Karlsruher Institut für Technologie

Modern Methods of Statistical Data Analysis

From parameter estimation to deep learning — A guided tour of probability

Lecture 5

Extended/Binned Likelihood & the χ^2 Method

P.-D. Dr. Roger Wolf

roger.wolf@kit.edu

Dr. Pablo Goldenzweig

pablo.goldenzweig@kit.edu

Dr. Jan Kieseler

jan.kieseler@kit.edu

Dr. Slavomira Stefkova

slavomira.stefkova@kit.edu

Program today

- Recap of lecture 4
- Answers to quiz 4
- Extended maximum Likelihood
- Binned Likelihood

5’ break

- χ^2 method
- Goodness of fit
- Quiz 5

Brief recap of last lecture

Review

- Estimators:
  - Use observations to construct functions that estimate properties of PDFs
    - E.g. mean, variance, covariance

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$\bar{x}$: sample mean; μ: population mean

- Introduced concept of **bias** and **consistency**

$$b = E[\hat{\theta}] - \theta \qquad \lim_{n\to\infty} \hat{\theta} = \theta$$

θ: **population parameter (e.g. population mean)** = true parameter in PDF

$\hat{\theta}$: **'hatted' sample parameter (e.g. sample mean)** = parameter we determine using observations

Brief recap of last lecture

Review

- Estimators:
  - Use observations to construct functions that estimate properties of PDFs
    - E.g. mean, variance, covariance

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$\bar{x}$: sample mean; μ: population mean

- Introduced concept of **bias** and **consistency**

$$b = E[\hat{\theta}] - \theta \qquad \lim_{n\to\infty} \hat{\theta} = \theta$$

'hatted' sample parameter (e.g. sample mean)

= parameter we determine using observations

Example last time of consistent but biased estimator:
Decay constant (see L04 slide 35)

$$\hat{\lambda} = 1/\hat{\tau} = n/\sum_{i=1}^{n} t_i.$$

$$E[\hat{\lambda}] = \lambda \frac{n}{n-1}$$

...biased for small n

Brief recap of last lecture

Review

- Method of maximum likelihood
  - Can find estimator for arbitrary parameters of interest of a PDF via maximizing the function
    - Given we have observations $x = (x_1, \ldots, x_n)$ distributed according to $f(x; \boldsymbol{\theta})$

$$\mathscr{L}(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$ ← PDF of x

parameter(s) we wish to estimate

Interpret: *$\mathscr{L}$ is a pure function of the parameters θ with the data x "baked in"*

    - Showed explicitly (= analytically) that this produces unbiased estimators for the mean decay time, if $f(x; \tau)$ is an exponential function.

- Discussed three methods to obtain variance of estimated values:

  1.) Analytical calculation

  2.) MC method

  3.) Graphical method / RCF

Brief recap of graphical / RCF

Review

- Key ingredient: Taylor expand log likelihood function around maximum:

$$\log\mathscr{L}(\theta) = \log\mathscr{L}(\hat{\theta}) + \left[\frac{\partial \log\mathscr{L}}{\partial\theta}\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \log\mathscr{L}}{\partial\theta^2}\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta})^2 + \ldots$$

$\log\mathscr{L}(\hat{\theta})$ = $\log\mathscr{L}_{\max}$ *by definition of* $\hat{\theta}$

$\left[\frac{\partial \log\mathscr{L}}{\partial\theta}\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta})$ = 0 *at maximum*

$\frac{1}{2!}\left[\frac{\partial^2 \log\mathscr{L}}{\partial\theta^2}\right]_{\theta=\hat{\theta}} (\theta-\hat{\theta})^2$ $\hat{\sigma^2}_{\hat{\theta}} = \left(-1 \Big/ \frac{\partial^2 \log\mathscr{L}}{\partial\theta^2}\right)\Big|_{\theta=\hat{\theta}}$ *RCF*

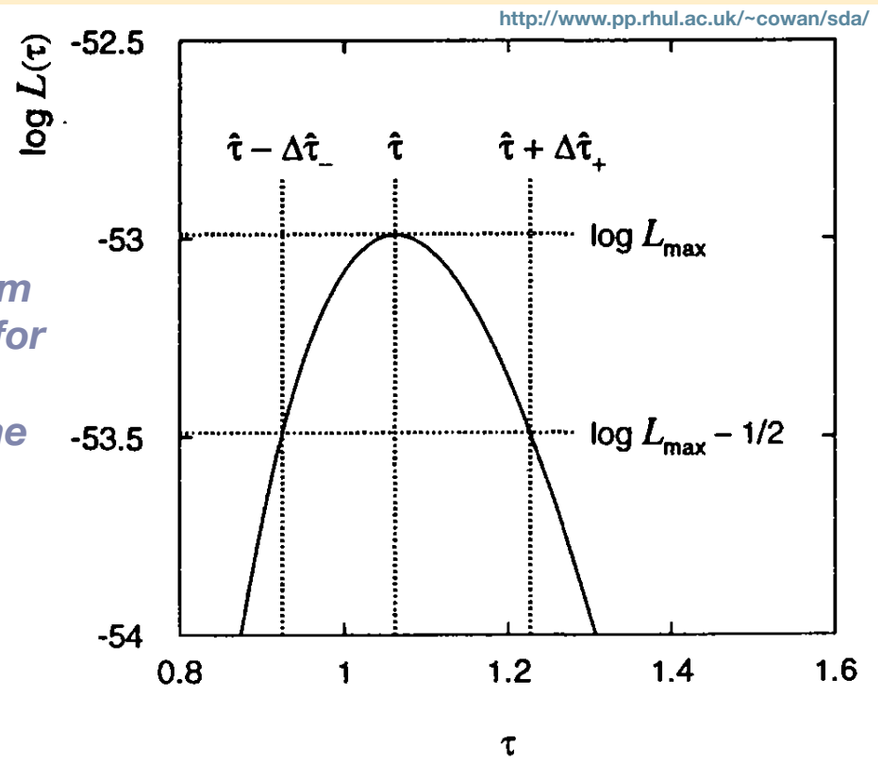
- For a single parameter of interest we identify
- Likelihood evaluated at $\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}$

$$\log\mathscr{L}(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log\mathscr{L}_{\max} - \frac{1}{2}$$

I.e., with known maximum, can determine estimator of variance by solving what value of σ_θ gives a likelihood value of $\log\mathscr{L}_{\max} - 1/2$

Example from last lecture for variance of mean lifetime

http://www.pp.rhul.ac.uk/~cowan/sda/



Answer Time: Quiz 4

Do your own Likelihood fit

2. Let us assume you have a simple PDF of the form

$$f(x;\lambda) = 1 + \lambda\,(x - 0.5)\ , \tag{2}$$

with a sample space S spanning the interval $[0, 1]$ such that $\int_S f(x;\lambda)\mathrm{d}x = \int_0^1 f(x;\lambda)\mathrm{d}x = 1$.

a) First sketch the PDF for $\lambda = 1, 0, -1$.

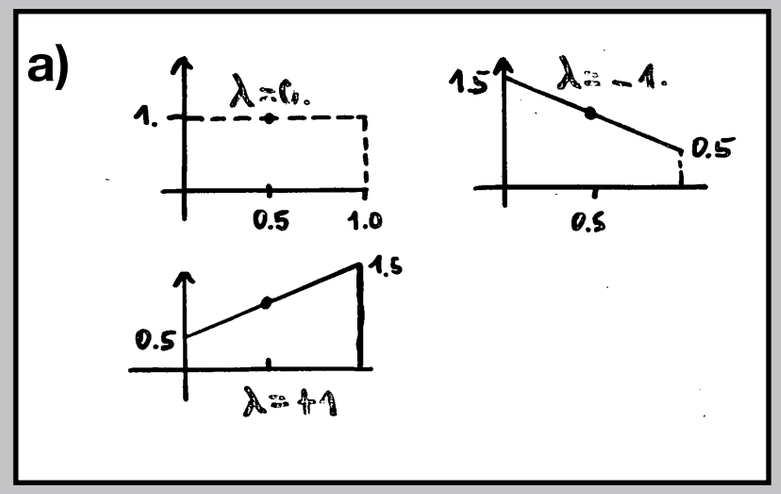
b) Five measurements were done giving $x = (0.89, 0.03, 0.50, 0.31, 0.49)$. Calculate the log-likelihood function for three different values of $\lambda = 1, -0.5, -1$ by hand.

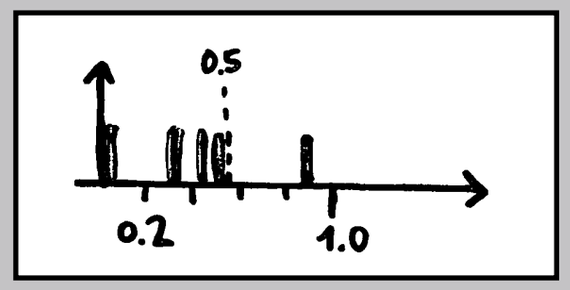
c) The log-likelihood function is in good approximation a parabolic function, i.e. can be described by a polynomial of second order as a function of the tested value λ. Calculate the coefficients a, b, c of $\log L(\lambda) = a\lambda^2 + b\lambda + c$. For what value of λ is $\log L(\lambda)$ maximal?

d) Sketch the log-likelihood function. Using the graphical method, i.e.

$$\log L(\hat{\lambda} \pm \hat{\sigma}_{\hat{\lambda}}) = \log L_{\max} - \frac{1}{2}\,, \tag{3}$$

determine the uncertainty $\hat{\sigma}_{\hat{\lambda}}$ of the estimated parameter $\hat{\lambda}$ (with $\hat{\lambda}$ denoting the value where $\log L(\hat{\lambda}) = \log L_{\max}$ and $L_{\max}$ is the maximal likelihood value).





b)

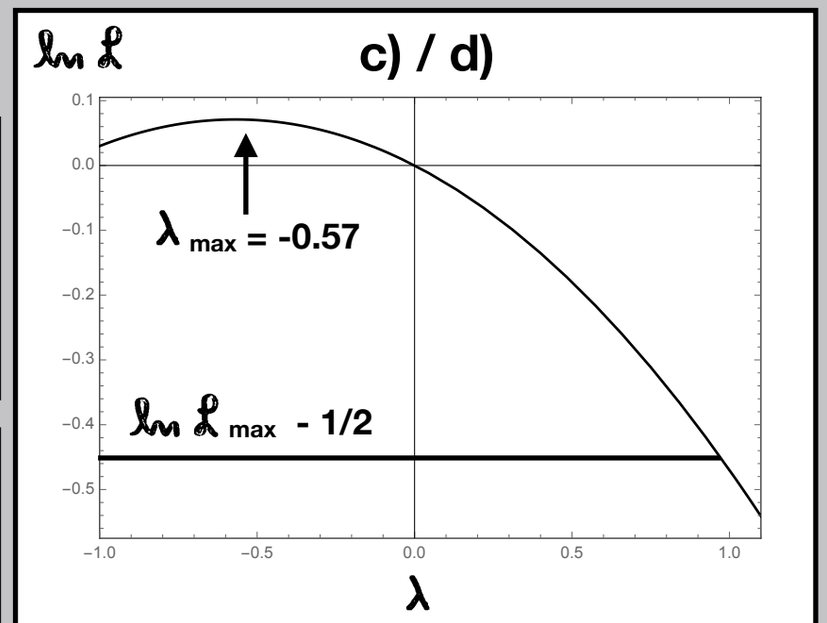
$\lambda = +1.0$ $\quad \ln \mathcal{L} = \ln 1.39 + \ln 0.53 + \underline{\ln 1.0} + \ln 0.86 + \ln 0.93 = -0.47$

$\lambda = -0.5$ $\quad \ln \mathcal{L} = \ln 0.81 + \ln 1.24 + \underline{\ln 1.0} + \ln 1.07 + \ln 1.01 = 0.07$

$\lambda = -1.0$ $\quad \ln \mathcal{L} = \ln 0.61 + \ln 1.47 + \underline{\ln 1.0} + \ln 1.14 + \ln 1.01 = 0.03$

c) **Solving the three equations you can construct with the three λ values you should find (using the rounded values from above)**

a= -0.22, b = -0.25, c = 0.0



$$\hat{\lambda} \pm \hat{\sigma}_{\hat{\lambda}} = -0.57 \pm 1.51$$

Extended ML

But first recall our old friend the Poisson RV

Poisson RV → *# of successes in an interval of time at a fixed rate*

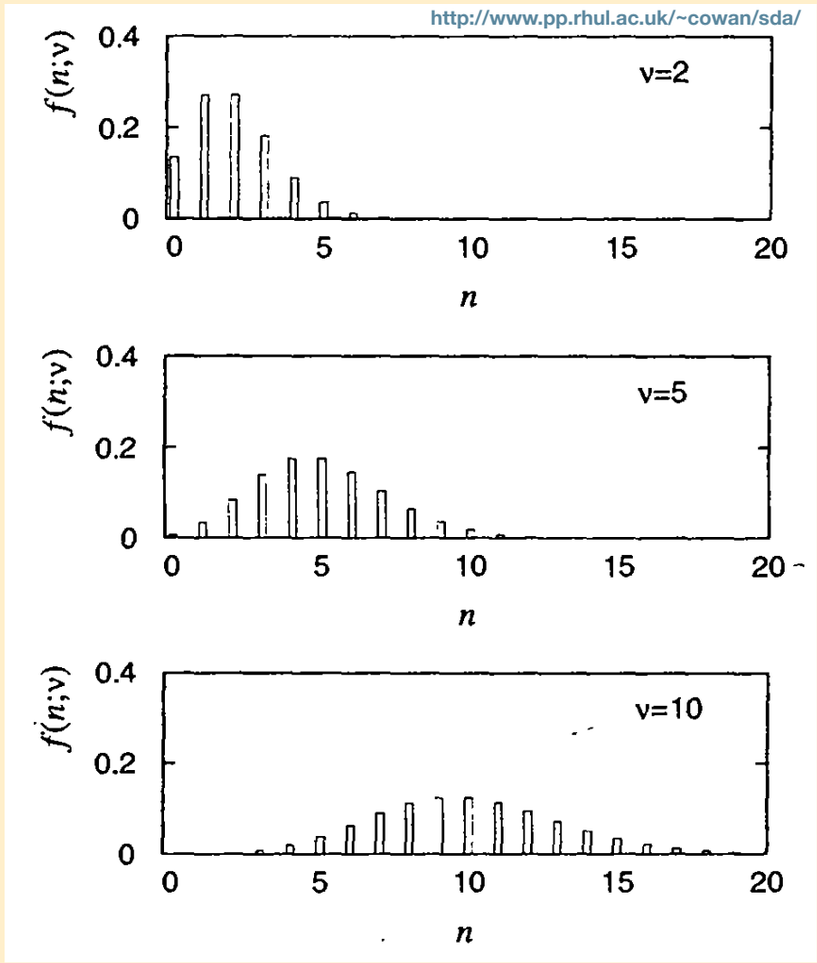
Consider an experiment that lasts a fixed interval of time

def: A Poisson random variable n is the # of successes over the experiment duration, assuming the time each success occurs is independent and the average # of successes over time is constant

$$f(n;\nu) = \frac{\nu^n}{n!}e^{-\nu}$$

$$E[n] = \sum_{n=0}^{\infty} n\frac{\nu^n}{n!}e^{-\nu} = \nu$$

$$V[n] = \sum_{n=0}^{\infty} (n-\nu)^2\frac{\nu^n}{n!}e^{-\nu} = \nu$$



Extended maximum likelihood

- Consider a random variable X distributed according to a PDF $f(x;\boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_1, \ldots, \theta_m)$
  - Further suppose we have data $x_1, \ldots, x_n$
  - So far we assumed that the number of observations n in the data sample always stays the same
  - **But** it is often the case that n itself is a **Poisson random variable** with some mean value ν
    - The likelihood function is then the product of the Poisson distribution probability to find n and the usual likelihood

$$\mathscr{L}(\nu, \boldsymbol{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) = \frac{e^{-\nu}}{n!} \prod_{i=1}^{n} \nu\, f(x_i;\boldsymbol{\theta})$$

Interpret: *the result of the experiment can be defined as the number of observations n in the sample* ***and*** *the n values $x_1, \ldots, x_n$*

Extended maximum likelihood

- **2 situations of interest:**
  - The Poisson parameter ν is given as a function of θ
  - The Poisson parameter ν is treated as an independent parameter

Extended maximum likelihood: Case 1

- **2 situations of interest:**

$$\mathscr{L}(\boldsymbol{\theta}) = \frac{e^{-\nu(\boldsymbol{\theta})}}{n!} \prod_{i=1}^{n} \nu(\boldsymbol{\theta})\, f(x_i; \boldsymbol{\theta})$$

  - The Poisson parameter ν is given as a function of $\boldsymbol{\theta}$

$$\log \mathscr{L}(\boldsymbol{\theta}) = n \log \nu(\boldsymbol{\theta}) - \nu(\boldsymbol{\theta}) + \sum_{i=1} \log f(x_i; \boldsymbol{\theta})$$

Additive terms not depending on θ have been dropped

$$= -\nu(\boldsymbol{\theta}) + \sum_{i=1}^{n} \log \left(\nu(\boldsymbol{\theta})\, f(x_i; \boldsymbol{\theta}) \right)$$

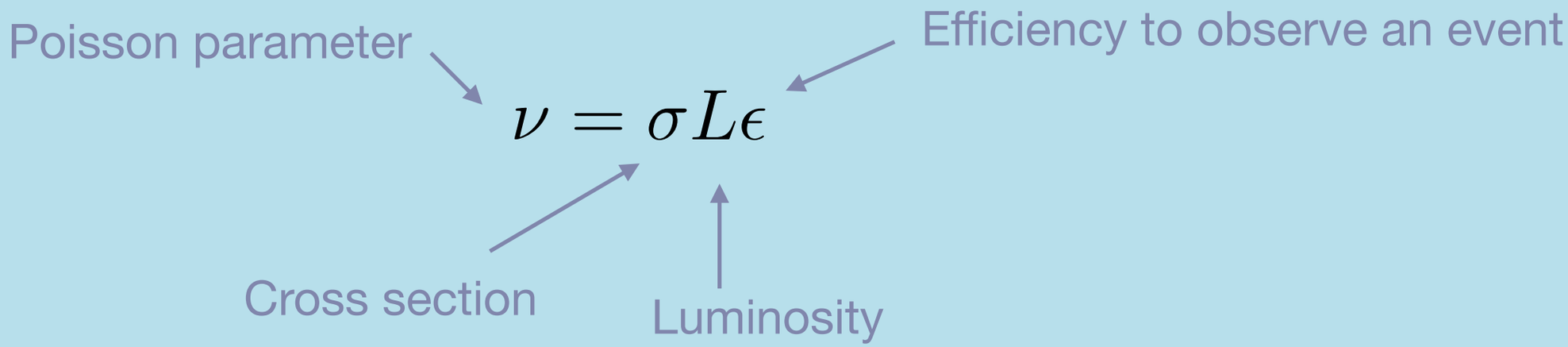
By including the Poisson term, the resulting estimators $\hat{\theta}$ exploit the information from n as well as from the variable x.

This leads to smaller variances for $\hat{\theta}$ than in the case where only the x values are used

Recall the particle scattering reaction

(L02, slide 54)

- Example:
  - number of decays of radioactive material in a fixed time period in the limit that the total number of decays is large
  - number of events of a certain type observed in a particle scattering experiment with a given integrated luminosity L. The **expectation value of the number of events** is



The statistical errors of the estimated parameters (e.g., cross section) depend on parameters such as particle masses and coupling constants

Will in general be smaller by including the information from the cross section

Extended maximum likelihood: Case 2

- **2 situations of interest:**
  - The Poisson parameter ν is given as a function of θ
  - The Poisson parameter ν is treated as an independent parameter

$$\mathscr{L}(\nu, \boldsymbol{\theta}) = \frac{e^{-\nu}}{n!} \prod_{i=1}^{n} \nu \, f(x_i; \boldsymbol{\theta})$$

Here, setting $\dfrac{\partial \log \mathscr{L}(\nu, \boldsymbol{\theta})}{\partial \theta_i} = 0$ one obtains the same estimators $\hat{\theta}_i$ as in the usual ML case

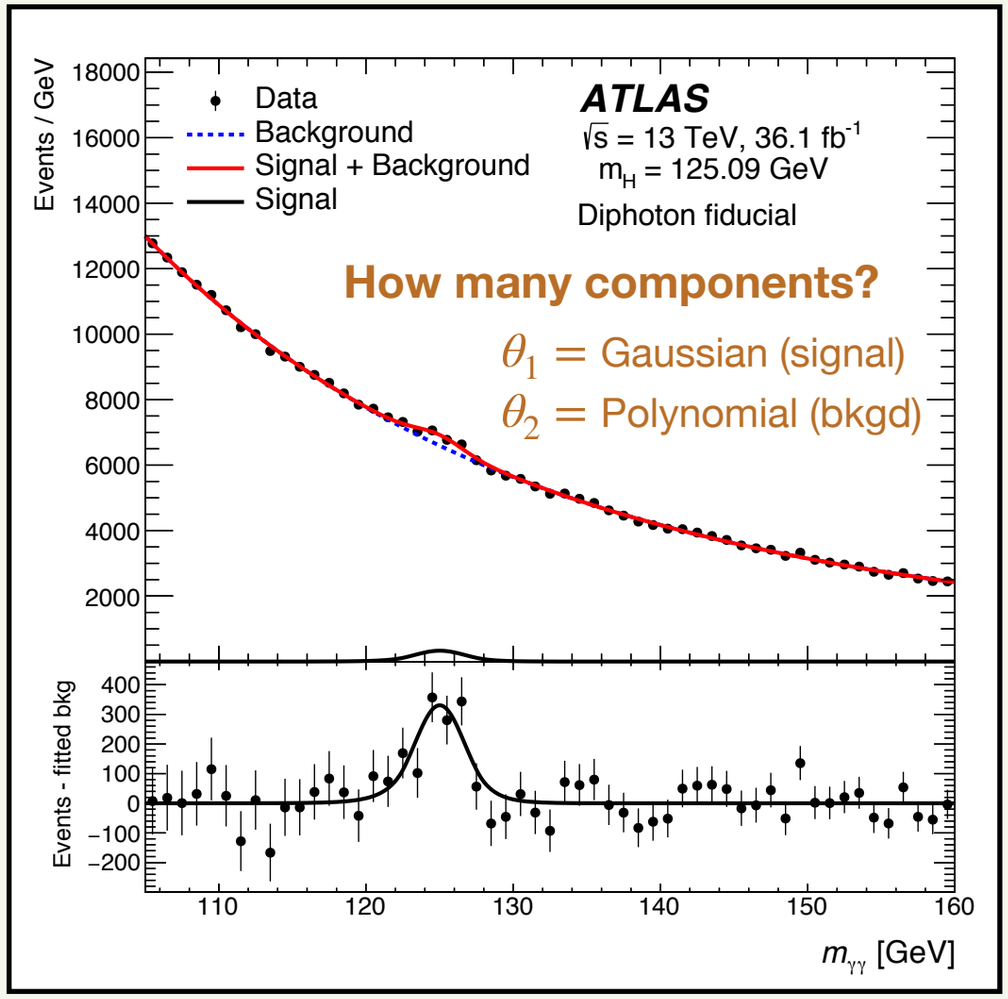
So why bother, since all you seem to have done is introduce an additional source of statistical fluctuation by regarding n as a RV?

Extended maximum likelihood: Case 2

Given: The PDF of some variable x (here $m_{\gamma\gamma}$) is the superposition of several components

$$f(x;\boldsymbol{\theta}) = \sum_{i=1}^{m} \theta_i f_i(x)$$

Task: Estimate the relative contribution of each component θ_i



$$\log \mathscr{L}(\boldsymbol{\nu},\boldsymbol{\theta}) = -\nu + \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} \nu\theta_i f_j(x_i)\right)$$

Define: $\mu_i = \theta_i \nu$

(the expected # of events of type i)

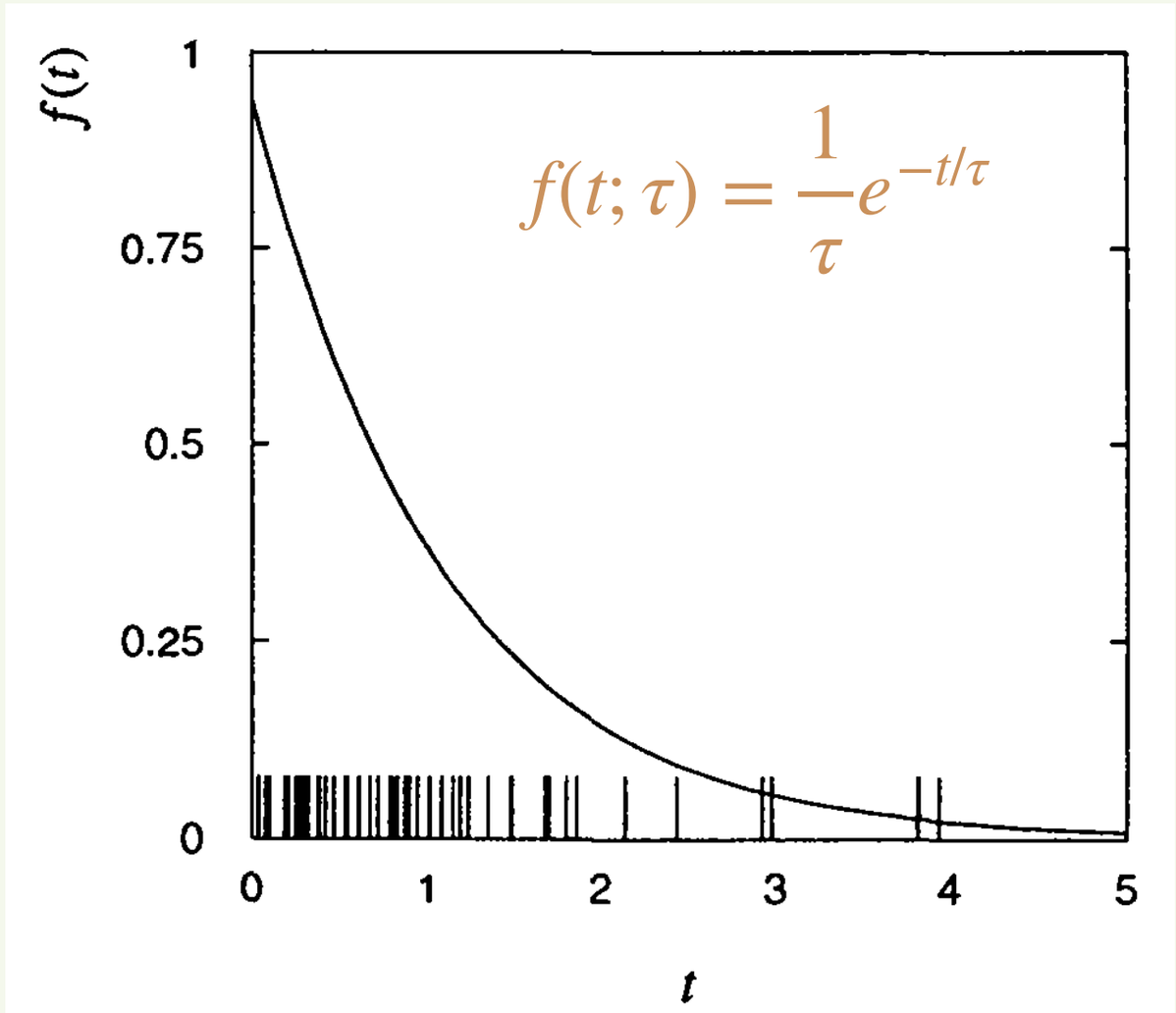
$$\log \mathscr{L}(\boldsymbol{\mu}) = -\sum_{i=1}^{m} \mu_j + \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} \mu_j f_j(x_i)\right)$$

Interpret: *The total number of events n is viewed as a sum of independent Poisson variables with means μ_i. Here the μ_i are more closely related to the production cross section for events of type i.*

Binned ML

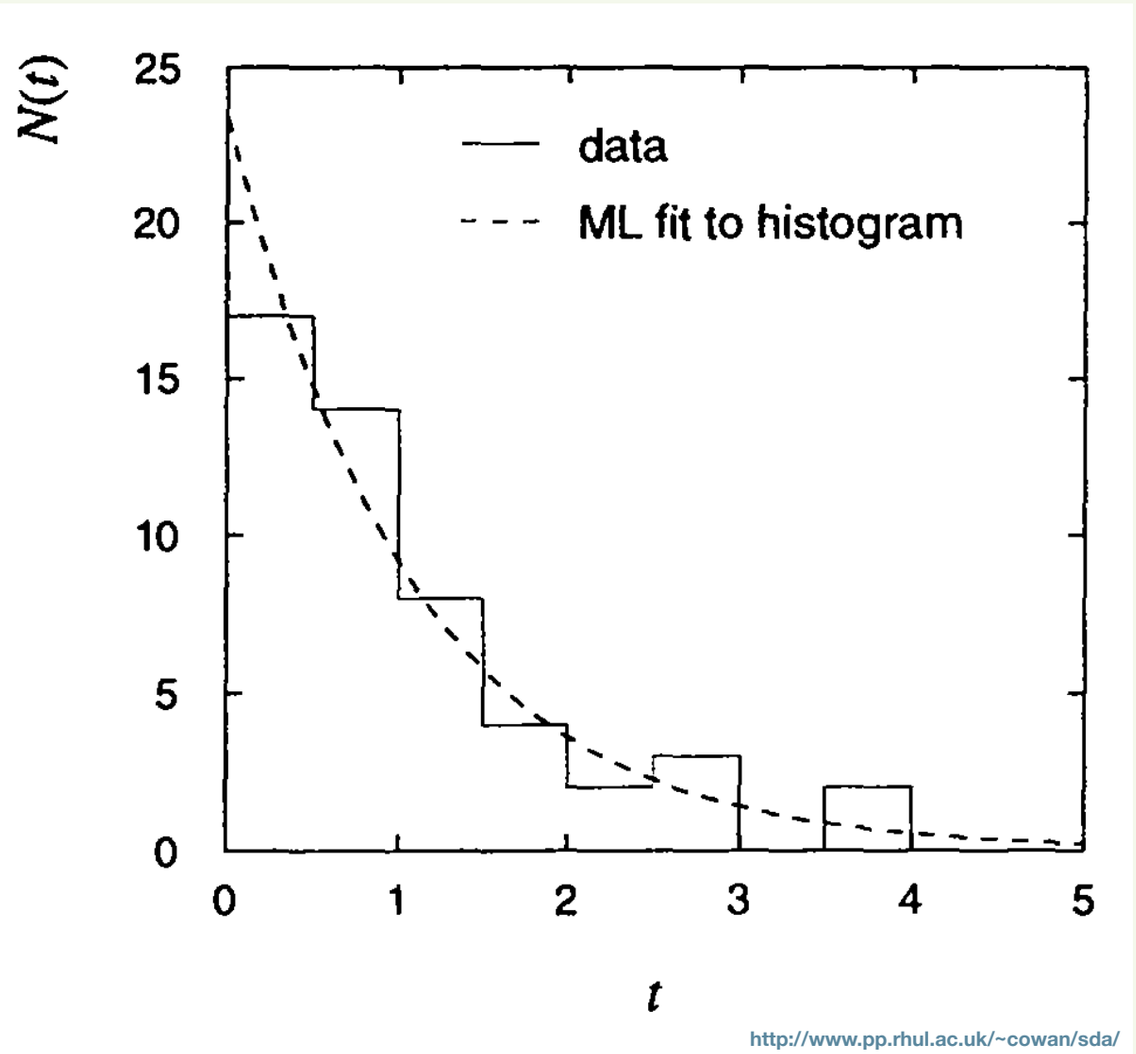
Binned Likelihood fit

- Consider again the sample with 50 measured particle decay times
  discussed in L04 (slide 34)



Same sample displayed as a histogram

$\Delta t = 0.5$



http://www.pp.rhul.ac.uk/~cowan/sda/

Binned Likelihood fit

Why do it?

For very large data sample, $\log \mathscr{L}$ is very expensive to compute

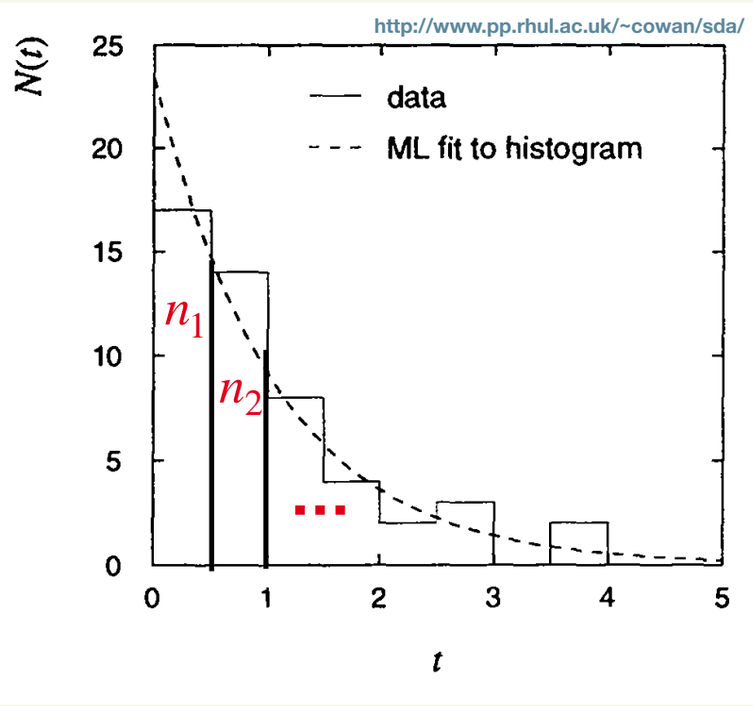
Must sum $\log f(x_i; \theta)$ for each x_i

Histogram: number of entries $\boldsymbol{n} = (n_1, \ldots, n_N)$ in N bins

$$\nu_i(\boldsymbol{\theta}) = n_{\text{tot}} \int_{x_i^{\min}}^{x_i^{\max}} f(x; \boldsymbol{\theta}) dx$$

Expectation values $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ of the number of entries

Bin limits



Binned Likelihood fit

as a single measurement of an N-dimensional random vector for which the joint PDF is given by a multinomial distribution

$$f_{\text{joint}}(\boldsymbol{n};\boldsymbol{\nu}) = \frac{n_{\text{tot}}!}{n_1!\ldots n_N!}\left(\frac{\nu_1}{n_{\text{tot}}}\right)^{n_1}\cdots\left(\frac{\nu_N}{n_{\text{tot}}}\right)^{n_N}$$

Prob to be in bin i is the expectation value (ν_i) divided by the total number of entries (n_{tot})

$$\log \mathscr{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} n_i \log \nu_i(\boldsymbol{\theta})$$

(Additive terms not depending on θ have been dropped)

In the limit of small bin size → nearly identical with unbinned ML.

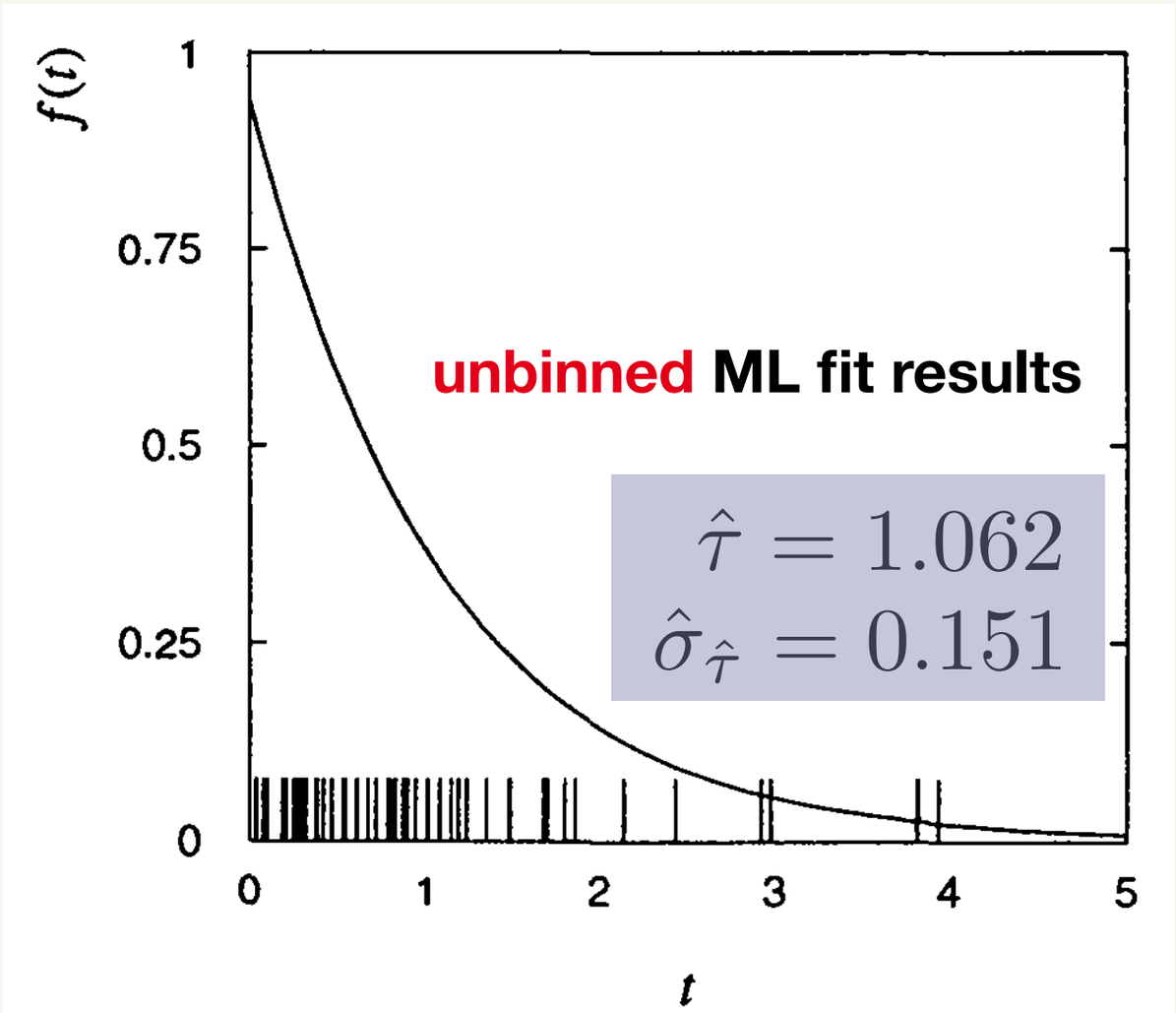
No problem if bins are empty or have few entries.

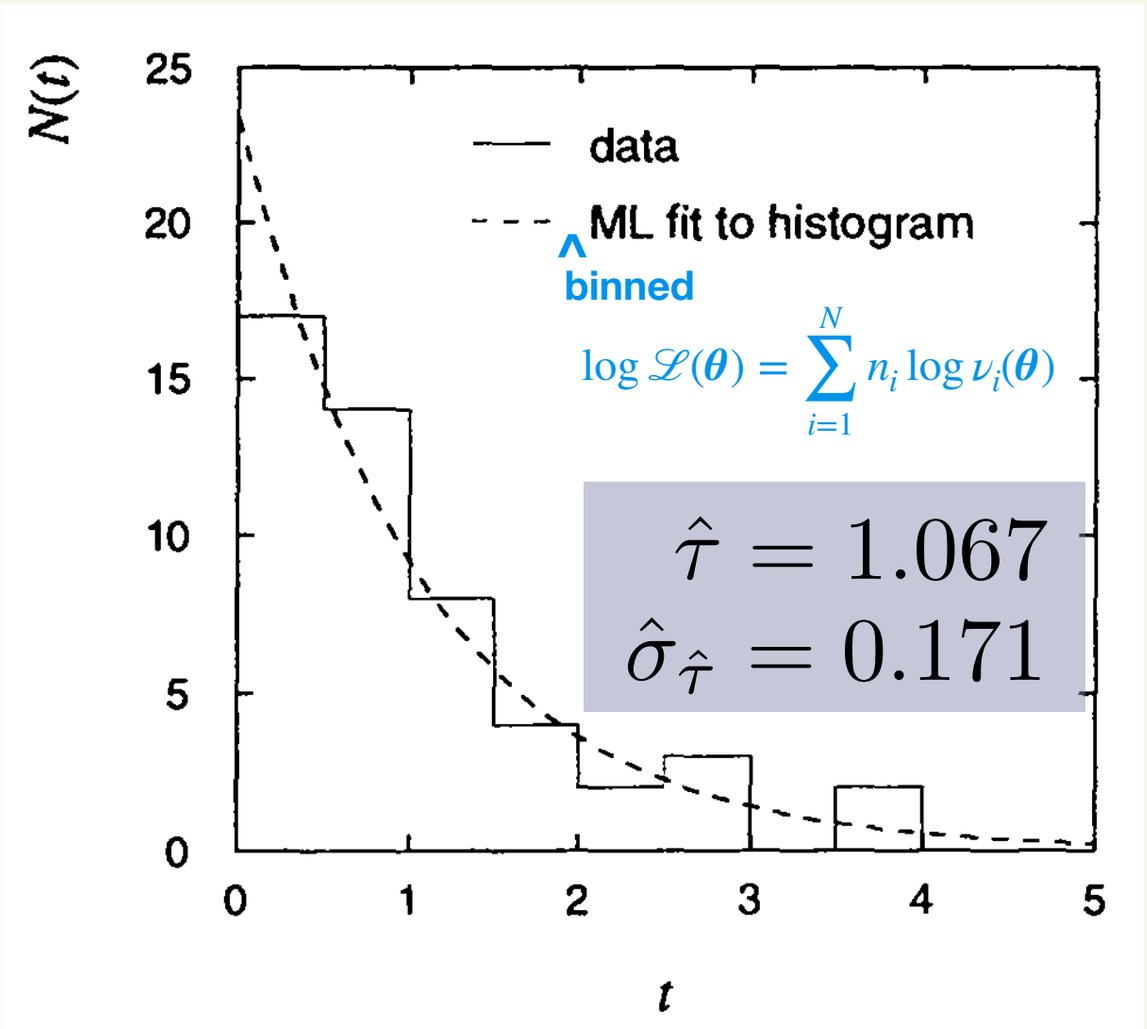
Binned Likelihood fit

- Now compare the results of binned vs. unbinned fits:

Both results consistent, but standard deviation larger for binned fit

In large sample limit this difference disappears





http://www.pp.rhul.ac.uk/~cowan/sda/

Read up on extended log-likelihood for binned data (Cowan pages 88-89)

Take 5

P-value

A simple example to warm up with

- Experiment:
  - **Assume** you have a fair coin (i.e., equal probability of heads or tails).
  - Toss the coin $N = 20$ times.
    - Expect $n_h = 10$ heads.

- Result:
  - Observe $n_h = 17$ heads.

- Question:
  - If this experiment is repeated many times, what is the probability of obtaining a result with the same level of discrepancy with the hypothesis (=fair coin) or higher?

- Answer:
  - P-value.

A simple example to warm up with

- Start with the binomial distribution for an experiment with N trials, characterized by n successes (RV) with probability p

$$f(n;N,p) = \frac{N!}{n!(N-n)!}p^n(1-p)^{N-n}$$

- Define the number of heads as success (n_h), and include your hypothesis *that the coin is fair* $p = 0.5$

$$f(n_h;N) = \frac{N!}{n_h!(N-n_h)!}\left(\frac{1}{2}\right)^{n_h}\left(\frac{1}{2}\right)^{N-n}$$

- **Question:** What is the probability of obtaining our result of $n_h = 17$ or an even larger discrepancy from the hypothesis of a fair coin?
  - **Answer:** Sum of the probabilities of $n_h = 0,1,2,3,17,18,19,20$.
    - Using our equation, we get P-value $= 0.0026$.
    - If this experiment ($N = 20$ coin tosses) were repeated many times under similar circumstances, there is $0.26\,\%$ probability of obtaining a result as compatible or less with our hypothesis (fair coin) than the one actually observed ($n_h = 17$ heads).

      Interpret: *The low value implies that there is a low level of agreement between the observed measurements and the assumption (prediction) we made.*

Is this a good fit?

Most famous answer: Pearson's χ^2

On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling

Sir Karl Pearson (1900),
Phil. Mag (5) 50, 157-175

Recall the χ^2 distribution

Chi-square distribution (i)

- The χ^2 (chi-square) distribution of the continuous variable z $(0 < z < \infty)$ is defined by

$$f(z;n) = \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{n/2-1} e^{-z/2}, \qquad n = 1, 2, \ldots,$$

- The parameter n is called **number of degrees of freedom** and the gamma function:

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

To calculate χ^2, need to know:
$\Gamma(n) = (n-1)!$ for integer n,
$\Gamma(x+1) = x\Gamma(x)$ and $\Gamma(1/2) = \sqrt{\pi}$

- Expectation value and variance:

$$E[z] = \int_0^\infty z \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{n/2-1} e^{-z/2}\, dz = n$$

Note that the expectation value is equal to the number of degrees of freedom

$$V[z] = \int_0^\infty (z-n)^2 \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{n/2-1} e^{-z/2}\, dz = 2n$$

Chi-square distribution (ii)

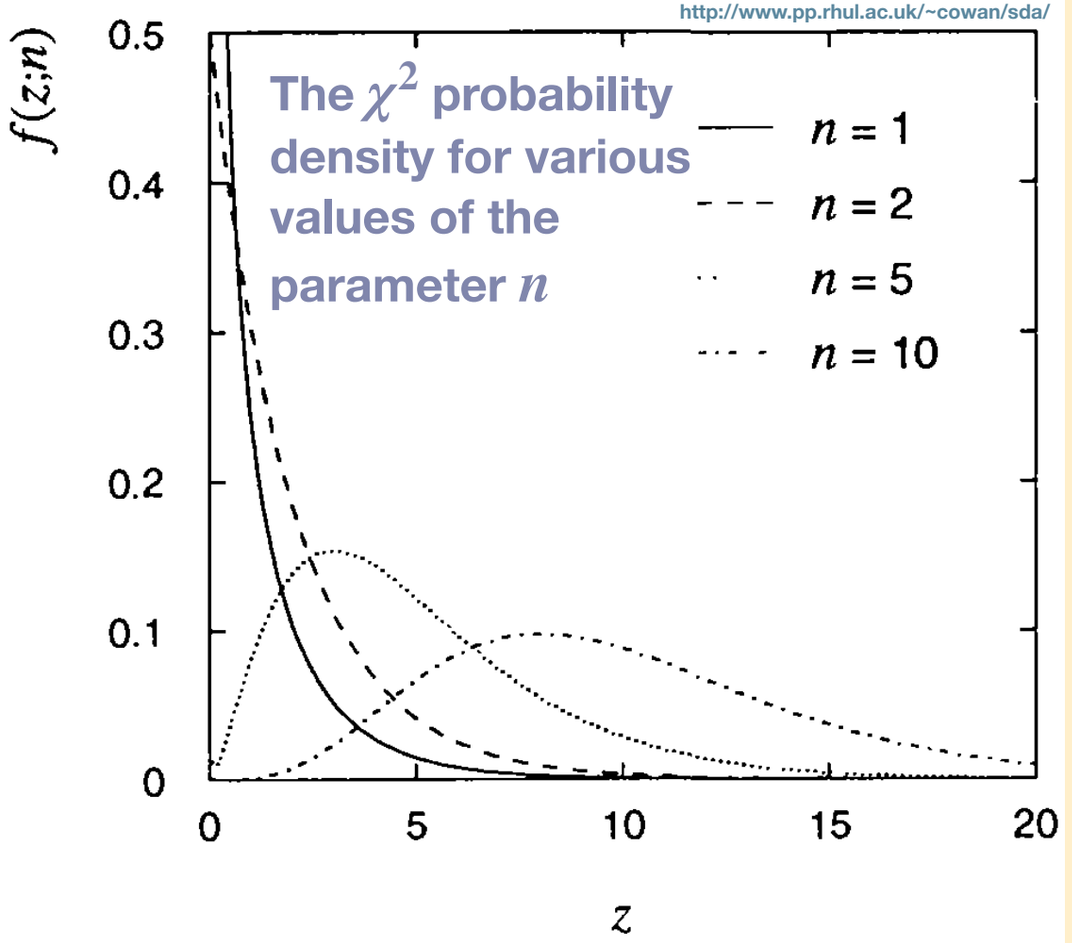
- The χ^2 **distribution** is important due to its **relation** to the **sum of squares of Gaussian distributed random variables.** Given N independent Gaussian random variables x_i with known means μ_i and variances σ_i^2, the variable

$$z = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

is distributed like a χ^2 **distribution with N degrees of freedom.**

Proof in Cowan Sec. 10.2

http://www.pp.rhul.ac.uk/~cowan/sda/



Also holds if x_i are **not independent** but are N-dimensionally Gaussian distributed

$$z = (\boldsymbol{x} - \boldsymbol{\mu})^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

Recall we said that variables following a χ^2 ***distribution will play an important role in tests of goodness-of-fits***

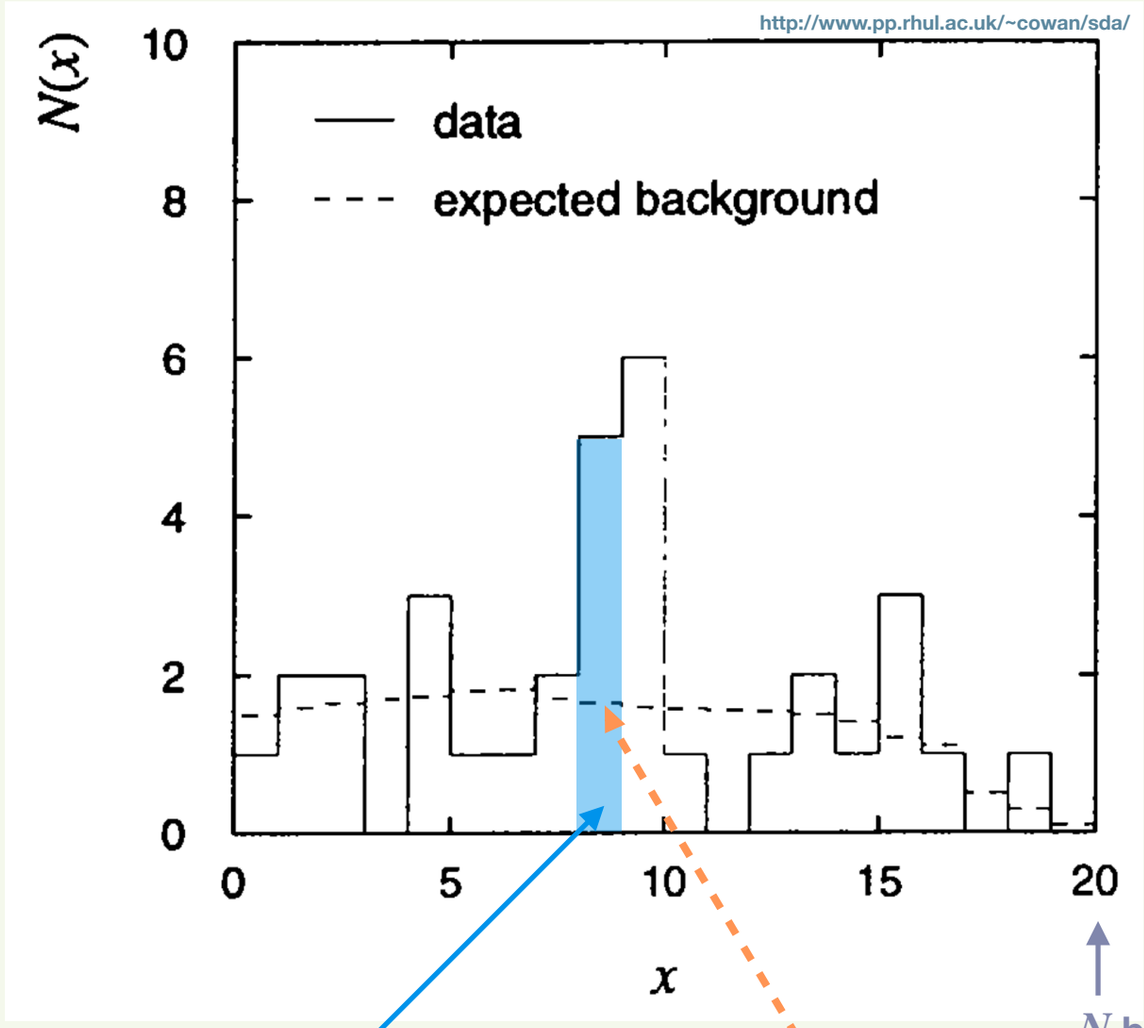
Pearson's χ^2 test

A goodness-of-fit test that can be applied to the distribution of a variable x

Construct a statistic which reflects the level of agreement between observed and expected histograms

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

- Data $\boldsymbol{n}$ are Poisson distributed with mean values $\boldsymbol{\nu}$.
- Since the σ of a Poisson RV with mean ν_i is $\sqrt{\nu_i}$, this statistic gives the sum of squares of the deviations between observed and expected values, measured in units of σ.

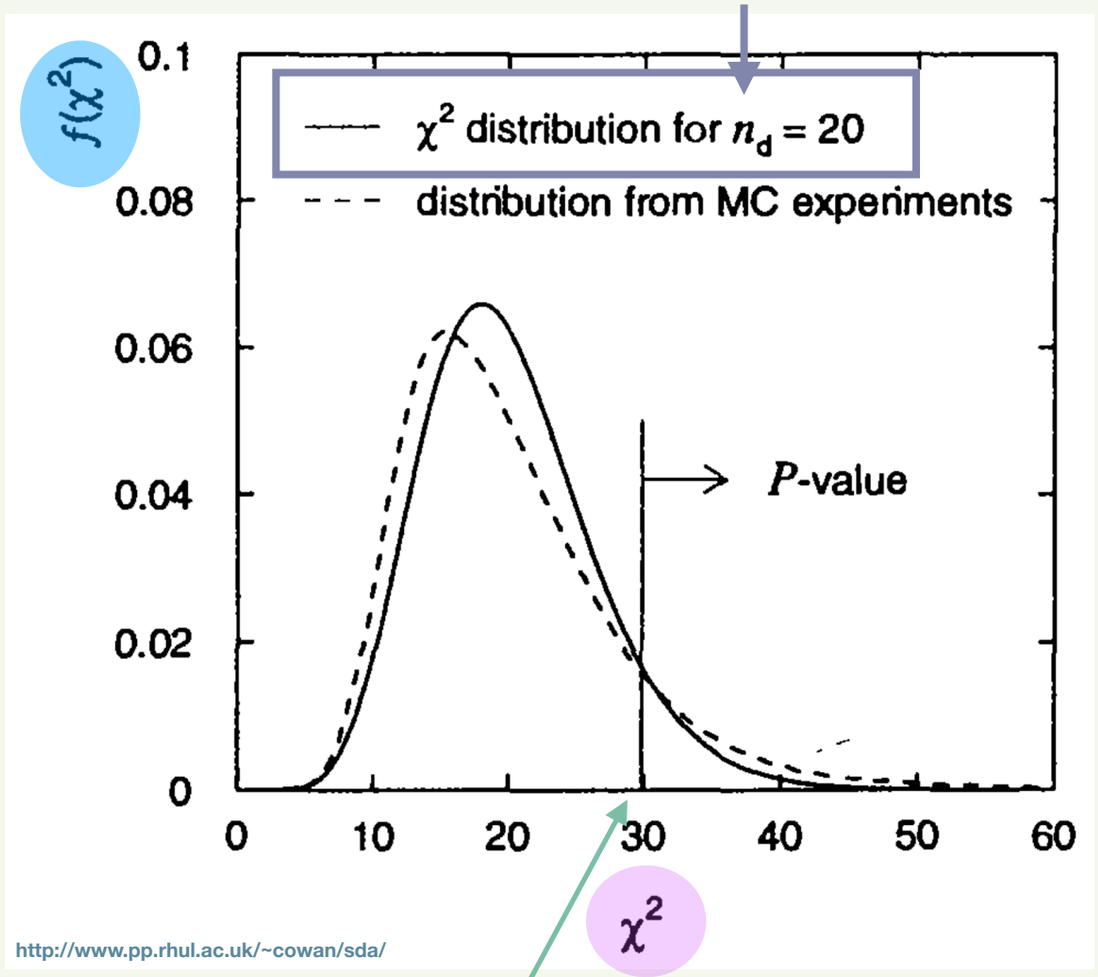


Obtain $\chi^2 = 29.8$ for $N = 20$ degrees of freedom

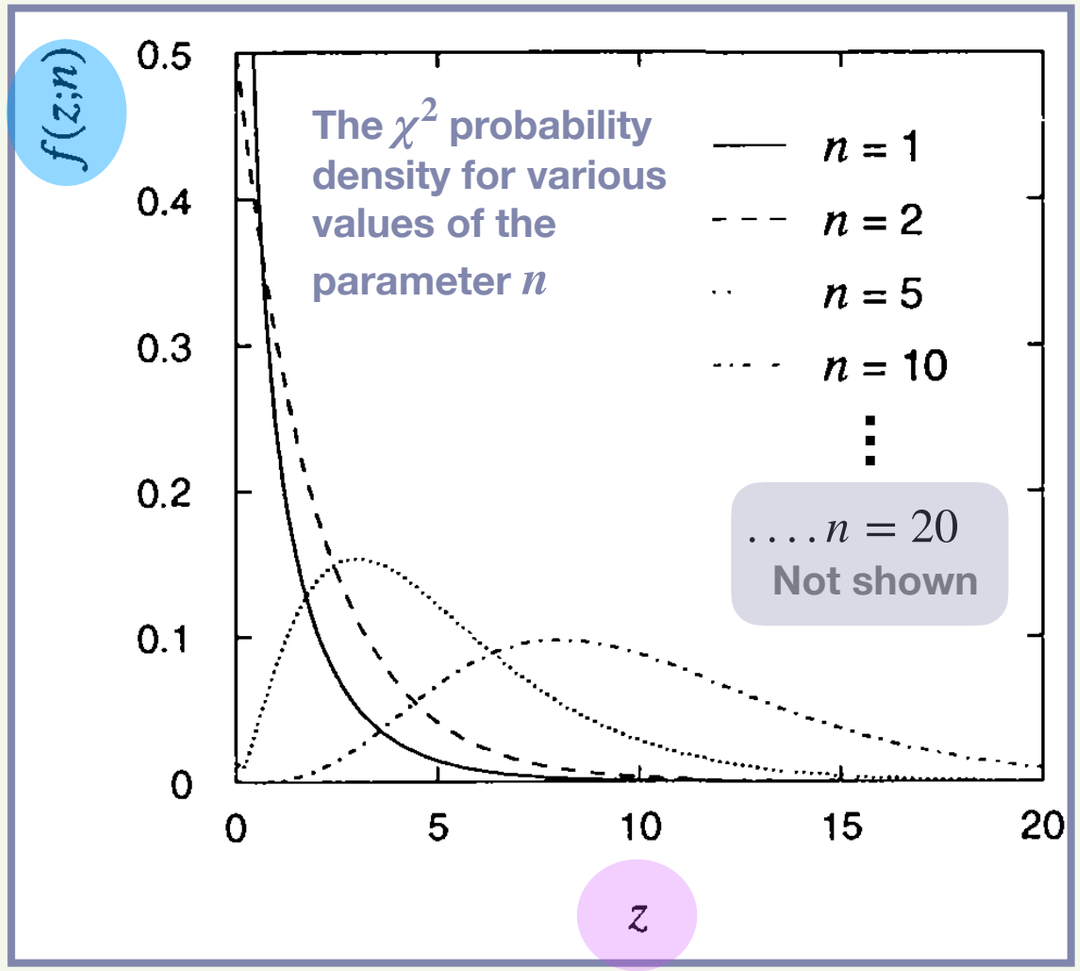
Pearson's χ^2 test

Find probability corresponding to $\chi^2 = 29.8$ by inspecting

distribution of $f(\chi^2)$ for 20 degrees of freedom



http://www.pp.rhul.ac.uk/~cowan/sda/



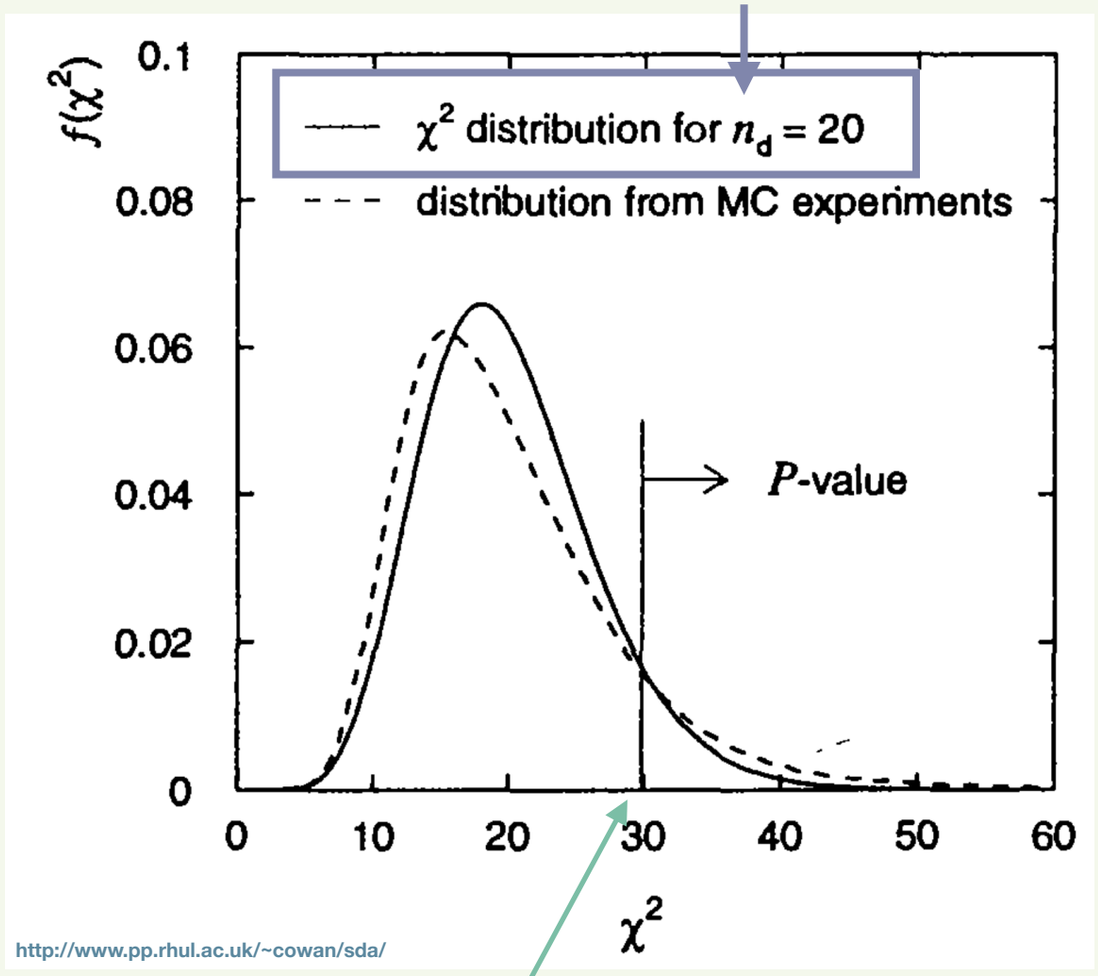
(Careful the difference in notation)

$\chi^2 = 29.8$

Pearson's χ^2 test

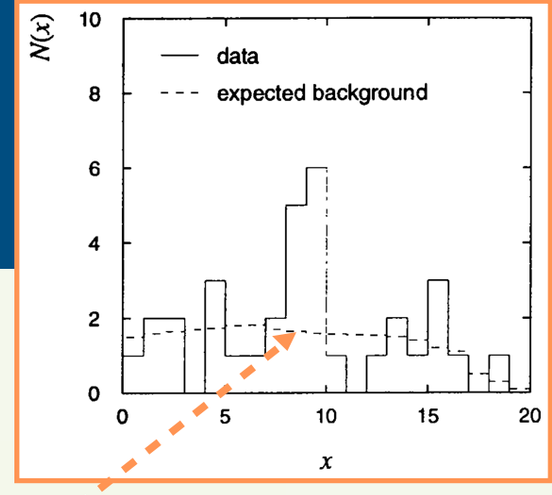
Find probability corresponding to $\chi^2 = 29.8$ by inspecting

distribution of $f(\chi^2)$ for 20 degrees of freedom



http://www.pp.rhul.ac.uk/~cowan/sda/

$\chi^2 = 29.8$



Then calculate the ***P*-value** ≡ the probability, under the hypothesis in question H_0, of obtaining a result as compatible or less with H_0 than the one actually observed

$$P = \int_{\chi^2}^{\infty} f(z; n_d)\, dz$$

i.e. $f(\chi^2)$ for $n_d = 20$

$= 0.073$ **Careful:** Is this reliable?

No! Too few entries / bin in our original histogram

$= 0.11$ **Obtained from MC program**

Done by generating Poisson values n_i for each bin based on the mean value ν_i, and then computing and recording the χ^2 value

Method of Least Squares

Method of least squares

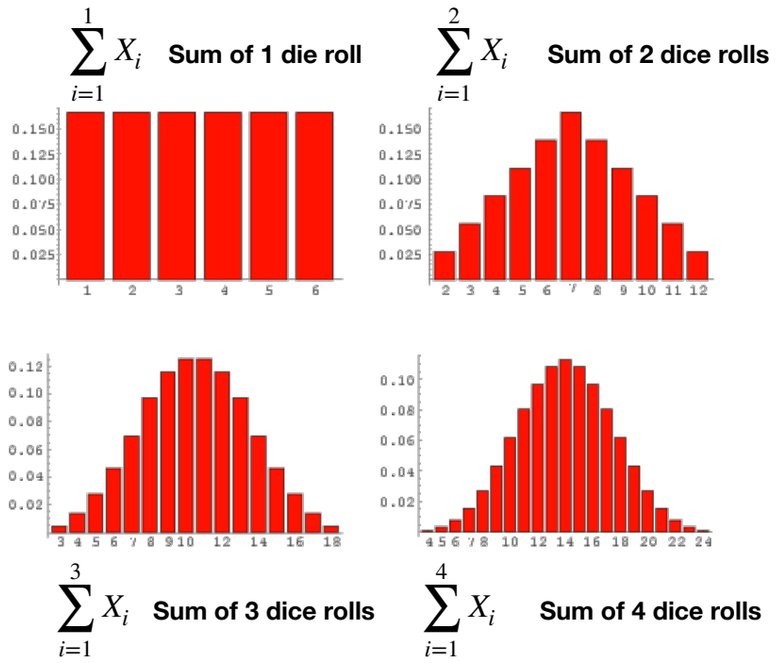
- In many situations, measured values can be regarded as Gaussian random variables
  - Consequence of the **Central Limit Theorem (CLT)**
    - Total error often sum of a large number of small contributions

Sum of dice rolls example we looked at in L03

Roll n independent dice

Let X_i be the outcome of roll i

X_i are i.i.d.



- Consider now a set of N <u>independent</u> Gaussian random variables y_i, $i = 1, \ldots, N$ each related to another variable x_i, which is assumed to be known without error

Method of least squares

- Further
  - Each value y_i has a different **unknown** mean λ_i
  - And each value y_i has **known** variance σ_i^2
- As before we can understand this set of measurements (e.g. from a single experiment) to be a random vector itself that changes if we repeat the experiment
- The joint PDF describing this vector is the product of N Gaussians:

$$g(y_1, \ldots, y_N; \lambda_1, \ldots, \lambda_N, \sigma_1^2, \ldots, \sigma_N^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(y_i - \lambda_i)^2}{2\sigma_i^2}\right)$$

  - Further assume that $\lambda_i = \lambda(x_i; \boldsymbol{\theta})$ **i.e. function of some parameters of interest we want to determine**

Method of least squares

$$g(y_1, \ldots, y_N; \lambda_1, \ldots, \lambda_N, \sigma_1^2, \ldots, \sigma_N^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(y_i - \lambda_i)^2}{2\sigma_i^2}\right)$$

Take the log (and drop additive terms that do not depend on the parameters)

$$\log \mathscr{L}(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{\left(y_i - \lambda(x_i; \boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

Maximize by finding the values of the parameters θ that minimize $\chi^2(\theta)$

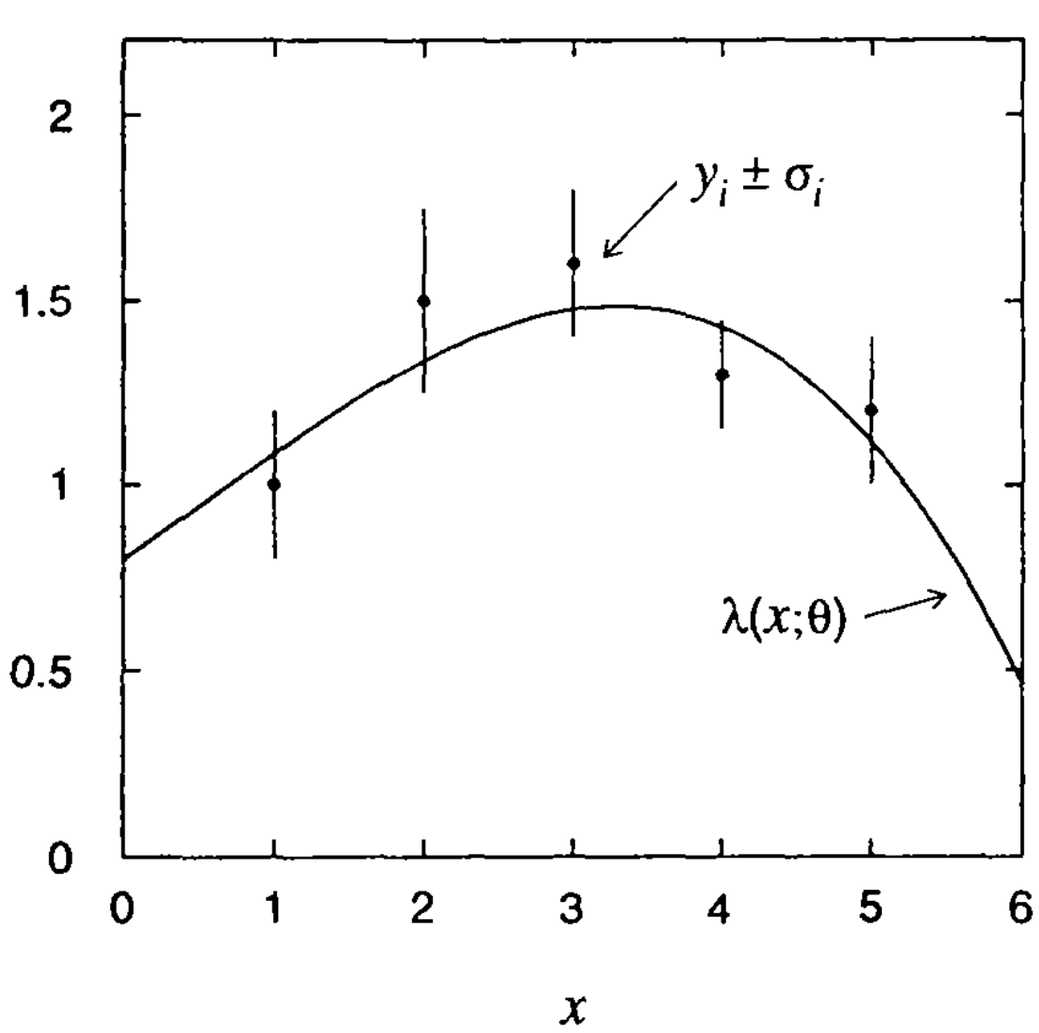
$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - \lambda(x_i; \boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

…the quadratic sum of the differences between the measured (y_i) and hypothesized (λ_i) values, weighted by the inverse of the variances

In other words…

- Ingredients to a least square problem:
  - N values $y_1, \ldots, y_N$ are measured with errors $\sigma_1, \ldots, \sigma_N$
  - The true value $\lambda_i = \lambda(x_i; \boldsymbol{\theta})$ depends on parameter(s) of interest $\boldsymbol{\theta}$

http://www.pp.rhul.ac.uk/~cowan/sda/



The value of $\boldsymbol{\theta}$ is adjusted to minimize

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - \lambda(x_i; \boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

*The parameters that minimize the χ^2 are called the **LS ('least square')** estimators $\hat{\theta}_1, \ldots, \hat{\theta}_m$*

One more thing…

- What if the measurements are **not independent** but described by an N-dimensional Gaussian PDF with known covariance matrix V but unknown mean values?

Start with the N-dimensional generalization of the Gaussian distribution (eqn 2.28)

Take the log (and drop additive terms that do not depend on the parameters)

$$\log \mathscr{L}(\boldsymbol{\theta}) = \ldots$$

Maximize by finding the values of the parameters θ that minimize

$$\chi^2(\boldsymbol{\theta}) = \ldots$$

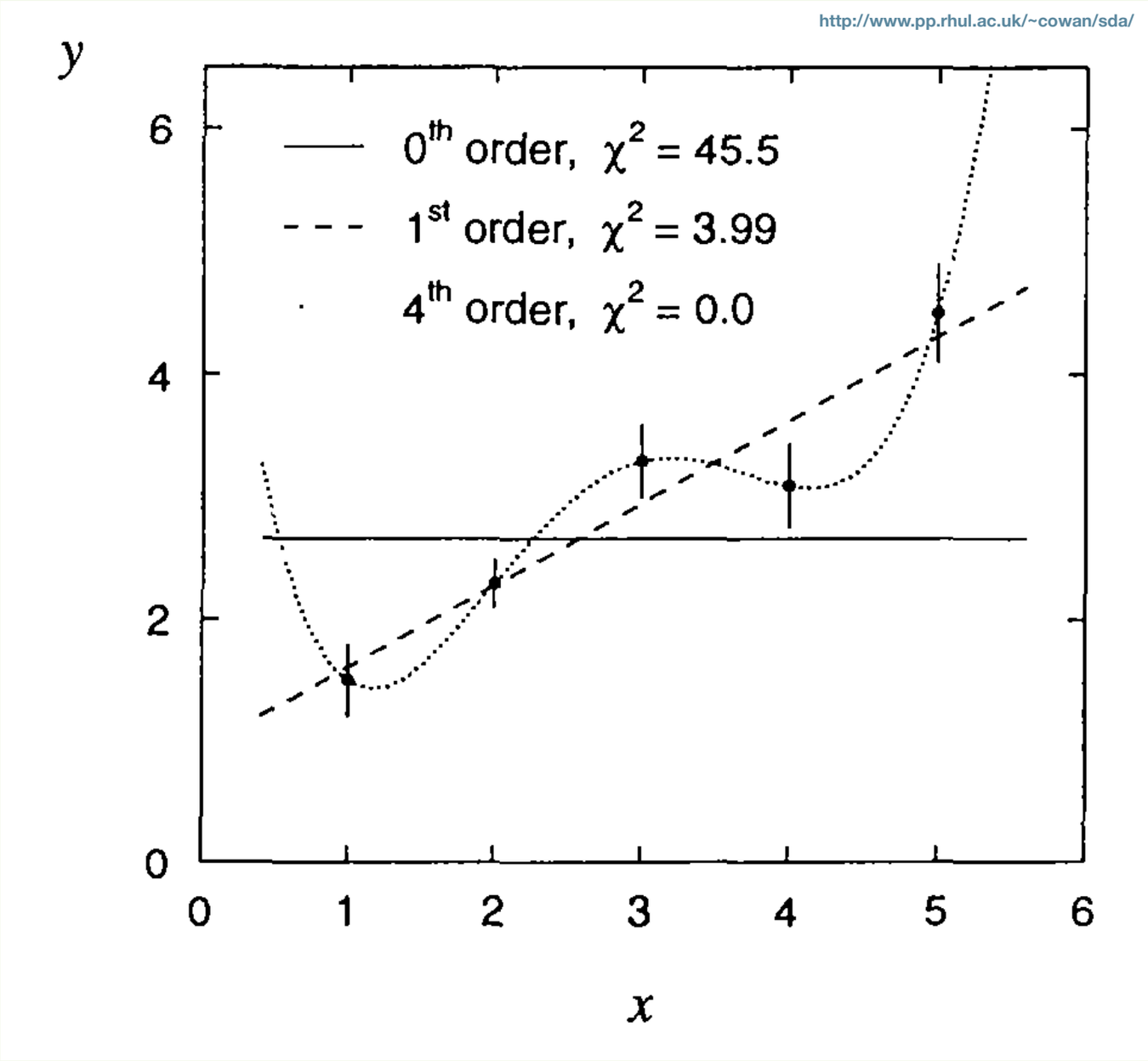
Reduces to 1D case (previous slide) if the covariance matrix is diagonal

See Cowan page 96

LS fit of a polynomial

Example: least square fit of a polynomial (i)

- As an example let us consider the following data:



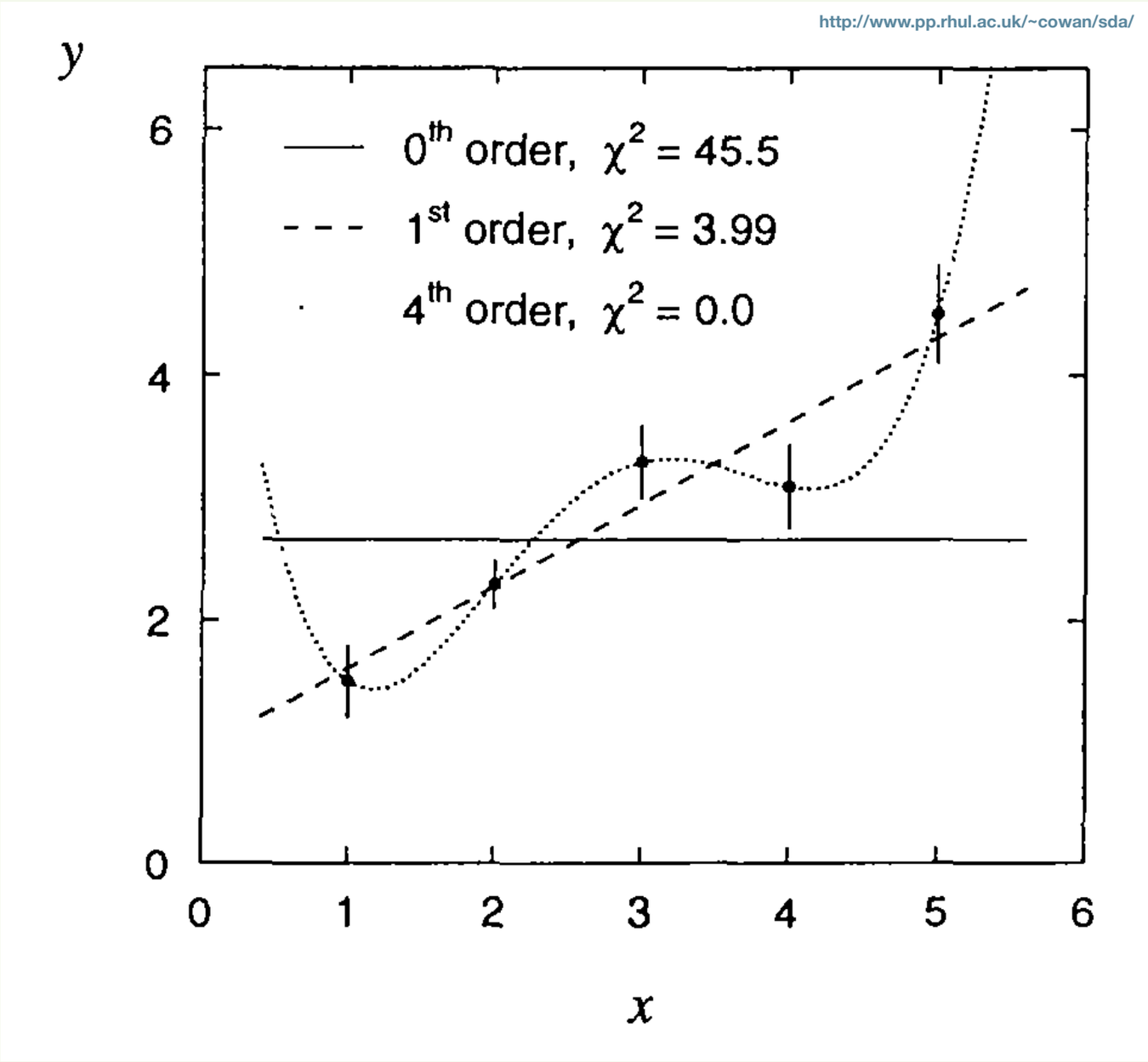
- Five measured values of a quantity y with errors Δy
- Let's assume the measured values y_i each come from a Gaussian distribution centered around (unknown) λ_i with a standard deviation of $\sigma_i = \Delta y_i$
- As hypotheses for $\lambda(x;\theta)$, we try fitting polynomials of order m:

$$\lambda(x;\theta_0, \ldots, \theta_m) = \sum_{j=0}^{m} x^j \theta_j$$

unknown parameters

Example: least square fit of a polynomial (ii)

- As an example let us consider the following data:



- With a 0^{th} order polynomial, we have a large χ^2 value

$$\chi^2 = 45.5 \qquad \hat{\theta}_0 = 2.66 \pm 0.13$$

- 1^{st} order polynomial:

$$\chi^2 = 3.99 \qquad \hat{\theta}_0 = 0.93 \pm 0.30$$

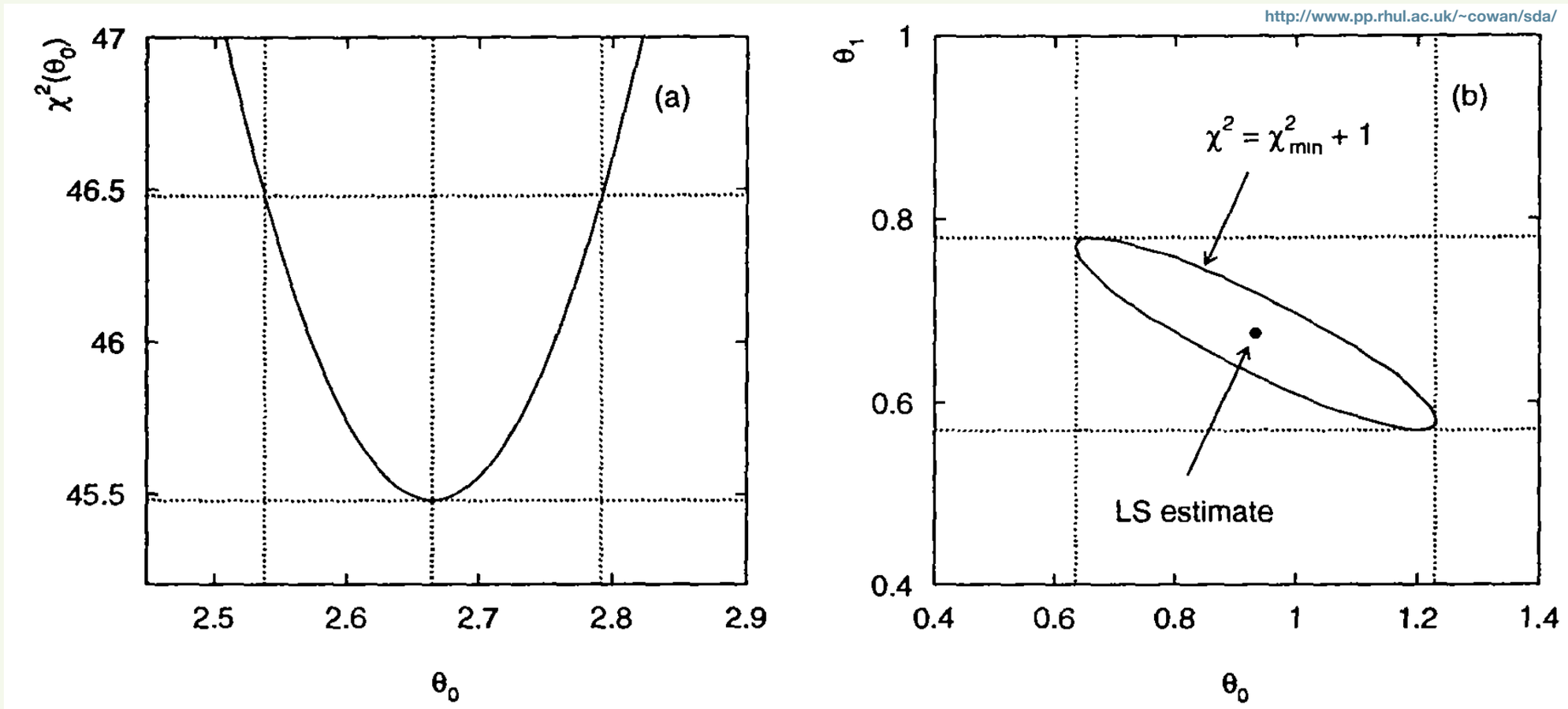
$$\hat{\theta}_1 = 0.68 \pm 0.10$$

- 4^{th} order polynomial fit: χ^2 of zero and goes exactly through all data points

Example: least square fit of a polynomial (iii)

- As for the ML method, the statistical errors and covariances can be estimated using several methods
  - 1) Analytically 2) MC method **3) graphical method**

0^{th} order polynomial fit **1^{st} order polynomial fit**



Example: least square fit of a polynomial (iv)

- As for the ML method, the statistical errors and covariances can be estimated using several methods
  - 1) Analytically 2) MC method **3) graphical method**

$$\hat{\sigma}_{\hat{\theta}_0} = \sqrt{\hat{U}_{00}} = 0.30$$

$$\hat{\sigma}_{\hat{\theta}_1} = \sqrt{\hat{U}_{11}} = 0.10$$

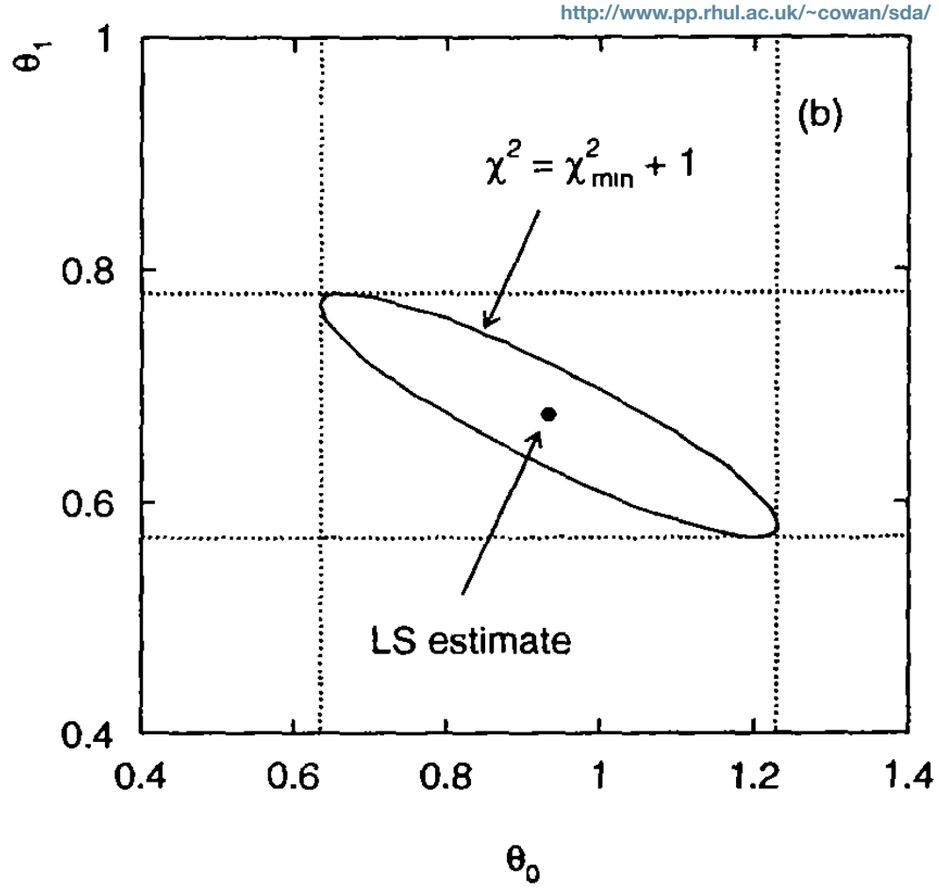
$$\widehat{\mathrm{cov}}[\hat{\theta}_0, \hat{\theta}_1] = \hat{U}_{01} = -0.028,$$

$$r = -0.90.$$

Strongly anti-correlated parameters: Very important to include the correlation coefficient when reporting such a fit.

1st order polynomial fit

http://www.pp.rhul.ac.uk/~cowan/sda/



Example: least square fit of a polynomial (v)

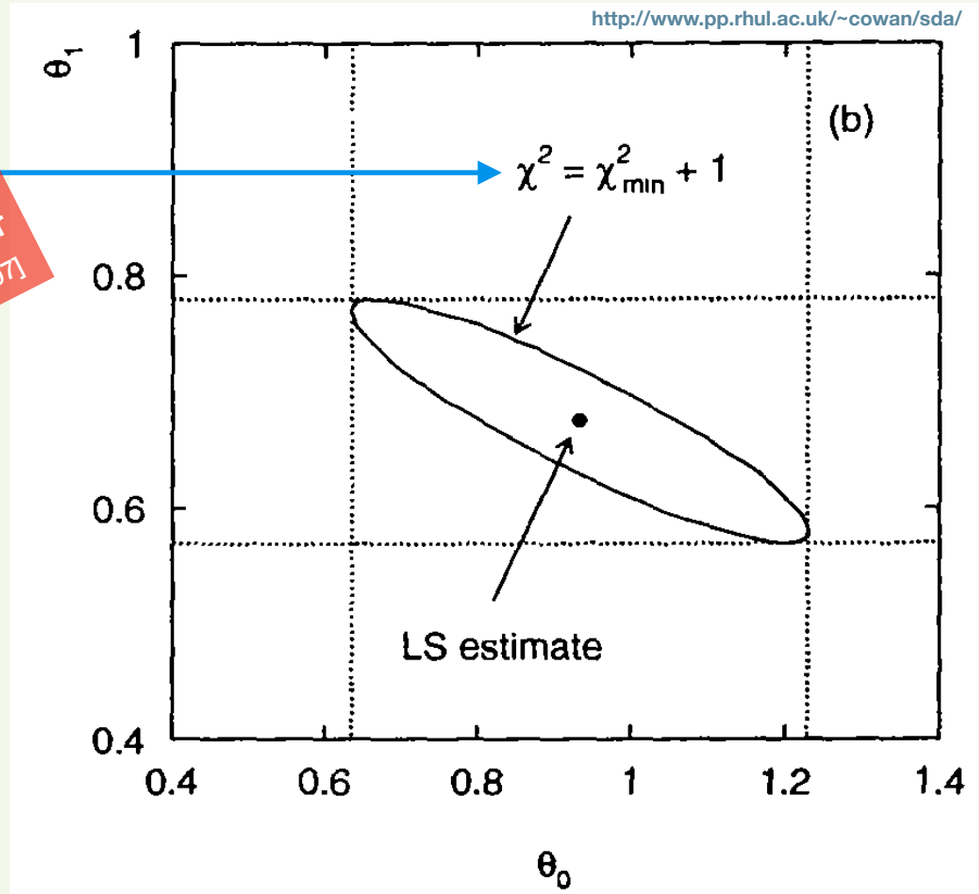
- As for the ML method, the statistical errors and covariances can be estimated using several methods
  - 1) Analytically 2) MC method **3) graphical method**

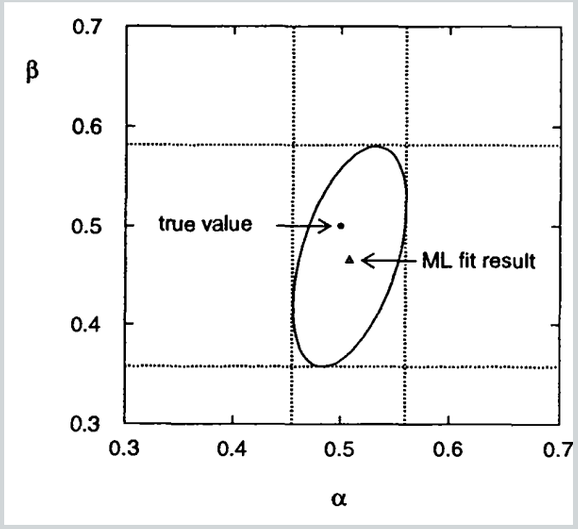
Contours in parameter space whose tangents are at $\hat{\theta}_i \pm \hat{\sigma}_i$, corresponding to a one standard deviation departure from the LS estimates:

$$\chi^2(\boldsymbol{\theta}) = \chi^2(\hat{\boldsymbol{\theta}}) + 1 = \chi^2_{\min} + 1$$

Careful! Only holds for the case of linear LS fits [Cowan, page 97]

1st order polynomial fit





You've seen this before for the case of **ML estimates!**

$$\log \mathscr{L} = \log \mathscr{L}_{\max} - \frac{1}{2}$$

Goodness of fit

Goodness of fit (i)

- If the measured values y_i are **Gaussian**, the resulting least-squared estimators coincide with the ML estimators.
  - Furthermore, the χ^2 values can be used to test how likely it is that if the hypothesis is true, you would measure the observed data
    - **Important but subtle point:** you can only make statements that how likely it is to observe the data given that the hypothesis is true, **BUT you cannot make a statement how probable it is that the hypothesis is true**

- The quantity $\dfrac{\left(y_i - \lambda(x_i;\boldsymbol{\theta})\right)}{\sigma_i}$ is a measure for the deviation between the i^{th} **measurement** y_i and the **hypothesized** λ **function**
  - So χ^2 is a measure of the total agreement between the observed data and hypothesis

Goodness of fit (ii)

- It can be shown that, if
  1. the y_i, $i = 1, \ldots, N$, are independent Gaussian RVs with known variances, σ_i^2 (or are distributed according to an N-dimensional Gaussian with known covariance matrix V);
  2. the hypothesis $\lambda(x; \theta_1, \ldots, \theta_m)$ is linear in the parameters θ_i; and
  3. the functional form of the hypothesis is correct,

- then the minimum value of χ^2 is distributed according to the χ^2 -distribution with $\boldsymbol{n}$ degrees of freedom ($\boldsymbol{n = N - m}$)

# of points ↑ N ; ↑ m # of free parameters in the fit

$$f(z; n) = \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{n/2-1} e^{-z/2}, \quad n = 1, 2, \ldots$$

$$E[z] = \int_0^\infty z \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{n/2-1} e^{-z/2}\, dz = n$$

$$V[z] = \int_0^\infty (z-n)^2 \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{n/2-1} e^{-z/2}\, dz = 2n$$

→ **Expectation value is $n = N - m$**
Thus often quote χ^2/n and speaks of a good fit if $\chi^2/n \sim 1$

Goodness of fit (iii)

- If $\chi^2/n \gg 1$, then there is some reason to doubt the hypothesis.
- If $\chi^2/n \ll 1$, fit is better than expected given the size of the measurement errors (but must check that the σ_i have not been overestimated or are not correlated).
  - One can calculate the probability that the hypothesis would lead to a χ^2 equal or worse (i.e. greater) than the actually one obtained:

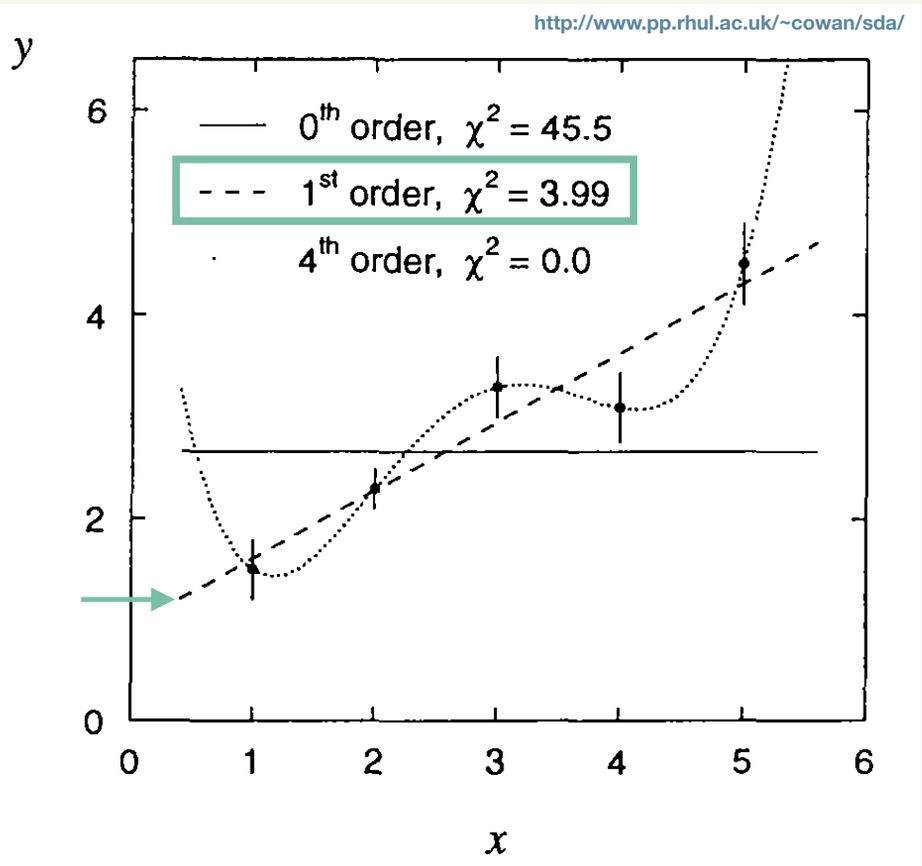
$$P = \int_{\chi^2}^{\infty} f(z; n_d)\, dz$$

- The P-value at which one decides to reject a hypothesis is subjective, but note that underestimated errors can cause a correct hypothesis to give a bad (i.e. large) χ^2

P**-value** also called **observed significance level** or **confidence level** *of the test*

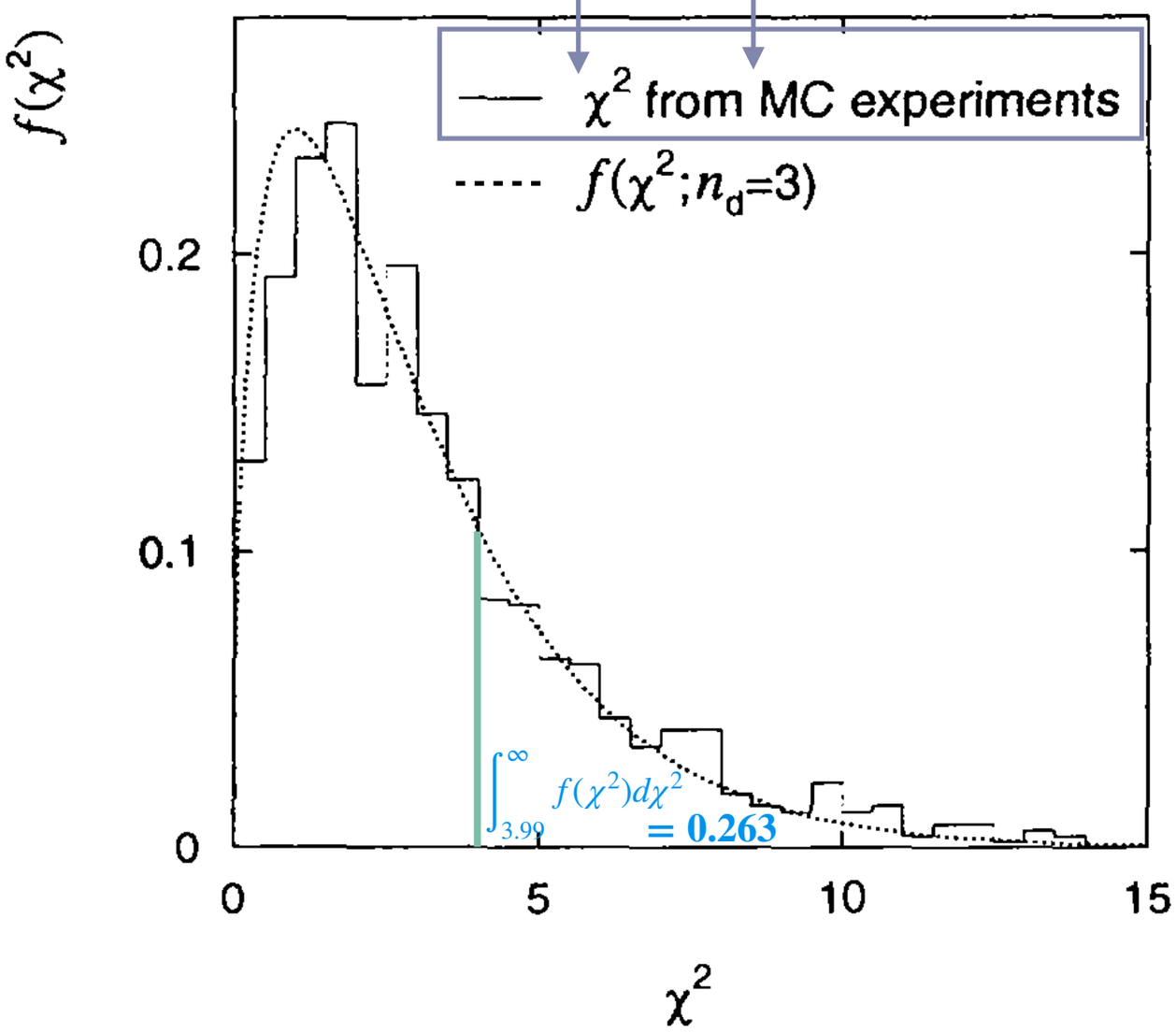
Goodness of fit: *Polynomial fit example* (i)

Revisit



- 1^{st} order Polynomial fit:
  - $\chi^2 = 3.99$
  - This gives $P = 0.263$, i.e. in 26.3% of all cases we expect an observed χ^2 value as large or greater

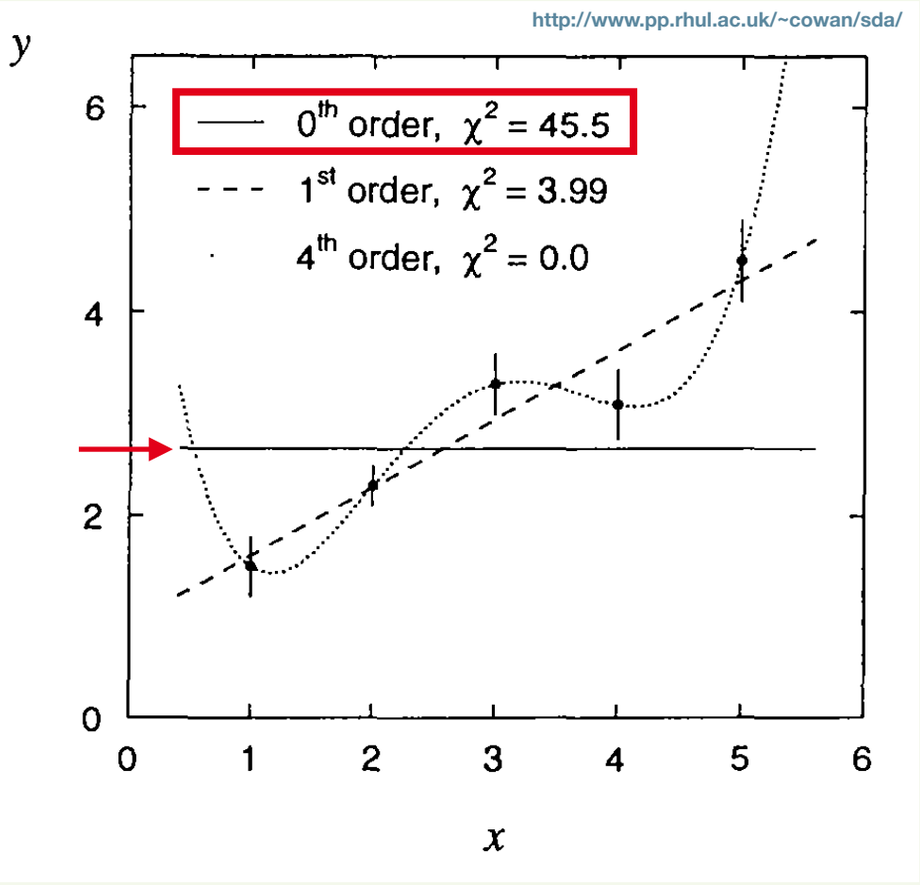
Normalized histogram of 1000 simulated



Can be checked with MC: 'true' parameters (θ_0, θ_1) are taken from the real experiment, and a 'measured' value for each data point is generated from a Gaussian of width σ given by the corresponding errors

Goodness of fit: *Polynomial fit example* (ii)

Revisit



- 0^{th} order Polynomial fit:
  - $\chi^2 = 45.5$
  - The corresponding significant level is $P = 3.1 \cdot 10^{-9}$, i.e. very small probability to observe such a data set if underlying hypothesis is true

Interpret: *If this horizontal-line hypothesis were true, one would expect a χ^2 as high or higher than the one obtained in only 3 out of a billion experiments*

LS with Binned Data

Least squares with binned data (i)

- As for the ML method, one can also carry out LS fits with binned data
  - So far the function relating "true" values λ to the variable x was not necessarily a PDF. This however can be remedied easily by making it proportional to one:

$$\lambda_i(\boldsymbol{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x;\boldsymbol{\theta})dx = np_i(\boldsymbol{\theta})$$

total number of events

PDF

probability to end up in i^{th} bin

- The parameters $\boldsymbol{\theta}$ are found by minimizing the quantity

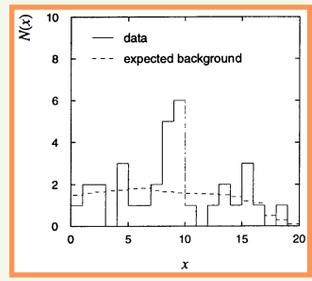
$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - \lambda_i(\boldsymbol{\theta})\right)^2}{\sigma_i^2}$$

Least squares with binned data (ii)

- If the mean number of entries in each bin is small compared to the total number of entries, the contents of each bin are approximatively Poisson distributed.
  - The variance thus becomes equal to the mean so that we recover

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - \lambda_i(\boldsymbol{\theta})\right)^2}{\lambda_i(\boldsymbol{\theta})} = \sum_{i=1}^{N} \frac{\left(y_i - np_i(\boldsymbol{\theta})\right)^2}{np_i(\boldsymbol{\theta})}$$

This takes us back to the original Pearson χ^2 we introduced at the beginning!



- An alternative method is to approximate the variance as the number of entries in ***bin i*** by the number of entries actually observed in y_i. This is called the **modified least-squares method:**

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\left(y_i - \lambda_i(\boldsymbol{\theta})\right)^2}{y_i} = \sum_{i=1}^{N} \frac{\left(y_i - np_i(\boldsymbol{\theta})\right)^2}{y_i}$$

Easier to handle, but errors maybe poorly estimated if any of the bins contain few events.

Bins with no events create a division by zero!

Quiz Time: 5th Round

Combining measurements with least squares

Based on Cowan Sec. 7.6

1. The χ^2 method offers an easy way to combine measurements. For uncorrelated measured quantities y_i with variances σ_i one can determine the LS estimator $\hat{\lambda}$ to be

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} y_i/\sigma_i^2}{\sum_{j=1}^{N} 1/\sigma_j^2} \tag{1}$$

or for correlated measurements with covariance V one has

$$\hat{\lambda} = \sum_{i=1}^{N} w_i y_i \qquad \text{with} \quad w_i = \frac{\sum_{j=1}^{N} (V^{-1})_{ij}}{\sum_{k,l=1}^{N} (V^{-1})_{kl}} \,. \tag{2}$$

Note that $\sum_{i=1}^{N} w_i = 1$ and the variance of $\hat{\lambda}$ is given by $V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j$.

a) Calculate the average of two measured quantities $y_1 = 5$ and $y_2 = 6$ with a covariance matrix

$$C = \begin{pmatrix} 0.5. & 0.2 \\ 0.2 & 0.7 \end{pmatrix} ,$$

first using the uncorrelated formula Eq. 1 and then using the proper expression Eq. 2. Remember that $\sigma_i^2 = V_{ii}$.

b) Calculate the correlated average of the same measured quantities but assume now that the covariance matrix is given by

$$C = \begin{pmatrix} 0.5. & 0.55 \\ 0.55 & 0.7 \end{pmatrix} .$$

Why is the average (you should get $\hat{\lambda} = 4.5$) not between 5 and 6? What is the variance of the average?

P-values

2. In the lecture we discussed that the obtained χ^2 value from a binned LS fit contains information of how probable an observation is given a presumed hypothesis. This probability is called the p-value and is given by

$$p = \int_{\chi^2_{\text{obs}}}^{\infty} f_{\chi^2}(x; \text{n.d.f.})\, \mathrm{d}x\,, \tag{3}$$

where χ^2_{obs} is the observed χ^2 value, f_{χ^2} is the χ^2-distribution, and n.d.f. denotes the numbers of degrees of freedom. If a LS fit has n bins and m free parameters, it is n.d.f. $= n - m$. You can calculate this integral easily using `ROOT` via the `TMath::Prob(x,ndf)` function.

a) What is the interpretation of a p-value and what does a low p-value imply?

b) Using the χ^2_{obs} table below decide if the following binned LS or χ^2 fits describe the observed data well:

$$(\chi^2_{\text{obs}}, n, m) = (6.2, 5, 3), \qquad (\chi^2_{\text{obs}}, n, m) = (1.2, 2, 1), \qquad (\chi^2_{\text{obs}}, n, m) = (2.2, 10, 3).$$

Degrees of freedom (df)	χ^2 value[20]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

For next time

- Required reading
  - Cowan textbook: chapters 4, 6.9-6.13, 7

- Suggested reading: /Reading material/ L05 /
  - Very nice Cambridge lecture series in 4 PDFs.

Next time

- Hypothesis testing
- Neyman-Pearson Lemma

Bibliography

- Part of the material presented in this lecture is taken from the following sources. See the active links (when available) for a complete reference
  - **Statistical Data Analysis** textbook by G. Cowan (U. London): all figures & equations with white background