the mean value function and the auto-correlation function are invariant to a shift in time

$$\left.\begin{aligned} &\mu_X(t) = cst \\ &R_{XX}(t_1, t_2) = f(t_2 - t_1) \end{aligned}\right\} \quad \textbf{Weakly stationary}$$

Random Processes

- **Stationarity and ergodicity**

  - A random process is said to be *strictly ergodic* if it is strictly stationary and in addition all its moments may be determined on the basis of one realization of the process.
  - A random process is said to be *weakly ergodic* if it is weakly stationary *and in addition* its first two moments may be determined on the basis of one realization of the process.

- **The assumptions in regard to stationarity and ergodicity are often very useful in engineering applications.**

  - If a random process is ergodic we can extrapolate probabilistic models of extreme events within short reference periods to any longer reference period.

Extreme Value Distributions

In engineering we are often interested in extreme values i.e. the smallest or the largest value of a certain quantity within a **certain time interval** e.g.:

The largest earthquake in 1 year

The highest wave in a winter season

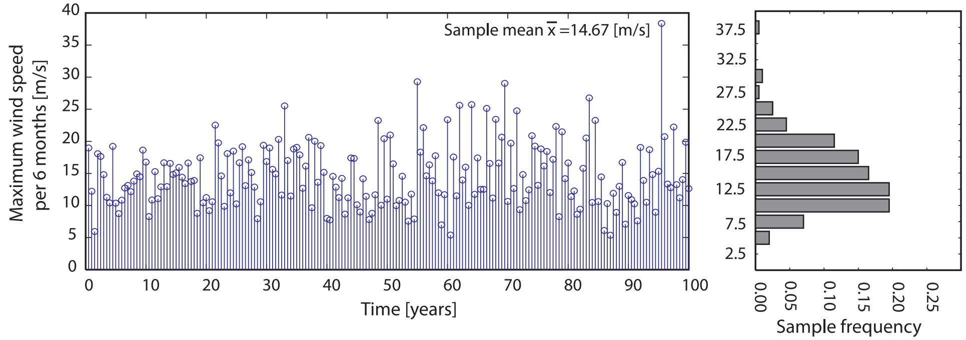
The largest rainfall in 100 years

Extreme Value Distributions

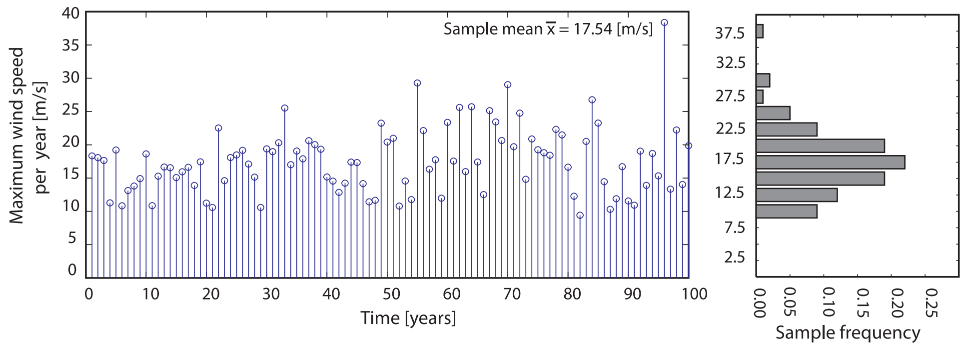
We could also be interested in the smallest or the largest value of a certain quantity within a **certain volume or area** unit e.g.:

- The largest concentration of pesticides in a volume of soil
- The weakest link in a chain
- The smallest thickness of concrete cover

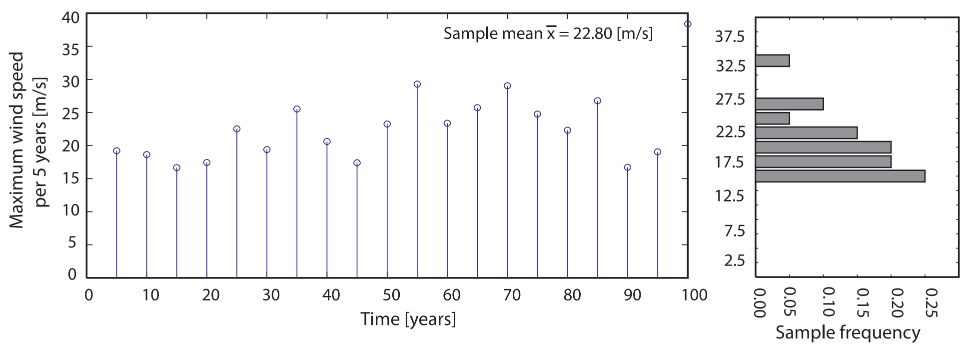
Extreme Value Distributions



Observed monthly extremes



Observed annual extremes



Observed 5-year extremes

Extreme Value Distributions

If the extremes within the period *T* of an ergodic random process *X*(*t*) are independent and follow the distribution:

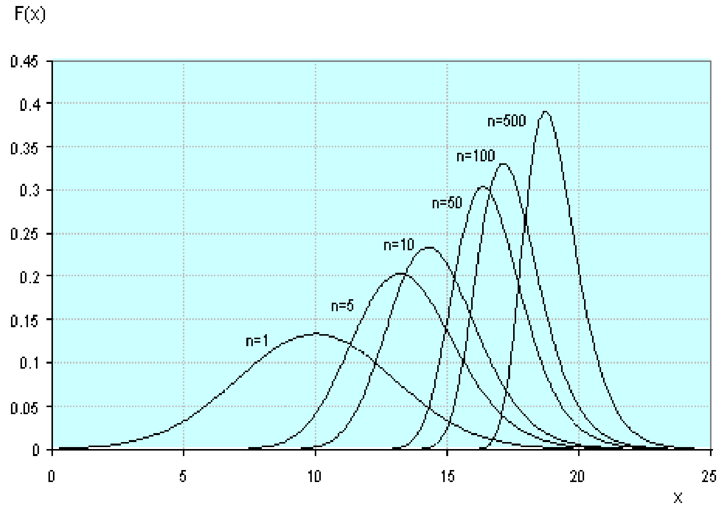
$$F_{X,T}^{\max}(x)$$

Then the extremes of the same process within the period:

$$n \cdot T$$

will follow the distribution:

$$F_{X,nT}^{\max}(x) = \left(F_{X,T}^{\max}(x)\right)^n$$



Extreme Value Distributions

Extreme Type I – Gumbel Max

When the upper tail of the probability density function falls off exponentially (exponential, Normal and Gamma distribution) then the maximum in the time interval *T* is said to be Type I extreme distributed

$$f_{X,T}^{\max}(x) = \alpha \exp(-\alpha(x-u) - \exp(-\alpha(x-u)))$$

$$F_{X,T}^{\max}(x) = \exp(-\exp(-\alpha(x-u)))$$

$$\mu_{X_T^{\max}} = u + \frac{\gamma}{\alpha} = u + \frac{0.577216}{\alpha}$$

$$\sigma_{X_T^{\max}} = \frac{\pi}{\alpha\sqrt{6}}$$

For increasing time intervals the variance is constant but the mean value increases as:

$$\mu_{X_{nT}^{\max}} = \mu_{X_T^{\max}} + \frac{\sqrt{6}}{\pi}\sigma_{X_T^{\max}} \ln(n)$$

Extreme Value Distributions

Extreme Type II – Frechet Max

When a probability density function is downwards limited at zero and upwards falls off in the form

$$F_X(x) = 1 - \beta\left(\frac{1}{x}\right)^k$$

then the maximum in the time interval *T* is said to be Type II extreme distributed

$$F_{X,T}^{\max}(x) = \exp\left(-\left(\frac{u}{x}\right)^k\right)$$

$$f_{X,T}^{\max}(x) = \frac{k}{u}\left(\frac{u}{x}\right)^{k+1} \exp\left(-\left(\frac{u}{x}\right)^k\right)$$

Mean value and standard deviation

$$\mu_{X_T^{\max}} = u\Gamma\left(1 - \frac{1}{k}\right)$$

$$\sigma_{X_T^{\max}}^2 = u^2\left[\Gamma\left(1 - \frac{2}{k}\right) - \Gamma^2\left(1 - \frac{1}{k}\right)\right]$$

Extreme Value Distributions

Extreme Type III – Weibull Min

When a probability density function is downwards limited at ε and the lower tail falls off towards ε in the form

$$F(x) = c(x-\varepsilon)^k$$

then the minimum in the time interval T is said to be Type III extreme distributed

$$F_{X,T}^{\min}(x) = 1 - \exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right)$$

$$f_{X,T}^{\min}(x) = \frac{k}{u-\varepsilon}\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{k-1}\exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k\right)$$

Mean value and standard deviation

$$\mu_{X_T^{\min}} = \varepsilon + (u-\varepsilon)\Gamma(1+\frac{1}{k})$$

$$\sigma_{X_T^{\min}}^2 = (u-\varepsilon)^2\left[\Gamma(1+\frac{2}{k}) - \Gamma^2(1+\frac{1}{k})\right]$$

Return Period

The **return period** for extreme events T_R may be defined as:

$$T_R = n \cdot T = \frac{1}{(1 - F_{X,T}^{\max}(x))}$$

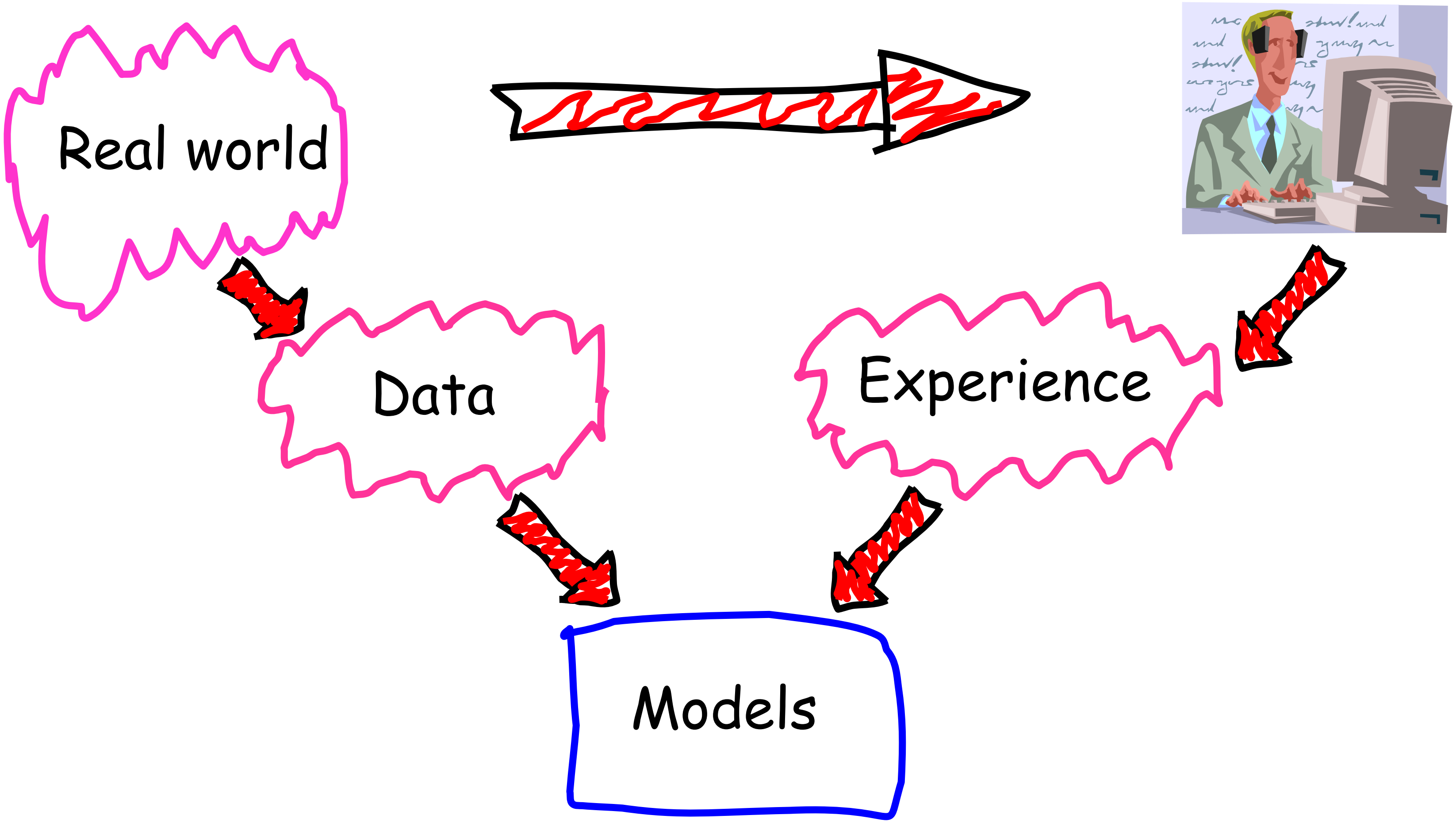
Example:

Let us assume that - according to the cumulative distribution function of the annual maximum traffic load - the annual probability that a truck load larger than 100 ton is equal to 0.02 – then the return period of such heavy truck events is:

$$T_R = n \cdot T = \frac{1}{0.02} \Rightarrow n = \frac{1}{1 \cdot 0.02} = 50 \text{ years}$$

Overview of Estimation and Model Building

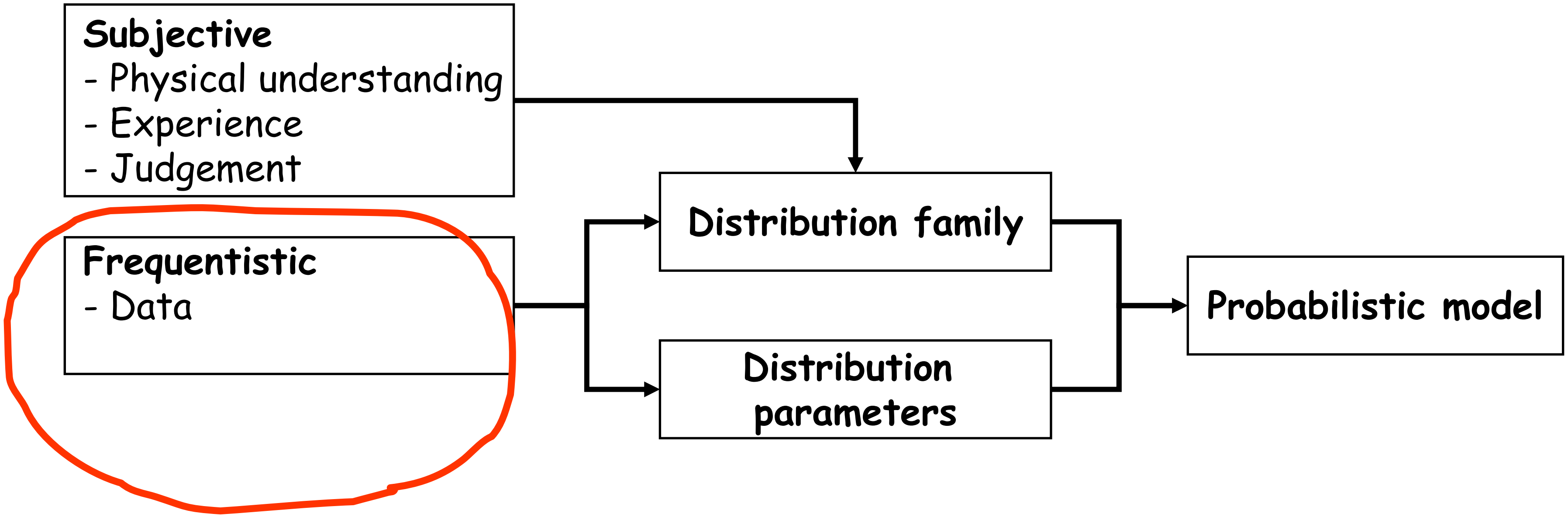
- **How do engineers establish knowledge**



Overview of Estimation and Model Building

Different types of information is used when developing engineering models

- subjective information
- frequentististic information



Overview of Estimation and Model Building

Model building may be seen to consist of five steps

1) **Assessment and statistical quantification of the available data**

2) **Selection of distribution function**

3) **Estimation of distribution parameters**

4) **Model verification**

5) **Model updating**

Probability Distribution Functions in Statistics

In the classical statistical theory a number of probability distribution functions which may all be derived from the **normal distribution function** are repeatedly used for assessment and testing purposes.

These **derived probability distribution** functions are the :

- Chi-square distribution
- Chi-distribution
- t-distribution
- F-distribution

Probability Distribution Functions in Statistics

The Chi-square (χ^2) distribution

When $X_i, i = 1,2,...n$

are standard Normal distributed and independent random variables then the sum of the squares of the random variables i.e.

$$Y_n = \sum_{i=1}^{n} X_i^2$$

is said to be **Chi-square distributed**

It is seen that the Chi square distribution is regenerative i.e. sums of Chi-square distributed random variables are also Chi-square distributed

Probability Distribution Functions in Statistics

The Chi-square (χ^2) distribution

Consider the simplest case with n=1, i.e. : $Y_1 = X^2$

Then we can write

$$F_{Y_1}(y) = P(Y_1 \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le +\sqrt{y})$$

$$= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = F_X(\sqrt{y}) - (1 - F_X(\sqrt{y})) =$$

$$= 2F_X(\sqrt{y}) - 1$$

and we get

$$f_{Y_1}(y) = \frac{dF_{Y_1}(y)}{dy} = \frac{d(2F_X(\sqrt{y}) - 1)}{dy} = y^{-\frac{1}{2}} f_X(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}} \exp(-\frac{1}{2} y)$$

Probability Distribution Functions in Statistics

The Chi-square probability density function is given as

$$f_{Y_n}(y_n) = \frac{y_n^{(n/2-1)}}{2^{n/2}\Gamma(n/2)} \exp(-y_n/2), \qquad y_n \geq 0$$

The mean value is $\mu_{Y_n} = n$ ← **Degrees of freedom**

The variance $\sigma_{Y_n}^2 = 2n$

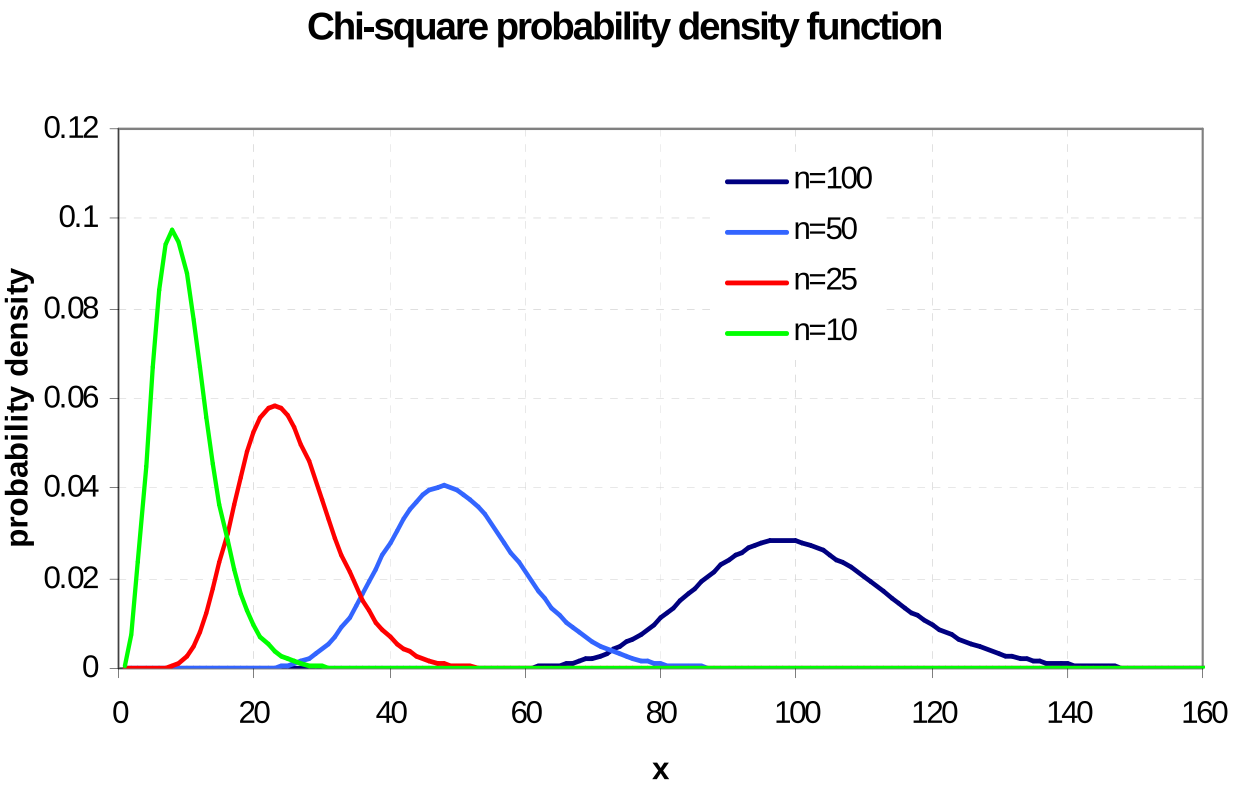
$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

is the complete Gamma function

for large *n* the Chi-square distribution converges to a Normal distribution – Central Limit Theorem

Probability Distribution Functions in Statistics

The Chi-square probability density function



Probability Distribution Functions in Statistics

The Chi (χ) distribution

When a random variable Z is given as the square root of a Chi-square distributed random variable Y_n i.e.

$$Z = \sqrt{Y_n}$$

it is said to be Chi-distributed witn *n* degrees of freedom

Probability Distribution Functions in Statistics

The Chi (χ) distribution

Assume that Y_n is Chi-square distributed with *n* degrees of freedom

If $Z = \sqrt{Y_n}$ **then we can write**

$$F_Z(z) = P(Z \le z) = P(\sqrt{Y_n} \le z) = P(Y_n \le z^2) = F_{Y_n}(z^2)$$

and we get

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{dF_{Y_n}(z^2)}{dz} = 2z f_{Y_n}(z^2) = \frac{z^{n-1}}{2^{n/2-1}\Gamma(n/2)} \exp(-\frac{1}{2}z^2)$$

Probability Distribution Functions in Statistics

The Chi probability density function is given as

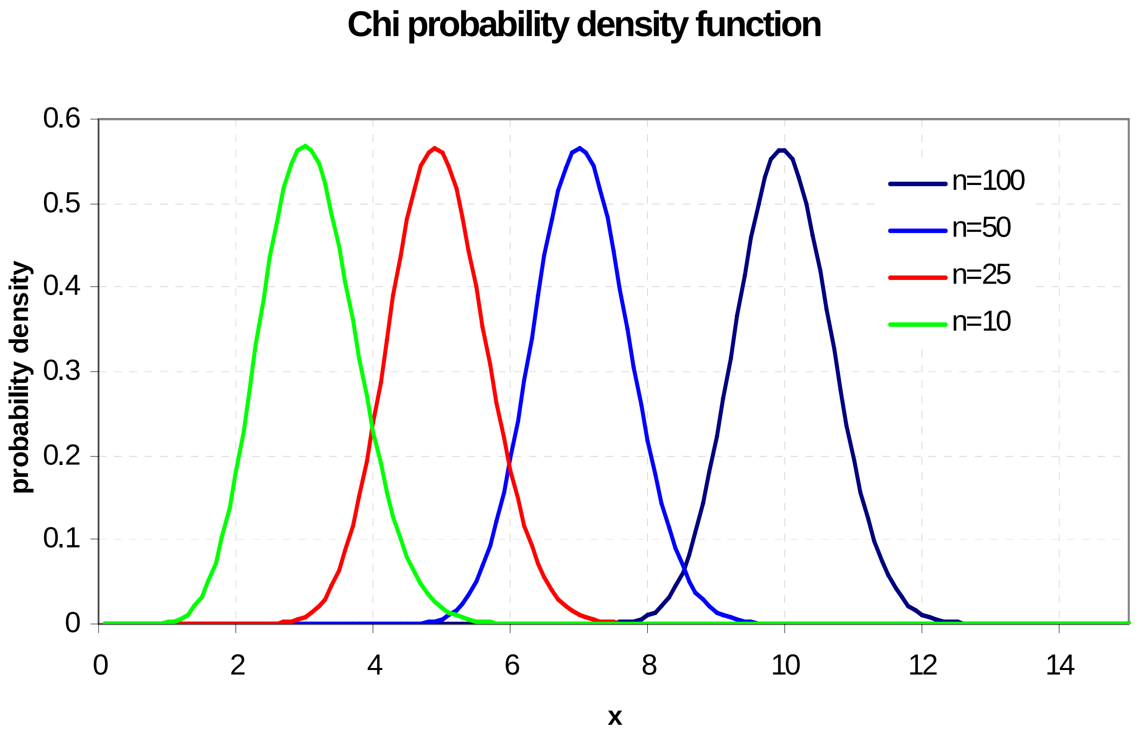
$$f_Z(z) = \frac{z^{(n-1)}}{2^{n/2-1}\Gamma(n/2)} \exp(-z^2/2), \qquad z \geq 0$$

The mean value is $$\mu_z = \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}$$

The variance $$\sigma_z^2 = n - 2\frac{\Gamma^2((n+1)/2)}{\Gamma^2(n/2)}$$

Probability Distribution Functions in Statistics

The Chi probability density function



Probability Distribution Functions in Statistics

The (Student's) *t* distribution

When a random variable S is given as standard Normal distributed, devided by a Chi distributed random variable i.e.

$$S = \frac{X}{\frac{\sqrt{\sum_{i=1}^{n} X_i^2}}{n}} = \frac{X}{\frac{\sqrt{Y_n}}{n}} = \frac{X}{\frac{Z}{n}} = \frac{nX}{Z}$$

it is said to be *t*-distributed witn *n* degrees of freedom

For large *n* the *t*-distribution converges to a Normal distribution.

Probability Distribution Functions in Statistics

The **(Student's) *t* probability density function** is given as

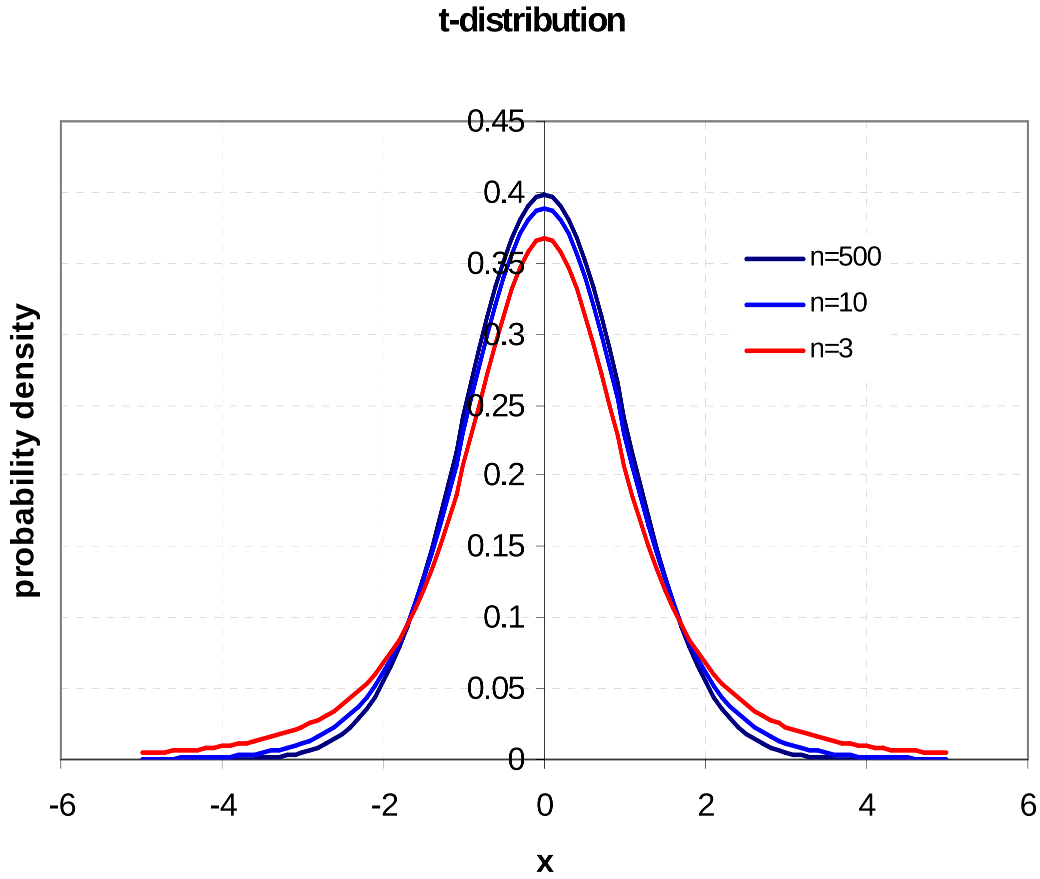
$$f_S(s) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\ \Gamma(n/2)} \left(1 + \frac{s^2}{n}\right)^{-(n+1)/2}, \qquad -\infty \leq s \leq \infty$$

The mean value is zero

The variance $\sigma_S^2 = \frac{n}{n-2}$

Probability Distribution Functions in Statistics

The (Student's) *t* probability density function



Probability Distribution Functions in Statistics

The *F* distribution

When a random variable Q is given as the ratio between two Chi-square distributed random variables i.e.

$$Q = \frac{Y_{n_1}}{Y_{n_2}}$$

it is said to be *F*-distributed witn parameters n_1, n_2

Probability Distribution Functions in Statistics

The *F* probability density function is given as

$$f_Q(q) = \frac{\Gamma((n_1+n_2)/2)\, q^{(n_1-2)/2}(1+q)^{-(n_1+n_2)/2}}{\Gamma(n_1/2)\Gamma(n_2/2)}, \qquad q \geq 0$$

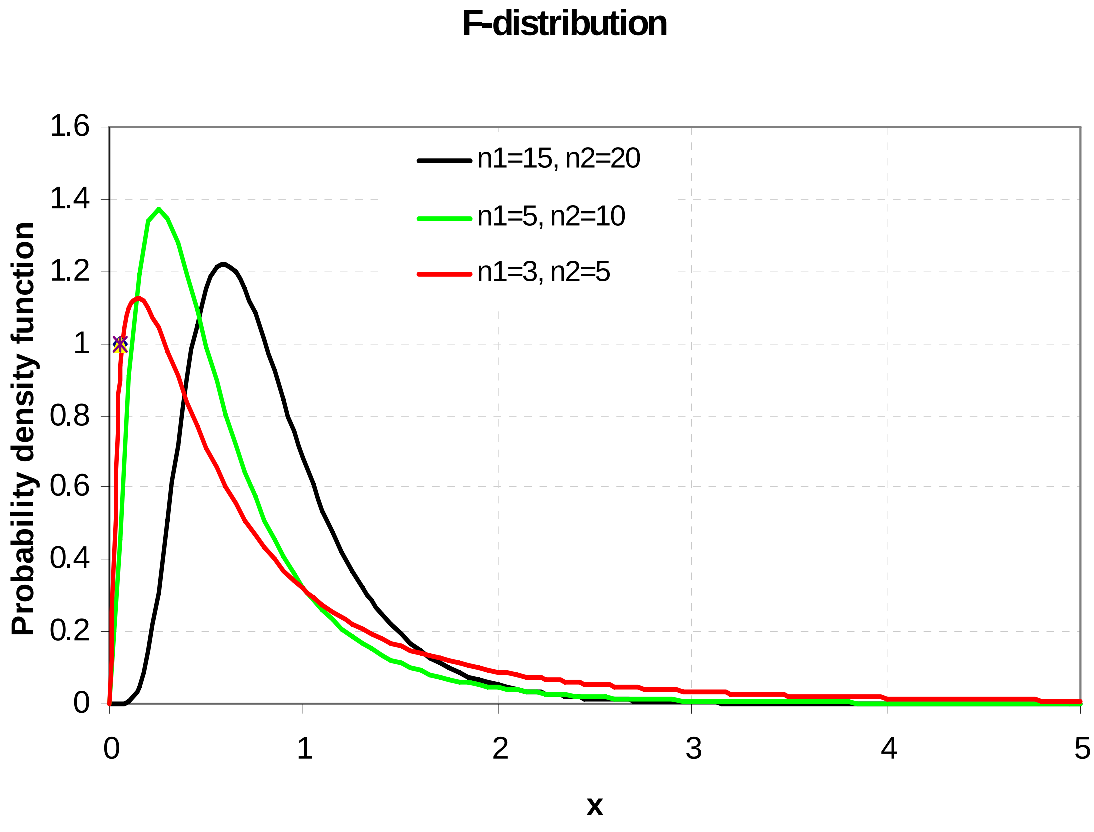
The mean value is $\mu_Q = \frac{n_2}{n_2-2}, \qquad n_2 > 2$

The variance

$$\sigma_Q^2 = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \qquad n_2 > 4$$

Probability Distribution Functions in Statistics

The *F* probability density function



Probability Distribution Functions in Statistics

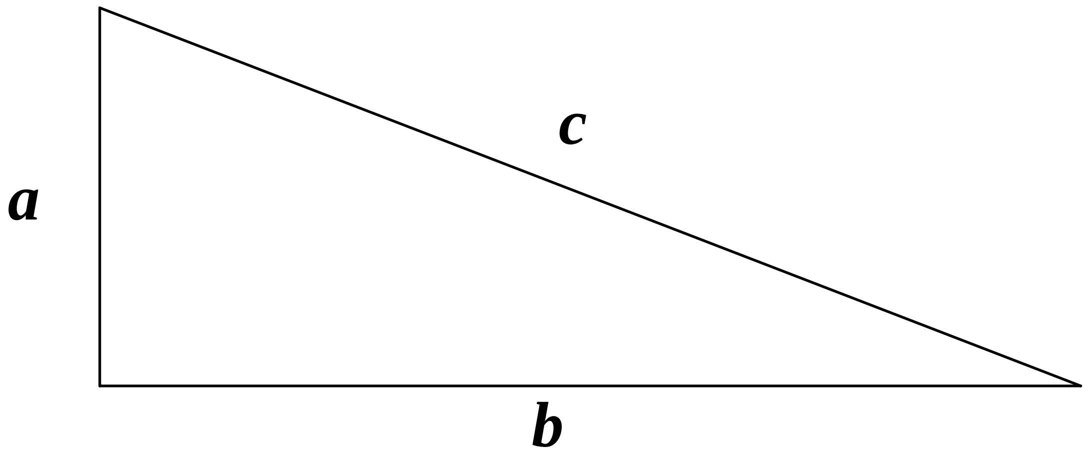
Summary of derived probability density functions:

Distribution Type	When
Chi-square distribution	sum of squared N(0;1)
Chi-distribution	square root of Chi-square
t-distribution	ratio of N(0;1) to Chi/n
F-distribution	ratio of two Chi-square

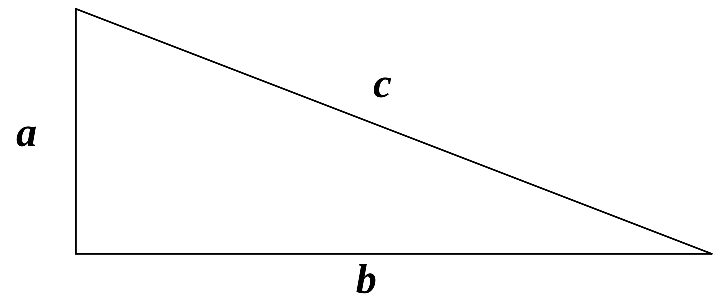
Probability Distribution Functions in Statistics

Example Chi distribution

In the field measurements have been performed of *a* and *b* with the purpose to assess *c*



Probability Distribution Functions in Statistics

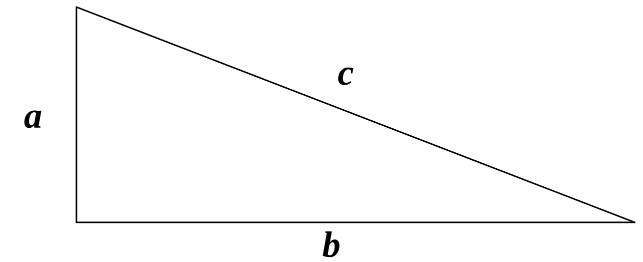


Example Chi distribution

It is assumed that the measurements of *a* and *b* are performed with the same absolute error e which is assumed to N(0; σ_e) i.e. Normal distributed, unbiased and with standard deviation σ_e.

Determine the statistical characteristics of the error in *c* when this is assessed using the measurements of *a* and *b*.

Probability Distribution Functions in Statistics



Example Chi distribution

Knowing that the error propagates according to

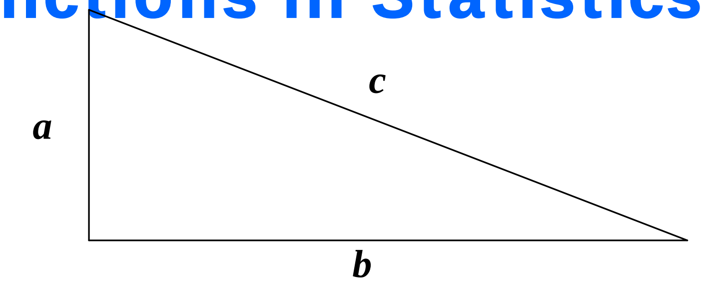
$$\varepsilon_c = \sqrt{\varepsilon_a^2 + \varepsilon_b^2}$$

we realize that

$$\frac{\varepsilon_c}{\sigma_\varepsilon} = \sqrt{\left(\frac{\varepsilon_a}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_b}{\sigma_\varepsilon}\right)^2}$$

is Chi distributed with 2 degrees of freedom

Probability Distribution Functions in Statistics



Example Chi distribution

The probability density function of $Z = \dfrac{\varepsilon_c}{\sigma_\varepsilon}$

can thus be determined from

$$f_Z(z) = z\exp(-0.5z^2), \qquad z \geq 0$$

yielding $f_{\varepsilon_c}(\varepsilon_c) = \dfrac{\varepsilon_c}{\sigma_\varepsilon}\exp(-0.5\varepsilon_c^2 / \sigma_\varepsilon^2), \qquad \varepsilon_c \geq 0$

where it was used that for $y = g(x)$ we have $f_y(y) = \left|\dfrac{dg^{-1}}{dy}\right| f_X(g^{-1}(y))$

Estimators for Sample Descriptors

The first step when new data are achieved is to assess the data

Data/observations

n	x_n	$F_X(x_n)$
1	24.4	0.047619048
2	27.6	0.095238095
3	27.8	0.142857143
4	27.9	0.19047619
5	28.5	0.238095238
6	30.1	0.285714286
7	30.3	0.333333333
8	31.7	0.380952381
9	32.2	0.428571429
10	32.8	0.476190476
11	33.3	0.523809524
12	33.5	0.571428571
13	34.1	0.619047619
14	34.6	0.666666667
15	35.8	0.714285714
16	35.9	0.761904762
17	36.8	0.80952381
18	37.1	0.857142857
19	39.2	0.904761905
20	39.7	0.952380952

Mean value

Variance

Median

⋮

etc

Any function of samples:

Sample characteristics

or

Sample statistics

Estimators for Sample Descriptors

We want to have a look at the statistical characteristics of such sample statistics – in order to better understand the information they contain

Assume we have a yet unknown sample of experiment outcomes

$$X_i, \; i = 1,2,..n$$

generated by the cumulative distribution functions

$$F_{X_i}(x_i, \mathbf{p}) = F_X(x, \mathbf{p}), i = 1,2,..n$$

then we can write the sample statistics for the

sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

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

Estimators for Sample Descriptors

The sample statistics are random variables because the experiment outcomes have not yet been realized – however we can evaluate the expected value and the variance of the sample statistics, i.e. for the sample mean we get :

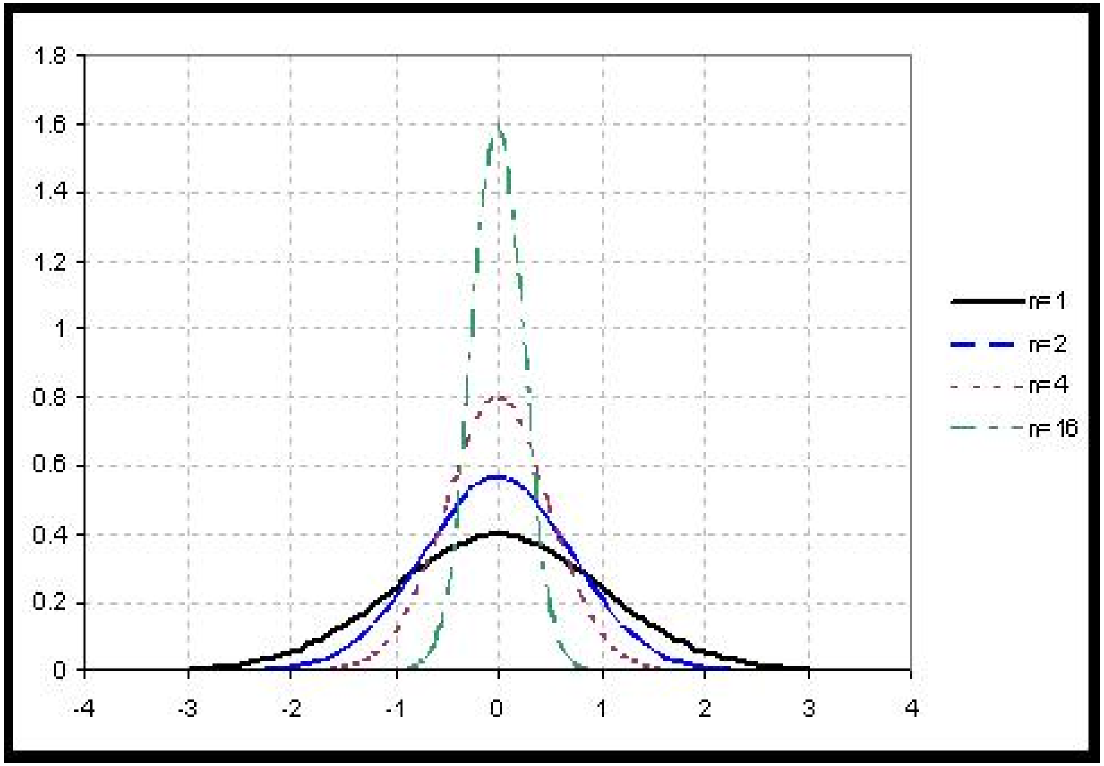
$$E\left[\overline{X}\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}E\left[X_i\right]=\frac{1}{n}n\cdot\mu_X=\mu_X$$

$$Var\left[\overline{X}\right]=Var\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=\frac{1}{n^2}Var\left[\sum_{i=1}^{n}X_i\right]=\frac{1}{n^2}\sum_{i=1}^{n}Var\left[X_i\right]$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}E\left[(X_i-\mu_X)^2\right]=\frac{1}{n}\sigma_X^2$$

Estimators for Sample Descriptors

The probability density function for the sample average can be assumed to be a Normal distribution – Central Limit Theorem



Estimators for Sample Descriptors

For the sample variance we get:

$$E\left[S^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}((X_i - \mu_X) - (\bar{X} - \mu_X))^2\right]$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}E\left[(X_i - \mu_X)^2\right] - nE\left[(\bar{X} - \mu_X)^2\right]\right)$$

$$= \frac{1}{n}\left(nE\left[(X_i - \mu_X)^2\right] - nE\left[(\bar{X} - \mu_X)^2\right]\right) =$$

$$= \frac{1}{n}\left(n\sigma_X^2 - n\frac{\sigma_X^2}{n}\right) = \sigma_X^2 - \frac{1}{n}\sigma_X^2 = \frac{(n-1)}{n}\sigma_X^2$$

The expected value of the sample variance is thus different from the variance – biased !

Estimators for Sample Descriptors

We can however easily identify an unbiased estimator for the variance as:

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

Not *n* as in the sample variance !

Estimators for Sample Descriptors

The goodness of an estimator cannot be judged upon whether it is biased or not alone – other properties are important such as

- efficiency	**least mean square error $E[(s^2-s^2)]$**
- invariance	$h(\bar{\theta}) = \overline{h(\theta)}$
- consistent	**converge to the true values**
- sufficiency	**make maximum use of the data**
- robustness	**sensitivity to omission of individual data**

we will not consider these in detail – just remember that these considerations may also be important

Confidence Intervals on Estimators

In the previous we have seen that estimators of e.g. the mean value are associated with uncertainty and we have established expressions to determine their mean value and variance –

Based on this information we are also able to determine so called **confidence intervals** on the estimators.

For the case where it is assumed that the **variance is known** and only the **mean value is uncertain** the so-called <u>double sided and symmetrical confidence interval on the mean value</u> is given by

$$P\left[-k_{\alpha/2} < \frac{\bar{X}-\mu_X}{\sigma_X \frac{1}{\sqrt{n}}} < k_{\alpha/2}\right] = P\left[-k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}} < \bar{X}-\mu_X < k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}}\right] = 1-\alpha$$

Confidence Intervals on Estimators

In words the confidence interval defines an interval within which e.g. the true mean value will lie with a probability 1-α

$$P\left[-k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}} < \bar{X} - \mu_X < k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}}\right] = 1-\alpha$$

For the case where α= 0.05, *n* = 16 and σ_X = 20 we get

$$k_{\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}\left(1-\frac{0.05}{2}\right) = 1.96$$

$$P\left[-9.8 < \bar{X} - \mu_X < 9.8\right] = 0.95$$

Confidence Intervals on Estimators

If we then observe that the sample mean is equal to e.g. 400 we know that with a probability equal to 0.95 the true mean will lie within the interval

$$P\left[-9.8 < \bar{X} - \mu_X < 9.8\right] = 0.95$$

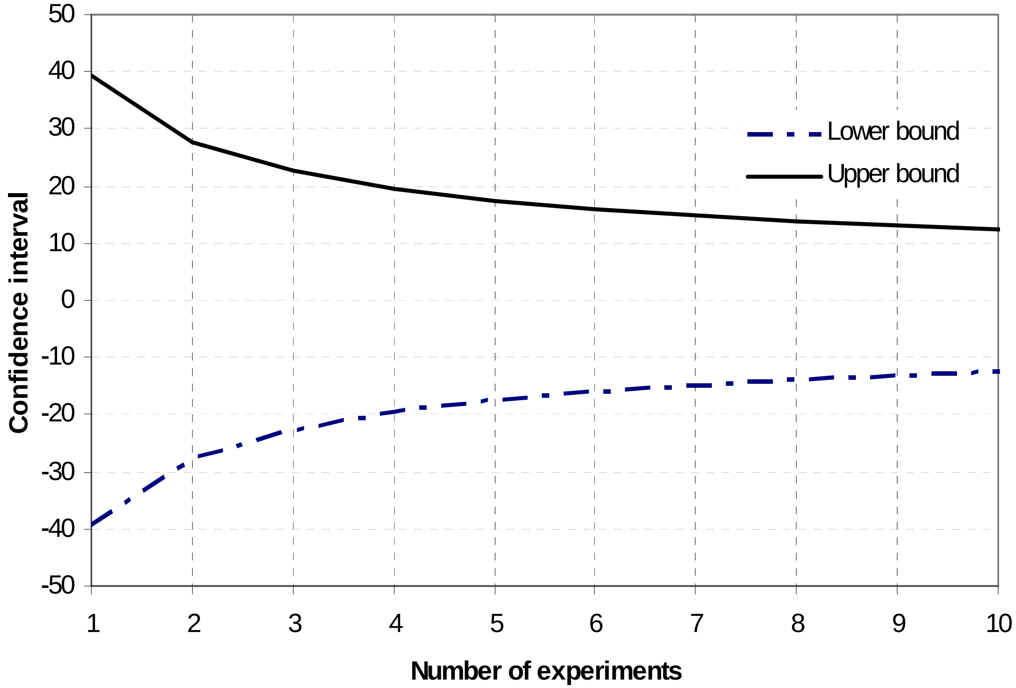
and so: $390.2 < \mu_X < 409.8$

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.

Confidence Intervals on Estimators

The number of available data has a significant importance for the confidence interval - using the same example as in the previous the confidence interval depends on *n* as shown below



Statistics and Probability Theory in Civil, Surveying and Environmental Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

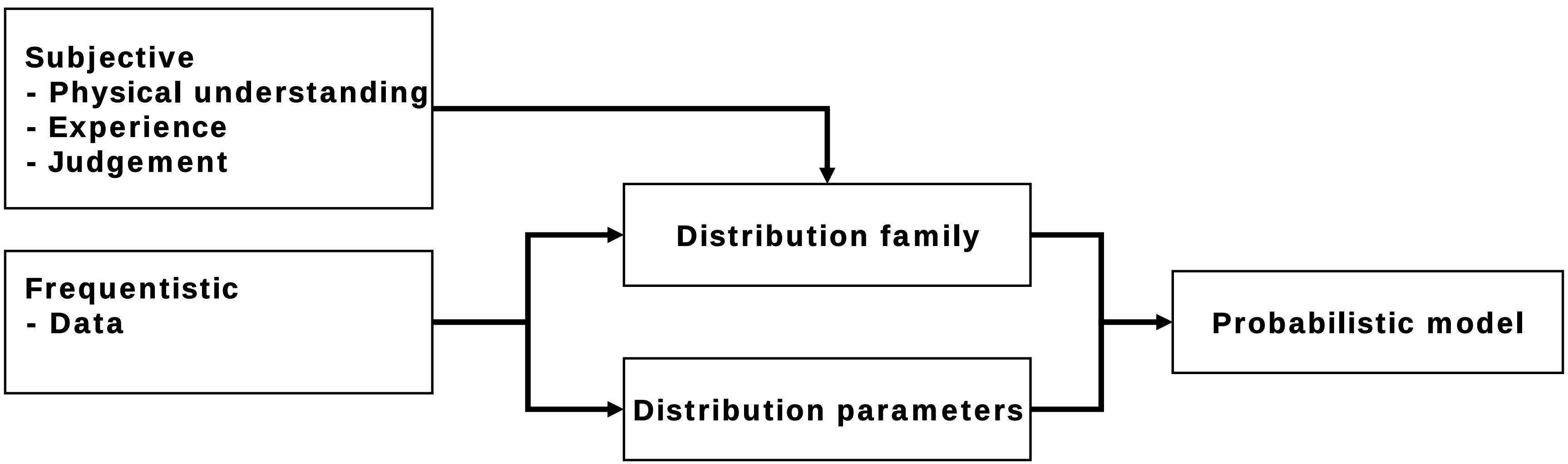
Contents of Todays Lecture

- Overview of Estimation and Model Building
- A short Summary of the Previous Lecture
- Estimators for Sample Descriptors
- Testing for Statistical Significance
  - The hypothesis testing procedure
  - Testing of the mean with known variance
  - Testing of the mean with unknown variance
  - Testing of the variance
  - Test of two or more data sets

Overview of Estimation and Model Building

Different types of information is used when developing engineering models

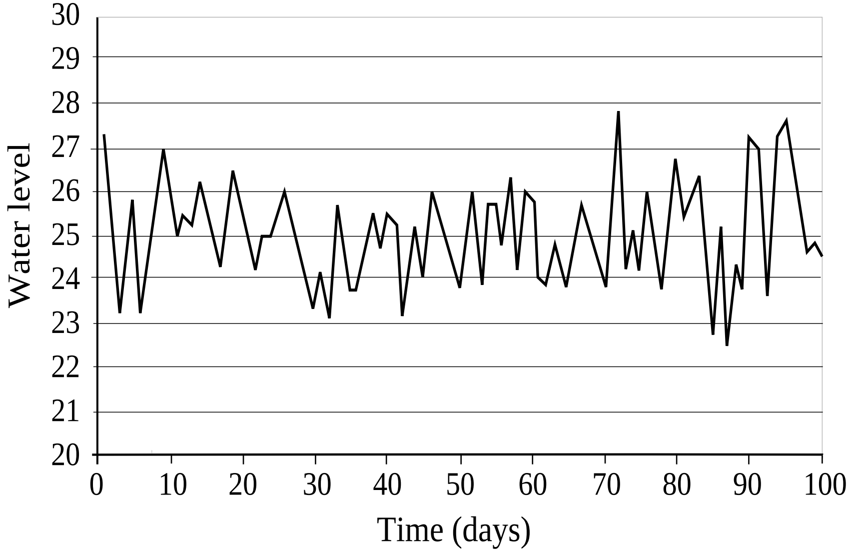
- subjective information
- frequentistic information



A Short Summary of the Previous Lecture

- **Continuous random processes**

  A continuous random process is a random process which has realizations continuously over time and for which the realizations belong to a continuous sample space.



Variations of;
water levels
wind speed
rain fall
.
.
.

Realization of continuous scalar valued random process

A Short Summary of the Previous Lecture

If the extremes within the period *T* of an ergodic random process *X*(*t*) are independent and follow the distribution:

$$F_{X,T}^{\max}(x) = P(\max_{T} X \le x)$$

then the extremes of the same process within the period:

$n \cdot T$ **will follow the distribution:**

$$F_{X,nT}^{\max}(x) = P\left(\left\{\max_{T_1} X \le x\right\} \bigcap \left\{\max_{T_2} X \le x\right\} \ldots \bigcap \left\{\max_{T_n} X \le x\right\}\right)$$

$$= P\left(\bigcap_{i=1}^{n} \left\{\max_{T_i} X \le x\right\}\right)$$

$$= \prod_{i=1}^{n} P\left(\max_{T_i} X \le x\right)$$

$$= \left(F_{X,T}^{\max}(x)\right)^{n}$$

A Short Summary of the Previous Lecture

If the extremes within the period *T* of an ergodic random process *X*(*t*) are independent and follow the distribution:

$$F_{X,T}^{\max}(x) = P(\max_{T} X \leq x)$$

then the extremes of the same process within the period:

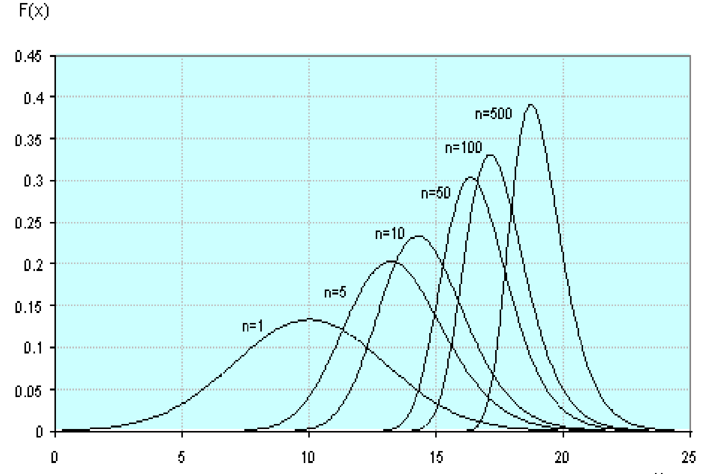
$n \cdot T$ **will follow the distribution:**

$$F_{X,nT}^{\max}(x) = P\left(\left\{\max_{T_1} X \leq x\right\} \bigcap \left\{\max_{T_2} X \leq x\right\} \ldots \bigcap \left\{\max_{T_n} X \leq x\right\}\right)$$

$$= P\left(\bigcap_{i=1}^{n}\left\{\max_{T_i} X \leq x\right\}\right)$$

$$= \prod_{i=1}^{n} P\left(\max_{T_i} X \leq x\right)$$

$$= \left(F_{X,T}^{\max}(x)\right)^{n}$$



A Short Summary of the Previous Lecture

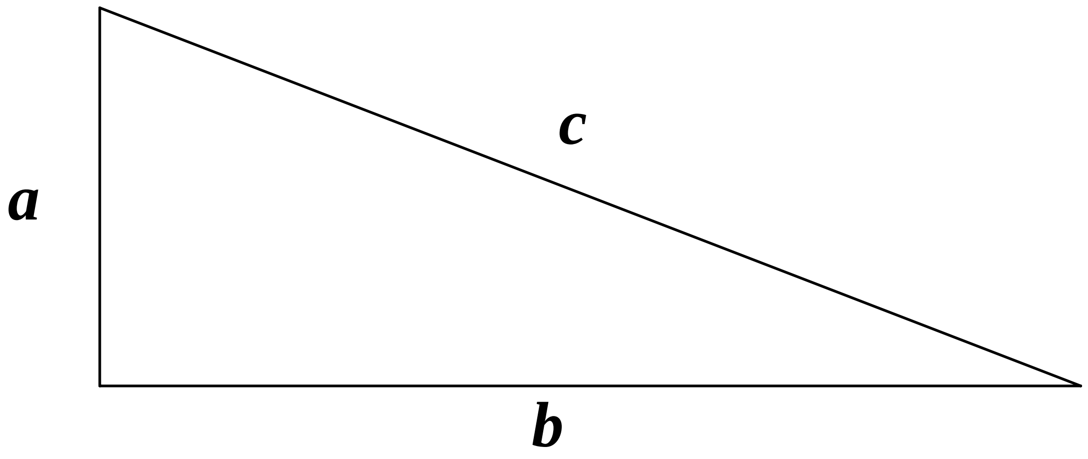
Based on independent Normal distributed random variables we could derive the following distributions:

Distribution Type	When
Chi-square distribution	sum of squared N(0;1)
Chi-distribution	square root of Chi-square
t-distribution	ratio of N(0;1) to Chi/*n*
F-distribution	ratio of two Chi-square

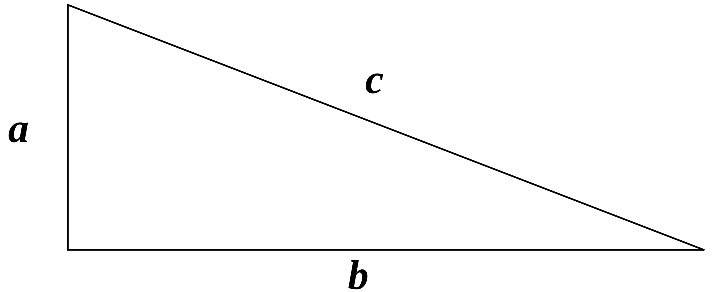
Probability Distribution Functions in Statistics

Example Chi distribution

In the field, measurements have been performed of *a* and *b* with the purpose to assess *c*



Probability Distribution Functions in Statistics

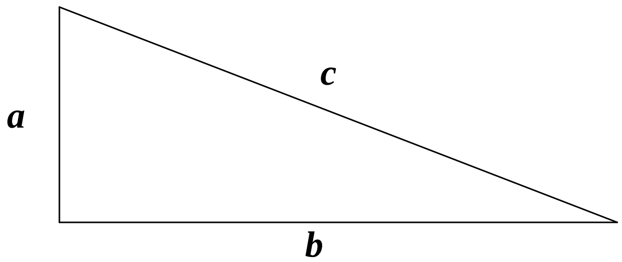


Example Chi distribution

It is assumed that the measurements of *a* and *b* are performed with the same absolute error ε which is assumed to N(0; σ_ε) i.e. Normal distributed, unbiased and with standard deviation σ_ε.

Determine the statistical characteristics of the error in *c* when this is assessed using the measurements of *a* and *b*.

Probability Distribution Functions in Statistics



Example Chi distribution

Knowing that the error propagates according to

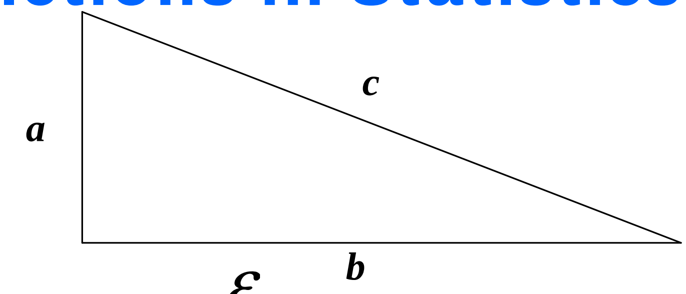
$$\varepsilon_c = \sqrt{\varepsilon_a^2 + \varepsilon_b^2}$$

we realize that

$$\frac{\varepsilon_c}{\sigma_\varepsilon} = \sqrt{\left(\frac{\varepsilon_a}{\sigma_\varepsilon}\right)^2 + \left(\frac{\varepsilon_b}{\sigma_\varepsilon}\right)^2}$$

is Chi distributed with 2 degrees of freedom

Probability Distribution Functions in Statistics



Example Chi distribution

The probability density function of $Z = \frac{\varepsilon_c}{\sigma_\varepsilon}$

can thus be determined from

$$f_Z(z) = z\exp(-0.5z^2), \qquad z \geq 0$$

yielding $$f_{\varepsilon_c}(\varepsilon_c) = \frac{\varepsilon_c}{\sigma_\varepsilon}\exp\left(-0.5\varepsilon_c^2 / \sigma_\varepsilon^2\right), \qquad \varepsilon_c \geq 0$$

Estimators for Sample Descriptors

The first step when new data are achieved is to assess the data

Data/observations

n	x_n	$F_X(x_n)$
1	24.4	0.047619048
2	27.6	0.095238095
3	27.8	0.142857143
4	27.9	0.19047619
5	28.5	0.238095238
6	30.1	0.285714286
7	30.3	0.333333333
8	31.7	0.380952381
9	32.2	0.428571429
10	32.8	0.476190476
11	33.3	0.523809524
12	33.5	0.571428571
13	34.1	0.619047619
14	34.6	0.666666667
15	35.8	0.714285714
16	35.9	0.761904762
17	36.8	0.80952381
18	37.1	0.857142857
19	39.2	0.904761905
20	39.7	0.952380952

Mean value

Variance

Median

⋮

etc

Any function of samples:

Sample characteristics

or

Sample statistics

Estimators for Sample Descriptors

We want to have a look at the statistical characteristics of such sample statistics – in order to better understand the information they contain

Assume we have a yet unknown sample of experiment

outcomes $X_i, \; i = 1,2,..n$

generated by the cumulative distribution functions

$$F_{X_i}(x_i, \mathbf{p}) = F_X(x, \mathbf{p}), i = 1,2,..n$$

then we can write the sample statistics for the

sample mean

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

sample variance

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

Estimators for Sample Descriptors

The sample statistics are random variables,

because the experiment outcomes have not yet been realized –

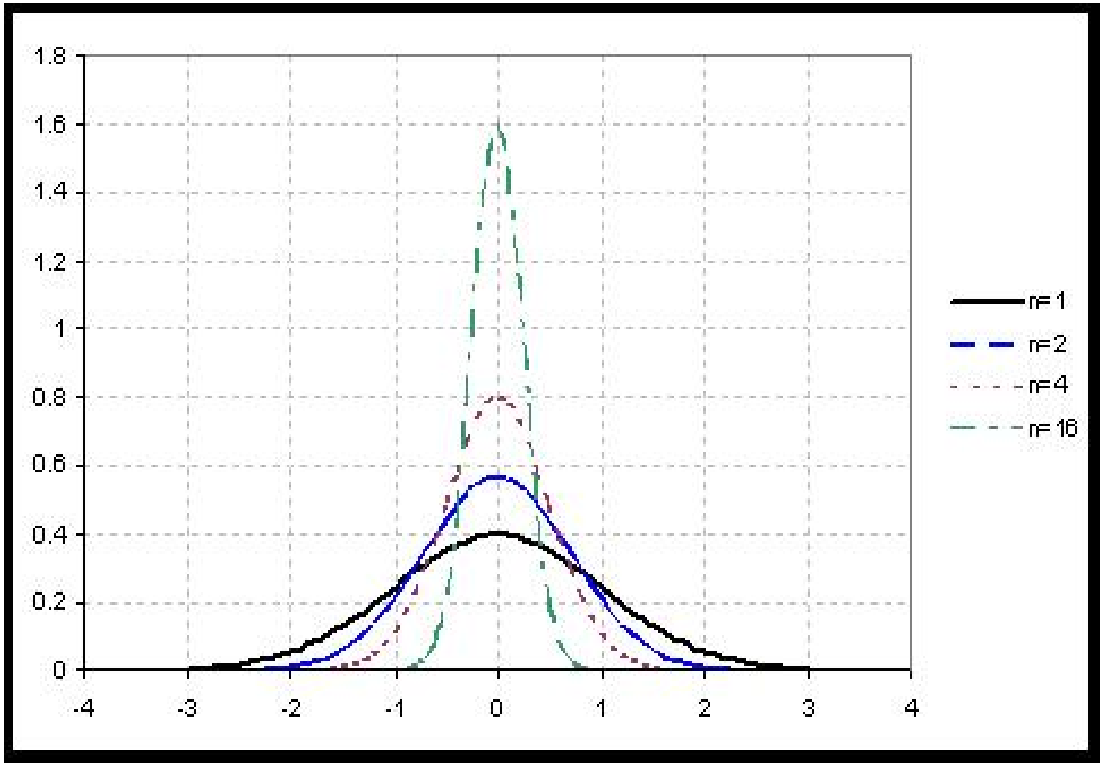
however we can evaluate the expected value and the variance of the sample statistics, i.e. for the sample mean we get :

$$E\left[\overline{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n} n \cdot \mu_X = \mu_X$$

$$Var\left[\overline{X}\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] = \frac{1}{n}\sigma_X^2$$

Estimators for Sample Descriptors

The probability density function for the sample average can be assumed to be a Normal distribution – Central Limit Theorem



Estimators for Sample Descriptors

For the sample variance we get:

$$E\left[S^2\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}((X_i-\mu_X)-(\bar{X}-\mu_X))^2\right]$$

$$=\frac{1}{n}\left(\sum_{i=1}^{n}E\left[(X_i-\mu_X)^2\right] - n\,E\left[(\bar{X}-\mu_X)^2\right]\right)$$

$$=\frac{1}{n}\left(n\cdot E\left[(X_i-\mu_X)^2\right] - n\,E\left[(\bar{X}-\mu_X)^2\right]\right)=$$

$$=\frac{1}{n}\left(n\cdot\sigma_X^2 - n\,\frac{\sigma_X^2}{n}\right)$$

$$=\sigma_X^2-\frac{1}{n}\sigma_X^2=\frac{(n-1)}{n}\sigma_X^2$$

The expected value of the sample variance is thus different from the variance – biased !

Estimators for Sample Descriptors

We can however easily identify an unbiased estimator for the variance as:

$$S^2_{unbiased} = \frac{n}{n-1} S^2$$

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

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

Confidence Intervals on Estimators

- In the previous we have seen that estimators of e.g. the mean value are associated with uncertainty and we have established expressions to determine their mean value and variance.

- Based on this information we are also able to determine so called **confidence intervals** on the estimators.

- Confidence intervals may be understood as intervals within which e.g. the mean value can be found

- Confidence is expressed in terms of probability

Confidence Intervals on Estimators

We may e.g. establish a confidence interval for the mean value.

For the case where it is assumed that the **mean value is uncertain** and the **variance is known** the so-called **double sided and symmetrical confidence interval on the mean value** is given by

Sample average

True mean

$$P\left[-k_{\alpha/2} < \frac{\bar{X}-\mu_X}{\sigma_X \dfrac{1}{\sqrt{n}}} < k_{\alpha/2}\right] = P\left[-k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}} < \bar{X}-\mu_X < k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}}\right] = 1-\alpha$$

Known std. dev.

Sample size

Significance level

Confidence Intervals on Estimators

In words: the confidence interval defines an interval within which the sample average will be located with a probability 1-α

$$P\left[-k_{\alpha/2}\sigma_X\frac{1}{\sqrt{n}}<\bar{X}-\mu_X<k_{\alpha/2}\sigma_X\frac{1}{\sqrt{n}}\right]=1-\alpha$$

Known std. dev. (σ_X), **Sample average** ($\bar{X}$), **True mean** (μ_X), **Sample size** (n)

The confidence interval may be determined using the assumption that the mean value is Normal distributed whereby there is:

$$k_{\alpha/2}=\Phi^{-1}\left(1-\frac{\alpha}{2}\right)=\Phi^{-1}\left(1-\frac{0.05}{2}\right)=1.96$$

Confidence Intervals on Estimators

For the case where α= 0.05, n = 16 and σ_X = 20 we get

$$P\left[-1.96 < \frac{\bar{X} - \mu_X}{20\dfrac{1}{\sqrt{n}}} < 1.96\right] = 1 - 0.05$$

$$P\left[-9.8 < \bar{X} - \mu_X < 9.8\right] = 0.95$$

Confidence Intervals on Estimators

- If we then observe that the sample mean is equal to e.g. 400 we know that with a probability equal to 0.95 the true mean will lie within the interval

$$P\left[-9.8 < \bar{X} - \mu_X < 9.8\right] = 0.95$$

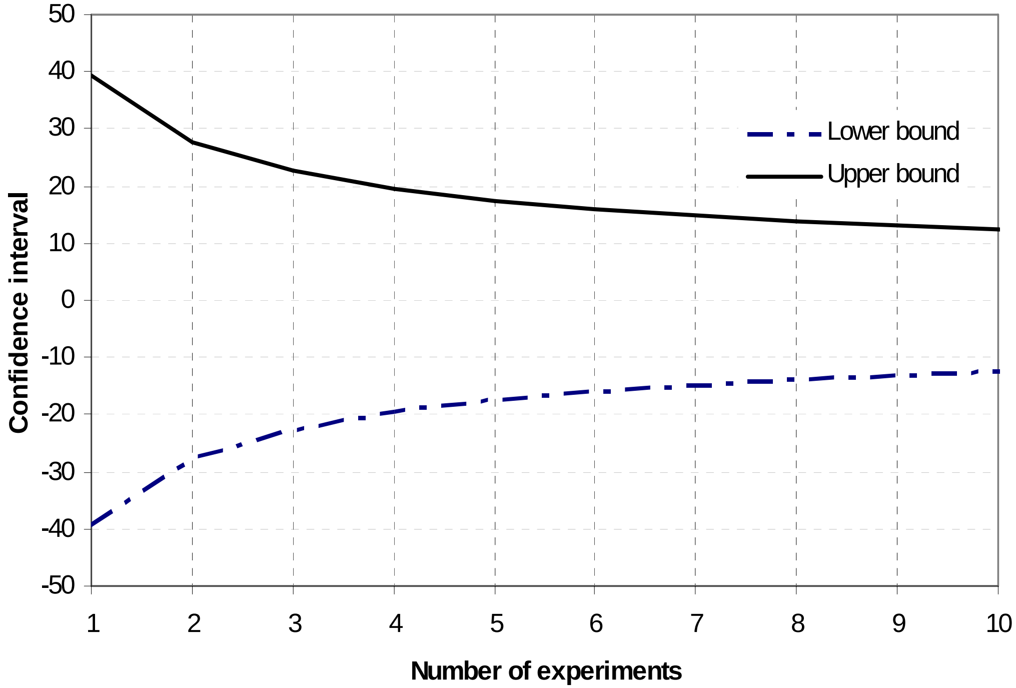
$$\text{and so:} \quad 390.2 < \mu_X < 409.8$$

- 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.

Confidence Intervals on Estimators

The number of available data has a significant importance for the confidence interval - using the same example as in the previous the confidence interval depends on *n* as shown below



Basic Statistics and Probability Theory in Civil, Surveying and Environmental Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Contents of Todays Lecture

- Short Summary of Previous Lecture
- Overview of Estimation and Model Building
- Testing for Statistical Significance
  - The hypothesis testing procedure
  - Testing of the mean with known variance
  - Testing of the mean with unknown variance
  - Testing of the variance
  - Test of two or more data sets
- Selection of Distribution Function
  - Model selection by use of probability paper

Short Summary of Previous Lecture

In the previous lecture we looked at:

Estimators for Sample Descriptors

Confidence Intervals on Estimators

Data/observations

n	x_n	$F_X(x_n)$
1	24.4	0.047619048
2	27.6	0.095238095
3	27.8	0.142857143
4	27.9	0.19047619
5	28.5	0.238095238
6	30.1	0.285714286
7	30.3	0.333333333
8	31.7	0.380952381
9	32.2	0.428571429
10	32.8	0.476190476
11	33.3	0.523809524
12	33.5	0.571428571
13	34.1	0.619047619
14	34.6	0.666666667
15	35.8	0.714285714
16	35.9	0.761904762
17	36.8	0.80952381
18	37.1	0.857142857
19	39.2	0.904761905
20	39.7	0.952380952

Mean value

Variance

Median

⋮

etc

Any function of samples:

Sample characteristics

or

Sample statistics

Short Summary of Previous Lecture

Sample descriptors are simply e.g.

The sample mean value

The sample variance

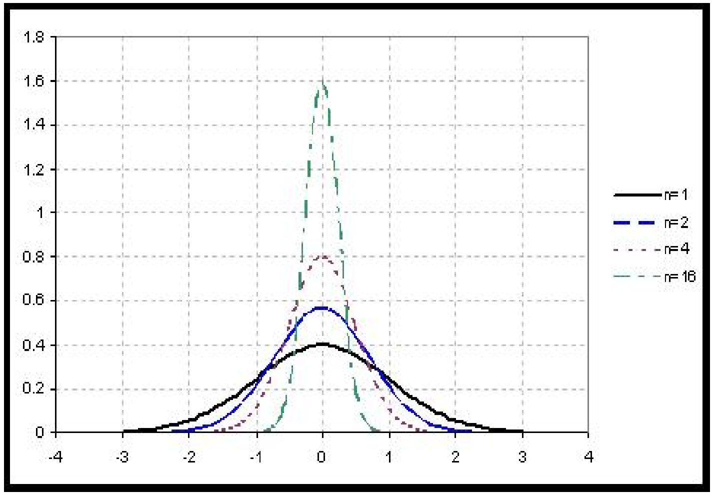
What did we learn?

The sample descriptors are associated with uncertainty due to statistical uncertainty (epistemical uncertainty)

Short Summary of Previous Lecture

The sample mean value is an unbiased descriptor

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n \cdot \mu_X = \mu_X$$



$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n}\sigma_X^2$$

Short Summary of Previous Lecture

The sample variance is biased !

$$E\left[S^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}((X_i - \mu_X) - (\bar{X} - \mu_X))^2\right]$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} E\left[(X_i - \mu_X)^2\right] - nE\left[(\bar{X} - \mu_X)^2\right]\right)$$

$$= \frac{1}{n}\left(nE\left[(X_i - \mu_X)^2\right] - nE\left[(\bar{X} - \mu_X)^2\right]\right) =$$

$$= \frac{1}{n}\left(n\sigma_X^2 - n\frac{\sigma_X^2}{n}\right) = \sigma_X^2 - \frac{1}{n}\sigma_X^2 = \frac{(n-1)}{n}\sigma_X^2$$

$$S^2_{unbiased} = \frac{n}{n-1}S^2$$

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

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

Short Summary of Previous Lecture

- Due to the uncertainty associated with the descriptors (e.g. sample mean) we dont know their exact value

- We can however determine intervals where we can find them with a given probability

These intervals we call confidence intervals!

Sample average

True mean

$$P\left[-k_{\alpha/2} < \frac{\overline{X}-\mu_X}{\sigma_X \frac{1}{\sqrt{n}}} < k_{\alpha/2}\right] = P\left[-k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}} < \overline{X}-\mu_X < k_{\alpha/2}\sigma_X \frac{1}{\sqrt{n}}\right] = 1-\alpha$$

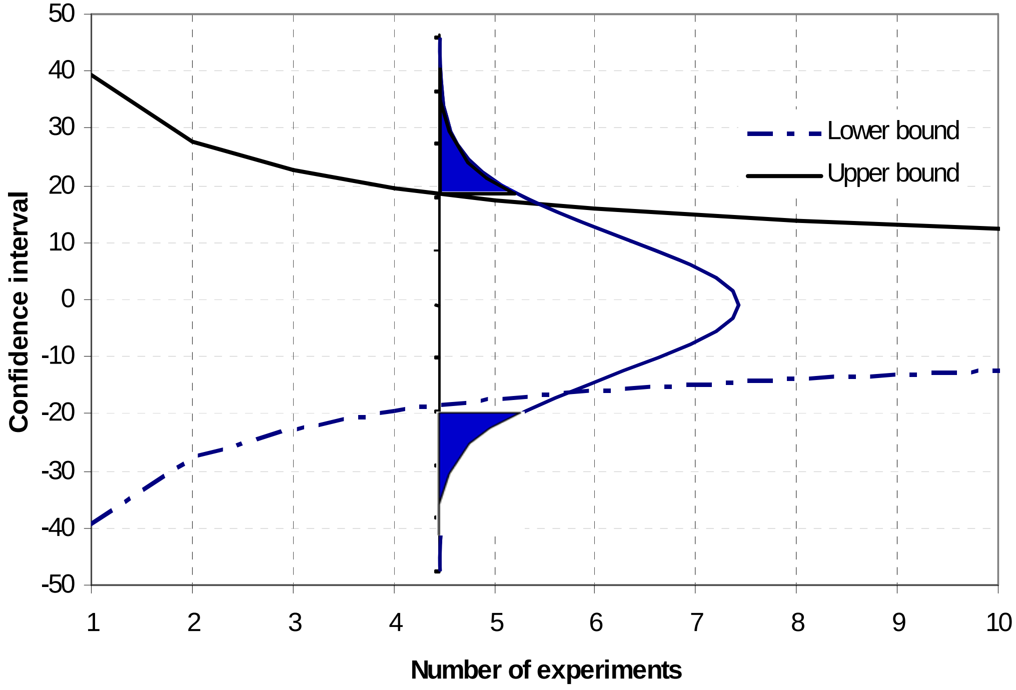
Known std. dev.

Sample size

Significance level

Short Summary of Previous Lecture

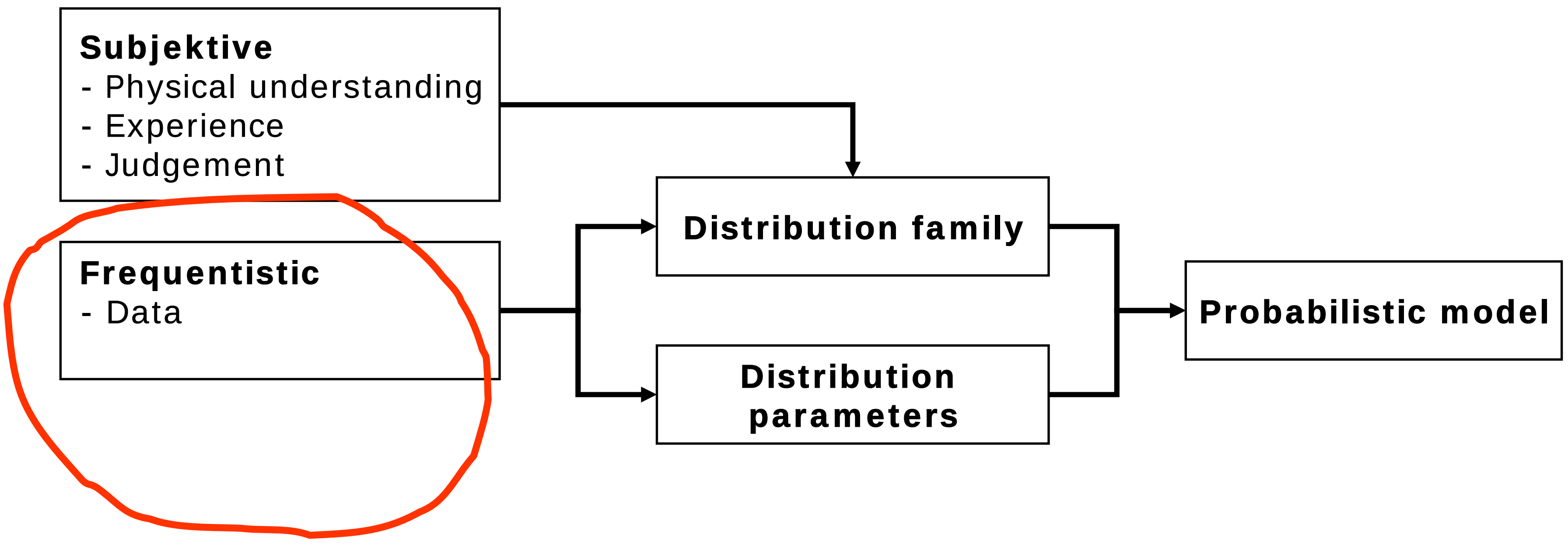
The number of available data has a significant importance for the confidence interval - using the same example as in the previous the confidence interval depends on *n* as shown below



Overview of Estimation and Model Building

Different types of information is used when developing engineering models

- subjektive information
- frequentististic information



Testing for Statistical Significance

Engineering dilemma :

Draw simple conclusions based on limited data with a high degree of variability –

E.g. : Make a few „on site“ tests to verify a calculation model of the soil strength characteristics

Use observations of traffic crossing a bridge to check if design traffic volume assumptions are valid

Collect ground water „samples“ to verify that the water is of drinking quality

Testing for Statistical Significance

It is important that such conclusions are drawn on a basis which is consistent and transparent – i.e. the conclusions should reflect the evidence (data) and a given formalism in regard to what evidence triggers which conclusions

One highly utilized and useful formalism for supporting such conclusions is to

1 Formulate hypothesis

2 Test hypothesis

We shall have a look into this approach is some detail in the following

Testing for Statistical Significance

1 The first step is to formulate a **null-hypothesis - H_0** e.g. postulating that a sample statistic (e.g. sample mean) is equal to a given value

2 The next step is to formulate an **operating rule** on the basis of which the null-hypothesis can either be accepted or rejected – given the evidence (test results) – such an operating rule is often defined by an interval D within which the observed sample statistic has to be in – for the null-hypothesis to be accepted - **rejecting the null-hypothesis H_0 corresponds to accepting the alternate H_1 hypothesis**

3 Select a **significance level a** for conducting the test – where a is the probability that the hypothesis will be rejected even though it is true (**Type I error**) – in this way a also influences the probability that the null-hypothesis is accepted even though it is false (**Type II error**)

Testing for Statistical Significance

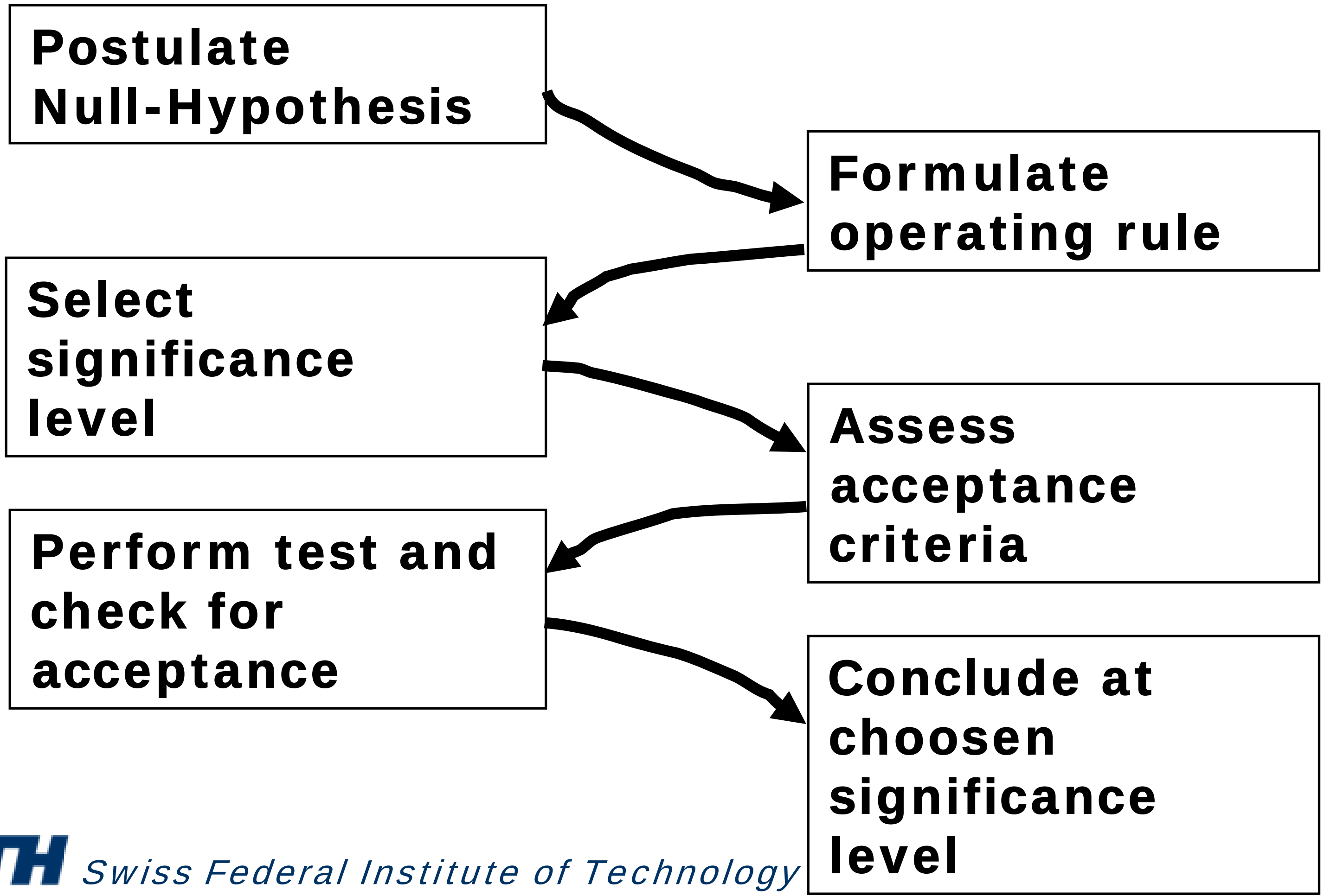
4 Calculate the value of D corresponding to a – calculate also if relevant the probability of performing a Type II error

5 Perform the planned tests and evaluate the observed sample statistic – check if the null-hypothesis should be rejected or accepted

6 Given that the null-hypothesis is not supported by the evidence (data) the null-hypothesis is rejected at significance level a – otherwise it is accepted.

Testing for Statistical Significance

The hypothesis testing procedure may be visualized as follows



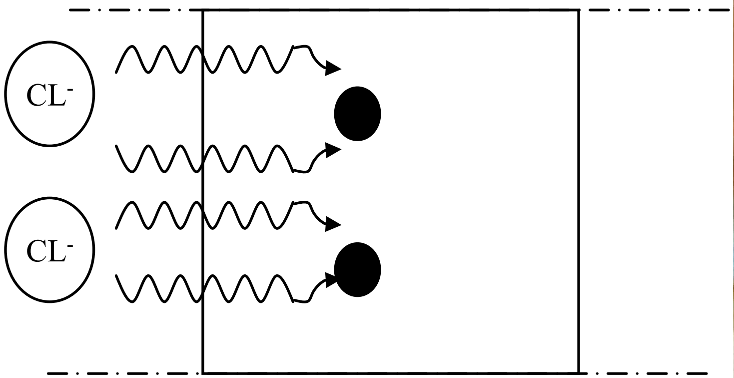
Testing for Statistical Significance

Typical Tests in Engineering

- Testing of the mean – with known variance
- Testing of the mean – with unknown variance
- Testing of the variance
- Test of two or more data sets

Testing for Statistical Significance

Example – chloride induced corrosion of concrete structures



Consider an example where we want to verify whether the chloride concentration on the surface of a concrete structure is in compliance with our design assumptions

Testing for Statistical Significance

Null-hypothesis

Testing of the mean – with known variance

The design assumptions: mean surface chloride concentration is 0.3 %

we assume that we know the std. dev. of the surface chloride concentration – equal to 0.04 %

The operating rule is formulated as:
Accept the Null-hypothesis at the α-level if

$$0.3 - \Delta \leq \bar{X} \leq 0.3 + \Delta$$

Testing for Statistical Significance

Testing of the mean – with known variance

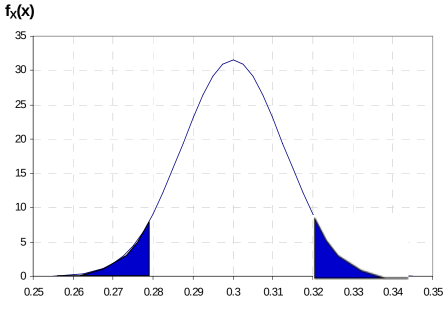
The acceptance criteria may be determined for given α by

$$P\left(0.3-\Delta \leq \overline{X} \leq 0.3+\Delta\right)=1-\alpha$$

Choosing α = 0.1, n= 10 experiments and assuming that the sample average is normal distributed we get

$$\Phi\left(\frac{x_U-\mu}{\sigma}\right)-\Phi\left(\frac{x_L-\mu}{\sigma}\right)=$$

$$\Phi\left(\frac{(0.3+\Delta)-0.3}{\frac{0.04}{\sqrt{10}}}\right)-\Phi\left(\frac{(0.3-\Delta)-0.3}{\frac{0.04}{\sqrt{10}}}\right)=0.9 \qquad \Rightarrow \qquad \Delta=0.0208$$



Testing for Statistical Significance

Testing of the mean – with known variance

If the sample average lies in the interval the Null-hypothesis H_0 should be accepted $[0.28 \leq \bar{x} \leq 0.32]$

Assume that 10 experiments are carried out and the following results are obtained

$$\mathbf{x} = (0.33, 0.32, 0.25, 0.31, 0.28, 0.27, 0.29, 0.3, 0.27, 0.28)^T$$

with sample average μ = 0.29 - it is concluded that the Null-hypothesis **should be accepted** at the 0.1 level.

Testing for Statistical Significance

Testing of the mean – with unknown variance

If now it is assumed that the variance is unknown the following sample statistic must be considered

$$T = \frac{\bar{X} - \mu}{\frac{S_{unbiased}}{\sqrt{n}}}$$

which may be realized to be t-distributed with *n*-1 degree of freedom

The operating rule is then: accept H_o if $-\Delta \leq T \leq \Delta$

The critical value can be calculated from: $P(-\Delta \leq T \leq \Delta) = 1 - \alpha$

from which Δ = 1.83 is determined using the *t*-distribution with 9 degrees of freedom

Testing for Statistical Significance

Testing of the mean – with unknown variance

Assuming the same experiment outcomes as before we get the same sample average but now the variance is given by

$$s_{unbiased} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.025$$

and the *t*-statistic becomes

$$t = \frac{(0.29 - 0.3)\sqrt{10}}{0.025} = -1.27$$

which is within the interval given by ± Δ (= ± 1.83)

Thus the Null-hypothesis should not be rejected

Testing for Statistical Significance

Testing of the variance

Consider as an example the case where the variance of the fatigue lifes of welded joints is attempted reduced by means of weld surface treatment.

As experiments are very expensive only a few data are available to verify the effect of the weld surface treatment.

Testing for Statistical Significance

Testing of the variance

We may as Null-hypothesis postulate that the variance of the fatigue lifes with the surface treatment is smaller that the variance before the surface treatment i.e. :

$$\sigma^2_{new} \leq \sigma^2_{old}$$

The operating rule is then to accept the Null hypothesis if

$$S^2 \leq \Delta$$

where Δ is determined from $P\left[S^2 \leq \Delta\right] = 1 - \alpha$

and it is used that S^2 is Chi-square distributed with n degrees of freedom

Testing for Statistical Significance

Testing of more than one data set

Typically we are in a situation where we have two or more data sets each not very large – and we would like to know how the data compare in terms of :

- mean values	Test for equal mean values
- variances	Test for equal variances
- correlation	Test for zero correlation

Testing for Statistical Significance

Testing for equal mean values

Here we assume that we have two data sets

$$\mathbf{x} = (x_1, x_2, .., x_k)^T \qquad \mathbf{y} = (y_1, y_2, .., y_l)^T$$

being realizations of the random variables *X* and *Y* both assumed to be normal distributed with mean values μ_X, μ_Y and variances σ_X, σ_Y

the statistic $T = \overline{X} - \overline{Y}$

is Normal distributed with mean value $\mu_{\overline{X}-\overline{Y}} = \mu_X - \mu_Y$

and variance $\sigma^2_{\overline{X}-\overline{Y}} = \dfrac{\sigma^2_X}{k} + \dfrac{\sigma^2_Y}{l}$

Testing for Statistical Significance

Testing for equal mean values

For α equal to 0.1 Δ can be calculated as

$$P\left(\overline{X}-\overline{Y}\leq\Delta\right)=1-\alpha \qquad \Rightarrow \qquad \Delta=1.28\sqrt{\frac{\sigma_X^2}{k}+\frac{\sigma_Y^2}{l}}$$

Testing for Statistical Significance

Testing for equal variances

A test for equal variances can be performed by considering the following statistic

$$T = \frac{S^2_{X,unbiased}}{S^2_{Y,unbiased}}$$

which is seen to be the ratio between two Chi-square distributed random variables – and *T* is thus *F*-distributed with parameters *k* and *l*.

The Null-hypothesis H_0 would be that $\sigma^2_X = \sigma^2_Y$

and the operating rule to accept H_0 if $T \leq \Delta$

where Δ is determined from $P(T \leq \Delta) = 1 - \alpha$

Testing for Statistical Significance

Some considerations regarding testing for significance

Test for statistical significance can be formulated for a variety of different types of problems

we must be very careful not to „over estimate“ the value of the significnace tests because the hypothesis can be formulated in different ways and using different significance levels a -
consequently it is in principle possible to prove anything –

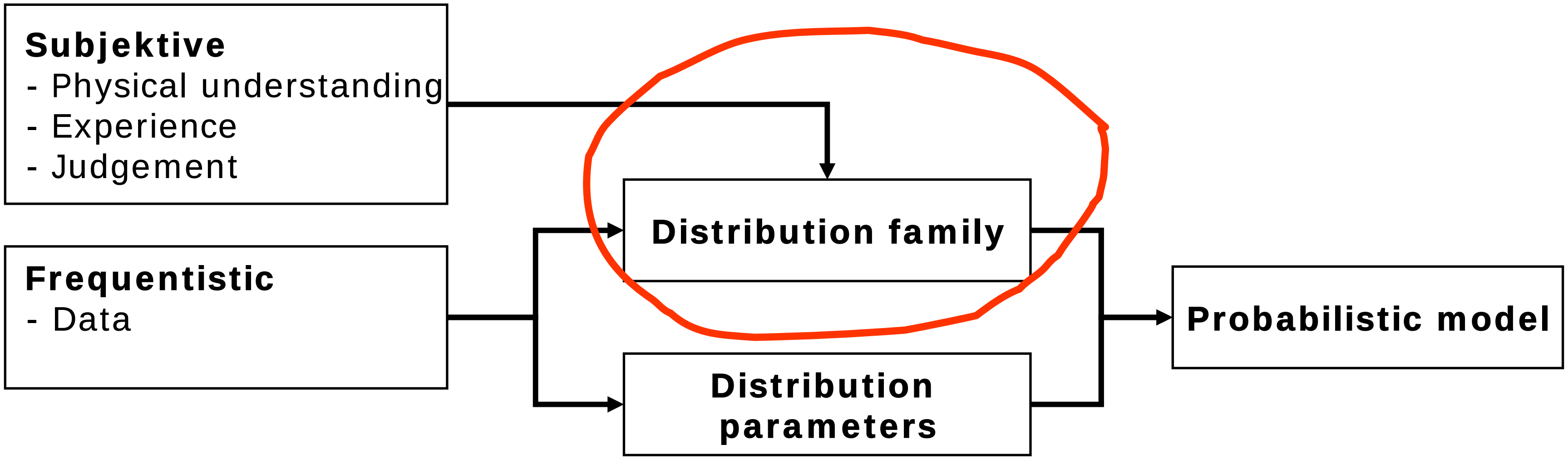
the different choises have direct effect on the probability of performing Type I and Type II errors – which may be related to significant economical consequences

the formulation of hypothesis and the choise of significance levels should be treated as a decision problem - which will be treated later.

Overview of Estimation and Model Building

Different types of information is used when developing engineering models

- subjektive information
- frequentististic information



Estimation and Model Building

Selection of probability distribution function

In general the distribution function for a given random variable or random process must be chosen on the basis of

Frequentistic information: Data
Physical arguments: Engineering understanding

A formalized classical approach is to

1. postulate a hypothesis for the probability distribution family
2. estimate the parameters of the postulated probability distribution
3. Perform a statistical test to reject/verify the hypothesis

Estimation and Model Building

Selection of probability distribution function

In engineering application it is often the case that

the available data is too sparse

to be able to support / reject the hypothesis of a given probability distribution – with a reasonable significance

Therefore it is necessary to use common sence i.e. :

First to consider physical reasons for selecting a given distribution

Thereafter to check if the available data are in gross contradiction with the selected distribution

Estimation and Model Building

Model selection by use of probability paper

Probability paper is constructed such that when a given probability distribution is plotted on the paper it will have the shape of a straight line

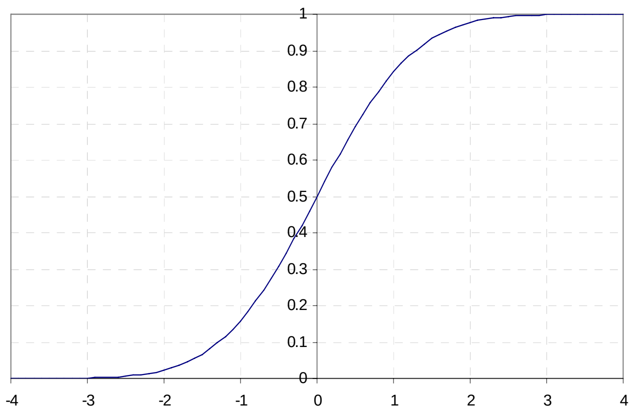
Estimation and Model Building

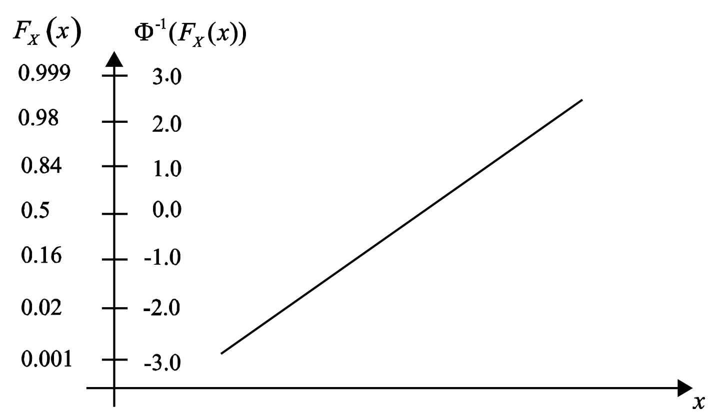
Model selection by use of probability paper

Example – probability paper for the normal probability distribution function

$$F_X(x) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right)$$

$$x = \Phi^{-1}(F_X(x)) \cdot \sigma_X + \mu_X$$



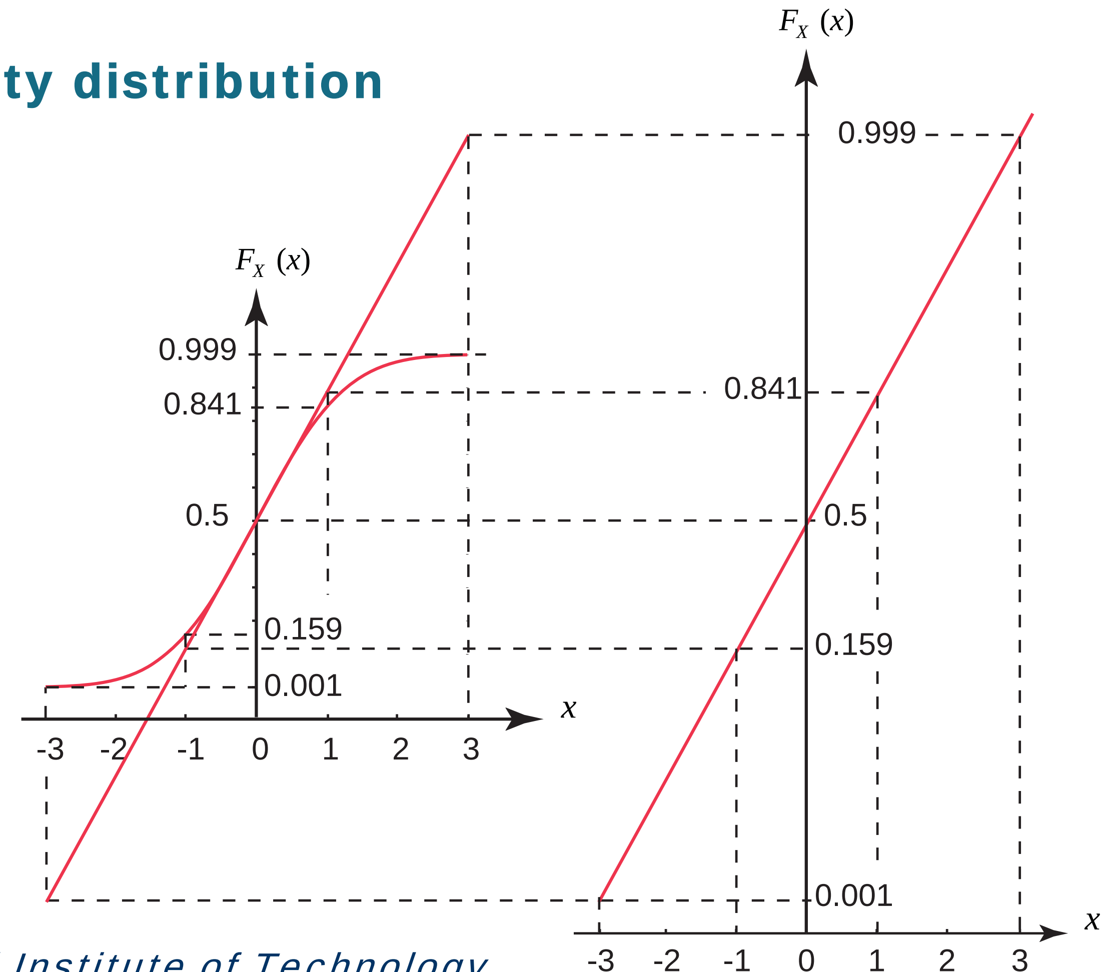


The y-axis scale is non-linear

Estimation and Model Building

Model selection by use of probability paper – graphical approach

Normal probability distribution



Estimation and Model Building

Model selection by use of probability paper

The sample probability distribution function may be established from the ordered sample as

$$F_X(x_i) = \frac{i}{N+1}$$

Example – concrete compression strength

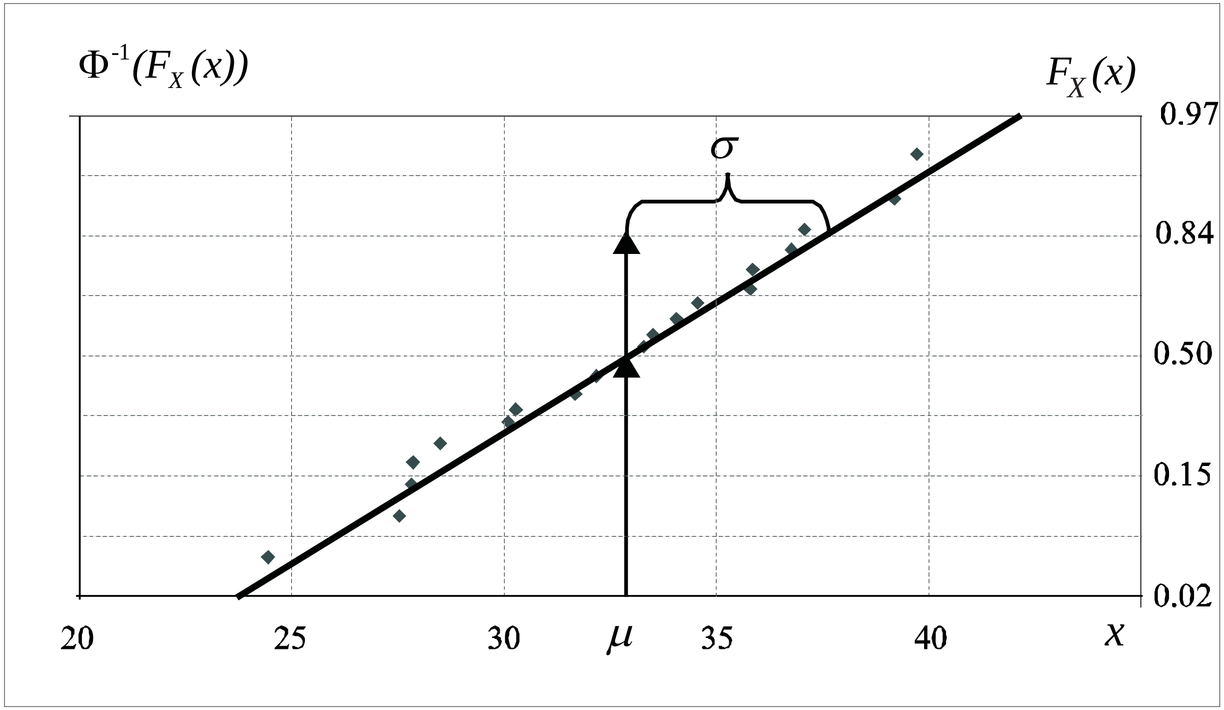
Normal probability paper

i	x_i	$F_x(x_i)$	$\Phi^{-1}(F(x_i))$
1	24.4	0.047619	-1.668391
2	27.6	0.095238	-1.309172
3	27.8	0.142857	-1.067571
4	27.9	0.190476	-0.876143
5	28.5	0.238095	-0.712443
6	30.1	0.285714	-0.565949
7	30.3	0.333333	-0.430727
8	31.7	0.380952	-0.302981
9	32.2	0.428571	-0.180012
10	32.8	0.47619	-0.059717
11	33.3	0.52381	0.059717
12	33.5	0.571429	0.180012
13	34.1	0.619048	0.302981
14	34.6	0.666667	0.430727
15	35.8	0.714286	0.565949
16	35.9	0.761905	0.712443
17	36.8	0.809524	0.876143
18	37.1	0.857143	1.067571
19	39.2	0.904762	1.309172
20	39.7	0.952381	1.668391

Estimation and Model Building

Model selection by use of probability paper

Plotting the sample probability distribution function in the probability paper yields



Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Contents of Todays Lecture

- The Results of the Assessment of the Lecture
- Short Summary of the Previous Lecture
- Overview of Estimation and Model Building
- Estimation of Distribution Parameters
  - The method of moments
  - The method of maximum likelihood

What did we Learn in the Previous Lecture

- In the previous lecture we introduced the concept of

  hypothesis testing

  - testing of the mean
  - testing of the variance
  - testing of more data sets

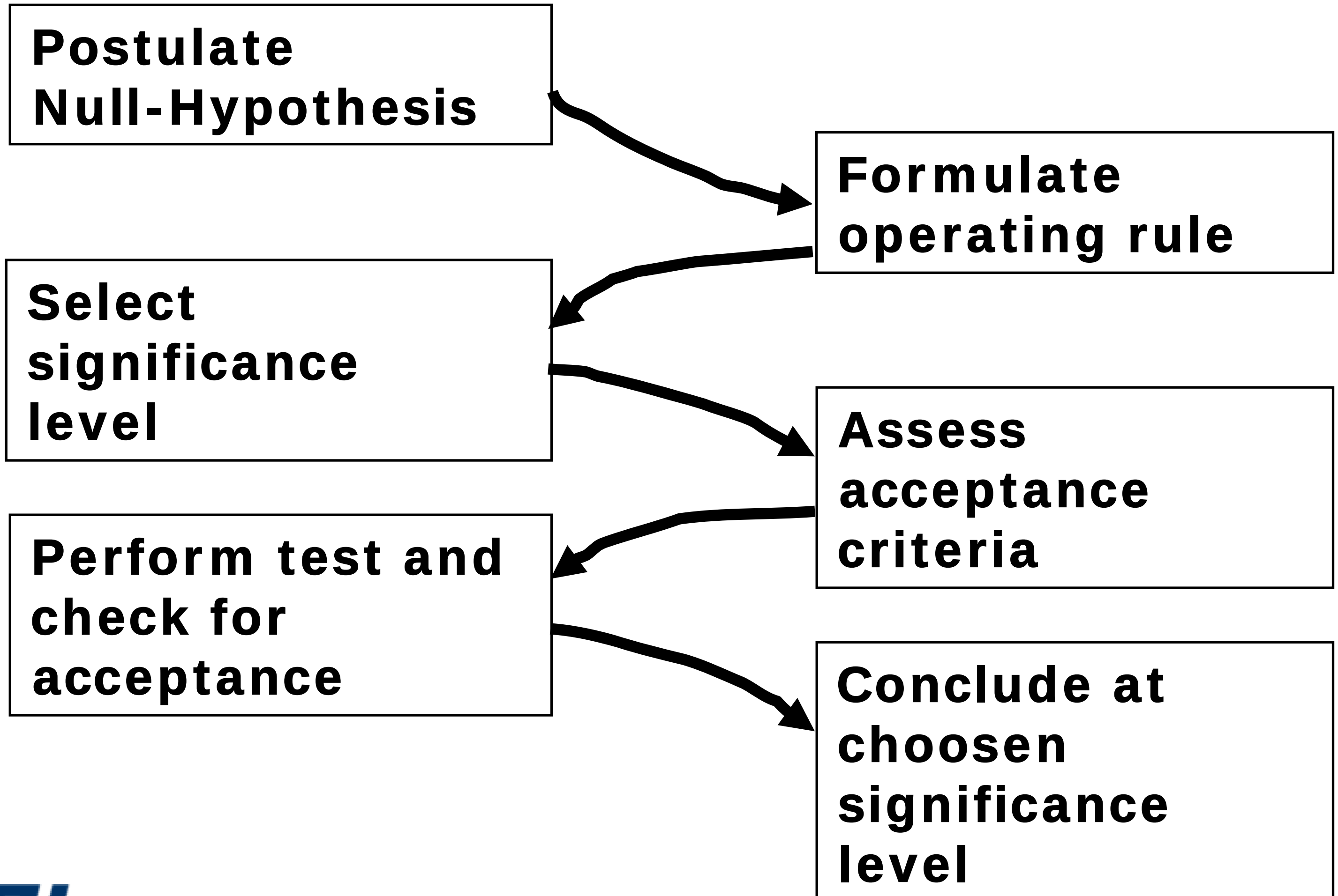
  and we also introduced the concept of

  probability paper

  - supporting the choice of a given probabilistic model based on data/observations

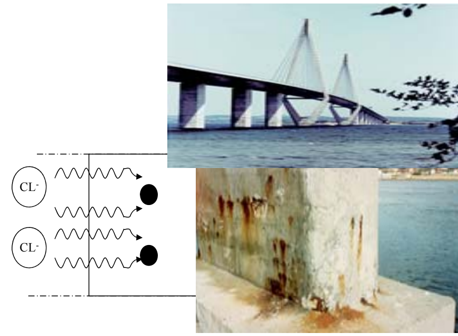
What did we Learn in the Previous Lecture

- **hypothesis testing – which are the steps!**



What did we Learn in the Previous Lecture

The design assumption:
The mean surface chloride concentration is 0.3 %



Knowledge:
Standard deviation of the surface chloride concentration – equal to 0.04 %

Hypothesis (H_0 hypothesis):
Design assumption is correct!

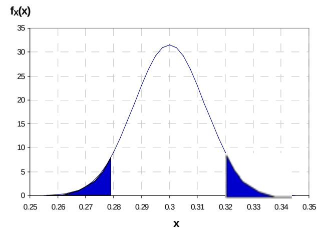
Operating rule / testing approach
Given that we know the standard deviation we know that the uncertain mean is normal distributed – we thus have a normal distributed test statistic *T*

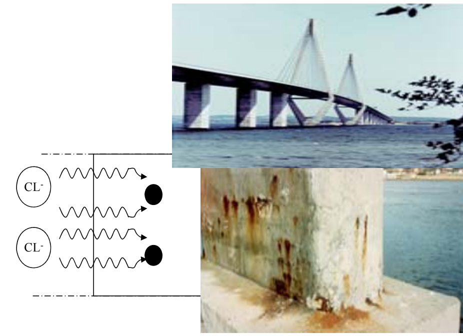
$$0.3 - \Delta \leq T \leq 0.3 + \Delta$$

What did we Learn in the Previous Lecture

The test acceptance criteria:

The operating rule must be fulfilled with a probability of 1-α.





$$P(0.3-\Delta \le T \le 0.3+\Delta) = 1-\alpha$$

Assessing acceptance criteria:

The interval for the operating rule is determined as:

$$\Phi\left(\frac{x_U-\mu}{\sigma}\right)-\Phi\left(\frac{x_L-\mu}{\sigma}\right)=\Phi\left(\frac{(0.3+\Delta)-0.3}{\frac{0.04}{\sqrt{10}}}\right)-\Phi\left(\frac{(0.3-\Delta)-0.3}{\frac{0.04}{\sqrt{10}}}\right)=0.9 \quad \Rightarrow \quad \Delta = 0.0208 \quad \Longrightarrow \quad [0.28 \le t \le 0.32]$$

Perform test and check for acceptance

Collect samples and calculte the mean value

$$\mathbf{x} = (0.33, 0.32, 0.25, 0.31, 0.28, 0.27, 0.29, 0.3, 0.27, 0.28)^T \Rightarrow t = 0.29$$

Conclusion

The validity of design assumtions cannot be rejected at the 0.1 significance level

What did we Learn in the Previous Lecture

- **Probability paper – what is the idea!**

Fundamentally what we want to do is to check whether data / observations follow a given cumulative distribution function

If they do we have support for assuming that the uncertain phenomenon which generated the data can be modelled by the given cumulative distribution function

The concept of probability paper provides us a standardized manner to perform this check

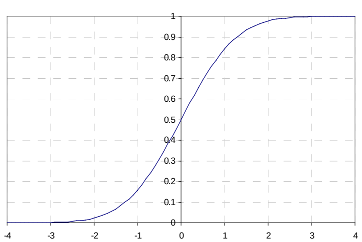
What did we Learn in the Previous Lecture

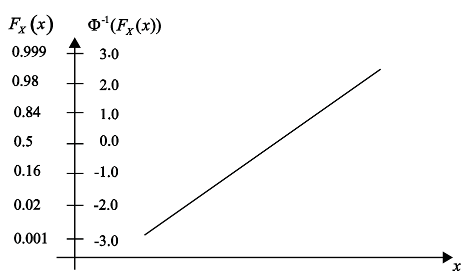
- **Probability paper – what is the idea!**

We construct probability paper for a given family of cumulative distribution functions such that a plot of the cumulative distribution follows a straight line in the paper

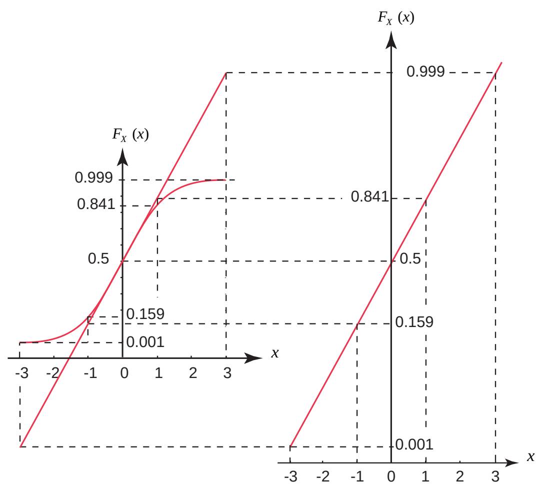
In order to do that we perform an non-linear transformation of the *y*-axis of the usual CDF plot

$$F_X(x) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right) \Rightarrow x = \Phi^{-1}(F_X(x)) \cdot \sigma_X + \mu_X$$





Analytically



Graphically

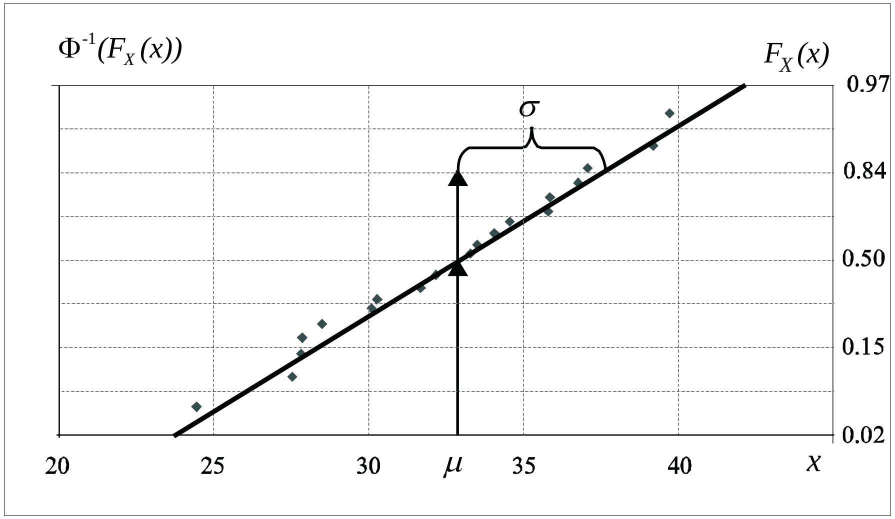
What did we Learn in the Previous Lecture

- **Probability paper – what is the idea!**

When we have the paper (we can construct it our selves or buy it in the book store ☺ we can plot observed values as a quantile-plot into the paper

$$F_X(x_i) = \frac{i}{N+1}$$

i	x_i	$F_X(x_i)$	$\Phi^{-1}(F(x_i))$
1	24.4	0.047619	-1.668391
2	27.6	0.095238	-1.309172
3	27.8	0.142857	-1.067571
4	27.9	0.190476	-0.876143
5	28.5	0.238095	-0.712443
6	30.1	0.285714	-0.565949
7	30.3	0.333333	-0.430727
8	31.7	0.380952	-0.302981
9	32.2	0.428571	-0.180012
10	32.8	0.47619	-0.059717
11	33.3	0.52381	0.059717
12	33.5	0.571429	0.180012
13	34.1	0.619048	0.302981
14	34.6	0.666667	0.430727
15	35.8	0.714286	0.565949
16	35.9	0.761905	0.712443
17	36.8	0.809524	0.876143
18	37.1	0.857143	1.067571
19	39.2	0.904762	1.309172
20	39.7	0.952381	1.668391

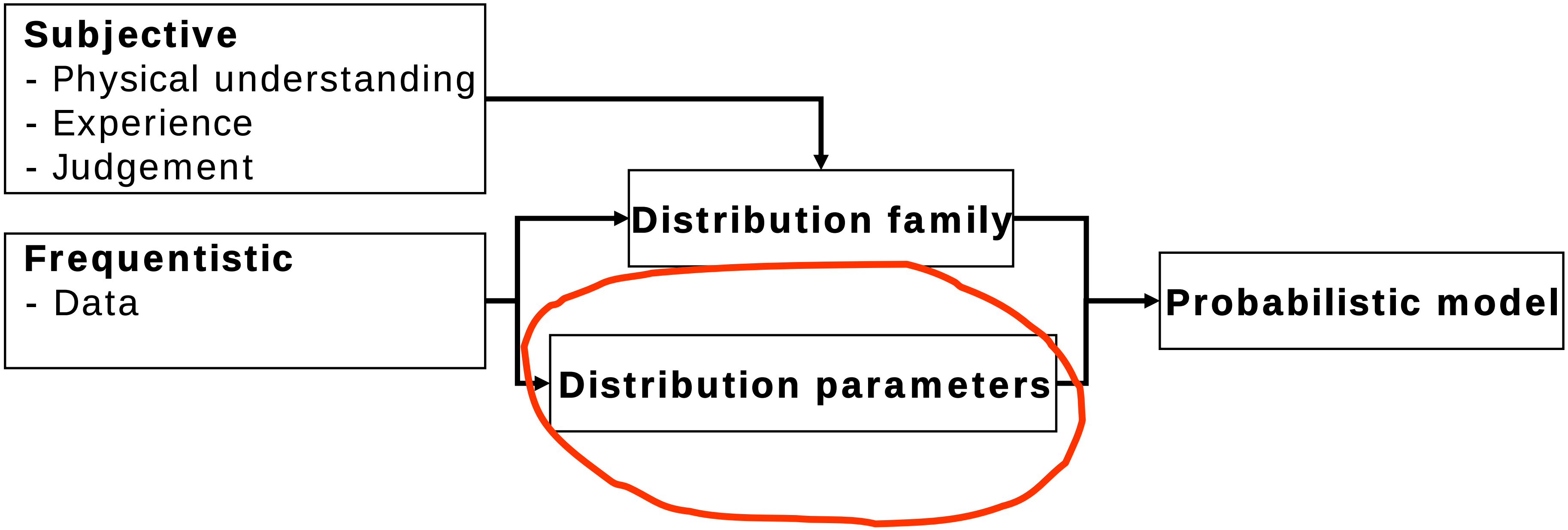


If the *q*-plot is close to straight in the important regions we have support for our model!

Overview of Estimation and Model Building

Different types of information is used when developing engineering models

- subjective information
- frequentististic information



Estimation of Distribution Parameters

We assume that we have identified a plausible family of probability distribution functions – as an example :

Normal Distribution

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

Weibull distribution

$$f_X(x) = \frac{k}{u-\varepsilon}\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^{k-1} \exp\left(-\left(\frac{x-\varepsilon}{u-\varepsilon}\right)^k \right)$$

and thus now need to determine – estimate - its parameters

$$\boldsymbol{\theta} = (\theta_1, \theta_2, .., \theta_k)^T$$

Estimation of Distribution Parameters

There are several methods for estimating the parameters of probability distribution functions, hereunder the so-called

- Point estimators
- Interval estimators

however, in the following we shall restrict ourselves to consider the

Method of moments

Method of maximum likelihood

Estimation of Distribution Parameters

- **The method of moments (MoM)**

  To start with we assume that we have data on the basis of which we can estimate the distribution parameters

  $$\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_{2,}, .., \hat{x}_n)^T$$

  The idea behind the **method of moments** is to determine the distribution parameters such that **the sample moments** (from the data) **and the analytical moments** (from the assumed distribution) **are identical.**

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

$$\lambda_j = \lambda_j(\theta_1, \theta_2, .., \theta_k) = \int_{-\infty}^{\infty} x^j \cdot f_X(x|\boldsymbol{\theta})dx$$

Sample moments **Analytical moments**

Estimation of Distribution Parameters

- **The method of moments (MoM)**

  If we assume that the considered probability distribution function has *n* parameters that we must estimate we thus need *n* equations, i.e:

$$m_j = \lambda_j(\boldsymbol{\theta}),\ j = 1,2,..,n$$

$$\Downarrow$$

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} x_i^j} = \boxed{\int_{-\infty}^{\infty} x^j \cdot f_X(x|\boldsymbol{\theta})dx},\ j = 1,2,..,n$$

Sample moment

Analytical moment

Estimation of Distribution Parameters

- **The method of moments (MoM)**

Consider as an example the data regarding the concrete compressive strength –

Again we assume that the concrete compressive strength is normal distributed – „the normal distribution family“

The normal distribution family has two parameters – we need thus to establish two equations

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i \qquad \lambda_1 = \int_{-\infty}^{\infty} x \cdot f_X(x|\mu,\sigma)dx$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i^2 \qquad \lambda_2 = \int_{-\infty}^{\infty} x^2 \cdot f_X(x|\mu,\sigma)dx$$

Estimation of Distribution Parameters

- **The method of moments (MoM)**

The sample moments are easily calculated as

$$m_1 = \frac{1}{20}\sum_{i=1}^{n}\hat{x}_i = 32.67 \qquad m_2 = \frac{1}{20}\sum_{i=1}^{n}\hat{x}_i^2 = 1083.36$$

The analytical moments can be established as function of the parameters

$$\lambda_1 = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}}\exp(-0.5\frac{(x-\mu)^2}{\sigma^2})dx \qquad \lambda_2 = \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sigma\sqrt{2\pi}}\exp(-0.5\frac{(x-\mu)^2}{\sigma^2})dx$$

Estimation of Distribution Parameters

- **The method of moments (MoM)**

  By formulating the following object function

$$g(\mu,\sigma)=(\lambda_1(\mu,\sigma)-m_1)^2+(\lambda_2(\mu,\sigma)-m_2)^2$$

  The parameters estimation problem can be solved numerically using Excel Solver finding the parameters minimizing the object function

  Let‘s have a look !

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

  The idea behind the method of maximum likelihood is that

  The parameters are determined such that the likelihood of the observations is maximized

  The likelihood can be understood as the probability of occurence of the observed data conditional on the model

  The Maximum Likelihood Method may seem more complicated that the MoM but has a number of attractive properties which we shall see later

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

Let us assume that we know that outcomes of experiments are generated according to the normal distribution, i.e.:

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

Then the likelihood *L* of one experiment outcome $\hat{x}$ is calculated as:

$$L = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\hat{x}-\mu}{\sigma}\right)^2 \right)$$

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

  Let us assume that we know that outcomes of experiments are generated according to the normal distribution, i.e.:

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

If we have n experiment outcomes $\hat{\mathbf{x}} = (x_1, \hat{x}_2, \ldots, x_n)^T$ **the likelihood *L* becomes:**

$$L(\boldsymbol{\theta} | \hat{\mathbf{x}}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\hat{x}_i-\mu}{\sigma}\right)^2 \right)$$

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

The parameters θ are estimated as those maximizing the likelihood function or equivalently minimizes the – likelihood function i.e.:

$$\min_{\boldsymbol{\theta}} \left(-L\left(\boldsymbol{\theta} \middle| \hat{\mathbf{x}}\right)\right)$$

It is avantageous to consider the log-likelihood function $l(\boldsymbol{\theta}|\hat{\mathbf{x}})$ **:**

$$l(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \log(f_X(\hat{x}_i|\boldsymbol{\theta}))$$

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

If the parameters θ are estimated as those minimizing the – log likelihood function i.e.:

$$\min_{\boldsymbol{\theta}} (-l(\boldsymbol{\theta} | \hat{\mathbf{x}}))$$

It can be shown that the estimated parameters are normal distributed with

mean values $\boldsymbol{\mu}_{\Theta} = (\theta_1^*, \theta_2^*, .., \theta_n^*)^T$

covariance matrix $\mathbf{C}_{\Theta\Theta} = \mathbf{H}^{-1}$

$$H_{ij} = -\frac{\partial^2 l(\boldsymbol{\theta} | \hat{\mathbf{x}})}{\partial \theta_i \partial \theta_j}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

not just point estimates – full distribution information!

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

Let us consider the concrete compressive strength example

The log-likelihood function can be written as

$$l(\boldsymbol{\theta}|\hat{\mathbf{x}}) = n \cdot \ln\left(\frac{1}{\sqrt{2\pi}\theta_1}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(\hat{x}_i - \theta_2)^2}{\theta_1^2}$$

the minimum of which may be found by the solution of the following equations

$$\frac{\partial l}{\partial \theta_1} = -\frac{n}{\theta_1} + \frac{1}{\theta_1^3}\sum_{i=1}^{n}(\hat{x}_i - \theta_2)^2 = 0$$

$$\frac{\partial l}{\partial \theta_2} = \frac{1}{\theta_1^2}\sum_{i=1}^{n}(\hat{x}_i - \theta_2) = 0$$

$\Rightarrow$

$$\theta_1 = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}_i - \theta_2)^2}{n}}$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i$$

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

Putting numbers into the solution we get:

$$\theta_1 = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{x}_i - \theta_2\right)^2}{n}} = \sqrt{\frac{367.19}{20}} = 4.05$$

Mean value of the standard deviation

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i = \frac{653.3}{20} = 32.67$$

Mean value of the mean value

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

As mentioned we may also determine the covariance matrix:

$$H = \begin{pmatrix} \dfrac{n}{\theta_1} - \dfrac{3\sum_{i=1}^{n}(x_i - \theta_2)^2}{\theta_1^4} & \dfrac{2\sum_{i=1}^{n}(x_i - \theta_2)}{\theta_1^3} \\ \dfrac{2\sum_{i=1}^{n}(x_i - \theta_2)}{\theta_1^3} & \dfrac{n}{\theta_1^2} \end{pmatrix}$$

$$C_{\Theta\Theta} = H^{-1} = \begin{pmatrix} 0.836 & 0 \\ 0 & 0.165 \end{pmatrix}$$

Variance of the mean value (0.165)

Variance of the standard deviation (0.836)

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

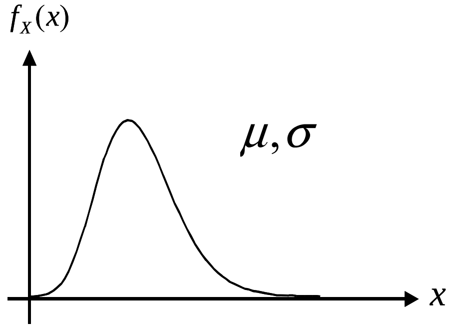
  We may also estimate the parameters completely numerically using Excel

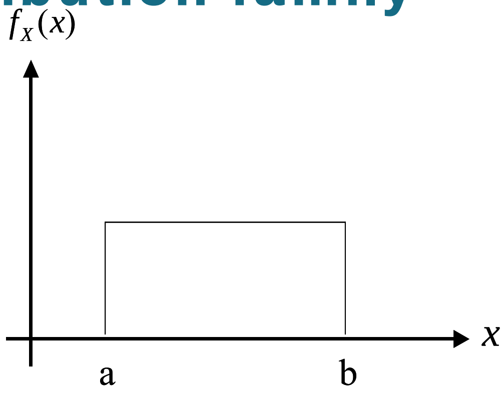
  <u>Lets take a look !</u>

Estimation of Distribution Parameters

- **Summary**

Given that we have selected a model for the distribution i.e. a distribution family





we have to estimate the distribution parameters

- Method of Moments
- Maximum Likelihood Method

Estimation of Distribution Parameters

- **Summary**

Method of Moments provide point estimates of the parameters

- No information about the uncertainty with which the parameter estimates are associated.

Maximum Likelihood Method provide point estimates of the estimated parameters

- Full distribution information – normal distributed parameters, mean values and covariance matrix.

Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

Prof. Dr. Michael Havbro

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Contents of Todays Lecture

- Short Summary of the Previous Lecture
- Overview of Estimation and Model Building
- Model Evaluation by Statistical Testing
  - The χ^2 goodness of fit test
  - The Kolmogorov-Smirnov goodness of fit test
  - Model comparison

Short Summary of the Previous Lecture

- **We considered the problem of assessing the parameters of distributions based on observations/data**

  What did we learn?

  We learned that parameters can be estimated using the

  - Method of Moments

  - Method of Maximum Likelihood

Short Summary of the Previous Lecture

- **The Method of Moments (MoM) – point estimates**

  The principle behind the MoM is that we estimate the parameters such that the moments we can calculate based on the analytical expressions become equal to the sample moments.

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i \qquad \lambda_1 = \int_{-\infty}^{\infty} x \cdot f_X(x|\mu,\sigma)dx$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i^2 \qquad \lambda_2 = \int_{-\infty}^{\infty} x^2 \cdot f_X(x|\mu,\sigma)dx$$

This leads to n equations which have to be solved simultaneously where n is the number of parameters

Short Summary of the Previous Lecture

- **The Method of Maximum Likelihood (MLM) –**
  full distribution estimates

  The principle behind the MLM is that we estimate the parameters such that the likelihood of the observations (data) is maximized)

$$L(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\hat{x}_i - \mu}{\sigma}\right)^2\right)$$

$$\min_{\boldsymbol{\theta}} \left(-L(\boldsymbol{\theta}|\hat{\mathbf{x}})\right) \qquad l(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \log(f_X(\hat{x}_i|\boldsymbol{\theta}))$$

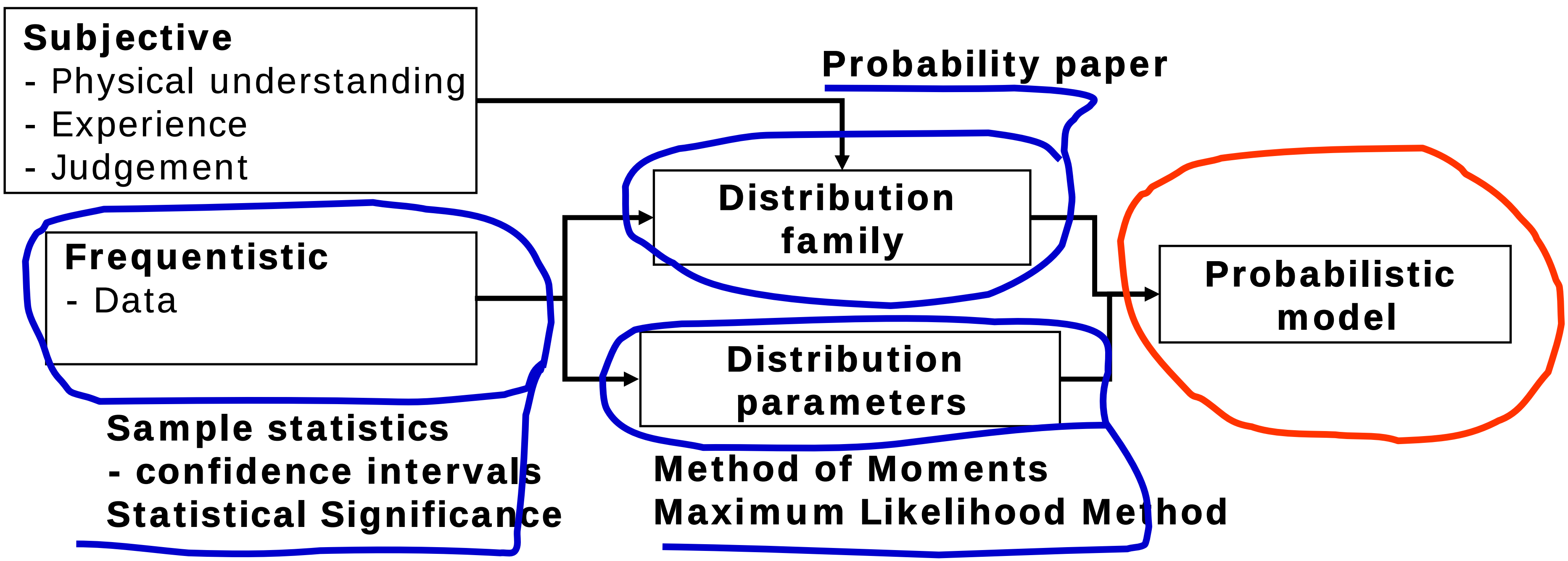
$$\Rightarrow \begin{cases} \boldsymbol{\mu}_{\Theta} = (\theta_1^*, \theta_2^*, .., \theta_n^*)^T \\ \mathbf{C}_{\Theta\Theta} = \mathbf{H}^{-1} \\ H_{ij} = -\dfrac{\partial^2 l(\theta|\hat{\mathrm{x}})}{\partial\theta_i \partial\theta_j}\Big|_{\theta=\theta^*} \end{cases}$$

The MLM provides an extremely strong statistical tool!

Overview of Estimation and Model Building

Different types of information is used when developing engineering models

- subjective information
- frequentististic information



Model Evaluation by Statistical Testing

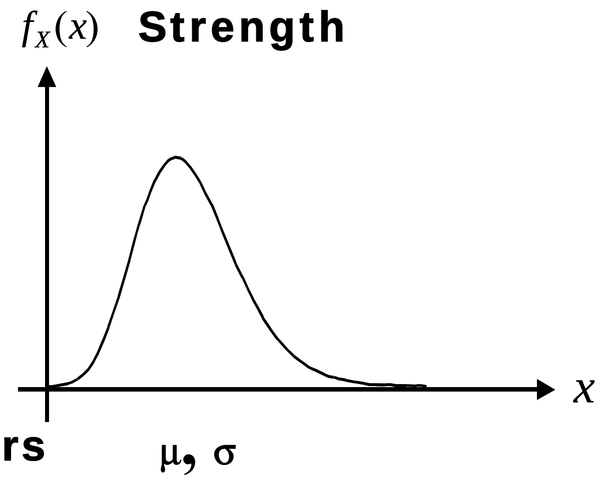
Let us assume that we have selected a distribution function as a model to describe an uncertain quantity

Distribution family

Concrete Compressive Strength

Data + Physics

Data



Distribution parameters

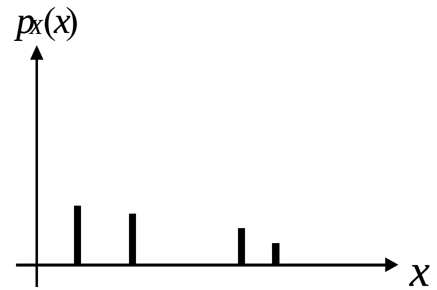
Now we want to validate our model selection – by means of statistical tests

Model Evaluation by Statistical Testing

Two different cases are considered – namely verification of

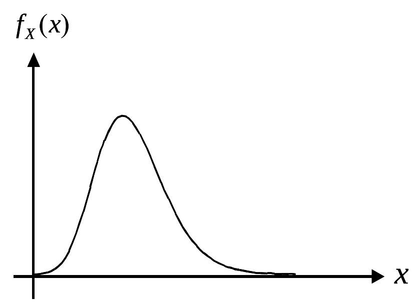
1: Discrete distribution functions

CHI-Square (χ^2) test



2: Continuous distribution functions

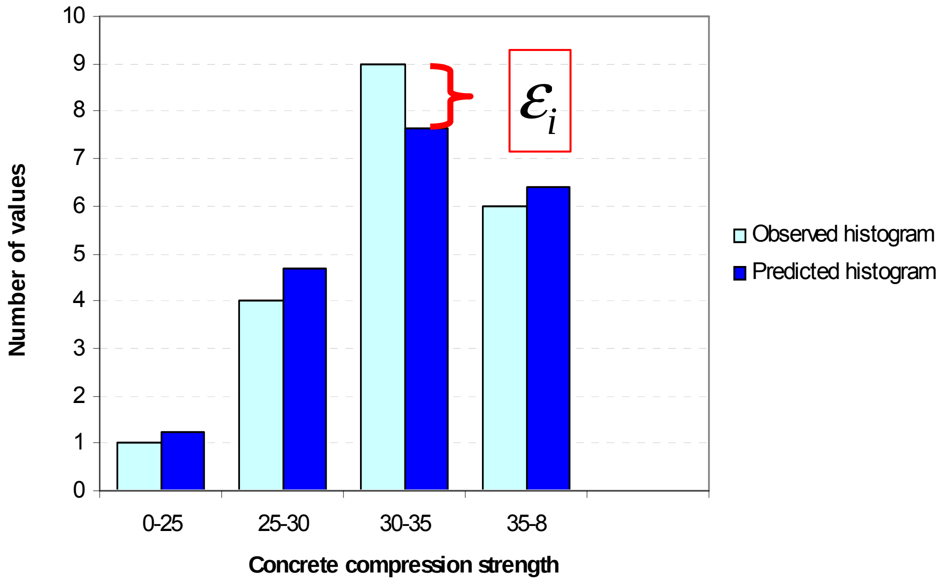
Kolmogorov Smirnov test



Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

The idea behind the CHI-Square goodness of fit test is that the difference between predicted and observed/sample histograms should be small

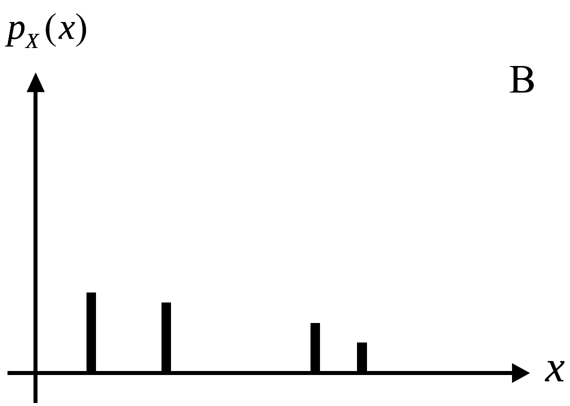


Model Evaluation by Statistical Testing

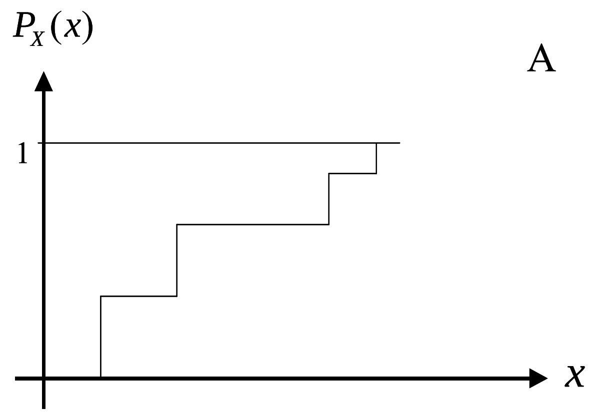
The CHI-square goodness of fit test

We remember that a discrete cumulative distribution is given by:

$$P(x_i) = \sum_{j=1}^{i-1} p(x_j), \qquad i \leq k$$



Probability density function



Cumulative distribution function

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

Assuming that we sample a discrete random variable X n times the number of realizations of $X=x_i$ i.e. N_i is a binomial distributed random variable with expected value and variance given as:

$$E[N_i] = np(x_i) = N_{p,i}$$

Predicted number of occurences at a given value

$$Var[N_i] = np(x_i)(1-p(x_i)) = N_{p,i}(1-p(x_i))$$

If the postulated model is correct and n large enough – Central Limit Theorem - the difference ε_i

$$\varepsilon_i = \frac{N_{o,i} - N_{p,i}}{\sqrt{N_{p,i}(1-p(x_i))}}$$

Observed number of occurences at a given value

will be standard Normal distributed

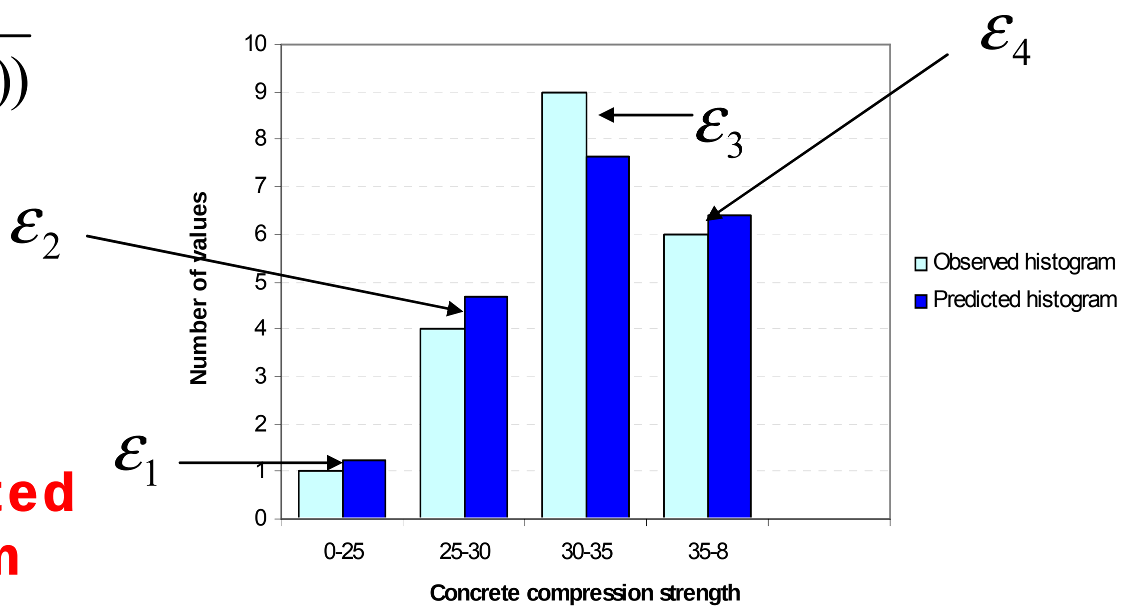
Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

By summing up the squared differences between the observed and the predicted histograms we get:

$$\varepsilon^2 = \sum_{i=1}^{k} \varepsilon_i^2 = \sum_{i=1}^{k} \frac{(N_{o,i} - N_{p,i})^2}{N_{p,i}(1 - p(x_i))}$$

$$\varepsilon_m{}^2 = \sum_{i=1}^{k} \frac{(N_{o,i} - N_{p,i})^2}{N_{p,i}}$$



CHI-Square distributed
k-1 degree of freedom

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

The idea is then to test – at a given significance level – α – if the sum of observed squared differences is plausible i.e.

Postulating the H_0 hypothesis that the assumed distribution function is not in gross contradiction with the observed data and formulating the operating rule **such as the null hypothesis cannot be accepted if $\varepsilon_m^2 \geq \Delta$. The critical value Δ can be estimated such as:**

$$P(\varepsilon_m^2 \geq \Delta) = \alpha$$

The alternate hypothesis H_1 is far less informative because it considers all other distribution functions than the assumed.

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

Consider as an example that we assume a Normal distribution with **parameters not estimated from the available data**

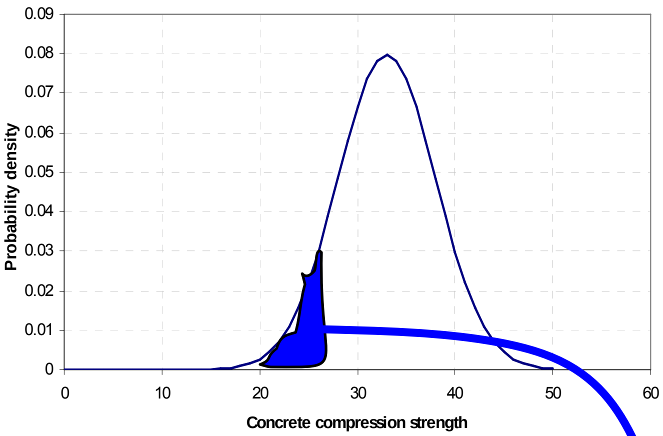
Mean:	33 Mpa
Standard deviation:	5 Mpa

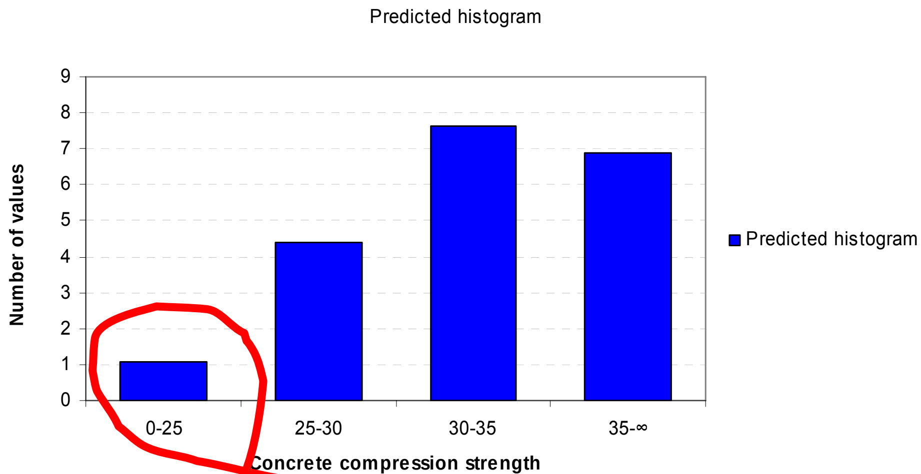
The **Normal distribution** is a continuous probability density function but **can easily be discretized**

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

The postulated probability density function is discretized:





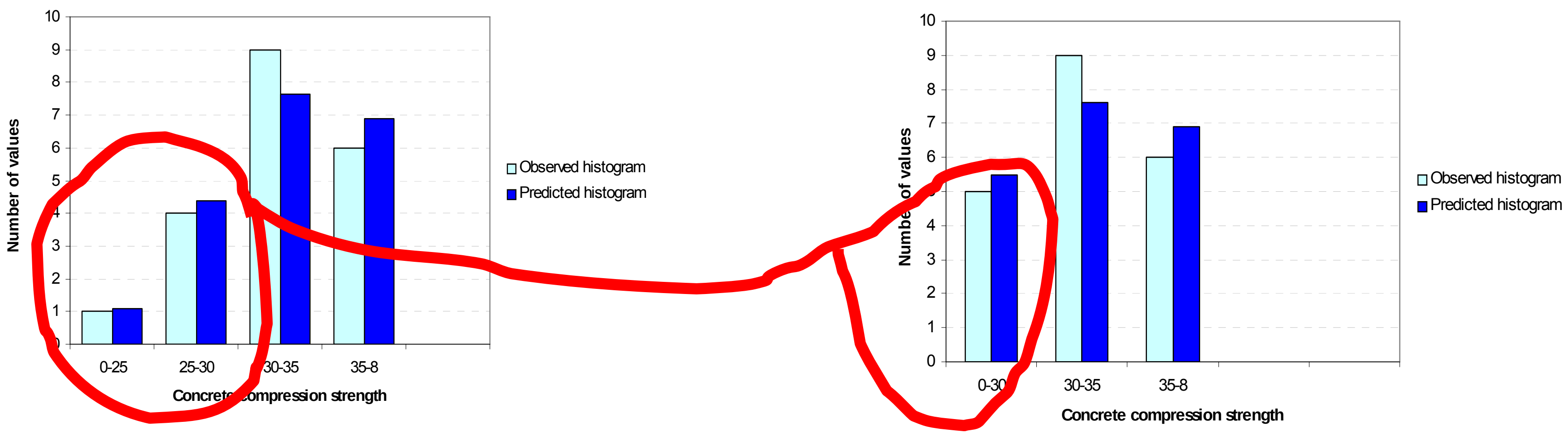
Total number of experiments

$$\textbf{0-25:}\quad 20\left[\Phi\left(\frac{25-33}{5}\right)-\Phi\left(\frac{-\infty-33}{5}\right)\right]=20\cdot 0.055=1.10$$

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

The observed and the predicted histograms may be compared



Due to a low number of samples in the lower interval the two lower intervals are „lumped“

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

$$\varepsilon_m^2 = \sum_{j=1}^{k} \frac{(N_{o,j} - N_{p,j})^2}{N_{p,j}}$$

The following calculation sheet may be produced

Interval - x_j	Number of observed values $N_{o,j}$	Predicted probability $p(x_j)$	Predicted number of observations $N_{p,j} = 20p(x_j)$	Sample statistic Equation (5.68)
0 -30	5	0.296671	5.933415	0.14684
30-35	9	0.381169	7.65443	0.236537
35-∞	6	0.344578	6.412155	0.026492
			Sum	0.40987

At the 5% significance level the CHI-Square distribution with 3-1=2 degree of freedom yields Δ = 5.99
As 0.40987 is smaller than 5.99 the H_0 hypothesis cannot be rejected !

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

If one or more (m) of the parameters of the postulated distribution function had been assessed using the same data as used for the testing we must reduce the number of degrees of freedom accordingly i.e. $n = k - 1 - m$

Assuming that we had estimated the variance from the data but not the mean value we would have $n = 3-1-1=1$

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

Assuming a postulated Normal distribution with

μ = 33.00

σ = 4.05

We get the following calculation sheet

Interval - x_j	Number of observed values $N_{o,j}$	Predicted probability $p(x_j)$	Predicted number of observations $N_{p,j} = 20p(x_j)$	Sample statistic Equation (5.26)
0 -30	5	0.274253	5.485061	0.042896
30-35	9	0.381169	7.623373	0.248591
35-∞	6	0.344578	6.891566	0.115342
			Sum	0.406829

At the 5% significance level the CHI-Square distribution with 3-1-1 = 1 degrees of freedom yields Δ = 3.84
As 0.406829 is smaller than 3.84 the H_0 hypothesis cannot be rejected !

Model Evaluation by Statistical Testing

The Kolmogorov-Smirnov goodness of fit test

The idea behind the Kolmogorov-Smirnov test is that

If the postulated cumulative distribution function is in accordance with the observed data then the maximal difference between the observed and the predicted cumulative distribution functions should be small

Model Evaluation by Statistical Testing

The Kolmogorov-Smirnov goodness of fit test

The observed cumulative distribution function may be calculated from

$$F_o(x_i) = \frac{i}{n}$$

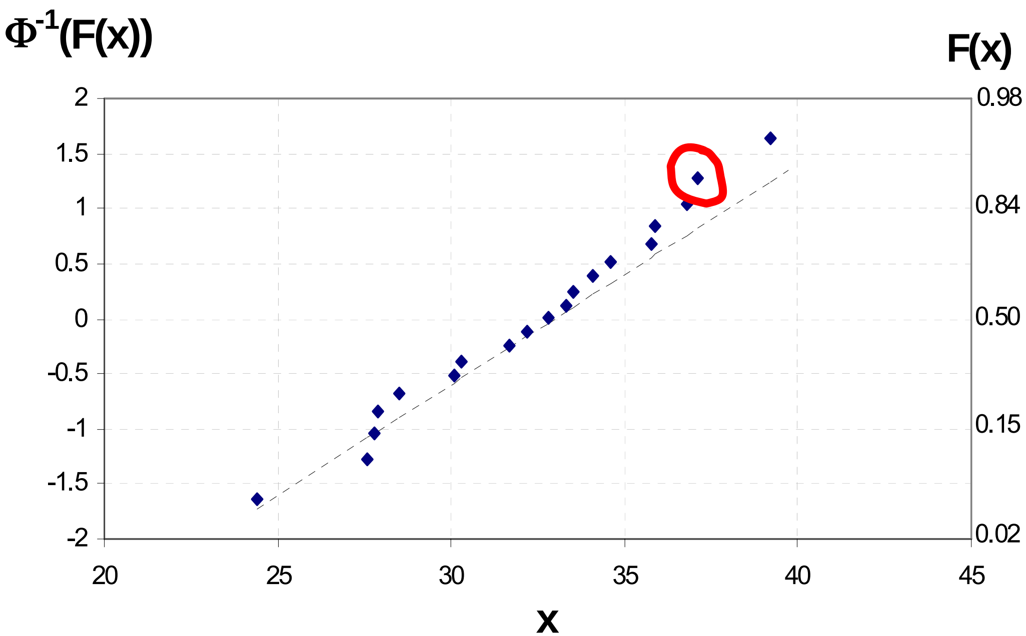
The following statistic has been proposed

$$\varepsilon_{\max} = \max_{i=1}^{n}\left[\left|F_o(x_i) - F_p(x_i)\right|\right] = \max_{i=1}^{n}\left[\left|\frac{i}{n} - F_p(x_i)\right|\right]$$

Model Evaluation by Statistical Testing

The Kolmogorov-Smirnov goodness of fit test

The Kolmogorov-Smirnov statistic may be assessed from



i	x_i	$F_{xo}(x_i)$	$F_{xp}(x_i)$	ε_i
1	24.4	0.05	0.042716	0.007284
2	27.6	0.1	0.140071	0.040071
3	27.8	0.15	0.14917	0.00083
4	27.9	0.2	0.153864	0.046136
5	28.5	0.25	0.18406	0.06594
6	30.1	0.3	0.280957	0.019043
7	30.3	0.35	0.294598	0.055402
8	31.7	0.4	0.397432	0.002568
9	32.2	0.45	0.436441	0.013559
10	32.8	0.5	0.484047	0.015953
11	33.3	0.55	0.523922	0.026078
12	33.5	0.6	0.539828	0.060172
13	34.1	0.65	0.587064	0.062936
14	34.6	0.7	0.625516	0.074484
15	35.8	0.75	0.71226	0.03774
16	35.9	0.8	0.719043	0.080957
17	36.8	0.85	0.776373	0.073627
18	37.1	0.9	0.793892	0.106108
19	39.2	0.95	0.892512	0.057488
20	39.7	1	0.909877	0.090123

Model Evaluation by Statistical Testing

The Kolmogorov-Smirnov goodness of fit test

The Kolmogorov-Smirnov statistic is tabulated

	n											
α	**1**	**5**	**10**	**15**	**20**	**25**	**30**	**40**	**50**	**60**	**70**	**80**
0.01	0.9950	0.6686	0.4889	0.4042	0.3524	0.3166	0.2899	0.2521	0.2260	0.2067	0.1917	0.1795
0.05	0.9750	0.5633	0.4093	0.3376	0.2941	0.2640	0.2417	0.2101	0.1884	0.1723	0.1598	0.1496
0.1	0.9500	0.5095	0.3687	0.3040	0.2647	0.2377	0.2176	0.1891	0.1696	0.1551	0.1438	0.1347
0.2	0.9000	0.4470	0.3226	0.2659	0.2315	0.2079	0.1903	0.1654	0.1484	0.1357	0.1258	0.1179

For n = 20 and a = 5% we get 0.2941

compared to observed statisic 0.1061

The H_0 hypothesis cannot be rejected at the 5% significance level.

Model Evaluation by Statistical Testing

Model comparison

Model verification by significance testing can be used to quantify the plausibility of a given model relative to given data (evidence)

Two cases have to be considered

1 it is shown that a model hypothesis cannot be rejected

2 it is shown that a model hypothesis can be rejected

What information is actually contained in these two cases ?

Model Evaluation by Statistical Testing

Model comparison

Given that the significance test shows that a model hypothesis cannot be rejected:

we must remember that other models could also be postulated – in fact it is often the case that several model hypothesis may pass testing !

Given that the significance test shows that a model hypothesis should be rejected:

it does not mean that the model necessary is bad – it may just say that the evidence is not strong enough to show it with significance – too little data !

Model Evaluation by Statistical Testing

Model comparison

If testing of two different model hypothesis both fall out positive i.e. both models are plausible we can compare the goodness of fit of the two models either by

- comparing the sample statistics directly
  could be misleading / inconclusive due to different number of degrees of freedom

- comparing the sample likelihoods

Model Evaluation by Statistical Testing

Model comparison

Consider the example with two different models

Model 1: N(33;5)

Parameters estimated not using data

n=3-1=2

CHI-Square sample statistic = 0.40987

Sample likelihood = 0.8151

Model 2: N(33;4.05)

Parameters estimated using data

n=3-1-1=1

CHI-Square sample statistic = 0.40683

Sample likelihood = 0.5236

Model Evaluation by Statistical Testing

Summary

The selection of appropriate probabilistic models may be supported by significance testing of the model hypothesis

The CHI-Square test is designed especially for discrete distribution functions

The Kolmogorov-Smirnov test is designed especially for continuous distribution functions

The goodness of fit of different model alternatives may be compared by comparing sample likelihood

Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

Contents of Todays Lecture

- **Basics of Reliability Analysis**

  - Short summary of previous lecture

  - The course at a glance

  - Failure events and basic random variables

  - Linear limit state functions and Normal distributed variables

  - Error propagation

  - Non-linear limit state functions

  - Monte-Carlo simulation

Summary of Previous Lecture

- **Testing for goodness of fit**
  - **The $\chi 2$ goodness of fit test**
  - **The Kolmogorov-Smirnov goodness of fit test**

- **Model comparison**

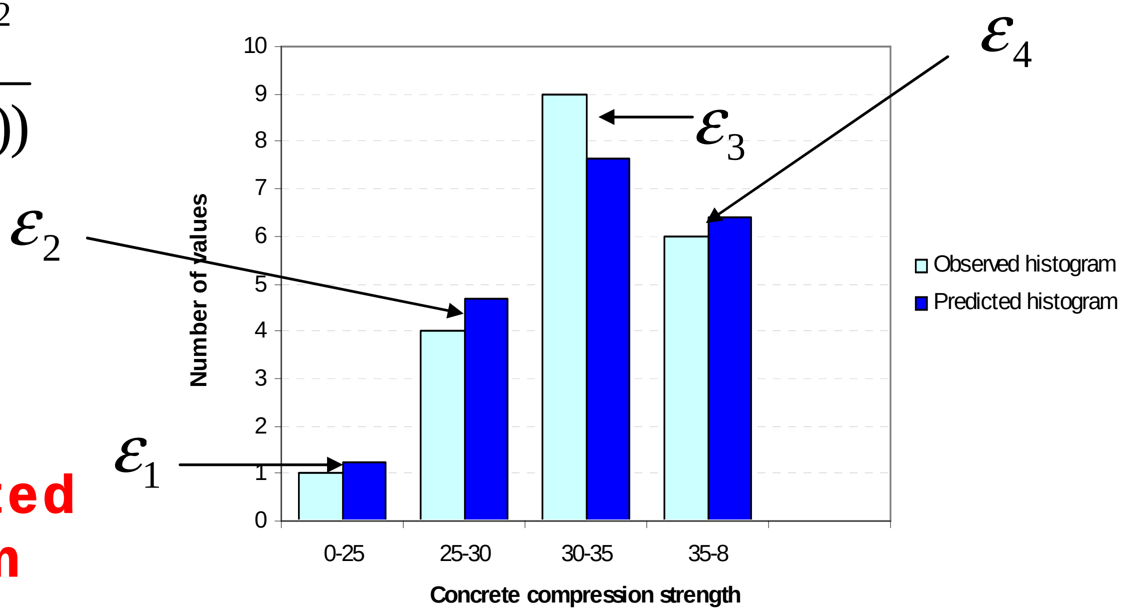
Summary of Previous Lecture

The CHI-square goodness of fit test

We test a ststistic constructed from the squared differences between the observed and the predicted histograms:

$$\varepsilon^2 = \sum_{i=1}^{k} \varepsilon_i^2 = \sum_{i=1}^{k} \frac{(N_{o,i} - N_{p,i})^2}{N_{p,i}(1 - p(x_i))}$$

$$\varepsilon_m^{\ 2} = \sum_{i=1}^{k} \frac{(N_{o,i} - N_{p,i})^2}{N_{p,i}}$$



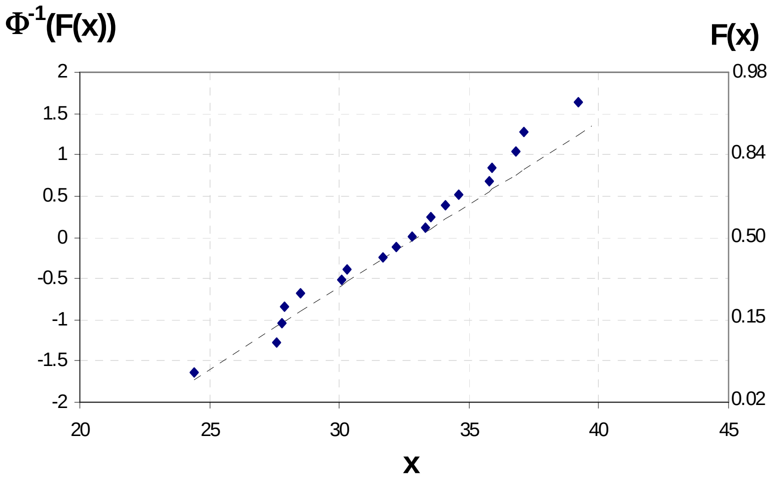
CHI-Square distributed
***k*-1 degree of freedom**

Summary of Previous Lecture

The Kolmogorov-Smirnov goodness of fit test

The observed cumulative distribution function may be calculated from:

$$F_o(x_i) = \frac{i}{n}$$



The following statistic is applied (tabularized):

$$\varepsilon_{\max} = \max_{i=1}^{n}\left[\left|F_o(x_i) - F_p(x_i)\right|\right] = \max_{i=1}^{n}\left[\left|\frac{i}{n} - F_p(x_i)\right|\right]$$

Summary of Previous Lecture

Model comparison

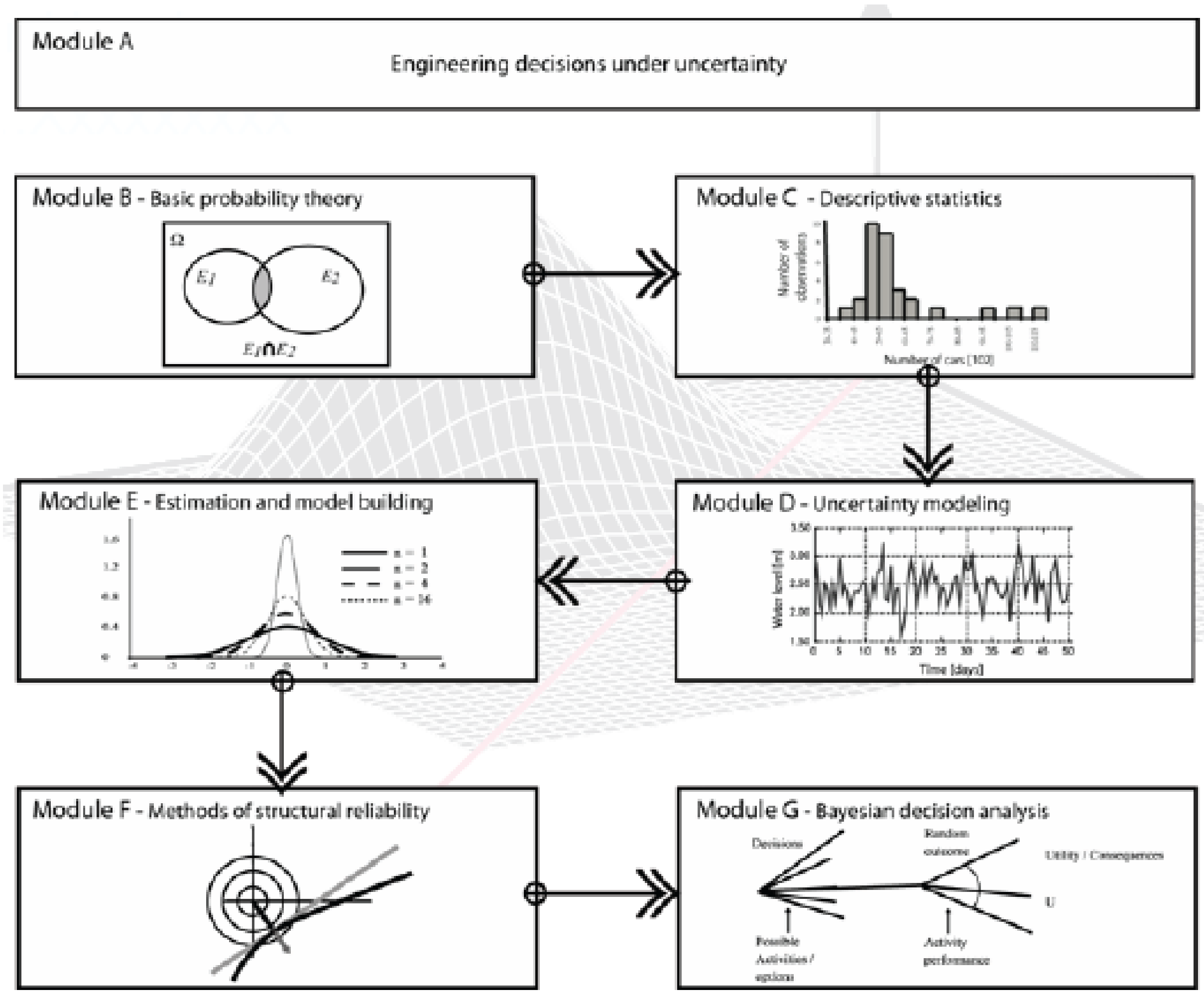
If testing of two different model hypothesis both fall out positive i.e. both models are plausible we can compare the goodness of fit of the two models either by

- comparing the sample statistics directly

  could be misleading/inconclusive due to different number of degrees of freedom

- comparing the sample likelihoods

The Course at a Glance



Basics of Reliability Analysis

- **Failure events and basic random variables**

By a failure event we associate in principle an event of special interest e.g. :

- Loss of functionality

- Costs

- Loss of lives

- Damage to the environment

Basics of Reliability Analysis

- **Failure events and basic random variables**

A failure event may conveniently be described in terms of a functional relationship

$$\mathbf{F} = \{g(\mathbf{x}) \leq 0\}$$

Such a functional relationship is denoted a limit state function

$$g(\mathbf{x})$$

↑

Realizations of basic random variables

Basics of Reliability Analysis

- **The probability of an event**

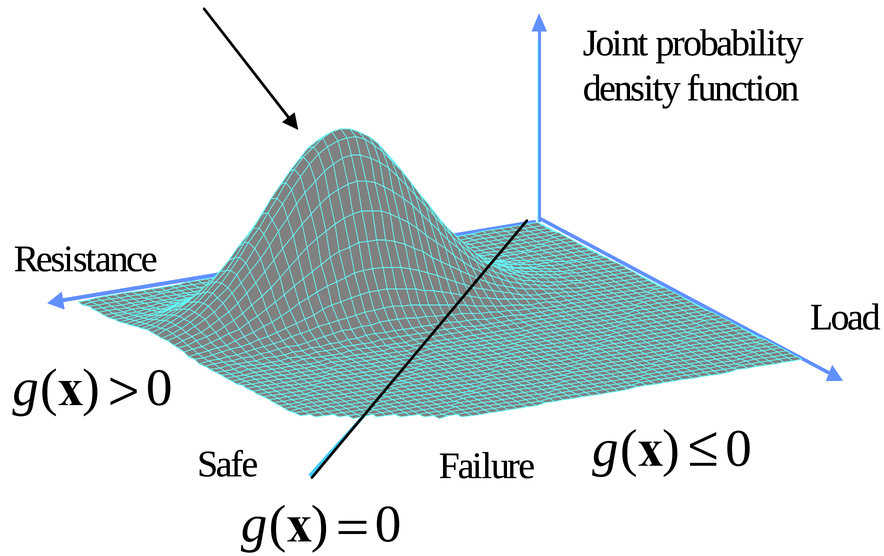
The probability of an event e.g. a failure event can be calculated by the following integral

$$P_f = \int_{g(\mathbf{x}) \leq 0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Joint probability density function of the basic random variables X

$$g(\mathbf{x}) = r - s$$

r : **Resistance**
s : **Load**



Basics of Reliability Analysis

- The probability of an event

  The probability integral is in general non-trivial – can be multi-dimensional and can have a complicated integration domain

$$P_f = \int_{g(\mathbf{x})\leq 0} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$$

Classical nummerical integration techniques such as e.g. Simpson, Gauss or Schebyschev integration are not computationally efficient for dimensions larger than 5-6. Other apporaches are needed – which we will study further -

Basics of Reliability Analysis

- **Linear limit state functions and normal distributed basic variables**

  First we consider the case where the limit state function is linear in the random variables and the random variables are normally distributed

$$g(x) = a_0 + \sum_{i=1}^{n} a_i x_i$$

For the case where the random variables X are normal distributed the safety margin M is also normal distributed

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

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

$$\sigma_M^2 = \sum_{i=1}^{n} a_i^2 \sigma_{X_i}^2 + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \rho_{ij} a_i a_j \sigma_{X_i} \sigma_{X_j}$$

Correlation coefficient

Basics of Reliability Analysis

- **Linear limit state functions and normal distributed basic variables**

  The probability of failure is then determined as

$$P_F = P(g(\mathbf{X}) \leq 0) = P(M \leq 0)$$

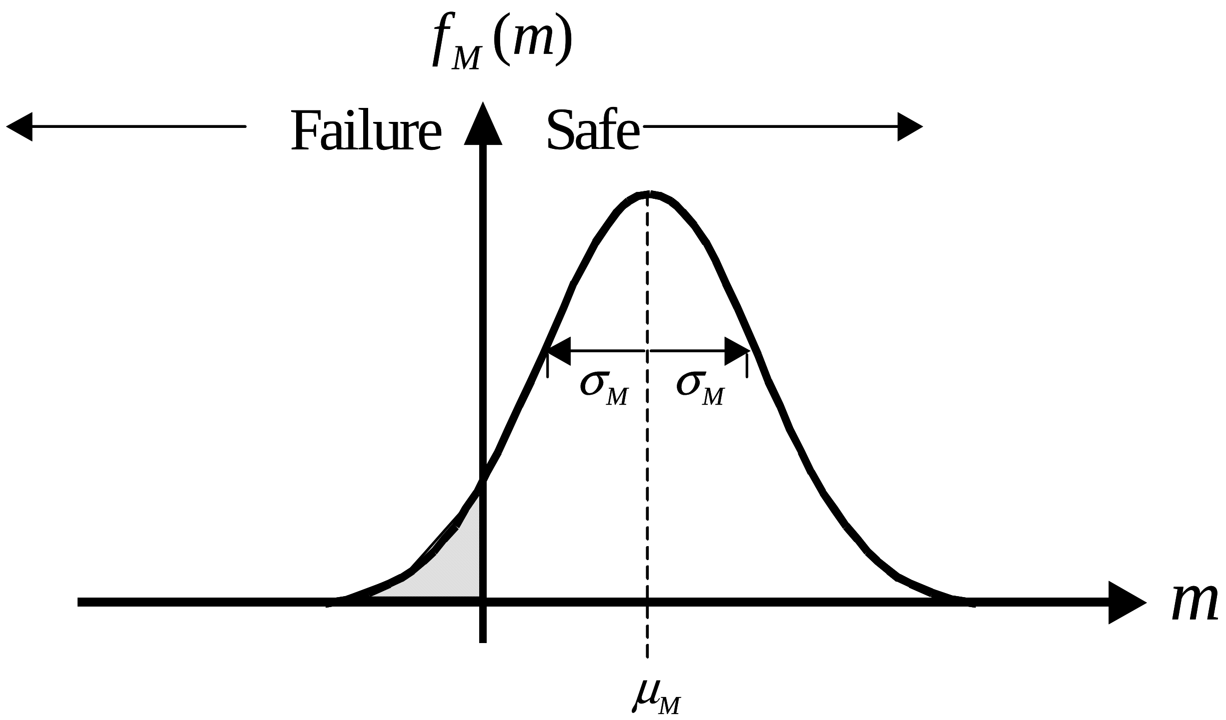
Which reduces to the determination of the standard normal probability distribution function

$$P_F = \Phi(\frac{0-\mu_M}{\sigma_M}) = \Phi(-\beta) \quad \textbf{with} \quad \beta = \frac{\mu_M}{\sigma_M}$$

Reliability or safety index

Basics of Reliability Analysis

- **Linear limit state functions and normal distributed basic variables**



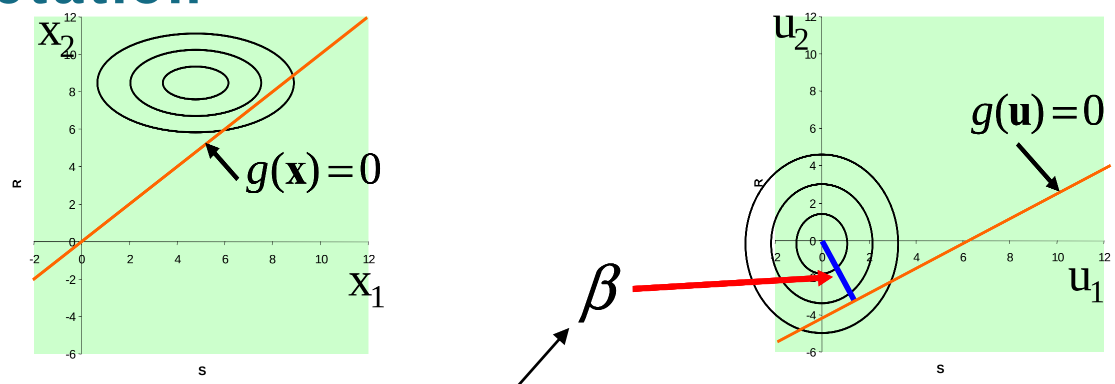
$$\beta = \frac{\mu_M}{\sigma_M}$$

The safety margin

Basics of Reliability Analysis

- **Linear limit state functions and normal distributed basic variables**

  The reliability index β has a geometrical interpretation



Smallest distance between the origin and the limit state function in standardized normal distributed space

$$U_i = \frac{X_i - \mu_{X_i}}{\sigma_{X_i}}$$

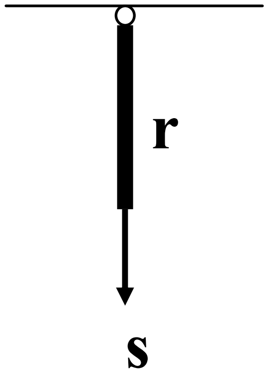
Zero mean and unit variance

Basics of Reliability Analysis

- **Linear limit state functions and normal distributed basic variables**

Example : Reliability of steel rod under tension loading

The resistance *R* and the max annual loading *S* are both assumed to be normal distributed



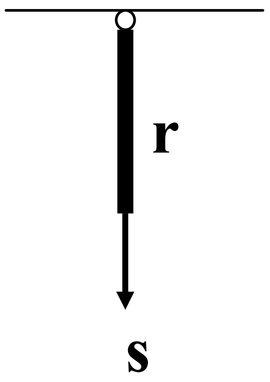
$$\mu_R = 350, \sigma_R = 35$$

$$\mu_S = 200, \sigma_S = 40$$

Basics of Reliability Analysis

- **Linear limit state functions and normal distributed basic variables**

Example : Reliability of steel rod under tension loading



The safety margin is thus normal distributed with parameters

$$\mu_M = 350 - 200 = 150$$

$$\sigma_M = \sqrt{35^2 + 40^2} = 53.15$$

The reliability index β becomes

$$\beta = \frac{150}{53.15} = 2.84$$

$$P_F = \Phi(-2.84) = 2.4 \cdot 10^{-3}$$

Basics of Reliability Analysis

- **The error accumulation law**

  In many engineering applications the accumulation of errors is a central question

  Examples are :

  - errors due to fabrication tolerances of building components

  - errors in connection with surveying

  - errors in connection with measurements performed in the laboratory

Basics of Reliability Analysis

- **The error propagation law**

Assume that the error ε can be written as a differentiable function of random variables i.e. :

$$\varepsilon = h(\mathbf{x}) \qquad \mathbf{x} = (x_1, x_2, .., x_n)^T$$

← **Vector of realization of basic random variables**

with parameters

$$\boldsymbol{\mu}_{\mathbf{X}} = (\mu_{X_1}, \mu_{X_2}, .., \mu_{X_n})^T \qquad Cov\left[X_i, X_j\right] = \rho_{ij}\sigma_{X_i}\sigma_{X_j}$$

Correlation coefficient (→ ρ_{ij})

Standard deviation (→ σ_{X_i})

The idea is to linearize *f(x)*

$$\varepsilon \cong h(\mathbf{x}_0) + \sum_{i=1}^{n} (x_i - x_{i,0}) \left.\frac{\partial f(\mathbf{x})}{\partial x_i}\right|_{\mathbf{x}=\mathbf{x}_0}$$

First order partial derivative taken in $\mathbf{x} = \mathbf{x}_0$

Basics of Reliability Analysis

- **The error propagation law**

If we linearize the error function around the mean value of the random variables its expected value and variance becomes :

$$\varepsilon \cong h(\boldsymbol{\mu}_{\mathbf{X}}) + \sum_{i=1}^{n} (x_i - \mu_{X_i}) \left. \frac{\partial h(\mathbf{x})}{\partial x_i} \right|_{\mathbf{x}=\boldsymbol{\mu}_{\mathbf{X}}}$$

$$E[\varepsilon] = h(\boldsymbol{\mu}_{\mathbf{X}})$$

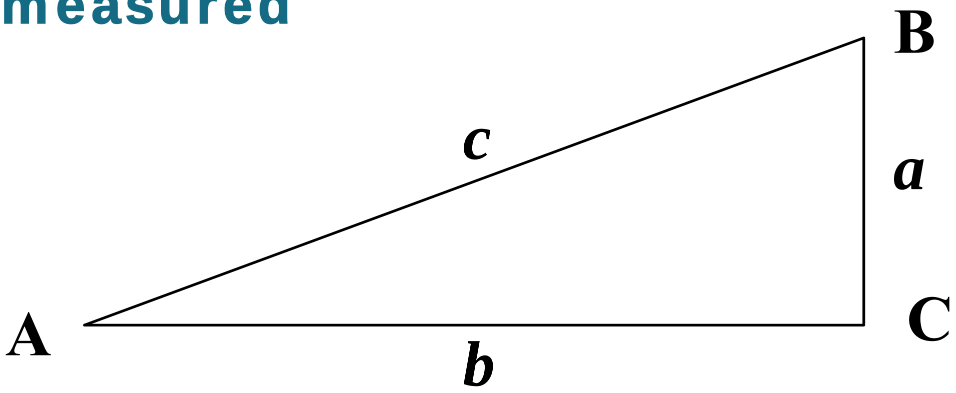
$$Var[\varepsilon] = \sum_{i=1}^{n} \left(\left. \frac{\partial h(\mathbf{x})}{\partial x_i} \right|_{\mathbf{x}=\boldsymbol{\mu}_{\mathbf{X}}} \right)^2 \sigma_{X_i}^2 + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(\left. \frac{\partial h(\mathbf{x})}{\partial x_i} \right|_{\mathbf{x}=\boldsymbol{\mu}_{\mathbf{X}}} \right) \left(\left. \frac{\partial h(\mathbf{x})}{\partial x_j} \right|_{\mathbf{x}=\boldsymbol{\mu}_{\mathbf{X}}} \right) \rho_{ij} \sigma_{X_i} \sigma_{X_j}$$

The mean value and the variance depends on the linearization point

Basics of Reliability Analysis

- **Example : Error propagation in measurements**

In order to estimate the length *c* i.e. the distance between the two points A and B the lengths *a* and *b* are measured



due to measurement uncertainty in assessing *a* and *b* also the length of *c* will be associated with uncertainty and it is of interest to know the probability that the length of *c* will exceed 13.5

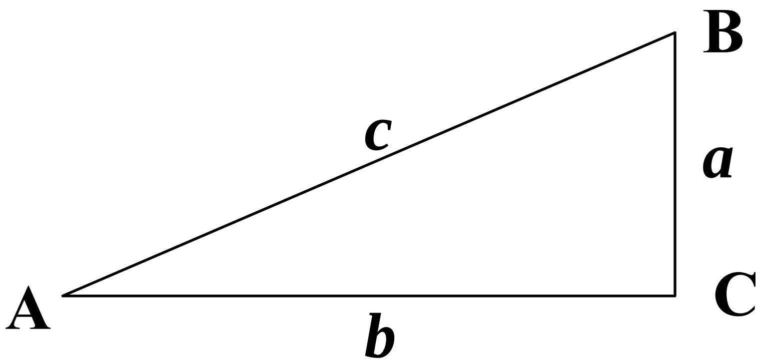
Basics of Reliability Analysis

- **Example : Error propagation in measurements**

It is assumed that *a* and *b* can be modeled as normal distributed random variables with parameters

$$\mu_a = 12.2 \quad \mu_b = 5.1$$

$$\sigma_a = 0.4 \quad \sigma_b = 0.3$$



Using that *c* can be given as

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

the statistical characteristics of *c* may be estimated through the error propagation law

Basics of Reliability Analysis

- **Example : Error propagation in measurements**

$$E[c] = \sqrt{\mu_a^2 + \mu_b^2}$$

$$Var[c] = \sum_{i=1}^{n} \left(\left. \frac{\partial h(\mathbf{x})}{\partial x_i} \right|_{\mathbf{x}=\boldsymbol{\mu}_\mathbf{X}} \right)^2 \sigma_{X_i}^2 = \frac{\mu_a}{\sqrt{\mu_a{}^2 + \mu_b{}^2}} \sigma_a^2 + \frac{\mu_b}{\sqrt{\mu_a{}^2 + \mu_b{}^2}} \sigma_b^2$$

⇓

$$E[c] = \sqrt{12.2^2 + 5.1^2} = 13.22$$

$$Var[c] = \frac{12.2^2}{12.2^2 + 5.1^2} 0.4^2 + \frac{5.1^2}{12.2^2 + 5.1^2} 0.3^2 = 0.15$$

$$P_f = P(13.5 - C \le 0) = \Phi(-\frac{13.5 - 13.22}{\sqrt{0.15}}) = 0.2349$$

Basics of Reliability Analysis

- **Non-linear limit state functions**

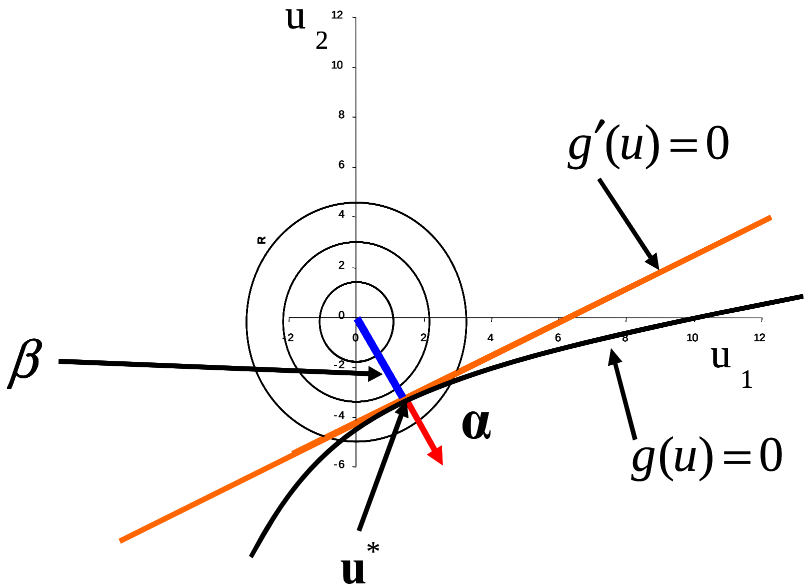
Limit state functions are often non-linear

As seen from the error propagation law it is possible to linearize such limit state functions but the results will depend on the linearization point and on the formulation of the limit state function

Basics of Reliability Analysis

- **Non-linear limit state functions**

Limit state functions are often non-linear



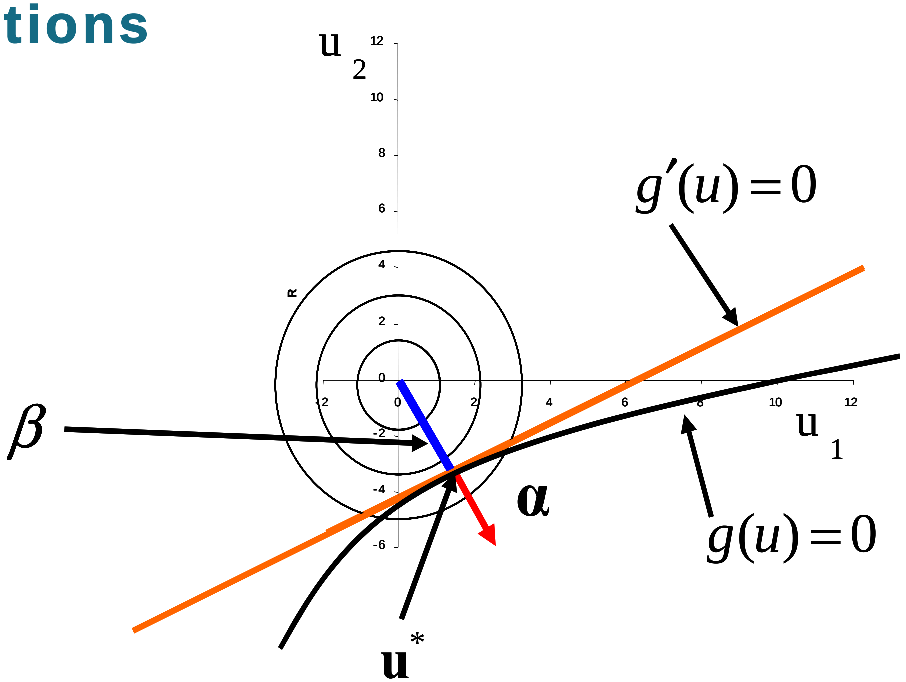
Hasofer and Lind suggested to linearize in the point where the limit state function is zero and closest to the origin in normal distributed space

Basics of Reliability Analysis

- **Non-linear limit state functions**

  The identification of the reliability index may be performed by solving an optimization problem

$$\beta = \min_{\mathbf{u}\in\{g(\mathbf{u})=0\}} \sqrt{\sum_{i=1}^{n} u_i^2}$$



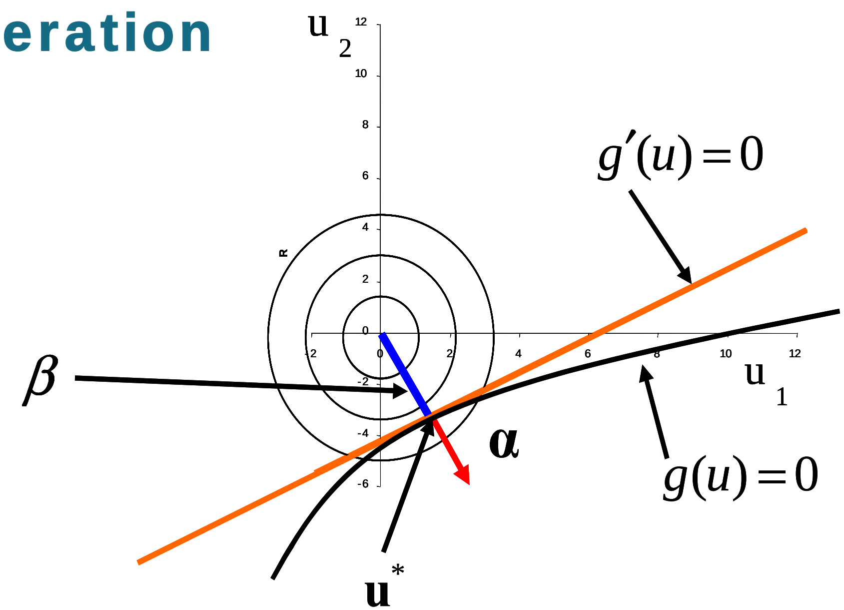
Basics of Reliability Analysis

- **Non-linear limit state functions**

The optimization problem may be solved using the following iteration scheme

$$\alpha_i = \frac{-\dfrac{\partial g}{\partial u_i}(\beta\boldsymbol{\alpha})}{\left[\sum_{j=1}^{n}\dfrac{\partial g}{\partial u_i}(\beta\boldsymbol{\alpha})^2\right]^{1/2}}, \quad i = 1, 2, ..n$$

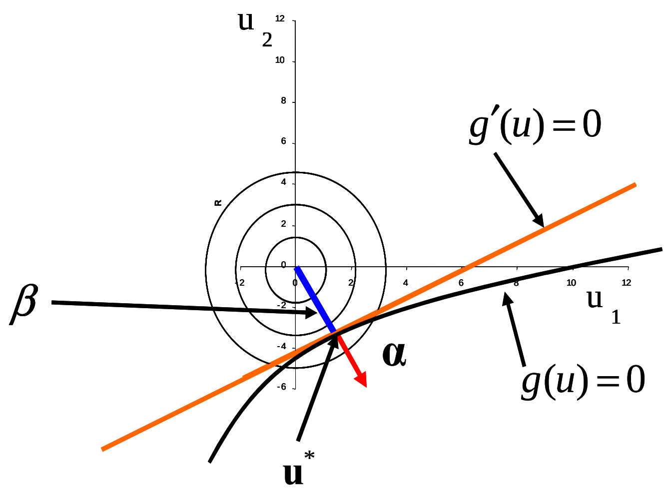
$$g(\beta\alpha_1, \beta\alpha_2, ...\beta\alpha_n) = 0$$



Provided that the limit state function is differentiable !

Basics of Reliability Analysis

In summary the iteration follows the following steps



1) the linearization point is chosen as $u^* = \beta\,\alpha$

2) the Normal vector to the limit state function is determined in the linearization point

$$\alpha_i = \frac{-\dfrac{\partial g}{\partial u_i}(\beta \boldsymbol{\alpha})}{\left[\displaystyle\sum_{j=1}^{n} \frac{\partial g}{\partial u_i}(\beta \boldsymbol{\alpha})^2\right]^{1/2}}, \quad i = 1, 2, ..n$$

3) the reliability index β is calculated from

$$g(\beta\alpha_1, \beta\alpha_2, ...\beta\alpha_n) = 0$$

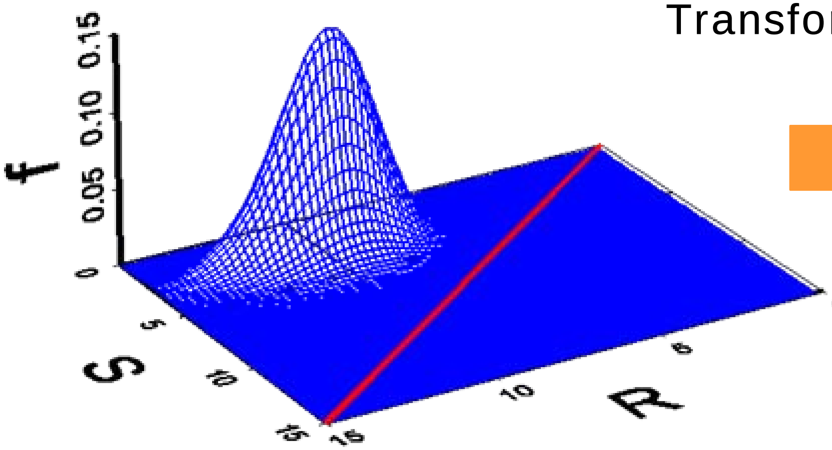
4) the new linearization point is

$$u^* = \left(\beta\alpha_1, \beta\alpha_2, ...\beta\alpha_n\right)^T$$

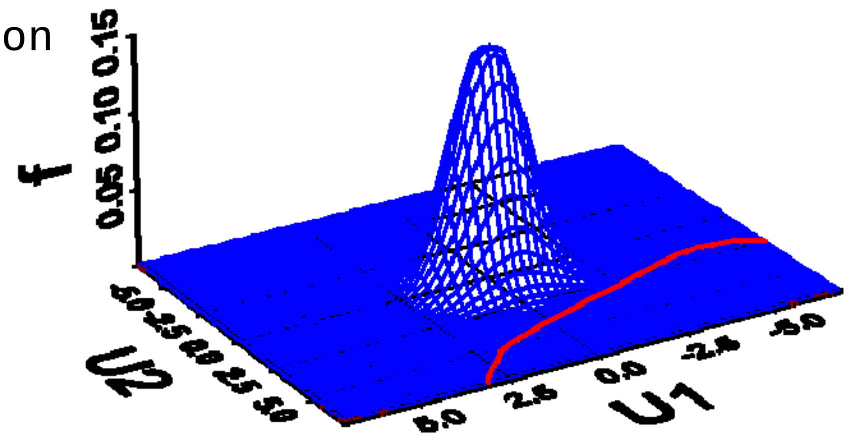
5) continue with step 2) until convergence in β

Basics of Reliability Analysis

Non-linear safety margins



Transformation



$g(Z)$: linear

$\mu_{Z1}, \mu_{Z2} \in R$

$\sigma_{Z1}, \sigma_{Z2} \in R$

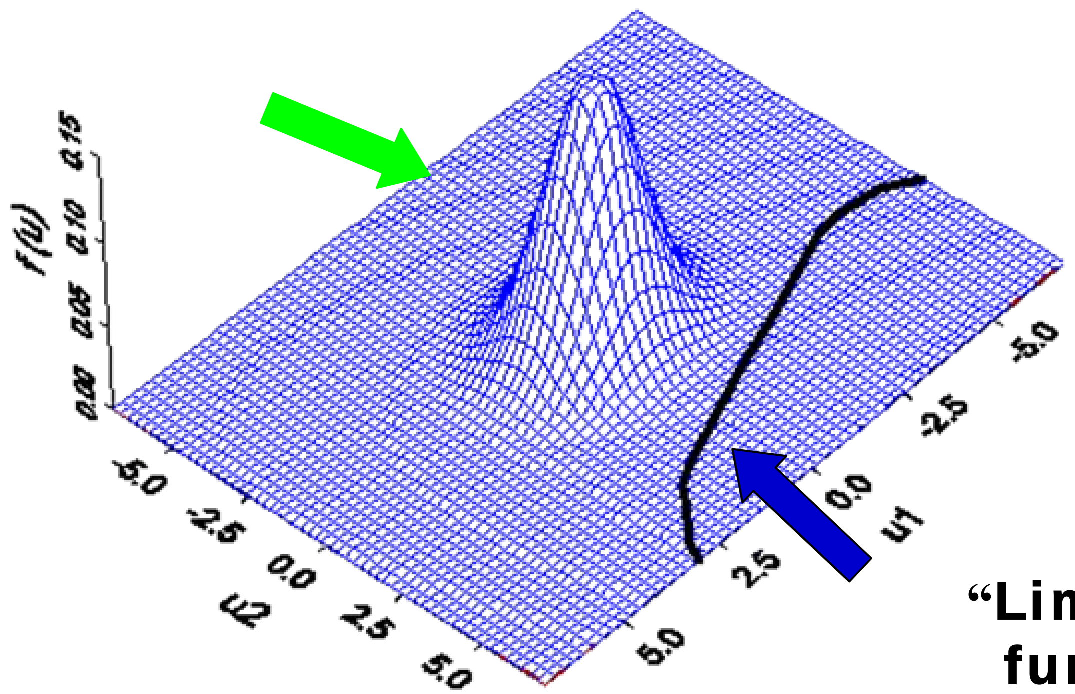
$g(U)$: non-linear

$\mu_{U1} = \mu_{U2} = 0$

$\sigma_{U1} = \sigma_{U2} = 1$

Basics of Reliability Analysis

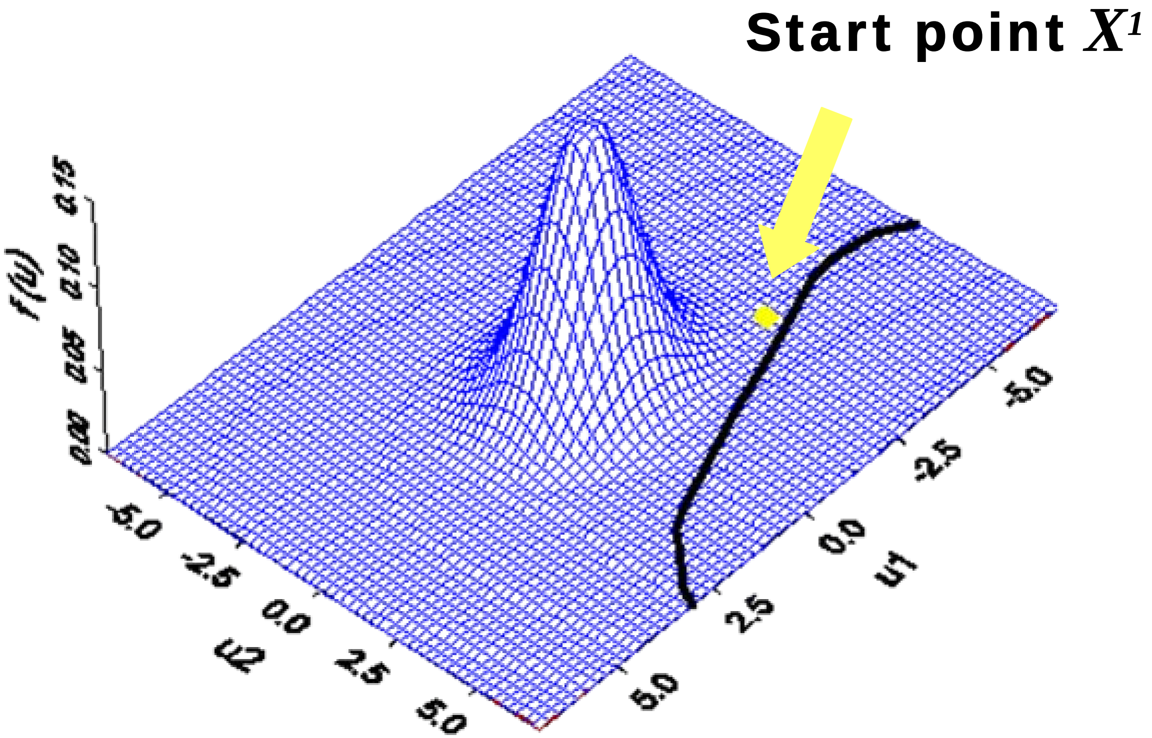
- **Non-linear safety margins**



"Limit state function"

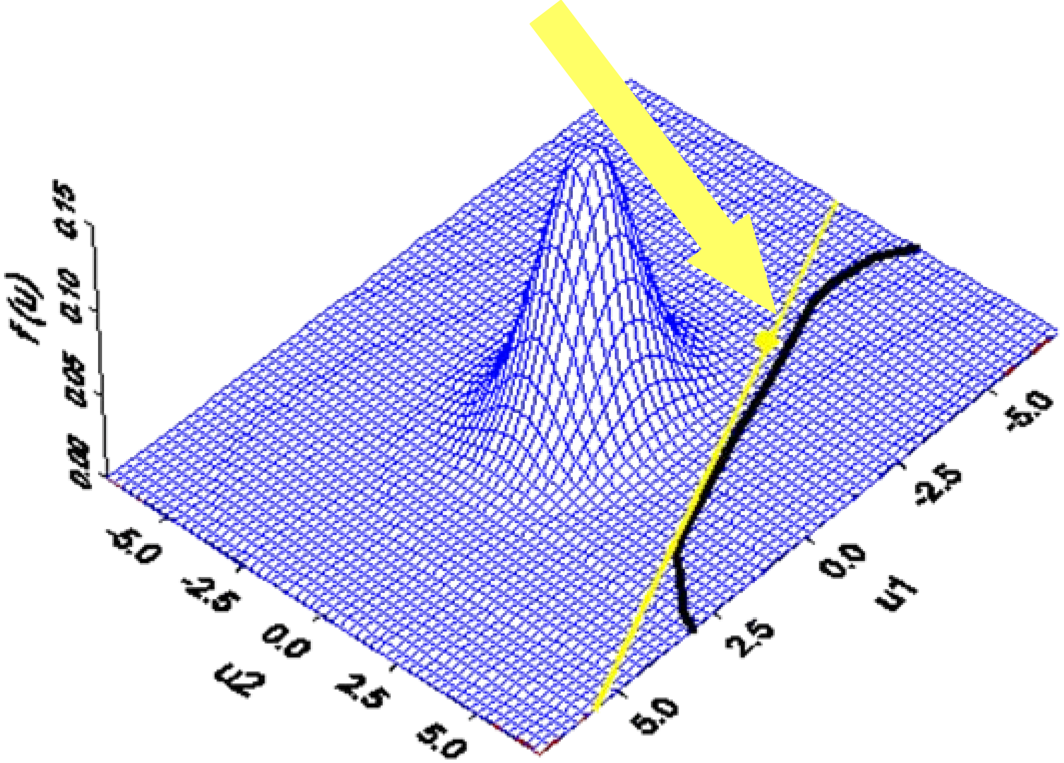
$$g(U) = R\text{-}S$$

Basics of Reliability Analysis



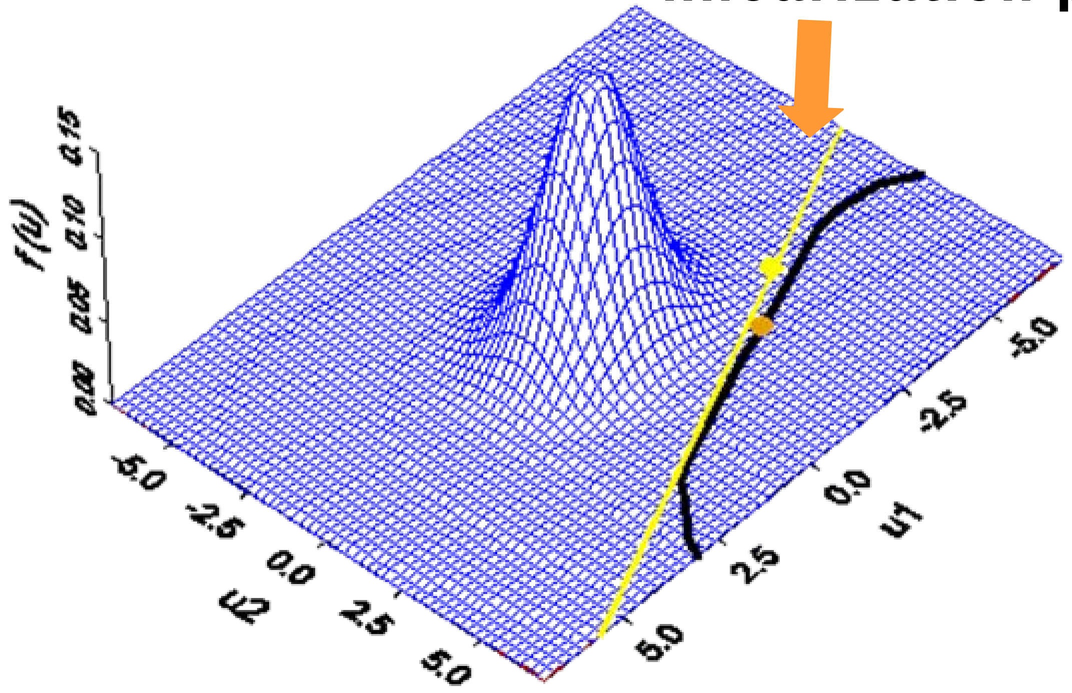
Basics of Reliability Analysis

Linearization of limit state function in *X*¹



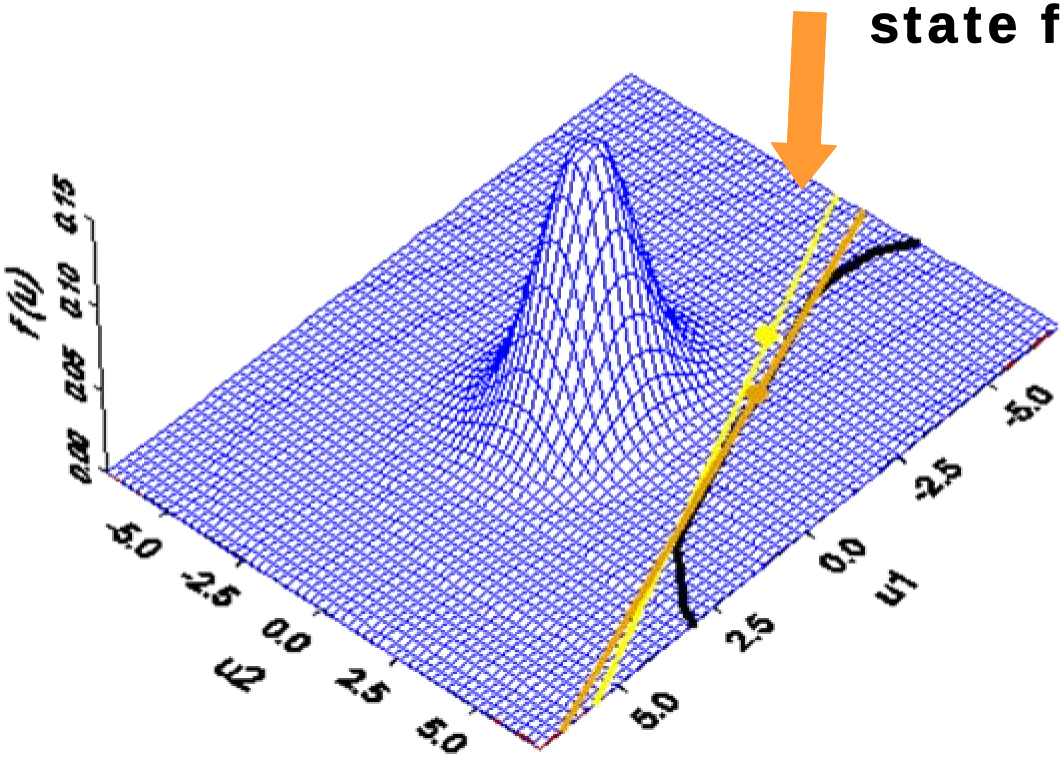
Basics of Reliability Analysis

Calculation of a new linearization point X^2



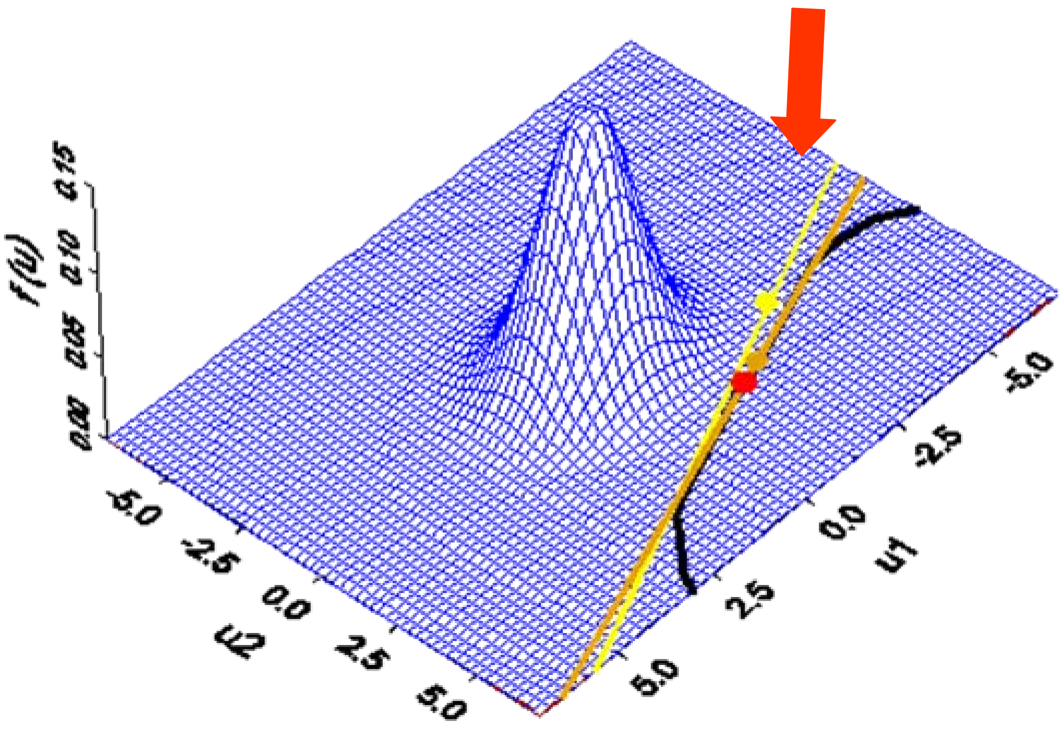
Basics of Reliability Analysis

Linearization of limit state function in X^2



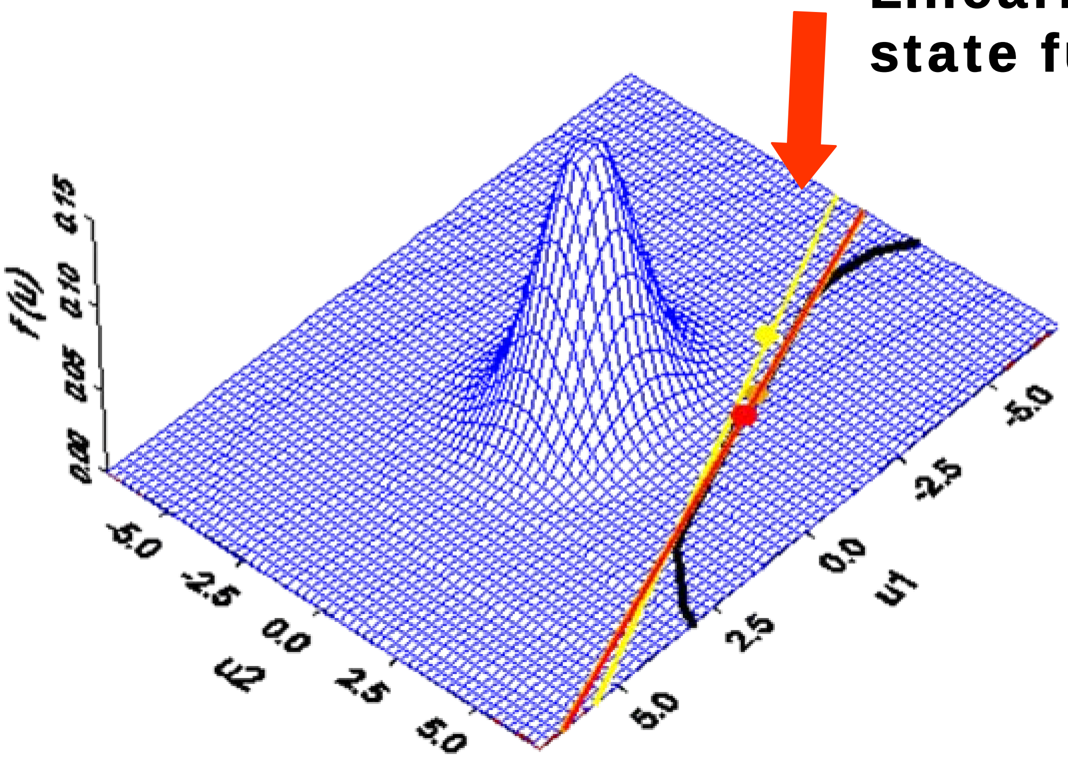
Basics of Reliability Analysis

Calculation of new linearization point X^3

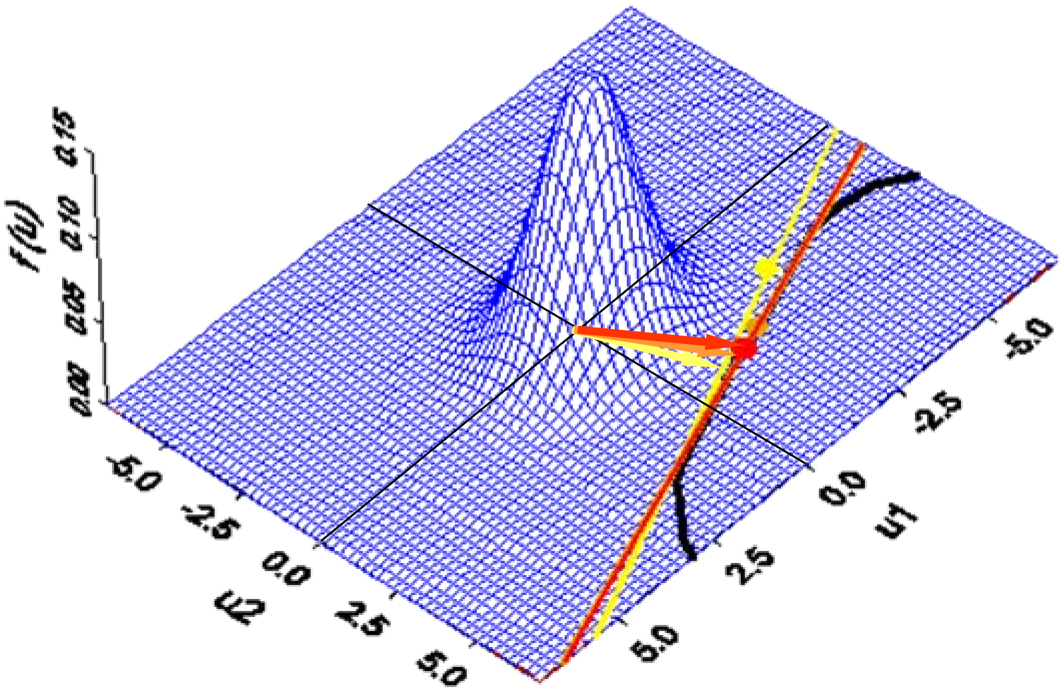


Basics of Reliability Analysis

Linearization of limit state function in X^3



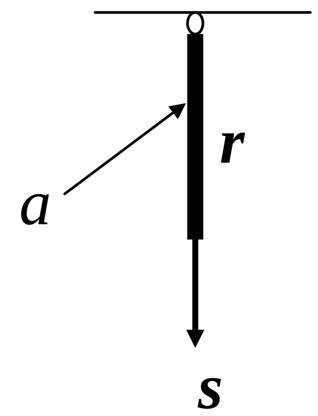
Basics of Reliability Analysis



β^1=3.556	
β^2=3.607	
β^3=3.608	
β^4=3.608	

Convergence criteria : $\Delta\beta = \left|\beta^{n+1} - \beta^{n}\right| \leq \varepsilon$

Basics of Reliability Analysis



- **Example : Reliability of steel rod**

Limit state function

Yield stress (r)

$$g(\mathbf{x}) = r \cdot a - s$$

← **Load** (s)

Cross sectional area (a)

it is assumed that R, S and A are normal distributed random variables

$$U_R = \frac{R - \mu_R}{\sigma_R} \qquad U_S = \frac{S - \mu_S}{\sigma_S} \qquad U_A = \frac{A - \mu_A}{\sigma_A}$$

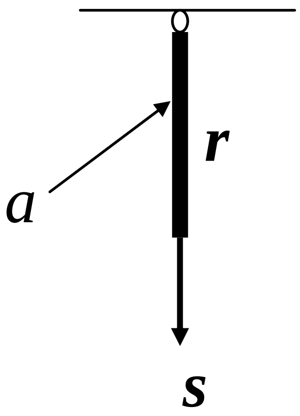
$$\mu_R = 350, \sigma_R = 35$$

$$\mu_S = 1500, \sigma_S = 300$$

$$\mu_A = 10, \sigma_A = 2$$

Basics of Reliability Analysis

- **Example : Reliability of steel rod**



We can now write the limit state function in terms of *u*-variables

$$\begin{array}{ccc} \boldsymbol{r} & \boldsymbol{a} & \boldsymbol{s} \end{array}$$

$$g(u) = (u_R\boldsymbol{\sigma}_R + \boldsymbol{\mu}_R)(u_A\boldsymbol{\sigma}_A + \boldsymbol{\mu}_A) - (u_S\boldsymbol{\sigma}_S + \boldsymbol{\mu}_S)$$

$$= (35u_R + 350)(u_A + 10) - (300u_S + 1500)$$

$$= 350\mathrm{u}_{\mathrm{R}} + 350u_A - 300u_S + 35u_Ru_A + 2000$$

Basics of Reliability Analysis

- **Example : Reliability of steel rod**

  The reliability index β may be found by iteration

$$\alpha_R = -\frac{1}{k}(350 + 35\beta\alpha_A)$$

$$\alpha_A = -\frac{1}{k}(350 + 35\beta\alpha_R) \qquad k = \sqrt{\alpha_R^2 + \alpha_A^2 + \alpha_S^2}$$

$$\alpha_S = \frac{300}{k}$$

$$\beta = \frac{-2000}{350\alpha_R + 350\alpha_A - 300\alpha_S + 35\beta\alpha_R\alpha_A}$$

Iteration	Start	1	2	3	4	5
β	3.0000	3.6719	3.7399	3.7444	3.7448	3.7448
α_R	-0.5800	-0.5701	-0.5612	-0.5611	-0.5610	-0.5610
α_A	-0.5800	-0.5701	-0.5612	-0.5611	-0.5610	-0.5610
α_S	0.5800	0.5916	0.6084	0.6086	0.6087	0.6087

$$\alpha_i = \frac{-\frac{\partial g}{\partial u_i}(\beta\boldsymbol{\alpha})}{\left[\sum_{j=1}^{n}\frac{\partial g}{\partial u_i}(\beta\boldsymbol{\alpha})^2\right]^{1/2}}, \quad i = 1, 2, ..n$$

$$g(\beta\alpha_1, \beta\alpha_2, ...\beta\alpha_n) = 0$$

$$g(u) = (u_R\boldsymbol{\sigma}_R + \boldsymbol{\mu}_R)(u_A\boldsymbol{\sigma}_A + \boldsymbol{\mu}_A) - (u_S\boldsymbol{\sigma}_S + \boldsymbol{\mu}_S)$$

$$= (35u_R + 350)(u_A + 10) - (300u_S + 1500)$$

$$= 350\mathrm{u}_R + 350u_A - 300u_S + 35u_Ru_A + 2000$$

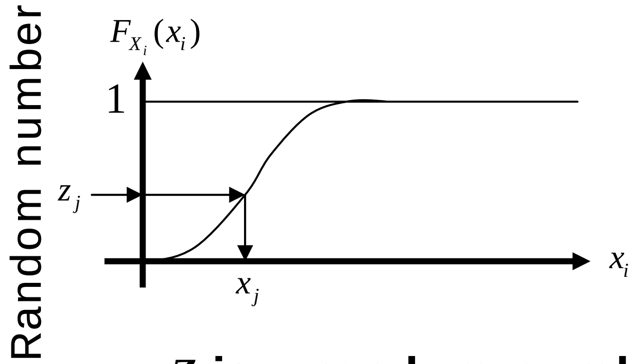
Basics of Reliability Analysis

- **Monte Carlo Simulation**

  The probability integration problem may be solved by Monte Carlo simulation

$$P_f = \int_{\Omega_f = \{g(\mathbf{x}) \le 0\}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

1) m **realizations of the vector** $\mathbf{X}$ **are produced**



2) for every realization the limit state function is calculated

Z is a random number uniformly distributed between 0 and 1

3) the realizations for which the limit state function is equal to or less than zero are counted

n_f

4) The probability of failure is estimated as

$$p_f = \frac{n_f}{m}$$

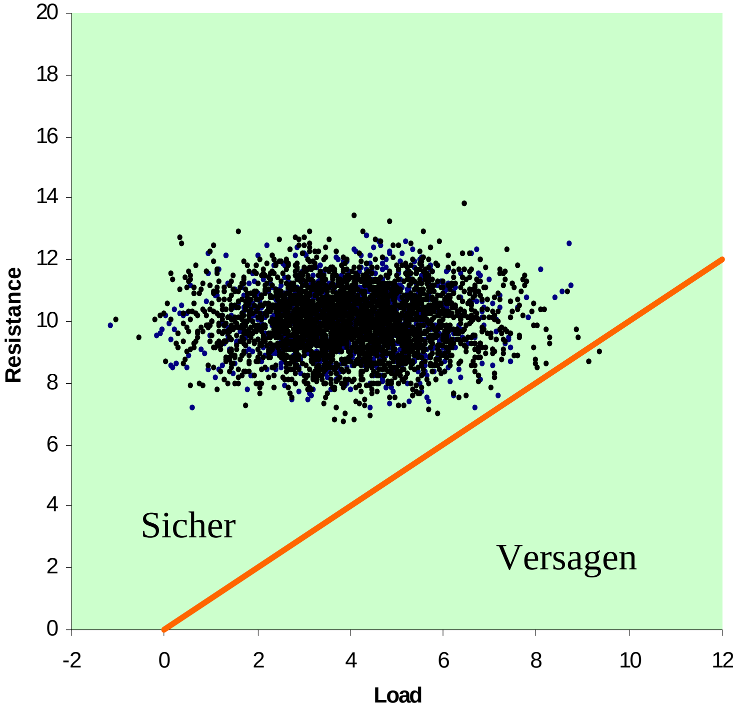
Basics of Reliability Analysis

- **Monte Carlo Simulation**

m random realizations of R and S are generated and the number of realizations n_f occuring in the failure space are counted n_f

The probability of failure p_f is then

$$p_f = \frac{n_f}{m}$$



Statistics and Probability Theory

in

Civil, Surveying and Environmental

Engineering

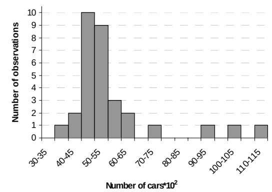
Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

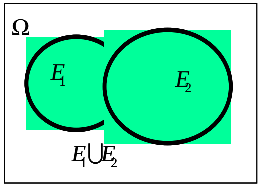
ETH Zurich, Switzerland

A Summary of the Lecture

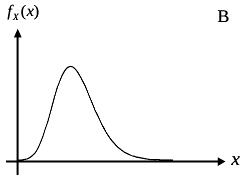
Graphical / numerical interpretation of data



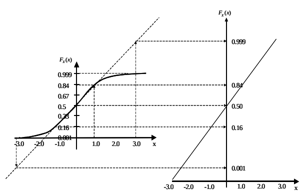
Basic probability theory



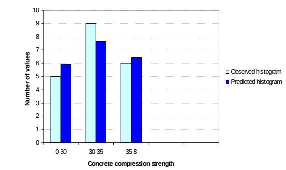
Distribution functions moments and extremes



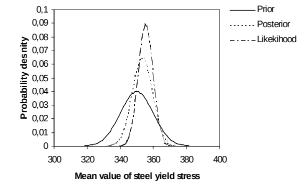
Modeling and Description of data



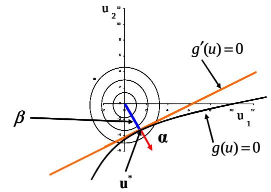
Verification and testing of models



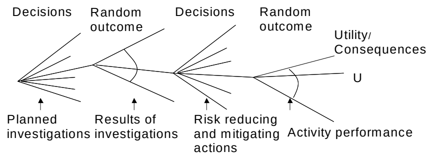
Bayesian modeling



Basic reliability analysis



Basic decision analysis



Decision Analysis in Engineering

- Introduction to Decision Theory
  - The problem
  - The decision tree
  - Prior decision analysis
  - Posterior decision analysis
  - Pre-posterior decision analysis

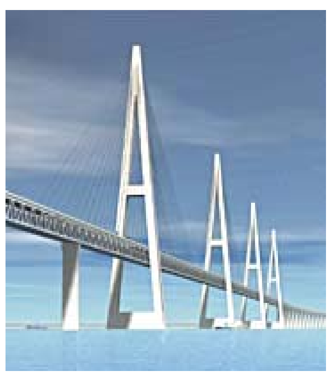
Decision Analysis in Engineering

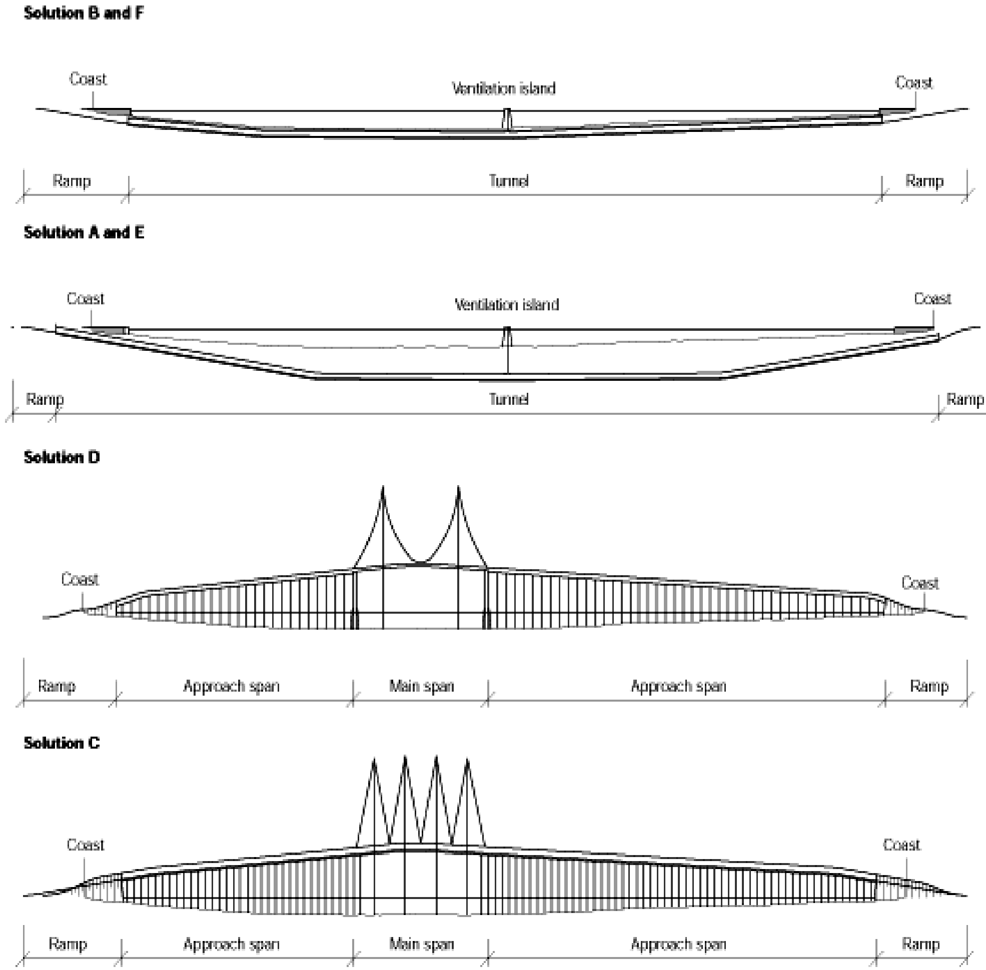
- **The basic engineering problem**

Several solutions may be identified

The available information is uncertain

A decision must be made
!





Decision Analysis in Engineering

Approach

- **Formulation of the decision problem**
  - The decision maker and the preferences of the decision maker must be identified
  - Mapping of the decision process
  - All the possible decision alternatives must be identified
  - Identification of the contributing uncertainties

- **Identification of potential consequences and their utility (cost / benefit)**

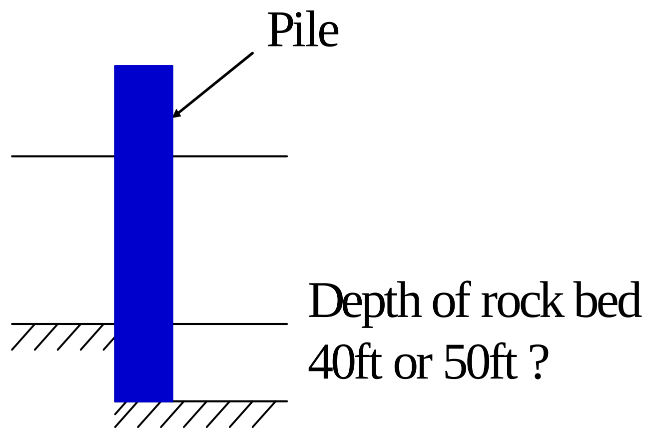
- **Assessment of the probabilities of the consequences**

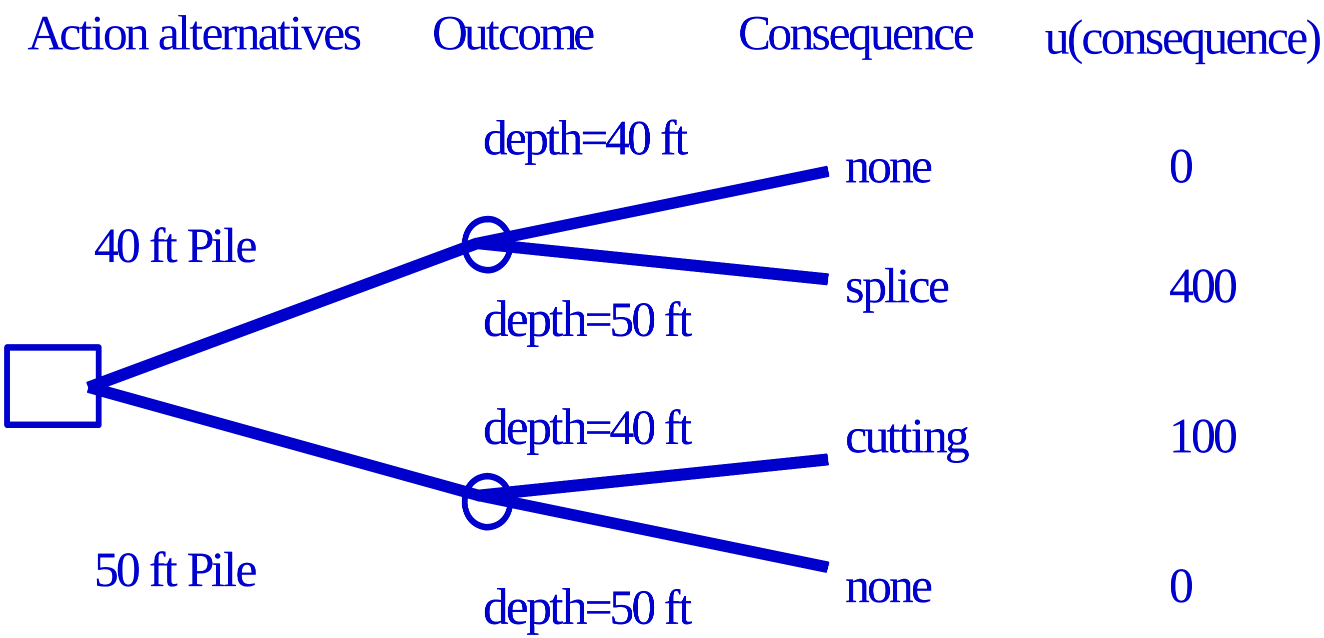
- **Comparison of the different decision alternatives based on their expected utilities**

- **Final decision making and reporting of the assumptions underlying the selected alternative**

Decision Analysis in Engineering

- **The decision tree**





Decision Analysis in Engineering

Assignment of utility

- The assignment of utility must reflect the preferences of the decision maker
- Utility functions may be defined as linear functions in monetary unity
- It is important to include all monetary consequences in the utility function

$$u(a_i) = \sum_{j=1}^{n} p_j \cdot u(K_j)$$

$u(a_i)\ldots$ Utility (cost/benefit) associated with action a_i

$p_j \cdot u(K_j)\ldots$ Expected utility associated with consequence K_j

$p_j\ldots$ Probability of the occurence of the consequence K_j

$u(K_j)\ldots$ Utility associated with the consequence K_j

$K_j\ldots$ A potential consequence associated with the action a_i

Decision Analysis in Engineering

The different types of decision analysis

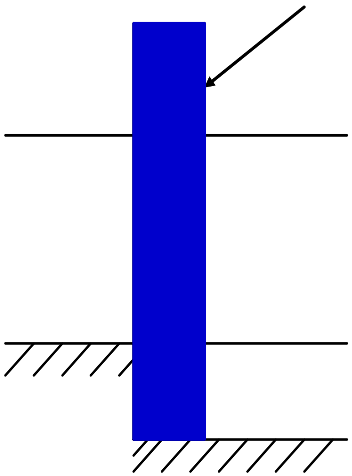
- **Prior**
- **Posterior**
- **Pre-posterior**

Illustrated on an example :

Question : What pile length should be applied ?

Alternatives :
a_0 : Choose a 40 ft pile
a_1 : Choose a 50 ft pile

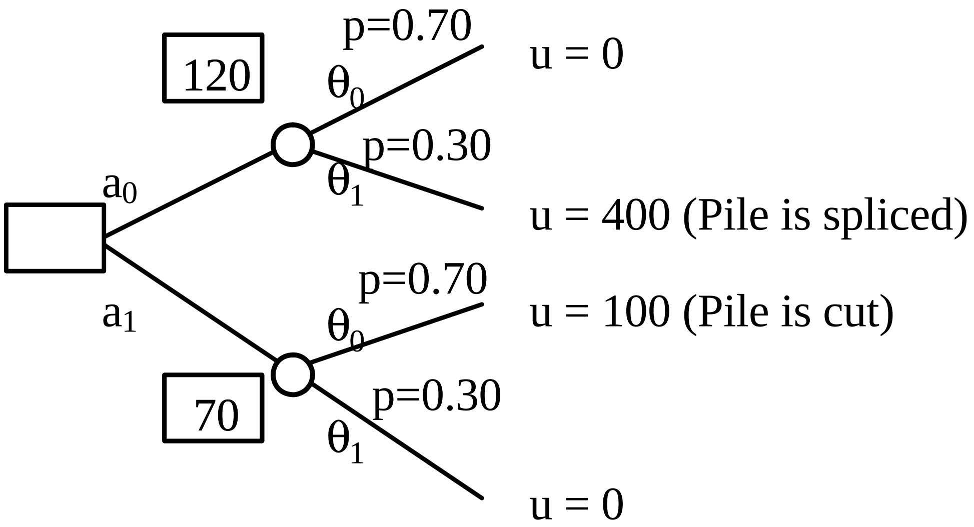
States of nature
(depth to rock bed)
0 : Rock bed in 40 ft
1 : Rock bed at 50 ft



Decision Analysis in Engineering

Prior Analysis

$P'[q_0] = 0.70$
$P'[q_1\,] = 0.30$



The expected utility is calculated to be equal to

$$E'[u] = \min\{u[a_0], u[a_1]\}$$

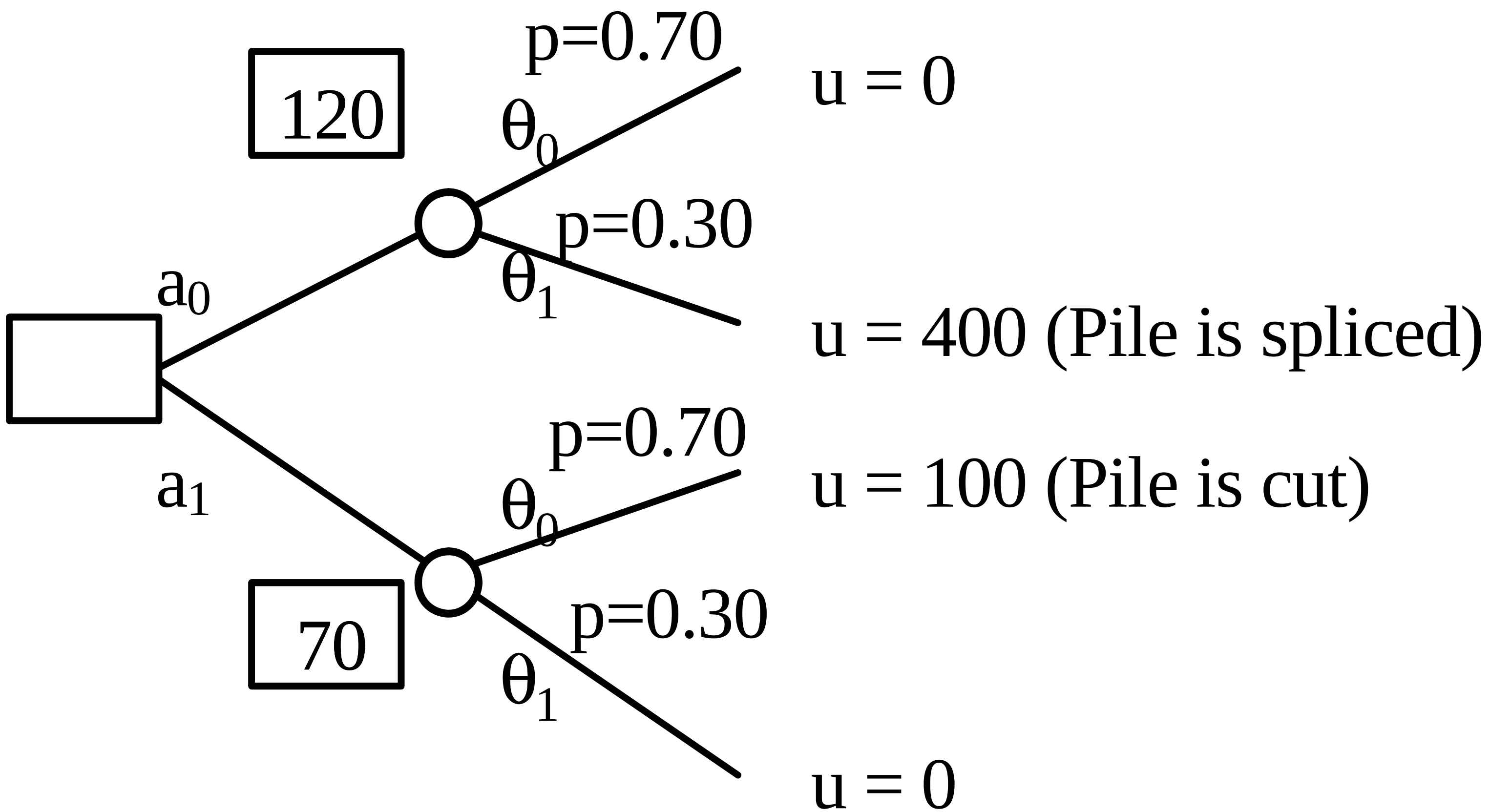
$$= \min\{P'[\theta_0]\times u[\theta_0|a_0] + P'[\theta_1]\times u[\theta_1|a_0],$$

$$P'[\theta_0]\times u[\theta_0|a_1] + P'[\theta_1]\times u[\theta_1|a_1]\}$$

$$= \min\{0.7\times 0 + 0.3\times 400,\ 0.7\times 100 + 0.3\times 0\}$$

$$= \min\{120,70\} = 70 \quad \Rightarrow \text{Decision for } a_1 \text{ (50ft Pile)}$$

Decision Analysis in Engineering



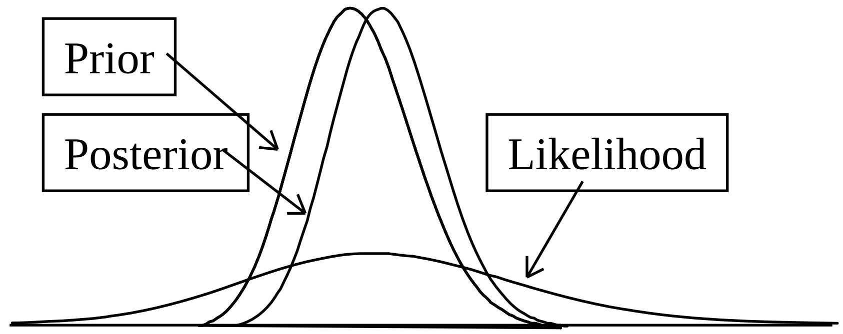
$\Rightarrow$ **Choice of pile a_1 (50ft Pfahl)**

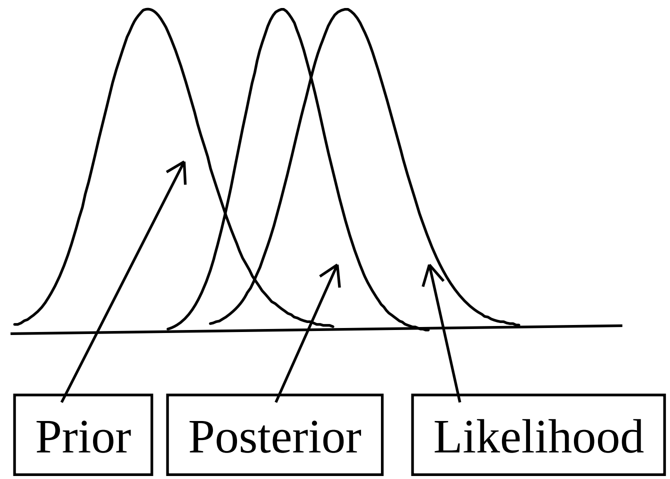
Posterior Analysis

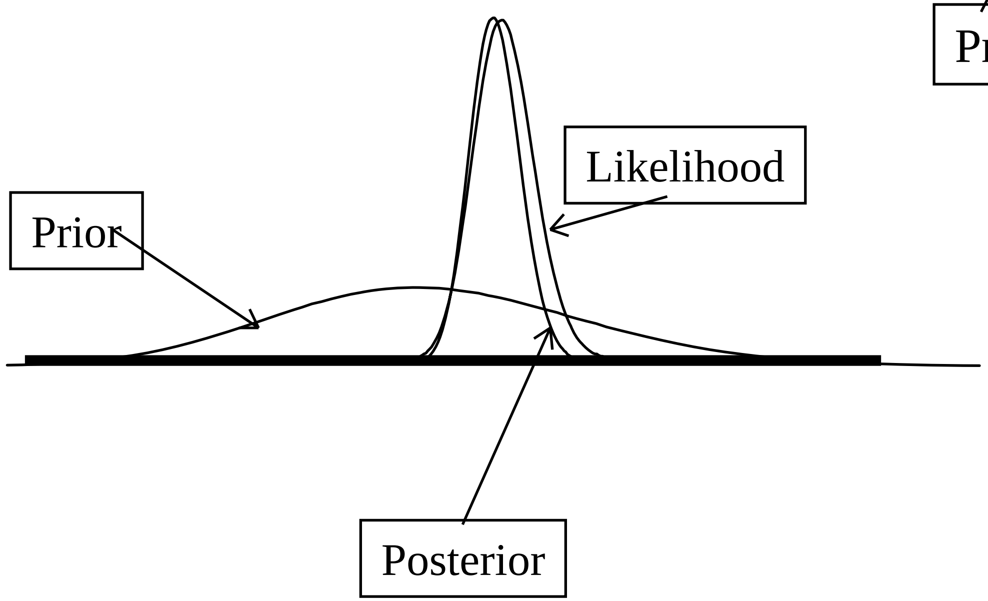
$$P''(\theta_i) = \frac{P\left[z_k \middle| \theta_i\right] P'\left[\theta_i\right]}{\sum_j P\left[z_k \middle| \theta_j\right] P'\left[\theta_j\right]}$$

$$\begin{pmatrix}\text{Posterior probability of } \theta_i \\ \text{with given sample outcome}\end{pmatrix} = \begin{pmatrix}\text{Normalizing} \\ \text{constant}\end{pmatrix} \text{x} \begin{pmatrix}\text{Sample likelihood} \\ \text{given } \theta\end{pmatrix} \text{x} \begin{pmatrix}\text{prior probability} \\ \text{of } \theta\end{pmatrix}$$

Posterior Analysis







Decision Analysis in Engineering

$$P''(\theta_i) = \frac{P[z_k|\theta_i]P'[\theta_i]}{\sum_j P[z_k|\theta_j]P'[\theta_j]}$$

Posterior Analysis

Ultrasonic tests to determine the depth to bed rock

True state / Test result	θ_0 40 ft – depth	θ_1 50 ft – depth
z_0 - 40 ft indicated	0.6	0.1
z_1 - 50 ft indicated	0.1	0.7
z_2 - 45 ft indicated	0.3	0.2

Likelihoods of the different indications/test results given the various possible states of nature – ultrasonic test methods

$$P[z_i|\theta_j]$$

Decision Analysis in Engineering

$$P''(\theta_i) = \frac{P[z_k|\theta_i]P'[\theta_i]}{\sum_j P[z_k|\theta_j]P'[\theta_j]}$$

Posterior Analysis

It is assumed that a test gives a 45 ft indication

$$P''[\theta_0] = P[\theta_0|z_2] \propto P[z_2|\theta_0]P[\theta_0] = 0.3\,x\,0.7 = 0.21$$

$$P''[\theta_1] = P[\theta_1|z_2] \propto P[z_2|\theta_1]P[\theta_1] = 0.2\,x\,0.3 = 0.06$$

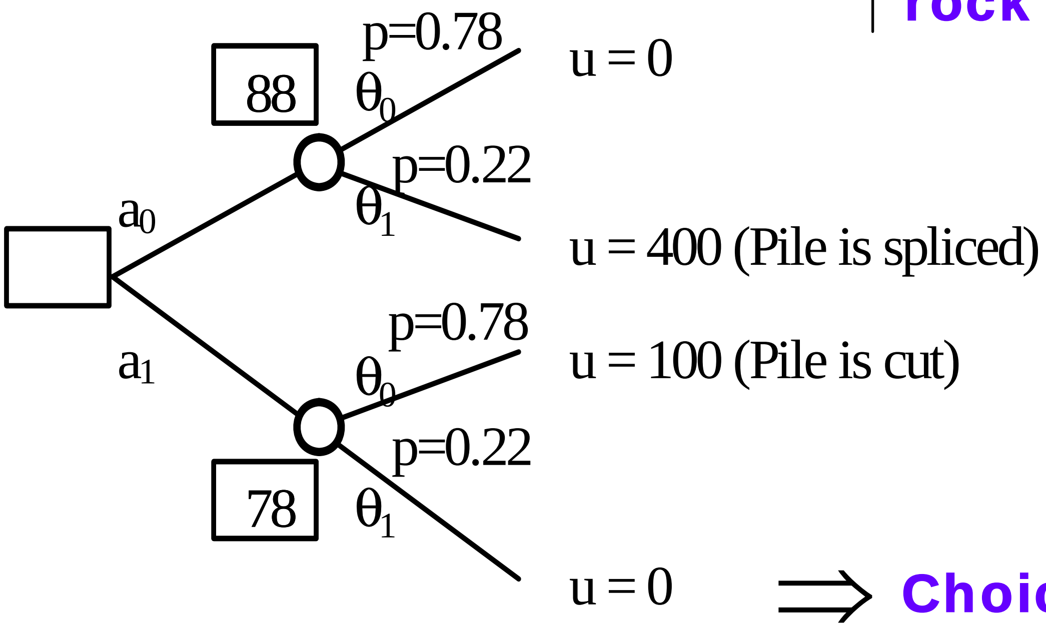
$$P''[\theta_0|z_2] = \frac{0.21}{0.21+0.06} = 0.78$$

$$P''[\theta_1|z_2] = \frac{0.06}{0.21+0.06} = 0.22$$

Decision Analysis in Engineering

Posterior Analysis

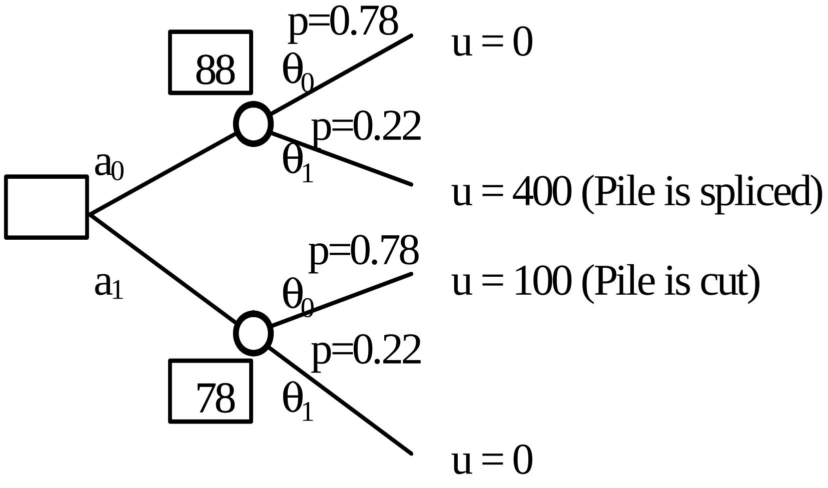
Test result indicates 45ft to rock bed



$\Longrightarrow$ **Choice of alternative a_1 (50ft Pile)**

Decision Analysis in Engineering

Posterior Analysis



$$E''\left[u \middle| z_2\right] = \min_j \{E''\left[u(a_j) \middle| z_2\right]\}$$

$$= \min\{P''[\theta_0] \times 0 + P''[\theta_1] \times 400, P''[\theta_0] \times 100 + P''[\theta_1] \times 0\}$$

$$= \min\{0.78 \times 0 + 0.22 \times 400, 0.78 \times 100 + 0.22 \times 0\}$$

$$= \min\{88, 78\} = 78$$

$\Rightarrow$ **Choice of alternative a_1 (50ft Pile)**

Pre-posterior Analysis

$$E[u] = \sum_{i=1}^{n} P'[z_i] \times E''[u|z_i] = \sum_{i=1}^{n} P'[z_i] \times \min_{j=1,m}\{E''[u(a_j)|z_i]\}$$

$$P'[z_i] = P[z_i|\theta_0] \times P'[\theta_0] + P[z_i|\theta_1] \times P'[\theta_1]$$

$$P'[z_0] = P[z_0|\theta_0] \times P'[\theta_0] + P[z_0|\theta_1] \times P'[\theta_1] = 0.6 \times 0.7 + 0.1 \times 0.3 = 0.45$$

$$P'[z_1] = P[z_1|\theta_0] \times P'[\theta_0] + P[z_1|\theta_1] \times P'[\theta_1] = 0.1 \times 0.7 + 0.7 \times 0.3 = 0.28$$

$$P'[z_2] = P[z_2|\theta_0] \times P'[\theta_0] + P[z_2|\theta_1] \times P'[\theta_1] = 0.3 \times 0.7 + 0.2 \times 0.3 = 0.27$$

Decision Analysis in Engineering

Pre-posterior Analysis

$$E''\left[u \middle| z_0\right] = \min_j \left\{ E''\left[u(a_j) \middle| z_0\right] \right\} =$$

$$\min\{\underbrace{\underbrace{P''\left[\theta_0 \middle| z_0\right]\times 0}_{\textbf{do nothing}} + \underbrace{P''\left[\theta_1 \middle| z_0\right]\times 400}_{\textbf{splicing}}}_{\mathbf{a_0}},\ \underbrace{\underbrace{P''\left[\theta_0 \middle| z_0\right]\times 100}_{\textbf{cutting}} + \underbrace{P''\left[\theta_1 \middle| z_0\right]\times 0}_{\textbf{do nothing}}}_{\mathbf{a_1}}\}$$

$$\min\{0.93\times 0 + 0.07\times 400,\ 0.93\times 100 + 0.07\times 0\} =$$

$$0.07\times 400 + 0.93\times 0 = 28$$

Pre-posterior Analysis

$$E''[u|z_1] = \min_j \{E''[u(a_j)|z_1]\} =$$

$$\min\{\overbrace{\underbrace{P''[\theta_0|z_1]\times 0}_{\text{do nothing}} + \underbrace{P''[\theta_1|z_1]\times 400}_{\text{splicing}}}^{a_0},\ \overbrace{\underbrace{P''[\theta_0|z_1]\times 100}_{\text{cutting}} + \underbrace{P''[\theta_1|z_1]\times 0}_{\text{do nothing}}}^{a_1}\}$$

$$\min\{0.25\times 0 + 0.75\times 400,\ 0.25\times 100 + 0.75\times 0\} =$$

$$0.25\times 100 + 0.75\times 0 = 25$$

Pre-posterior Analysis

The minimum expected costs based on pre-posterior decision analysis – not including costs of experiments

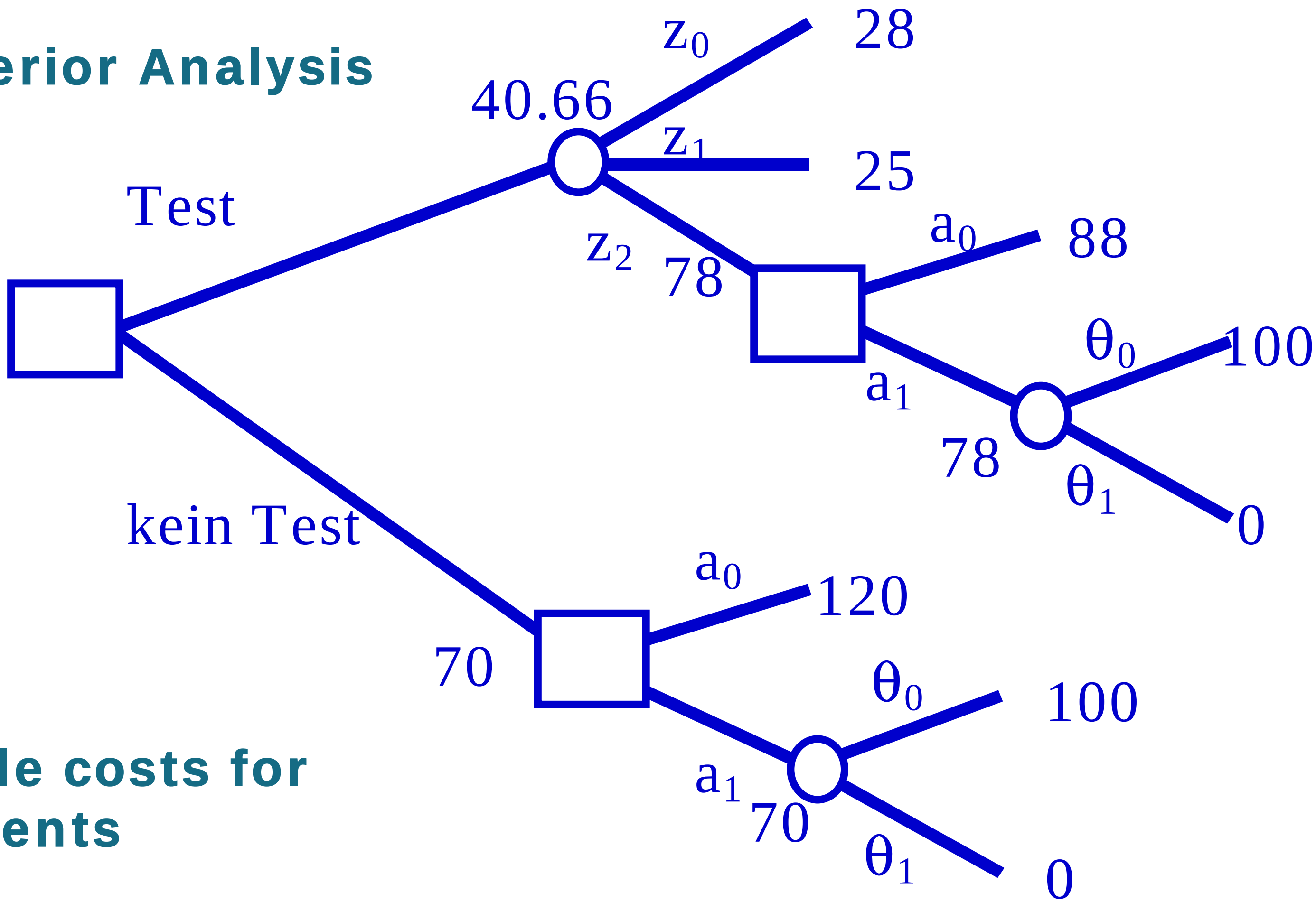
$$E[u] = \sum_{i=1}^{n} P'[z_i] \times E''[u|z_i] = 28 \times 0.45 + 25 \times 0.28 + 78 \times 0.27 = 40.66$$

Allowable costs for the experiment

$$E'[u] - E[u] = 70.00 - 40.66 = 29.34$$

Decision Analysis in Engineering

Pre-posterior Analysis



Allowable costs for experiments

$$E'[u] - E[u] = 70.00 - 40.66 = 29.34$$