Statistics – a very short course

Peter Križan, Ljubljana

Analysis of data

If we have N independent (unbiased) measurements x_i of some unknown quantity μ with a common, but unknown, variance σ^2, then

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2$$

are unbiased estimates of μ and σ^2. The uncertainties of these estimates are

- for μ: σ/sqrt(N)
- for σ: σ/sqrt(**2**N) (for Gaussian distributed x_i and large N)

Analysis of data 2: unbinned likelihood fit

Assume now that we have N independent (unbiased) measurements x_i that come from a probability density function (p.d.f.) f(x; θ), where θ **= (**θ_1**, ... **θ_m**)** is a set of m parameters whose values are unknown. The method of maximum likelihood takes the estimators θ **to be those values of** θ **that maximize the likelihood function,**

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i; \boldsymbol{\theta}) \ .$$

It is easier to maximize ln*L* (same minimum, but product → sum)

Also – from practical reasons: max (ln*L*) → min(- ln*L*) (minimisation algorithms)

→ Solve a set of m equations

$$\frac{\partial \ln L}{\partial \theta_i} = 0 \ , \qquad i = 1, \ldots, n \ .$$

Analysis of data 3

The errors and correlations between parameters $\theta=(\theta_1,....\theta_m)$ can be found from the inverse of the covariance matrix

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right|_{\widehat{\theta}}$$

The variance σ^2 on the paramter θ_i is V_{ii}

Analysis of data 4: binned likelihood fit

If the sample is large (large n), data can be grouped in a histogram. The content of each bin, n_i, is distributed according to the Poisson distribution with mean $\nu_i(\theta)$,

$$f(\nu_i(\theta), n_i) = \nu_i(\theta)^{n_i} \exp(-\nu_i(\theta)) / n_i!$$

The parameters θ are determined by maximizing a properly normalized likelihood function

$$-2 \ln \lambda(\boldsymbol{\theta}) = 2 \sum_{i=1}^{N} \left[\nu_i(\boldsymbol{\theta}) - n_i + n_i \ln \frac{n_i}{\nu_i(\boldsymbol{\theta})} \right]$$

In the limit of zero bin width, maximizing this expression is equivalent to maximizing the unbinned likelihood function.

N.B. In the expression above we have assumed n_i to be large such that the Stirling approximation can be used, $\ln n! \sim n \ln n - n$

Analysis of data 5: least squares method

If we have N independent measurements of variable **y_i** at points **x_i** , and if y_i are Gaussian distributed around a mean $F(x_i, \theta)$ with variance σ_i^2 , the log likelihood function yields

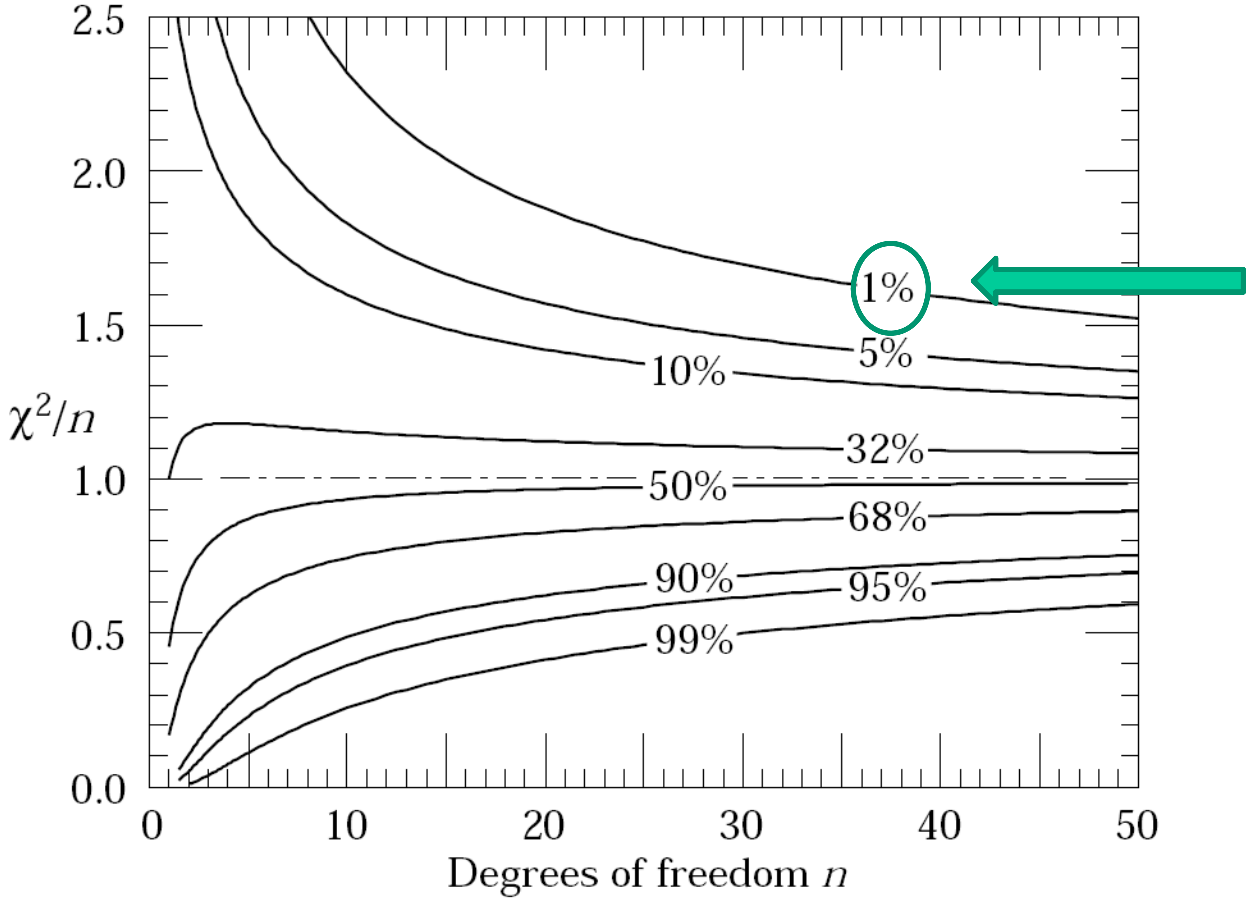
$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - F(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2}$$

and the parameters θ are determined by minimizing this expression.

This weighted sum of squares can be used in a general case of a non-Gaussian distribution → Least squares method

Analysis of data 6: least squares method

The value of χ^2 at the minimum is an indication of the goodness of fit. The mean of χ^2 should be roughly equal to the number of degrees of freedom, n = N-m, where m is the number of parameters. Popular use: for each fit to the data quote χ^2/n



Probability that the fit would give χ^2/n bigger than the observed value