Statistics and Probability Theory

in

Civil, Surveying and Environmental Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

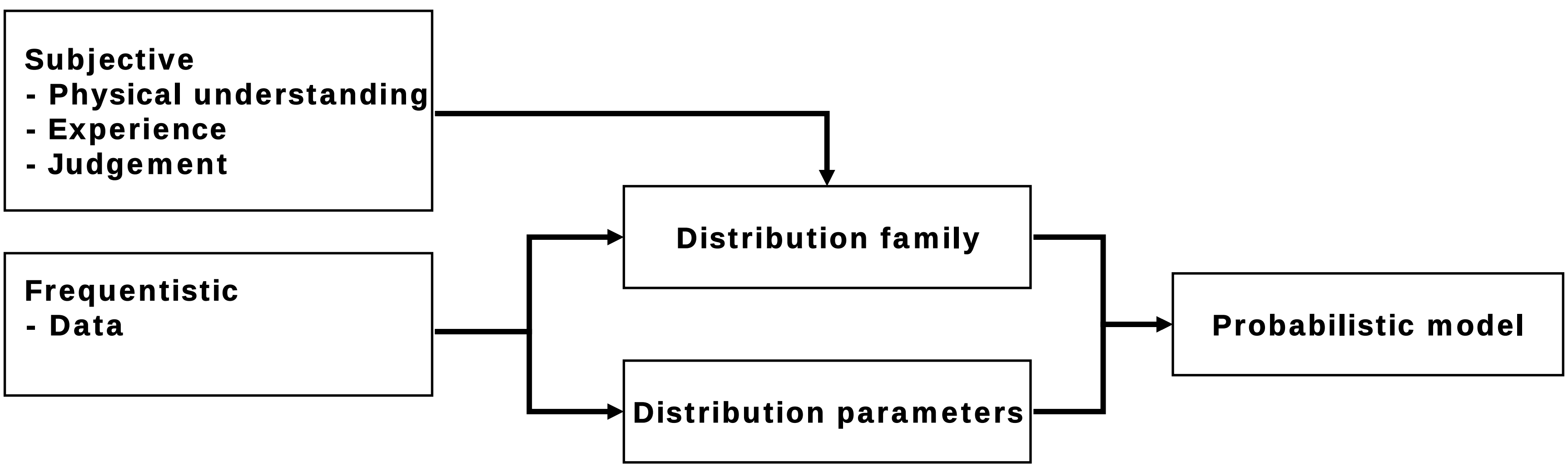
Contents of Todays Lecture

- 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

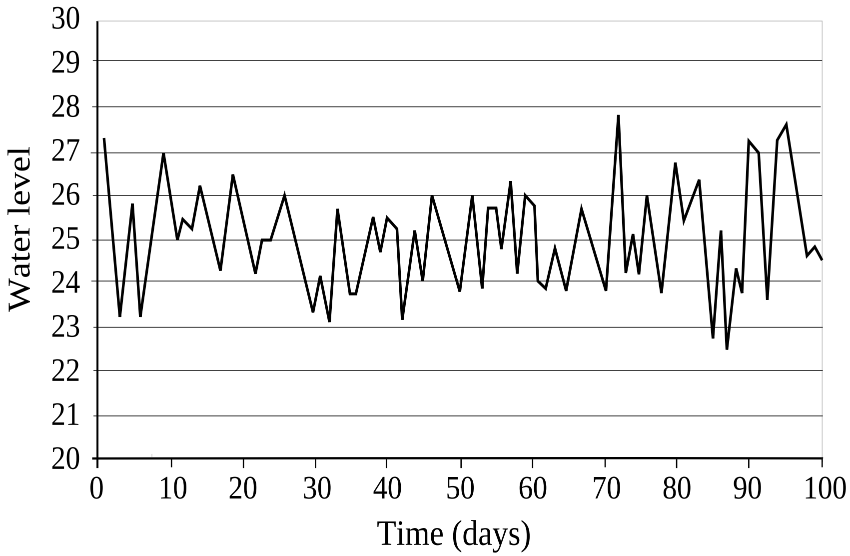
- subjective information
- frequentistic information



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:

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

then the extremes of the same process within the period:

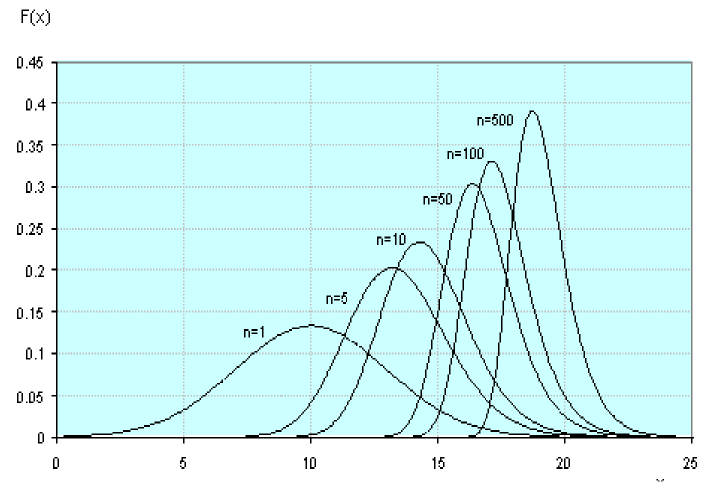
$n \cdot T$ **will follow the distribution:**

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

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

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

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



A Short Summary of the Previous Lecture

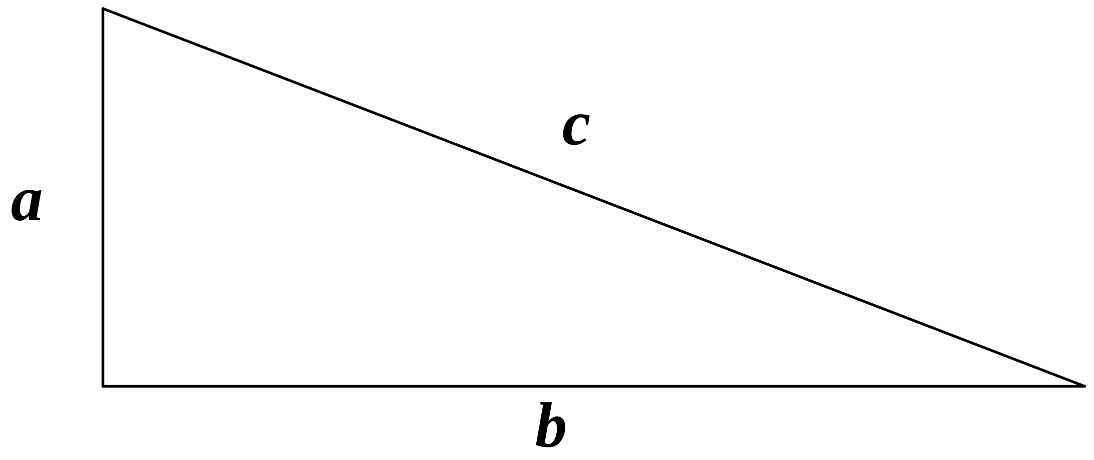
Based on independent Normal distributed random variables we could derive the following distributions:

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

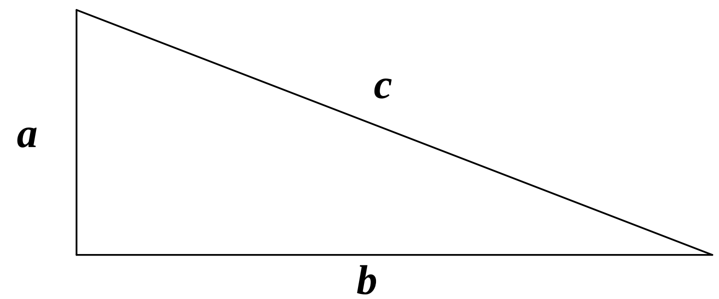
Probability Distribution Functions in Statistics

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

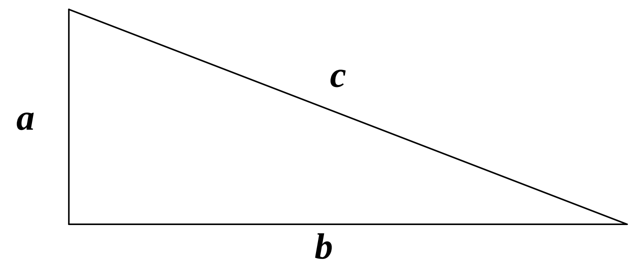


Example Chi distribution

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

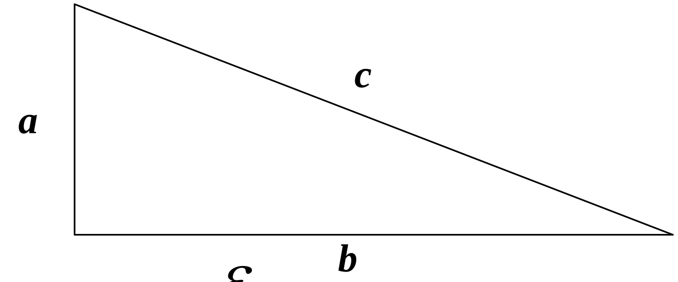
$$\varepsilon_c = \sqrt{\varepsilon_a^2 + \varepsilon_b^2}$$

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 $Z = \dfrac{\varepsilon_c}{\sigma_\varepsilon}$

can thus be determined from

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

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

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 –

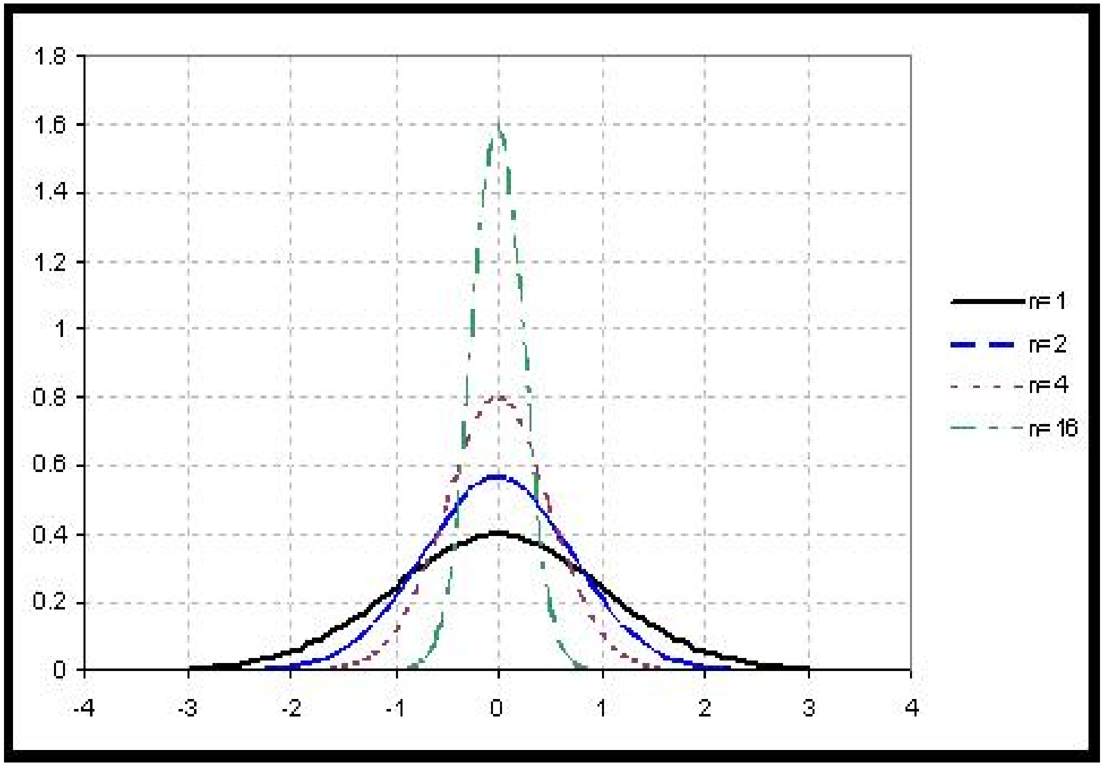
however we can evaluate the expected value and the variance of the sample statistics, i.e. for the sample mean we get :

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

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

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:

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

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

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

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

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

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

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

Known std. dev.

Sample size

Significance level

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

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

Known std. dev.

Sample average

True mean

Sample size

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

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

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

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

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

$$P\left[390.2 < \mu_X < 409.8\right] = 0.95$$

- 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