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

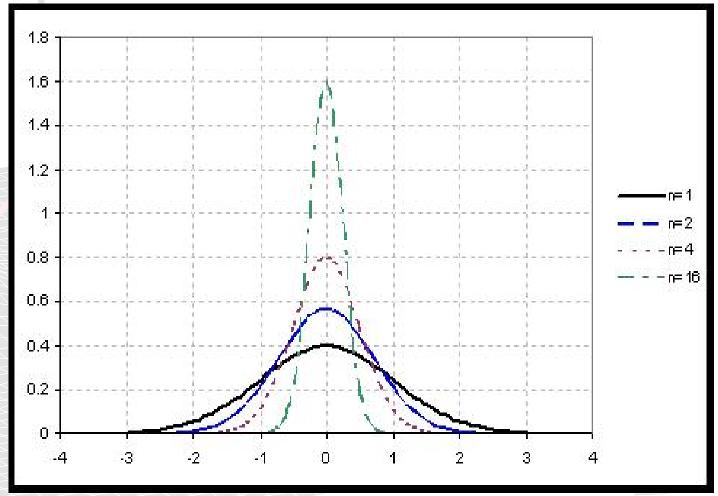
The sample variance

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



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

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

Short Summary of Previous Lecture

The sample variance is biased !

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

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

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

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

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

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

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

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

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 are called confidence intervals!

Sample average

True mean

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

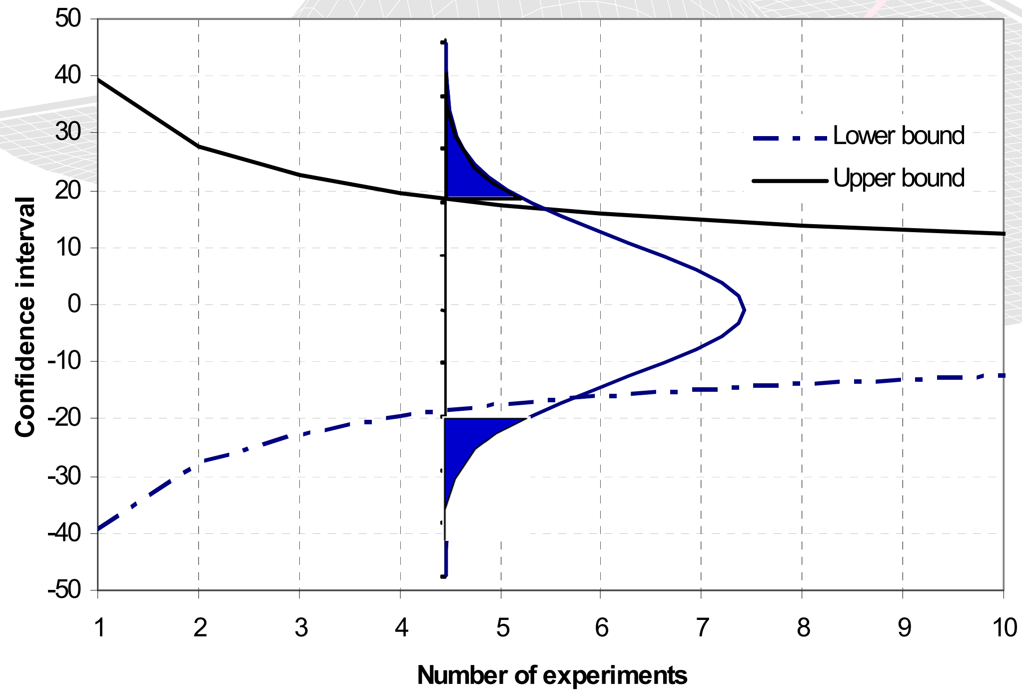
Known std. dev.

Sample size

Significance level

Short Summary of Previous Lecture

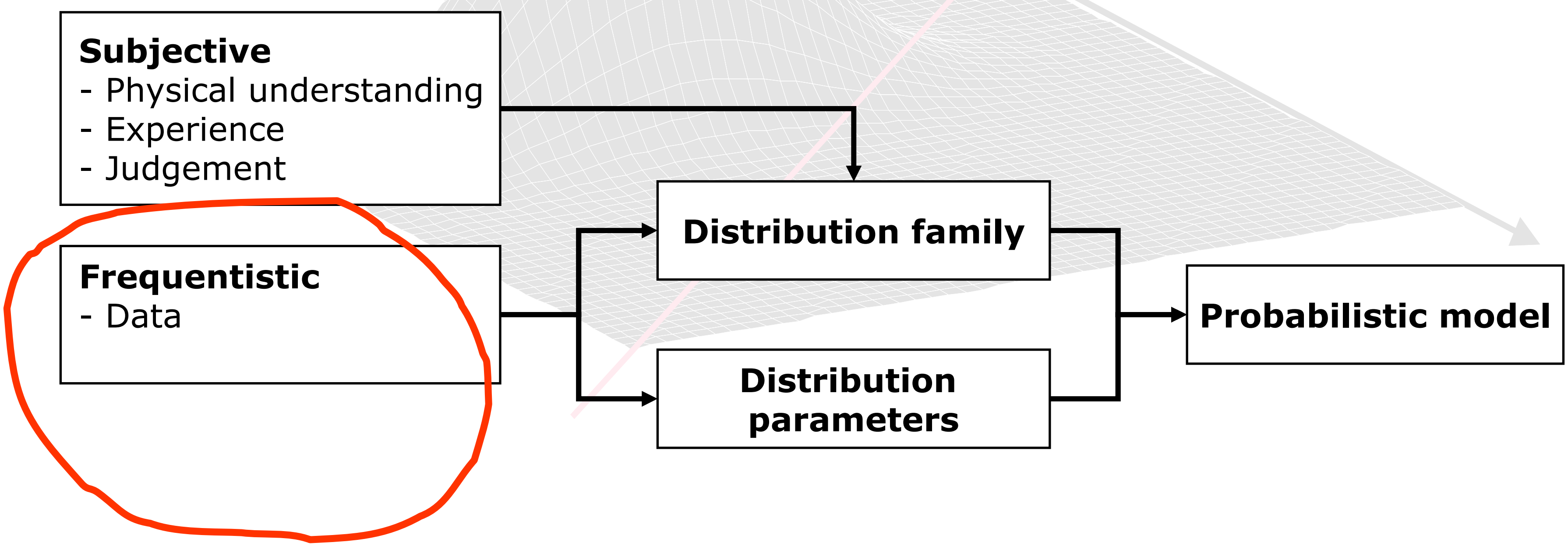
The number of available data has a significant importance for the confidence interval - using the same example as in the previous the confidence interval depends on *n* as shown below



Overview of Estimation and Model Building

Different types of information is used when developing engineering models

- subjective 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** α 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

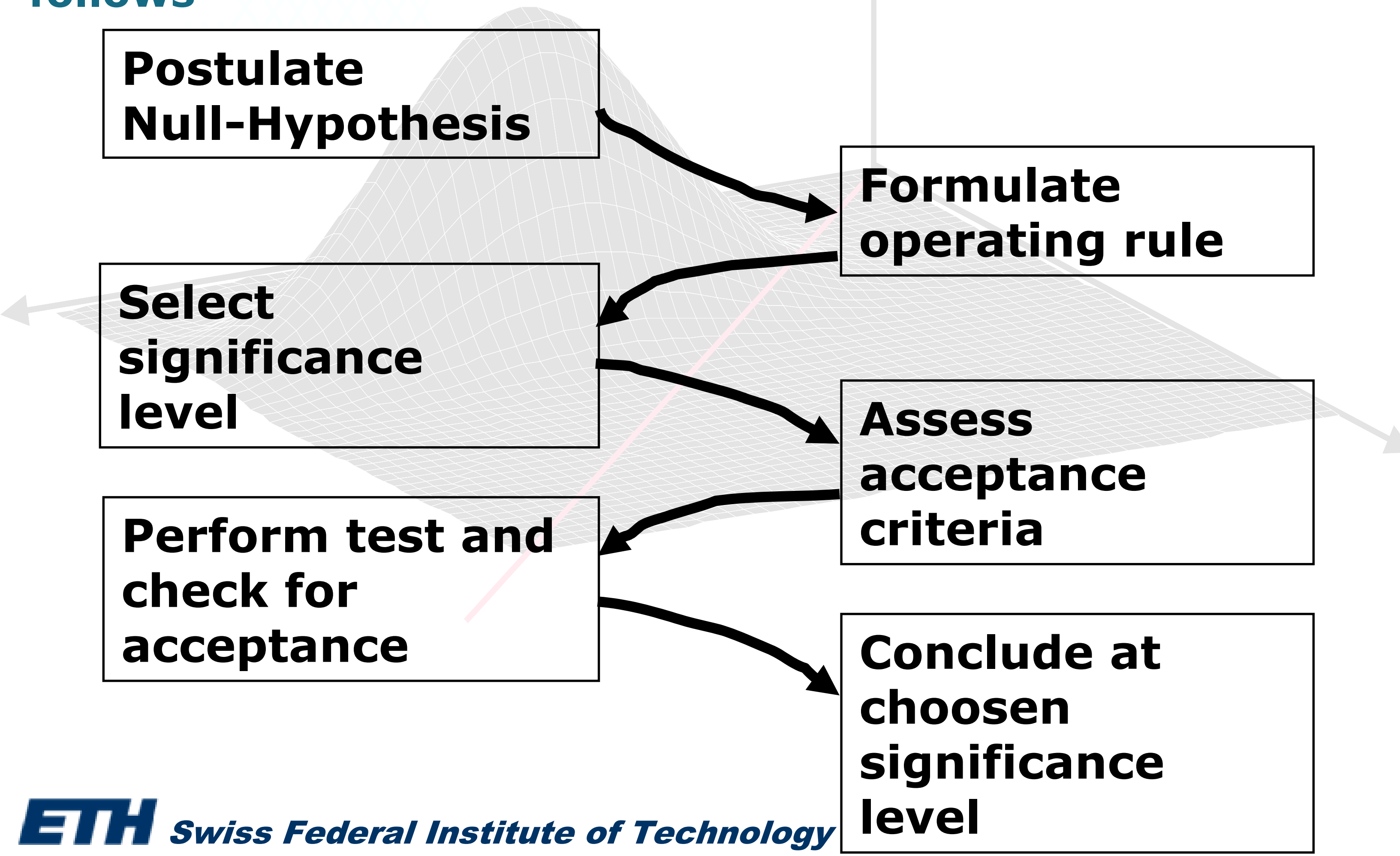
4 Calculate the value of Δ corresponding to α – calculate also if relevant the probability of performing a Type II error

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 the α significance level – otherwise it is accepted.

Testing for Statistical Significance

The hypothesis testing procedure may be visualized as follows



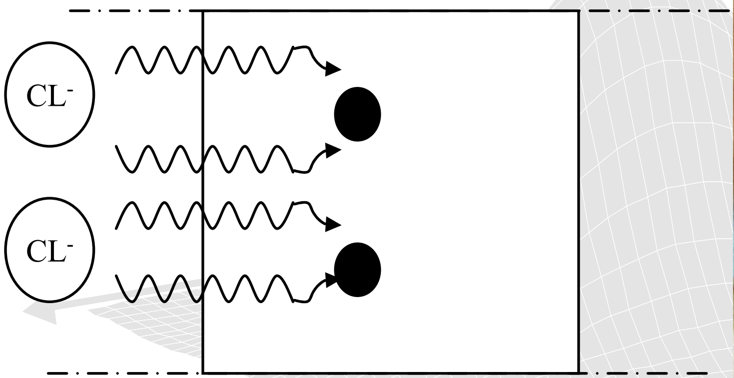
Testing for Statistical Significance

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

Testing of the mean – with known variance

The design assumptions: **mean surface chloride concentration is 0.3%**

Null-hypothesis

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

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

Testing for Statistical Significance

Testing of the mean – with known variance

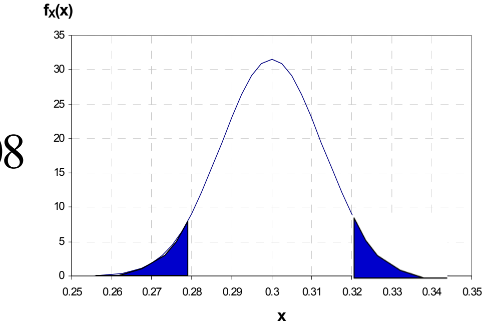
The acceptance criteria may be determined for given a by

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

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

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

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



Testing for Statistical Significance

Testing of the mean – with known variance

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

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 $P\left(-\Delta \leq T \leq \Delta\right) = 1 - \alpha$

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

$$P\left[S^2 \geq \Delta\right] = 1 - \alpha$$

where Δ is determined from $S^2 \geq \Delta$

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

$$\mathbf{x}=\left(x_1,x_2,\ldots,x_k\right)^T \qquad \mathbf{y}=\left(y_1,y_2,\ldots,y_l\right)^T$$

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 $T=\bar{X}-\bar{Y}$

is normal distributed with mean value $\mu_{\bar{X}-\bar{Y}}=\mu_X-\mu_Y$

and variance $\sigma^2_{\bar{X}-\bar{Y}}=\dfrac{\sigma_X^2}{k}+\dfrac{\sigma_Y^2}{l}$

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 $\sigma_X^2 = \sigma_Y^2$

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 α - consequently **it is in principle possible to prove anything –**

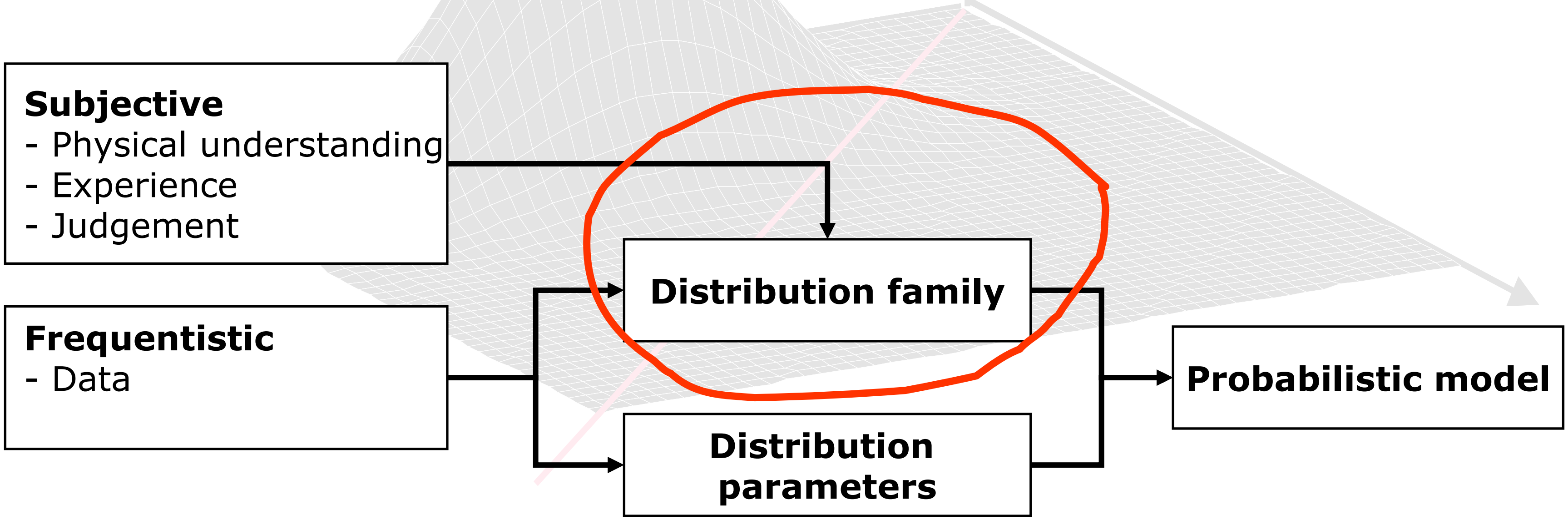
the different choises have direct effect on the probability of performing Type I and Type II errors – which may be related to significant economical consequences

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

- subjective 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 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 cumulative distribution is plotted on the paper it will have the shape of a straight line

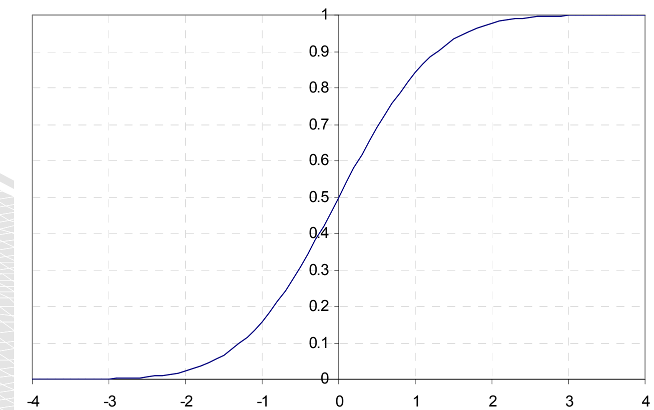
Estimation and Model Building

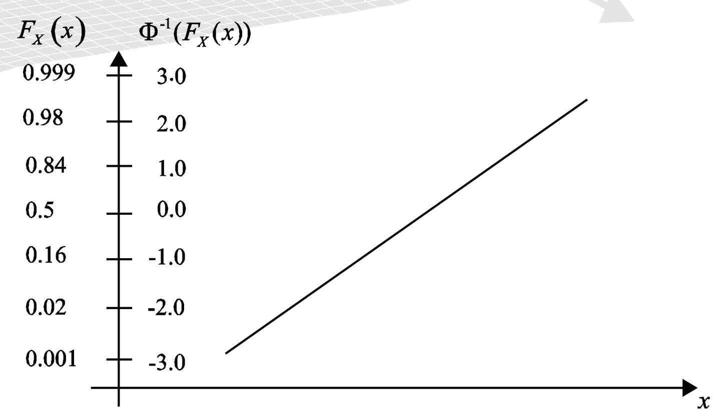
Model selection by use of probability paper

Example – probability paper for the Normal cumulative distribution function

$$F_X(x) = \Phi\left(\frac{x-\mu_X}{\sigma_X}\right)$$

$$x = \Phi^{-1}(F_X(x)) \cdot \sigma_X + \mu_X$$



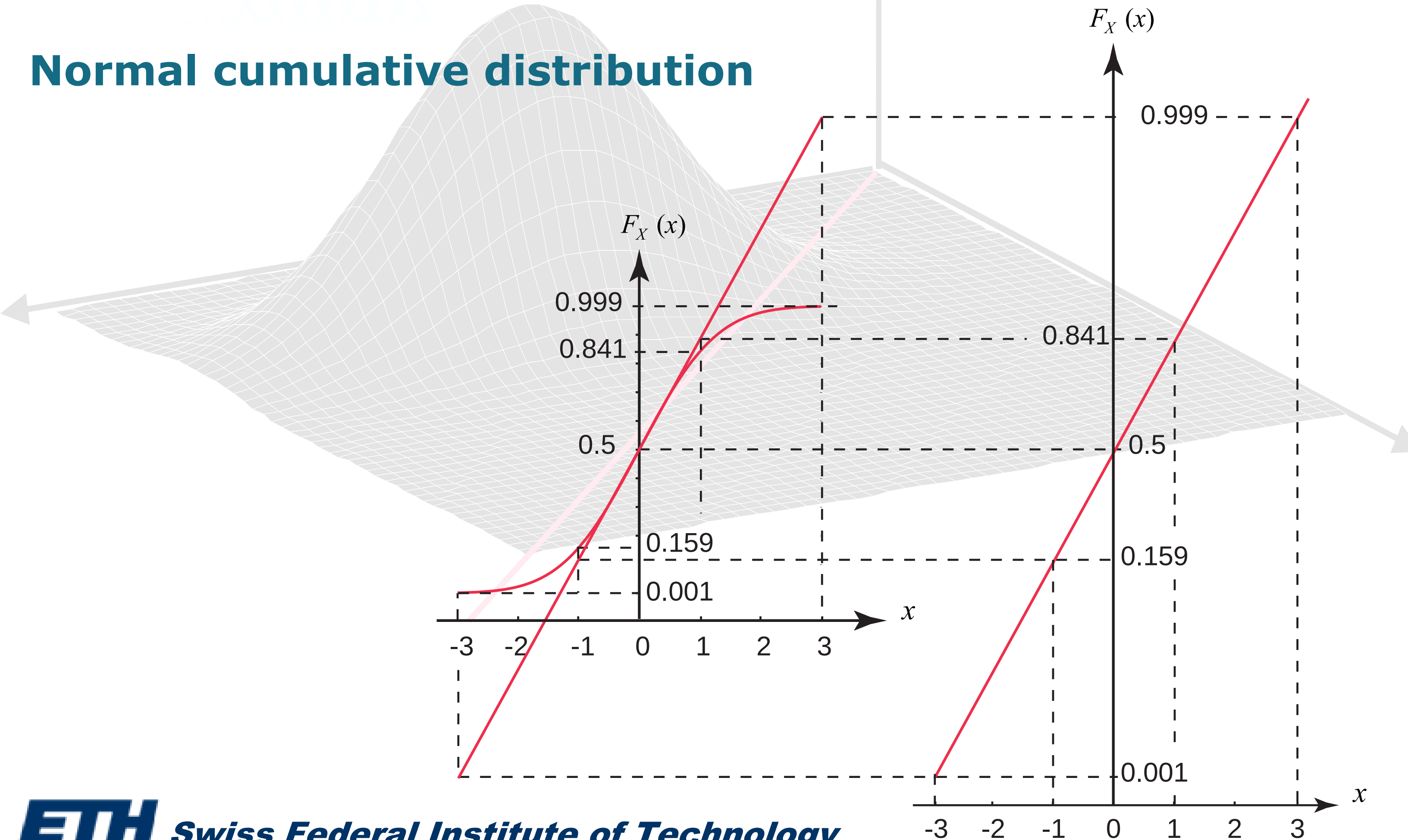


The y-axis scale is non-linear

Estimation and Model Building

Model selection by use of probability paper – graphical approach

Normal cumulative distribution



Estimation and Model Building

Model selection by use of probability paper

The sample cumulative distribution function may be established from the ordered sample as

$$F_X(x_i) = \frac{i}{N+1}$$

Example – concrete compressive 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 cumulative distribution function in the probability paper yields