Statistics and Probability Theory

in

Civil, Surveying and Environmental Engineering

Prof. Dr. Michael Havbro Faber

Swiss Federal Institute of Technology

ETH Zurich, Switzerland

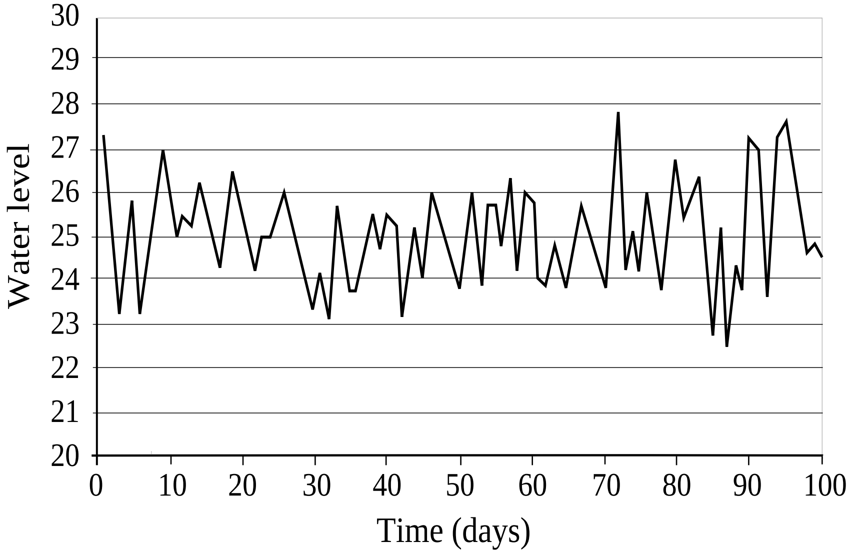
Contents of Today‘s Lecture

- 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:**

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

$i = j$ **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; $X(t_1), X(t_2),\dots,X(t_j)$

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

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

$$\left.\begin{aligned}&\mu_X(t)=cst\\[2ex]&R_{XX}(t_1,t_2)=f(t_2-t_1)\end{aligned}\right\}\quad \textbf{Weakly stationary}$$

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

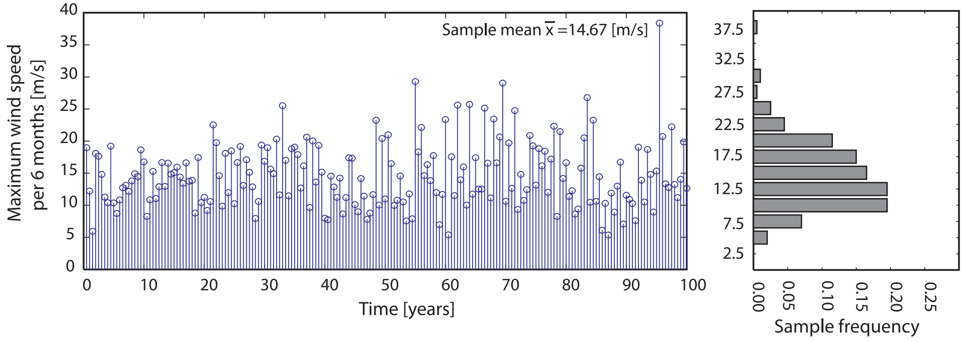
We could also be interested in the smallest or the largest value of a certain quantity within a certain volume or area unit e.g.:

The largest concentration of pesticides in a volume of soil

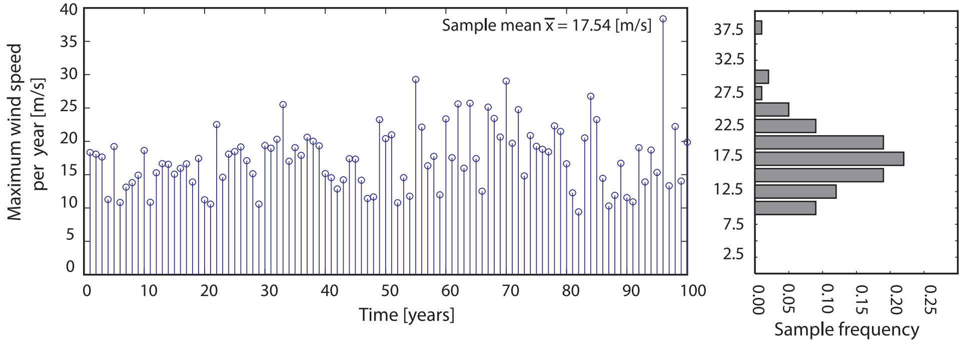
The weakest link in a chain

The smallest thickness of concrete cover

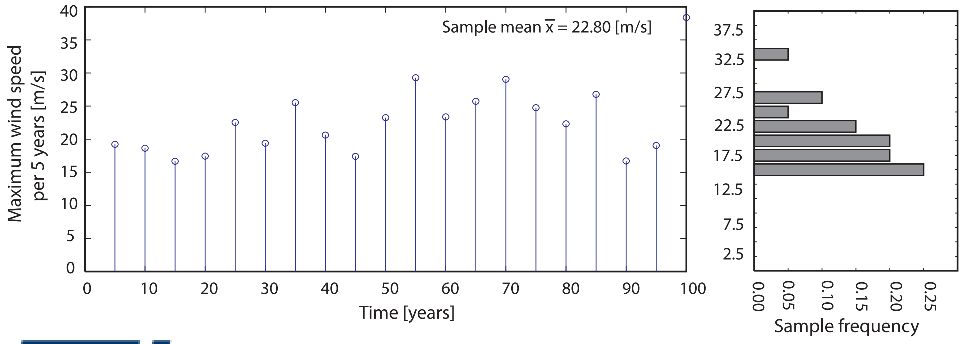
Extreme Value Distributions



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:

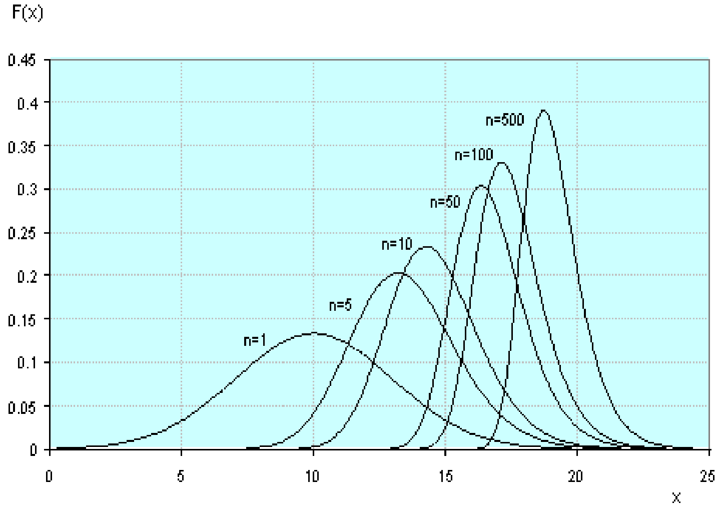
$$F_{X,T}^{\max}(x)$$

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

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

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

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

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

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

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

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

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

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

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 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

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

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

Return Period

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

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

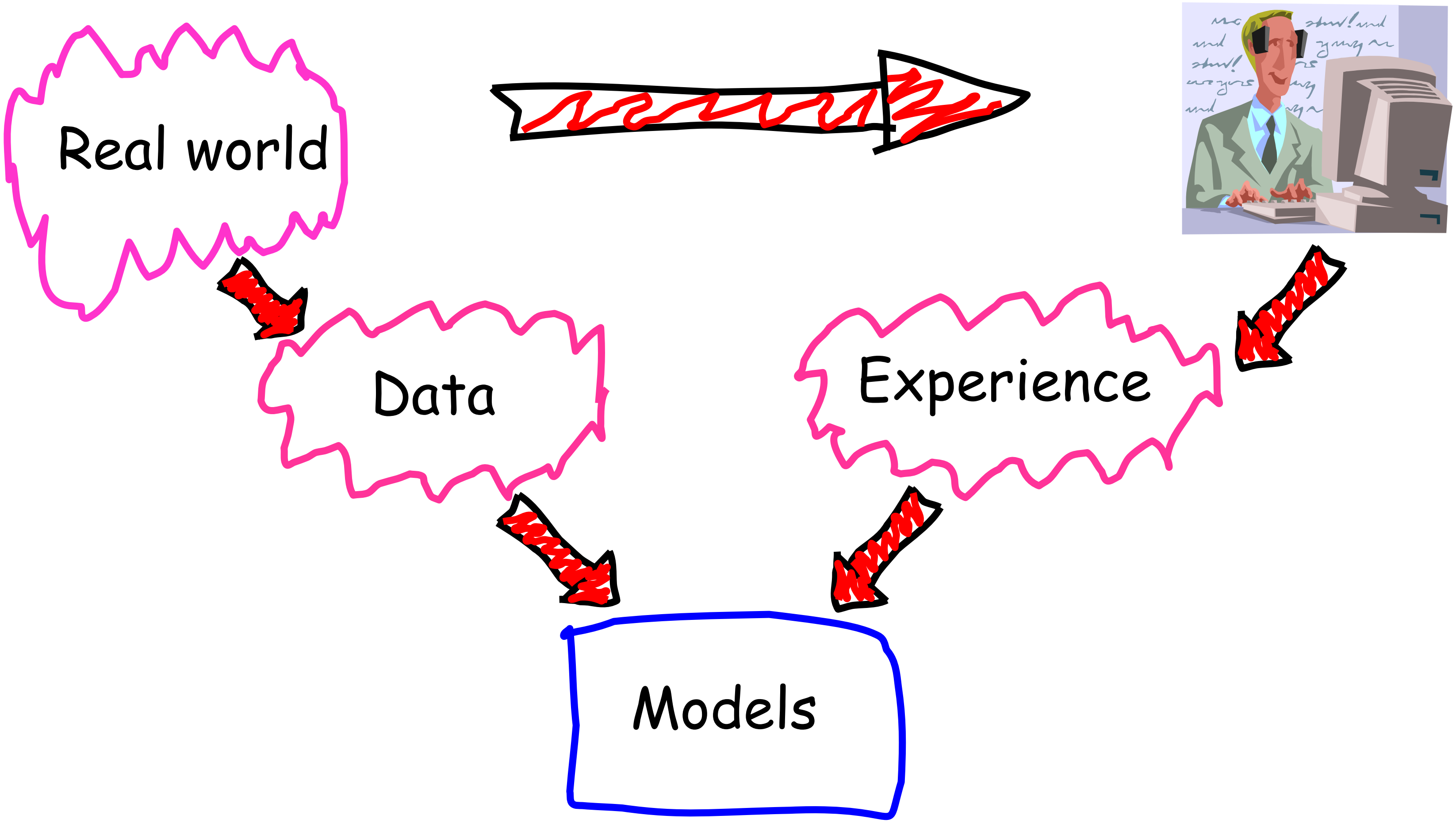
Example:

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



Real world

Data

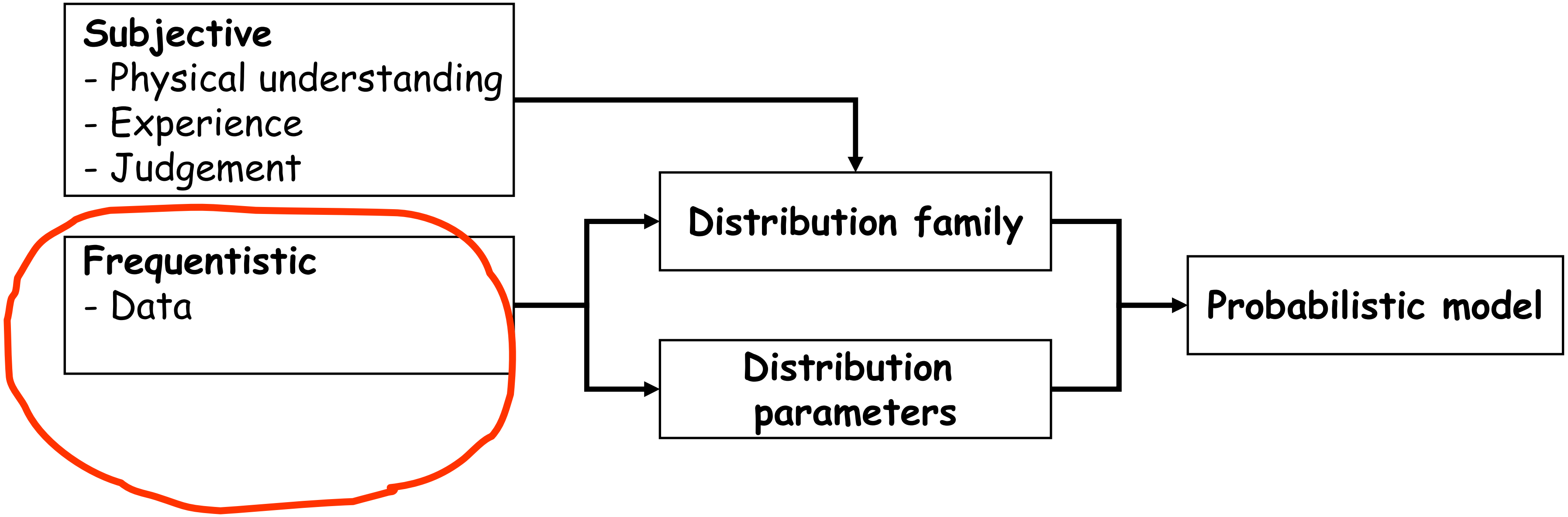
Experience

Models

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

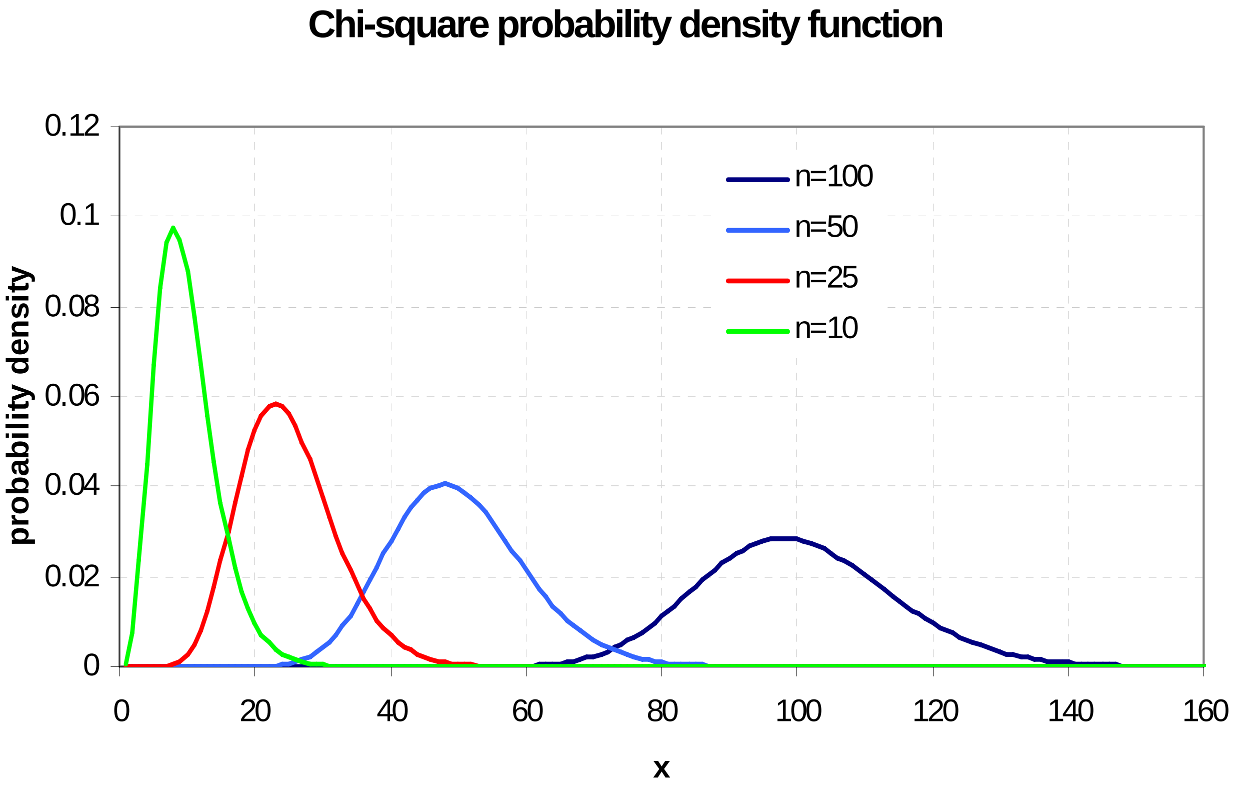
$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

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

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

Probability Distribution Functions in Statistics

The **Chi probability density** function is given as

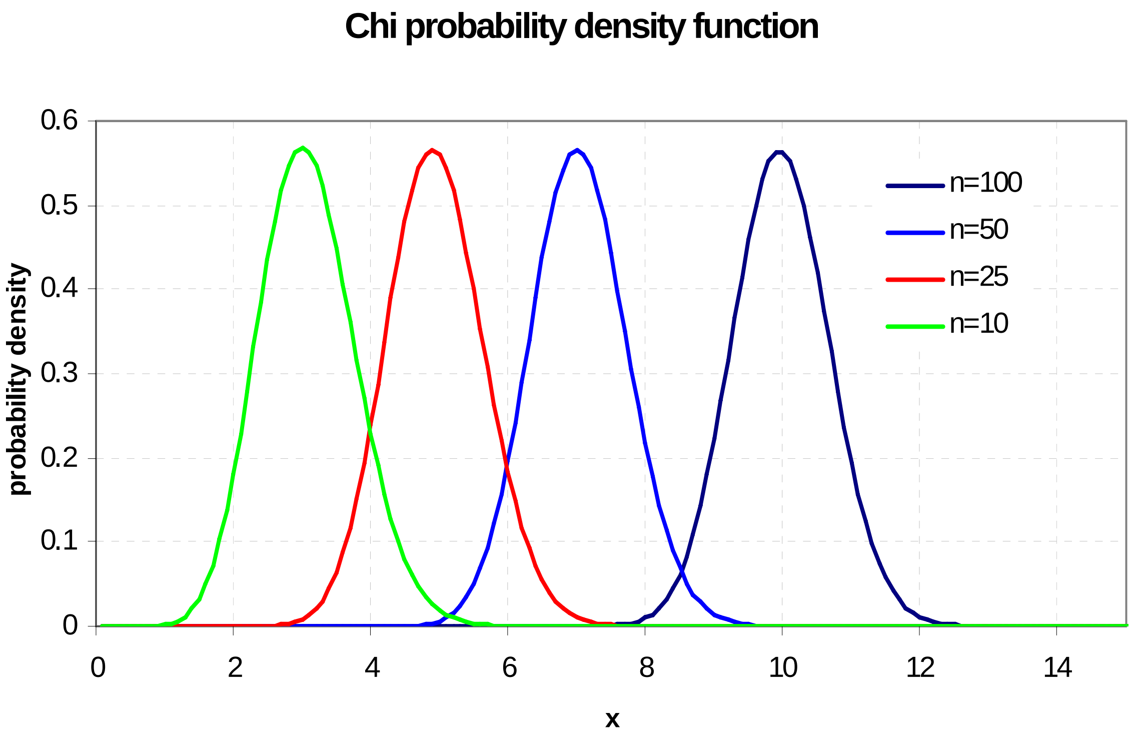
$$f_Z(z) = \frac{z^{(n-1)}}{2^{n/2-1}\Gamma(n/2)}\exp(-z^2/2), \qquad z \geq 0$$

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

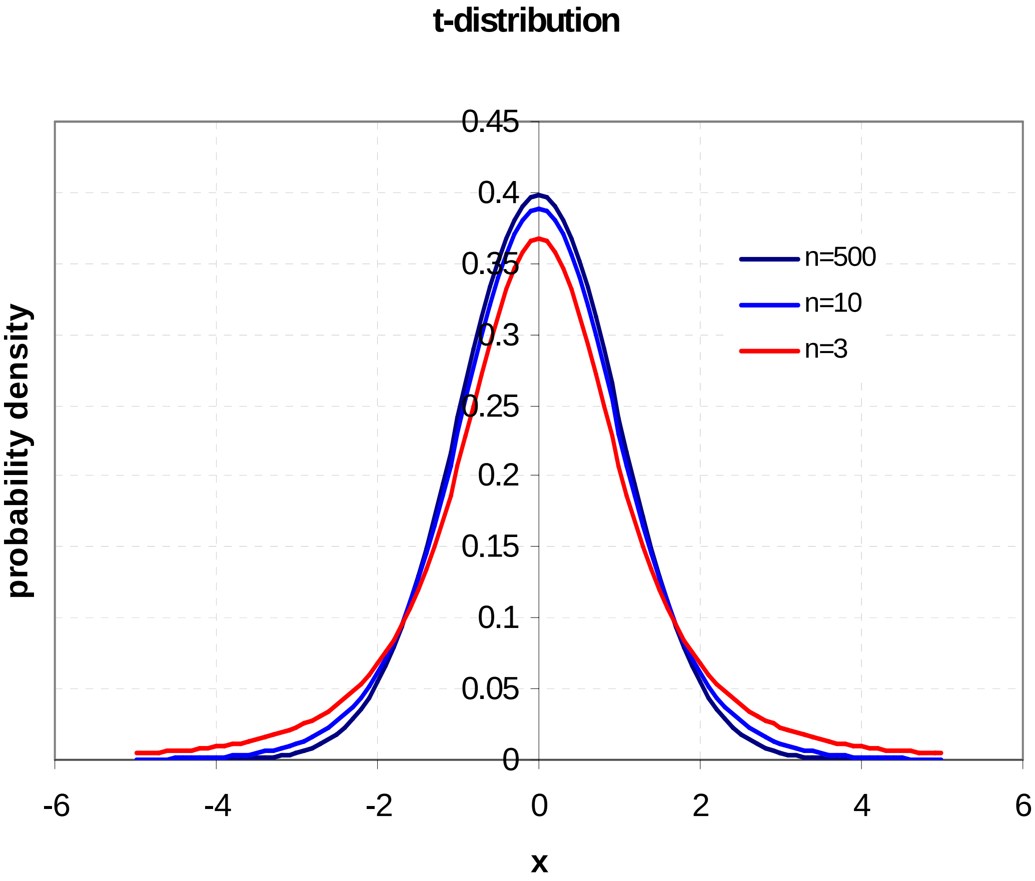
$$f_S(s) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\ \Gamma(n/2)} \left(1 + \frac{s^2}{n}\right)^{-(n+1)/2} , \qquad -\infty \leq s \leq \infty$$

The mean value is zero

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

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

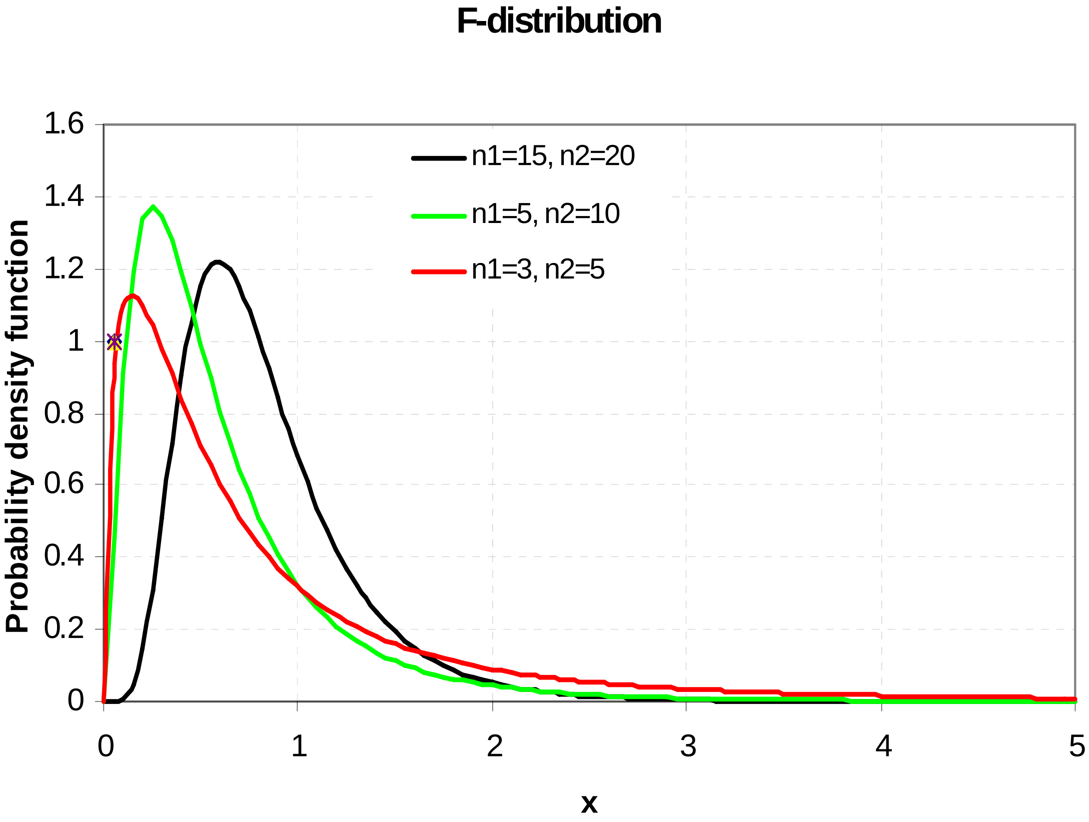
$$f_Q(q) = \frac{\Gamma((n_1+n_2)/2)q^{(n_1-2)/2}(1+q)^{-(n_1+n_2)/2}}{\Gamma(n_1/2)\Gamma(n_2/2)}, \qquad q \geq 0$$

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

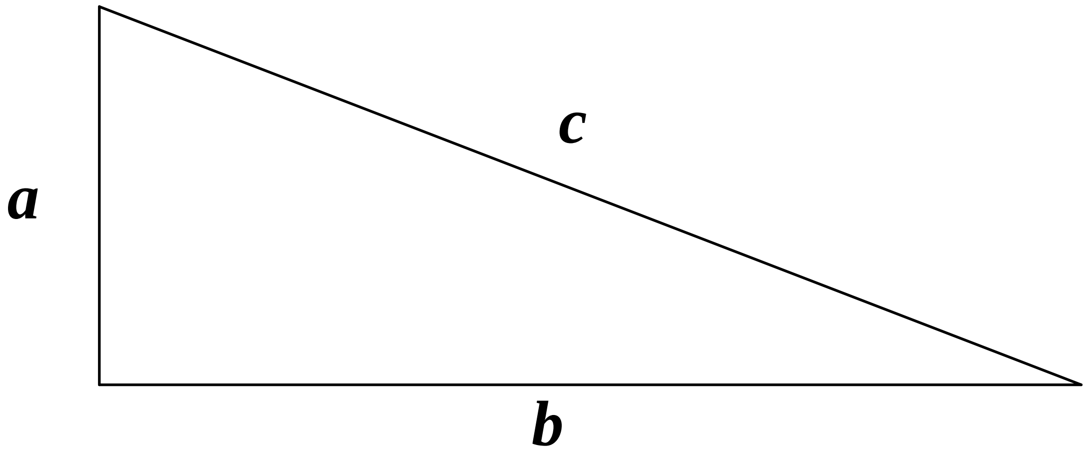
Summary of derived probability density functions:

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

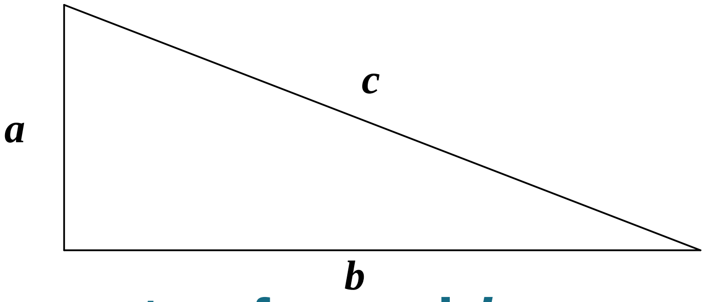
Example Chi distribution

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



Probability Distribution Functions in Statistics

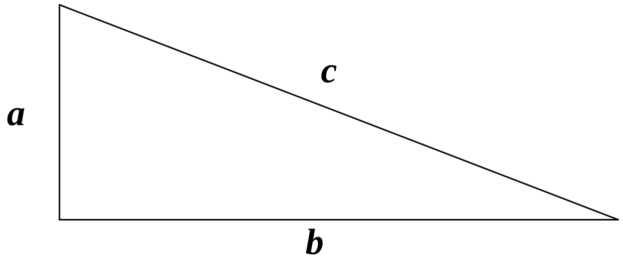
Example Chi distribution



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

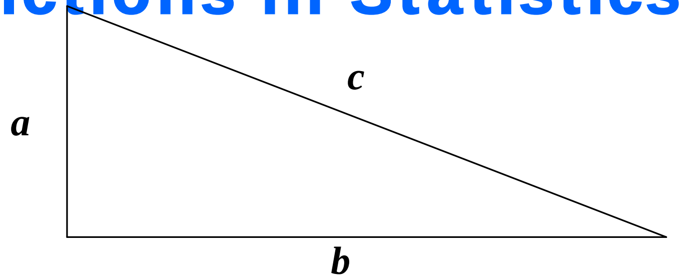
$$\varepsilon_c = \sqrt{\varepsilon_a^2 + \varepsilon_b^2}$$

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

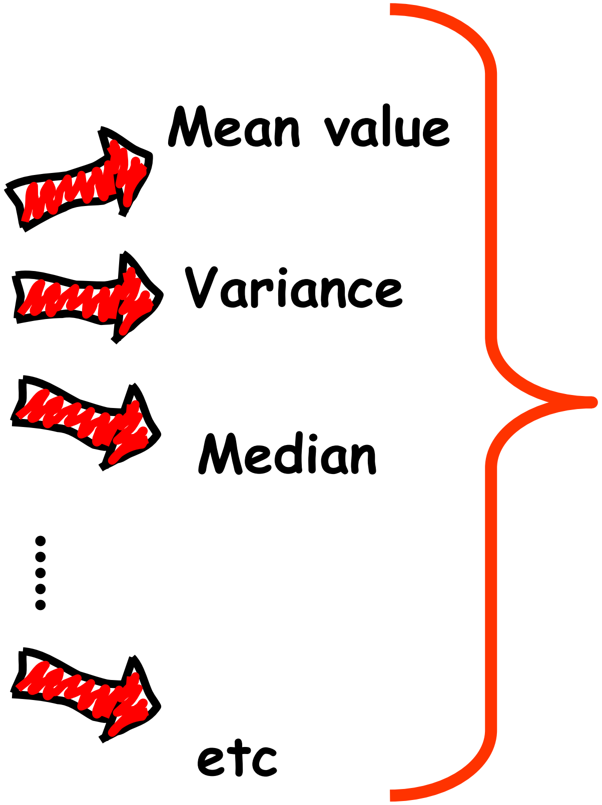
where it was used that for $y = g(x)$ we have $f_y(y) = \left|\dfrac{dg^{-1}}{dy}\right| f_X(g^{-1}(y))$

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 :

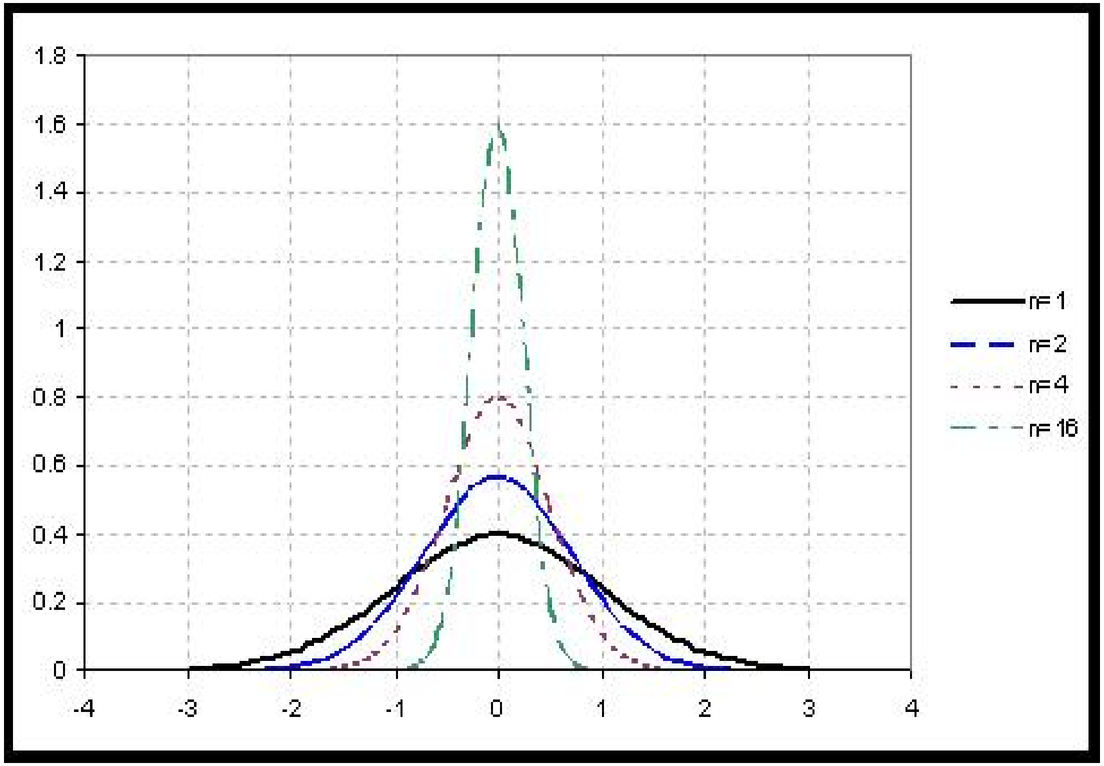
$$E[\overline{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n \cdot \mu_X = \mu_X$$

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

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

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:

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

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

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

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

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

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	$h(\overline{\theta})=\overline{h(\theta)}$
- 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

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

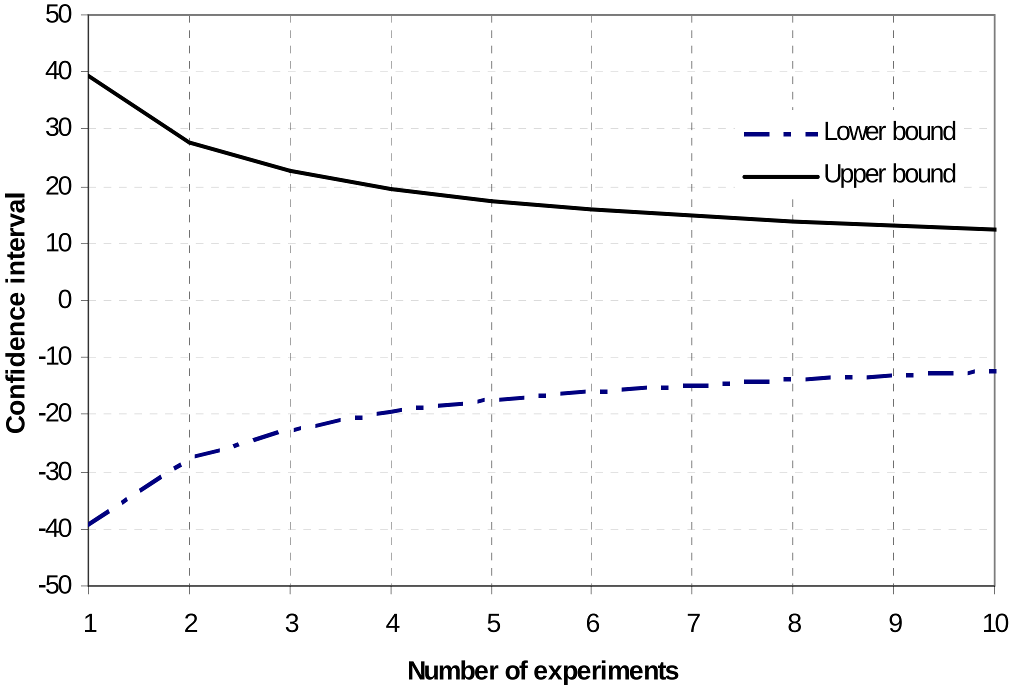
$$P\left[390.2 < \mu_X < 409.8\right] = 0.95$$

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


Confidence interval
Number of experiments
Lower bound
Upper bound