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)

$$L(\boldsymbol{\theta}|\hat{\mathbf{x}}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\hat{x}_i - \mu}{\sigma}\right)^2\right)$$

$$\min_{\boldsymbol{\theta}} \; (-L(\boldsymbol{\theta}|\hat{\mathbf{x}})) \qquad l(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \log(f_X(\hat{x}_i|\boldsymbol{\theta}))$$

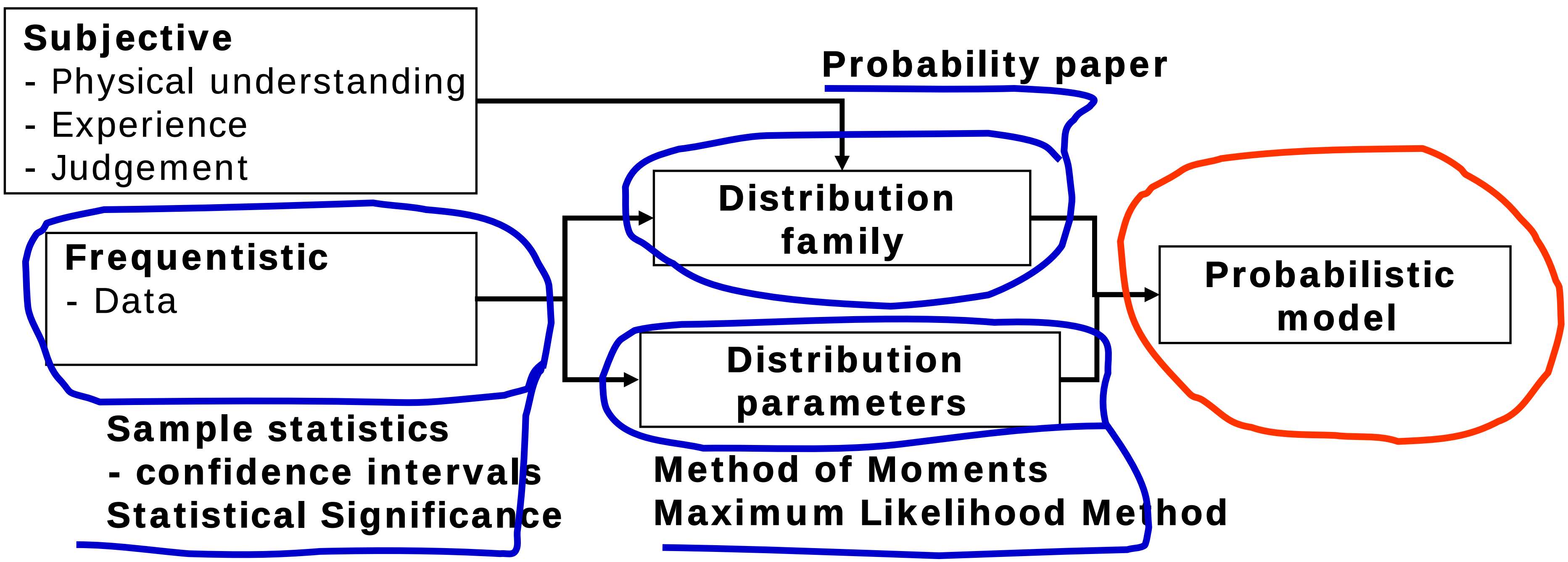
$$\Rightarrow \begin{cases} \boldsymbol{\mu}_\Theta = (\theta_1^*, \theta_2^*, .., \theta_n^*)^T \\ \mathbf{C}_{\Theta\Theta} = \mathbf{H}^{-1} \\ H_{ij} = \dfrac{\partial^2 - l(\boldsymbol{\theta}|\hat{\mathbf{x}})}{\partial\theta_i \partial\theta_j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \end{cases}$$

The MLM provides an extremely strong statistical tool!

Overview of Estimation and Model Building

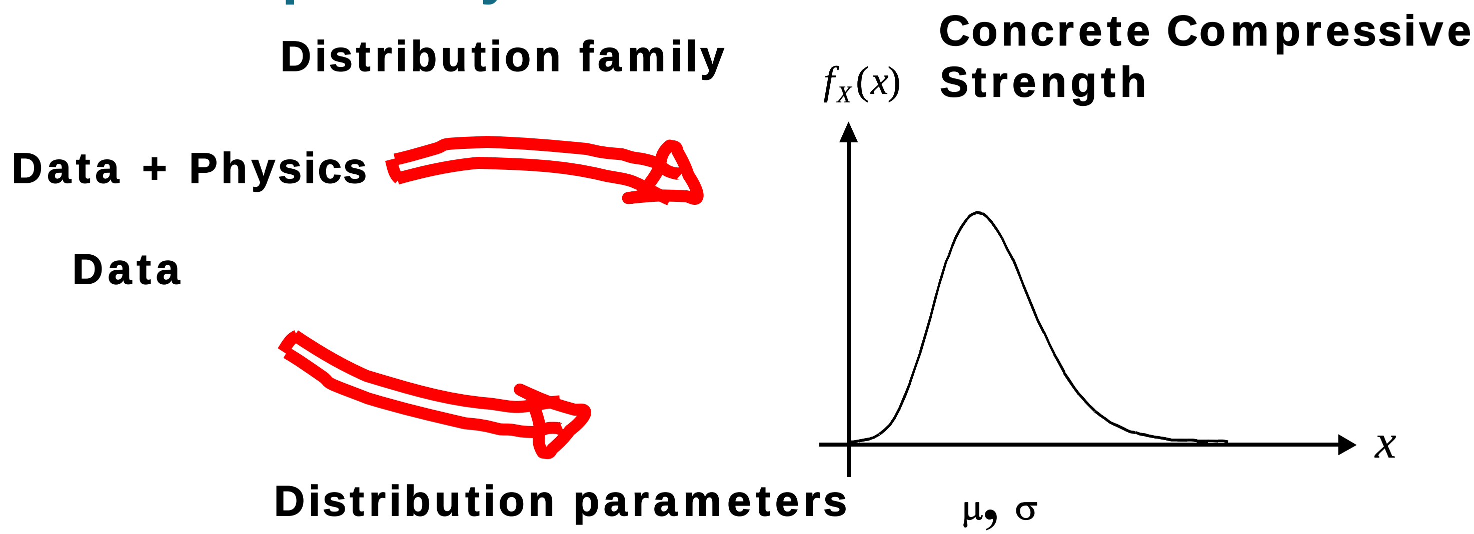
Different types of information is used when developing engineering models

- 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



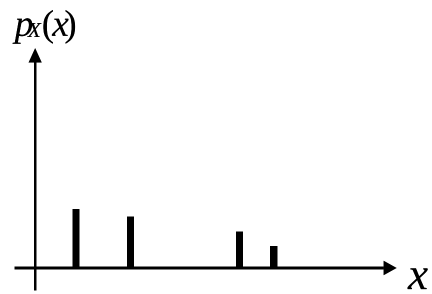
Now we want to validate our model selection – by means of statistical tests

Model Evaluation by Statistical Testing

Two different cases are considered – namely verification of

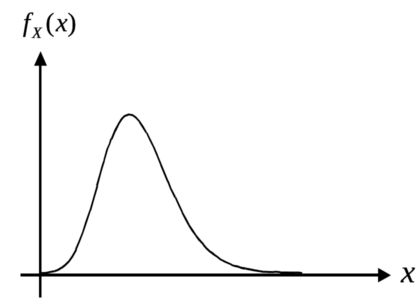
1: Discrete distribution functions

CHI-Square (χ^2) test



2: Continuous distribution functions

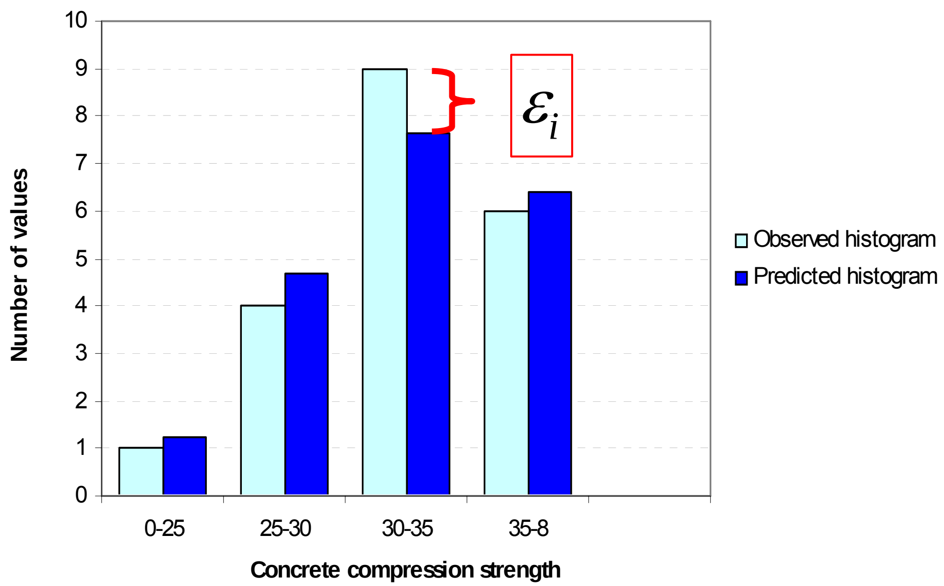
Kolmogorov Smirnov test



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

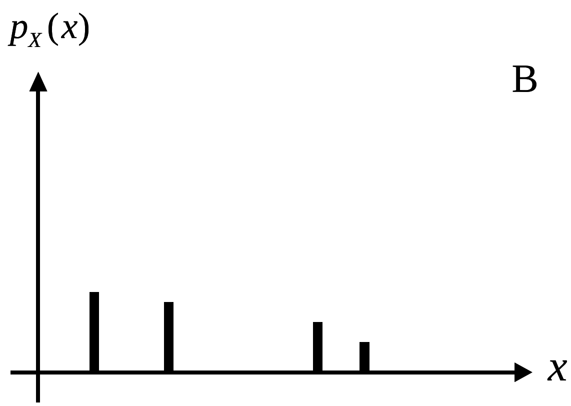


Model Evaluation by Statistical Testing

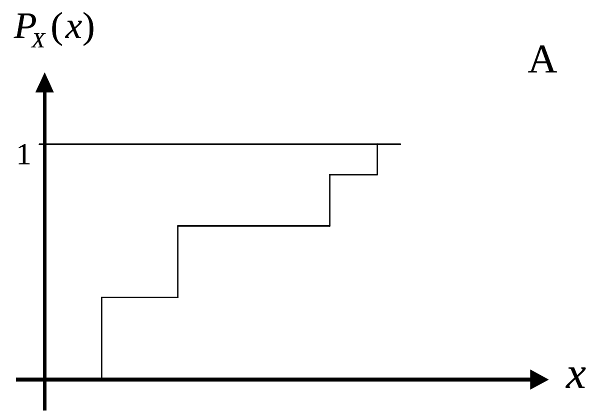
The CHI-square goodness of fit test

We remember that a discrete cumulative distribution is given by:

$$P(x_i) = \sum_{j=1}^{i-1} p(x_j), \qquad i \le k$$



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

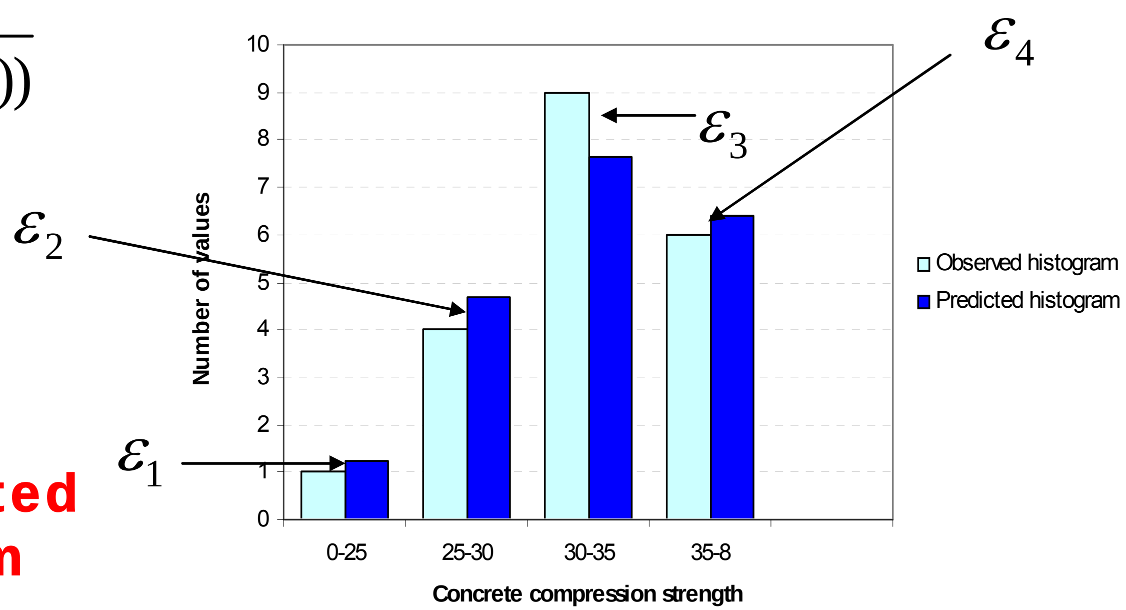
Model Evaluation by Statistical Testing

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

$$P(\varepsilon_m^2 \geq \Delta) = \alpha$$

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**

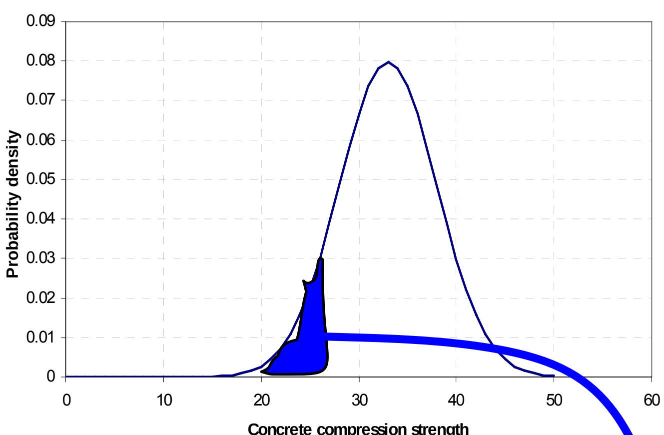
Mean:	33 Mpa
Standard deviation:	5 Mpa

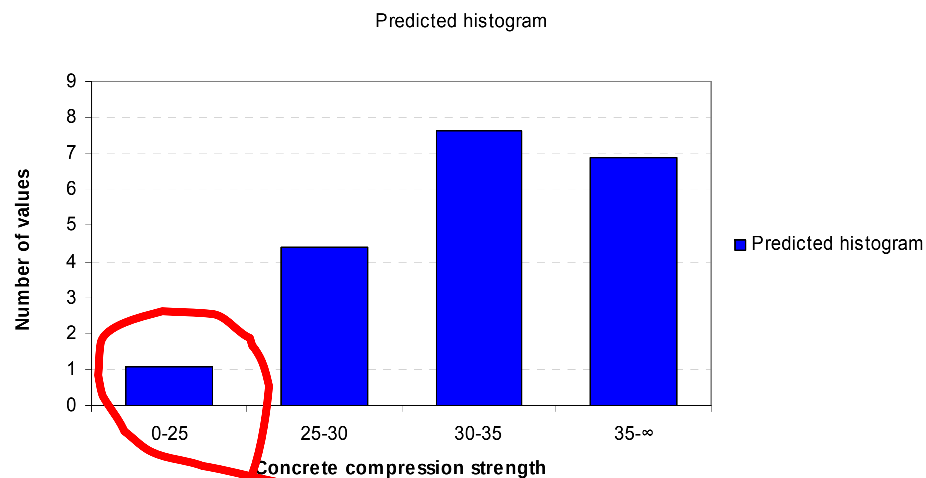
The **Normal distribution** is a continuous probability density function but **can easily be discretized**

Model Evaluation by Statistical Testing

The CHI-square goodness of fit test

The postulated probability density function is discretized:





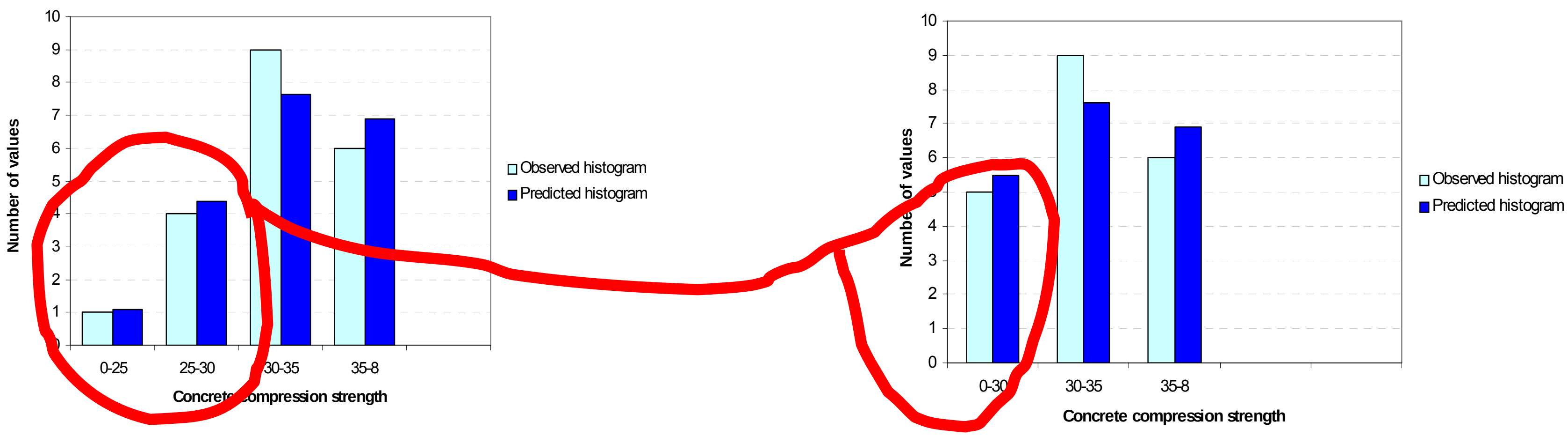
Total number of experiments

0-25: $$20\left[\Phi\left(\frac{25-33}{5}\right)-\Phi\left(\frac{-\infty-33}{5}\right)\right]=20\cdot 0.055=1.10$$

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 degree 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

$$F_o(x_i) = \frac{i}{n}$$

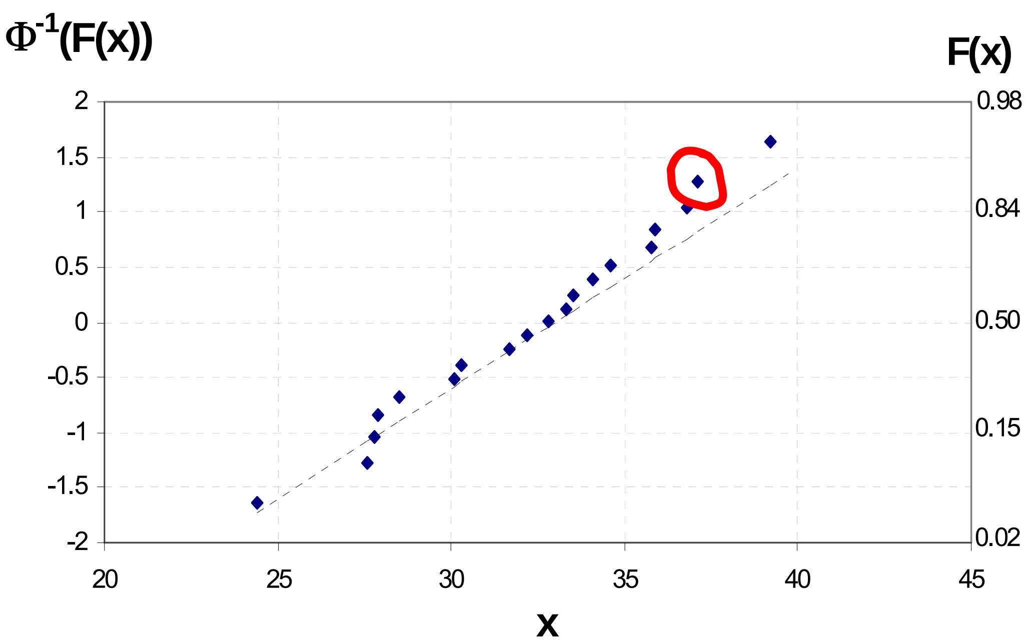
The following statistic has been proposed

$$\varepsilon_{\max} = \max_{i=1}^{n}\left[\left|F_o(x_i) - F_p(x_i)\right|\right] = \max_{i=1}^{n}\left[\left|\frac{i}{n} - F_p(x_i)\right|\right]$$

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