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

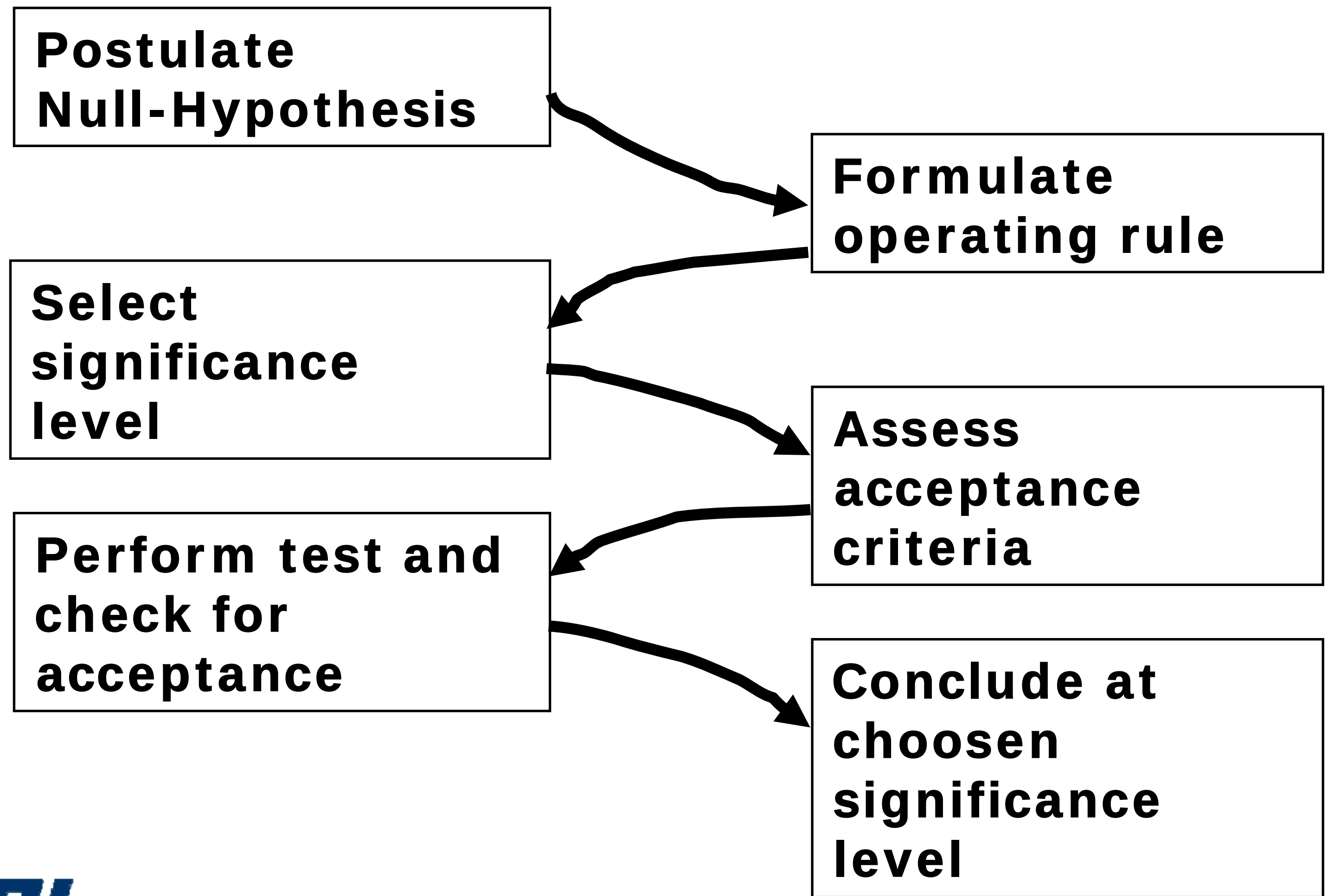
  and we also introduced the concept of

  probability paper

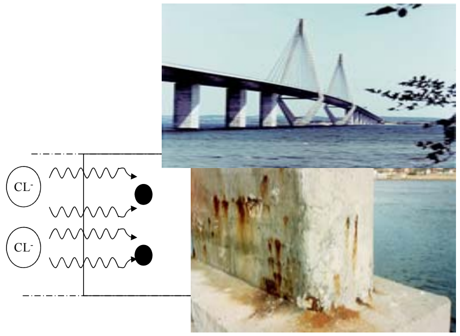
  - supporting the choice of a given probabilistic model based on data / observations

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*

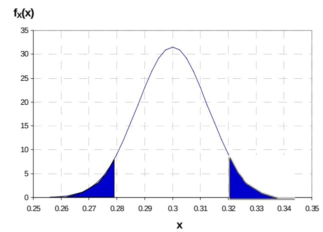
$$0.3-\Delta \leq T \leq 0.3+\Delta$$

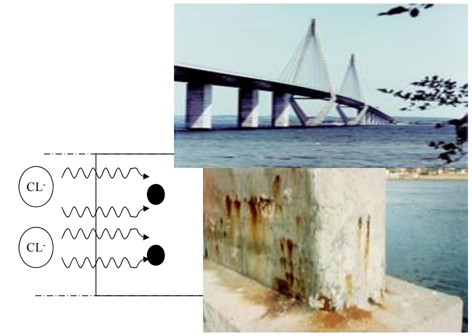
What did we Learn in the Previous Lecture

The test acceptance criteria:

The operating rule must be fulfilled with a probability of 1-α.

$$P\left(0.3-\Delta \le T \le 0.3+\Delta\right)=1-\alpha$$





Assessing acceptance criteria:

The interval for the operating rule is determined as:

$$\Phi\left(\frac{x_U-\mu}{\sigma}\right)-\Phi\left(\frac{x_L-\mu}{\sigma}\right)=\Phi\left(\frac{(0.3+\Delta)-0.3}{\frac{0.04}{\sqrt{10}}}\right)-\Phi\left(\frac{(0.3-\Delta)-0.3}{\frac{0.04}{\sqrt{10}}}\right)=0.9 \quad \Rightarrow \quad \Delta=0.0208 \quad \Longrightarrow \quad \left[0.28\le t\le 0.32\right]$$

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

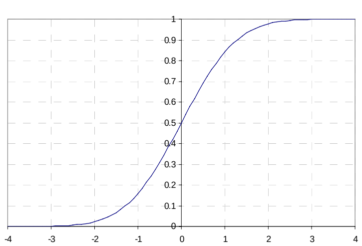
What did we Learn in the Previous Lecture

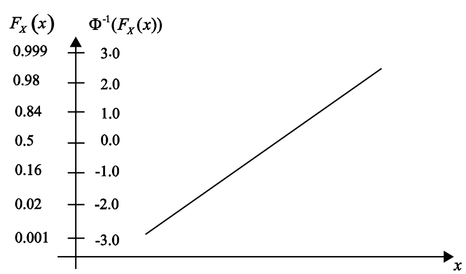
- **Probability paper – what is the idea!**

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

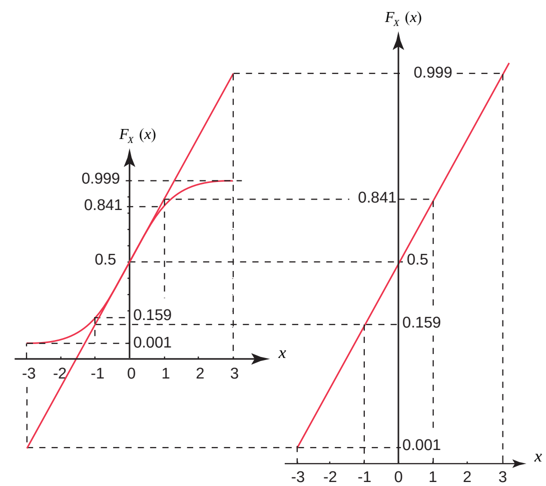
In order to do that we perform an non-linear transformation of the *y*-axis of the usual CDF plot

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





Analytically



Graphically

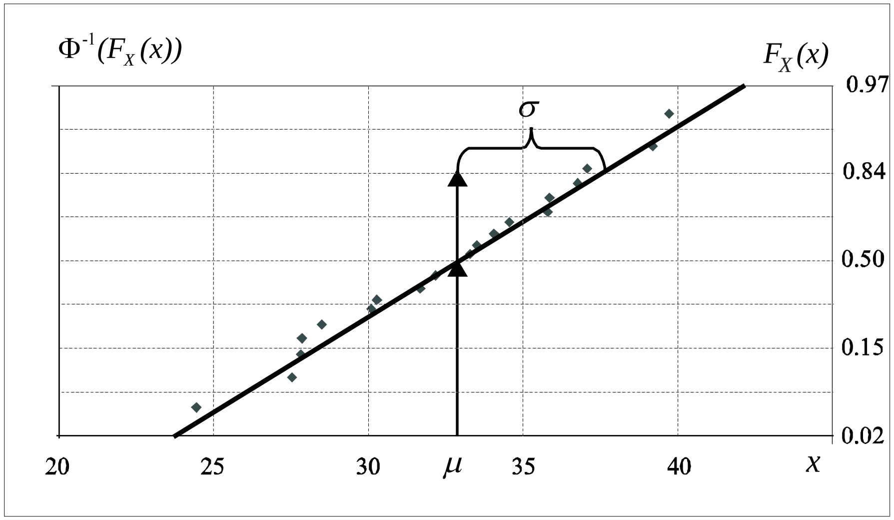
What did we Learn in the Previous Lecture

- **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

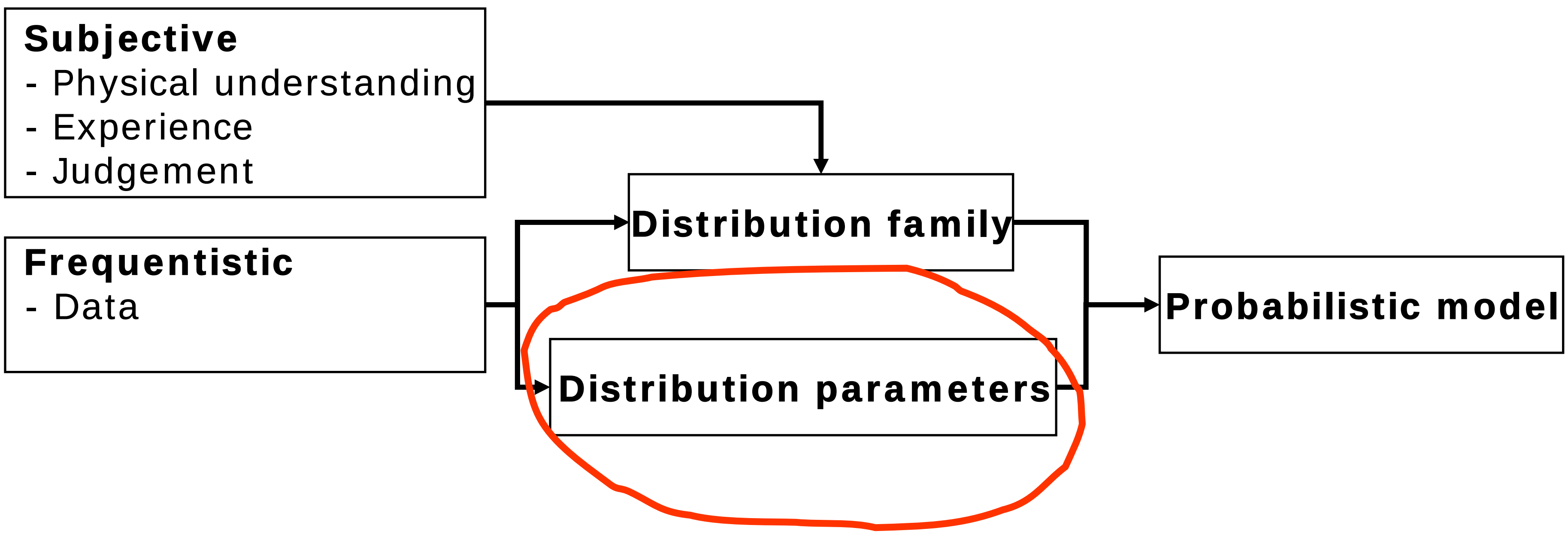


If the *q*-plot is close to straight in the important regions we have support for our model!

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.

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$$

Sample moments

$$\lambda_j = \lambda_j(\theta_1, \theta_2, .., \theta_k) = \int_{-\infty}^{\infty} x^j \cdot f_X(x|\boldsymbol{\theta})dx$$

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 $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)^T$ **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.:

$$\min_{\boldsymbol{\theta}} \left(-L(\boldsymbol{\theta} \,|\, \hat{\mathbf{x}})\right)$$

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 $\mathbf{C}_{\Theta\Theta} = \mathbf{H}^{-1}$ $H_{ij} = \left.\dfrac{\partial^2 - l(\boldsymbol{\theta}|\hat{\mathbf{x}})}{\partial\theta_i \partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$

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

$$l(\boldsymbol{\theta}|\hat{\mathbf{x}}) = n \cdot \ln\left(\frac{1}{\sqrt{2\pi}\theta_1}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left(\hat{x}_i - \theta_2\right)^2}{\theta_1^2}$$

  the minimum of which may be found by the solution of the following equations

$$\frac{\partial l}{\partial \theta_1} = -\frac{n}{\theta_1} + \frac{1}{\theta_1^3}\sum_{i=1}^{n}\left(\hat{x}_i - \theta_2\right)^2 = 0$$

$$\frac{\partial l}{\partial \theta_2} = \frac{1}{\theta_1^2}\sum_{i=1}^{n}\left(\hat{x}_i - \theta_2\right) = 0$$

$$\Rightarrow$$

$$\theta_1 = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{x}_i - \theta_2\right)^2}{n}}$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i$$

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

Putting numbers into the solution we get:

$$\theta_1 = \sqrt{\frac{\sum_{i=1}^{n} (\hat{x}_i - \theta_2)^2}{n}} = \sqrt{\frac{367.19}{20}} = 4.05$$

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:

$$H = \begin{pmatrix} \frac{n}{\theta_1} - \frac{3\sum_{i=1}^{n}(x_i - \theta_2)^2}{\theta_1^4} & \frac{2\sum_{i=1}^{n}(x_i - \theta_2)}{\theta_1^3} \\ \frac{2\sum_{i=1}^{n}(x_i - \theta_2)}{\theta_1^3} & \frac{n}{\theta_1^2} \end{pmatrix}$$

$$C_{\Theta\Theta} = H^{-1} = \begin{pmatrix} 0.836 & 0 \\ 0 & 0.165 \end{pmatrix}$$

Variance of the mean value (0.165)

Variance of the standard deviation (0.836)

Estimation of Distribution Parameters

- **The Maximum Likelihood Method (MLM)**

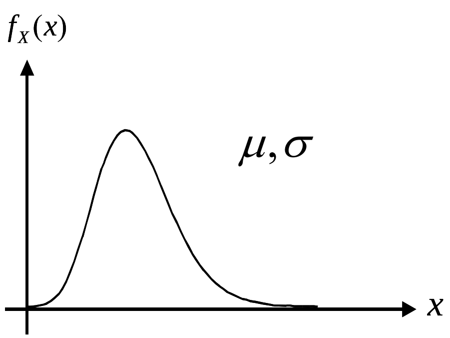
  We may also estimate the parameters completely numerically using Excel

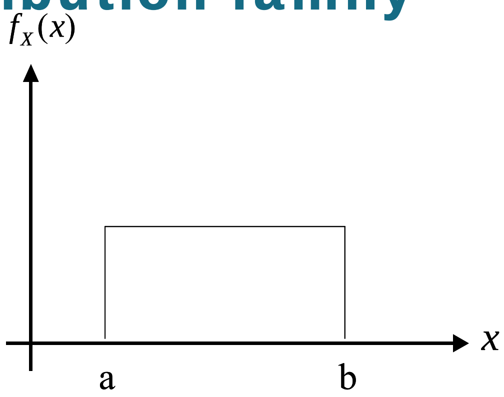
  Lets take a look !

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.