Powerweek Data Analysis

Helmholtz Research School
for Quark Matter Studies
in Heavy Ion Collisions

Klaus Reygers
Physikalisches Institut
Universität Heidelberg

Contents

- Part I: Invariant mass analyses (example: π^0 analysis)
  - Kinematics
  - Acceptance and efficiency
  - Bin-shift correction
  - Effects of energy scale uncertainties
- Part II: Statistics
  - Error propagation
  - Maximum likelihood method
  - Least squares method

Part II: Statistics

Statistics books:

G. Cowan, Statistical Data Analysis, Clarendon, Oxford, 1998
see also www.pp.rhul.ac.uk/~cowan/sda

R.J. Barlow, Statistics, A Guide to the Use of Statistical
in the Physical Sciences, Wiley, 1989
see also hepwww.ph.man.ac.uk/~roger/book.html

Slides taken from:

Glen Cowan,
Introduction to Statistical Methods for High Energy Physics,
2008 CERN Summer Student Lectures

1. Error Propagation

Expectation values

Consider continuous r.v. x with pdf $f(x)$.

Define expectation (mean) value as $E[x] = \int x\, f(x)\, dx$

Notation (often): $E[x] = \mu$ ~ "centre of gravity" of pdf.

For a function $y(x)$ with pdf $g(y)$,

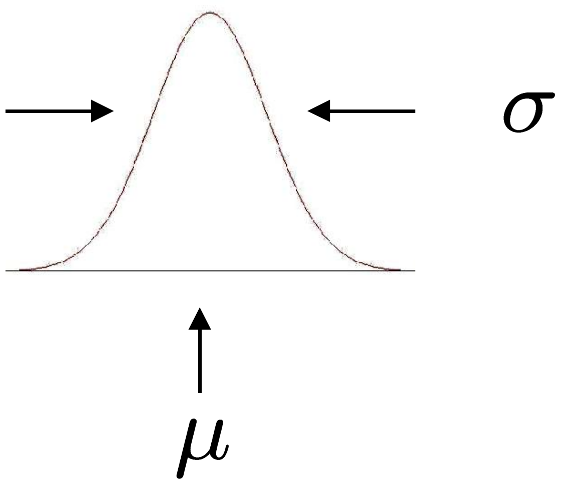
$$E[y] = \int y\, g(y)\, dy = \int y(x) f(x)\, dx$$ (equivalent)

Variance: $V[x] = E[x^2] - \mu^2 = E[(x-\mu)^2]$

Notation: $V[x] = \sigma^2$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

σ ~ width of pdf, same units as x.



Covariance and correlation

Define covariance cov[x,y] (also use matrix notation V_{xy}) as

$$\mathrm{cov}[x,y] = E[xy] - \mu_x \mu_y = E[(x-\mu_x)(y-\mu_y)]$$

Correlation coefficient (dimensionless) defined as

$$\rho_{xy} = \frac{\mathrm{cov}[x,y]}{\sigma_x \sigma_y}$$

If x, y, independent, i.e., $f(x,y) = f_x(x) f_y(y)$, then

$$E[xy] = \int\int xy\, f(x,y)\, dx dy = \mu_x \mu_y$$

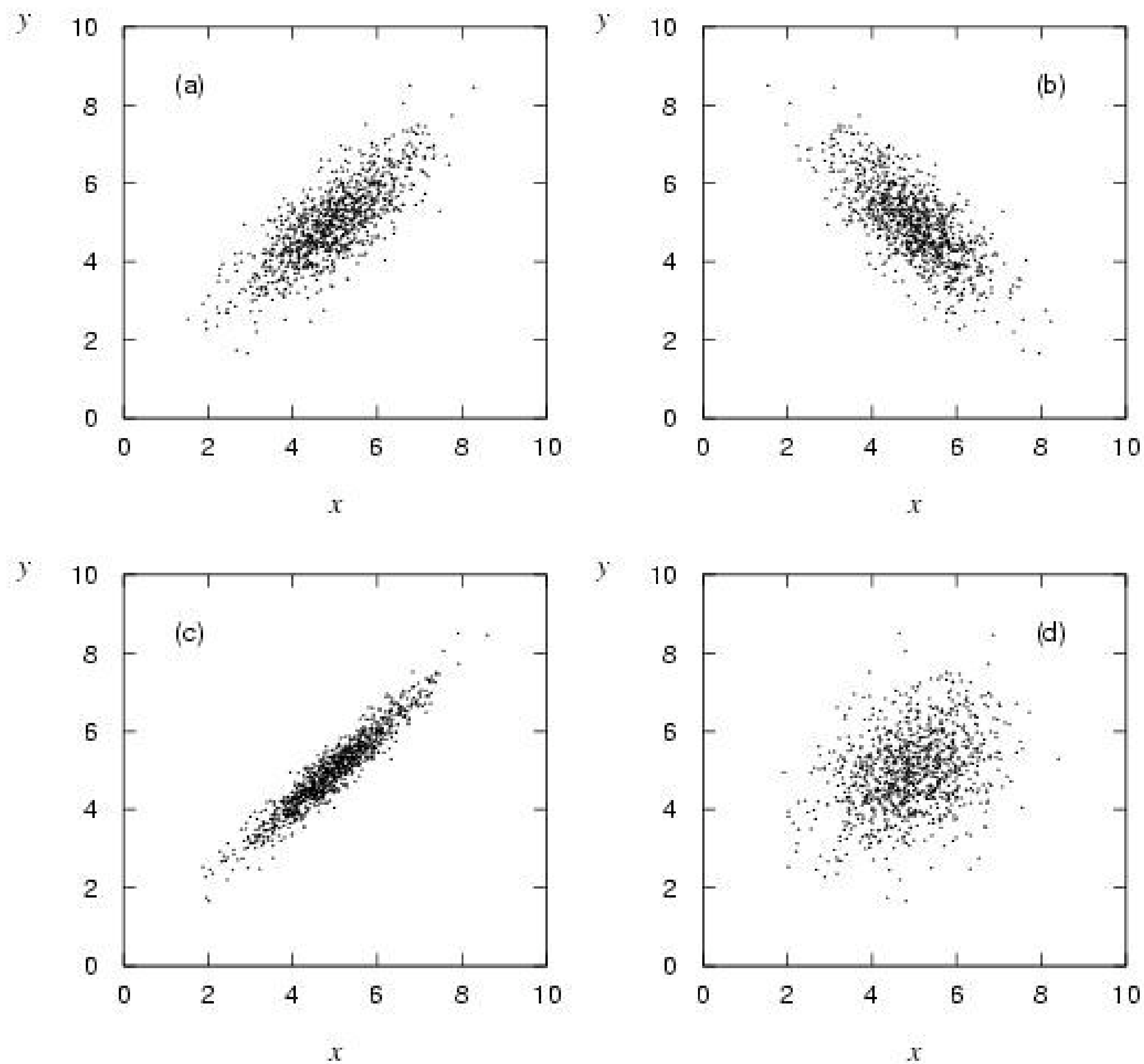
$\rightarrow$ $\mathrm{cov}[x,y] = 0$ x and y, ‘uncorrelated’

N.B. converse not always true.

Correlation (cont.)

$\rho = 0.75$

$\rho = -0.75$



$\rho = 0.95$

$\rho = 0.25$

Error propagation

Suppose we measure a set of values $\vec{x} = (x_1, \ldots, x_n)$

and we have the covariances $V_{ij} = \text{cov}[x_i, x_j]$

which quantify the measurement errors in the x_i.

Now consider a function $y(\vec{x})$.

What is the variance of $y(\vec{x})$?

The hard way: use joint pdf $f(\vec{x})$ to find the pdf $g(y)$,

then from $g(y)$ find $V[y] = E[y^2] - (E[y])^2$.

Often not practical, $f(\vec{x})$ may not even be fully known.

Error propagation (2)

Suppose we had $\vec{\mu} = E[\vec{x}]$

in practice only estimates given by the measured $\vec{x}$

Expand $y(\vec{x})$ to 1st order in a Taylor series about $\vec{\mu}$

$$y(\vec{x}) \approx y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$

To find $V[y]$ we need $E[y^2]$ and $E[y]$.

$$E[y(\vec{x})] \approx y(\vec{\mu}) \quad \text{since} \quad E[x_i - \mu_i] = 0$$

Error propagation (3)

$$E[y^2(\vec{x})] \approx y^2(\vec{\mu}) \; + \; 2y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i]$$

$$+E\left[\left(\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) \right) \left(\sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j) \right) \right]$$

$$= y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Putting the ingredients together gives the variance of $y(\vec{x})$

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Error propagation (4)

If the x_i are uncorrelated, i.e., $V_{ij} = \sigma_i^2 \delta_{ij}$, then this becomes

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]^2_{\vec{x}=\vec{\mu}} V_{ij}$$

Similar for a set of m functions $\vec{y}(\vec{x}) = (y_1(\vec{x}), \ldots, y_m(\vec{x}))$

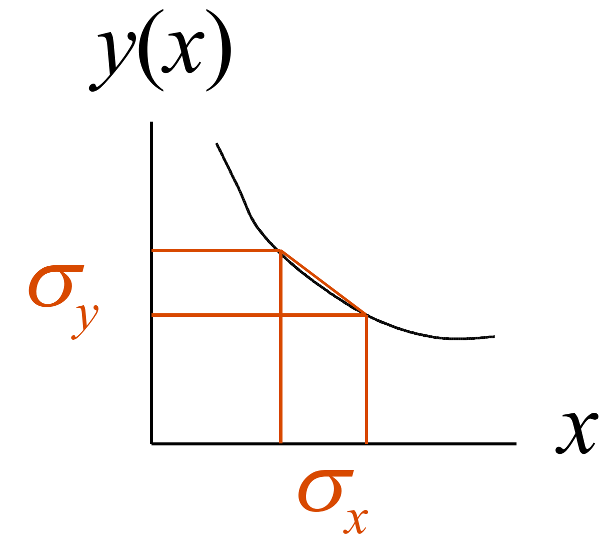
$$U_{kl} = \mathrm{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

or in matrix notation $U = AVA^T$, where

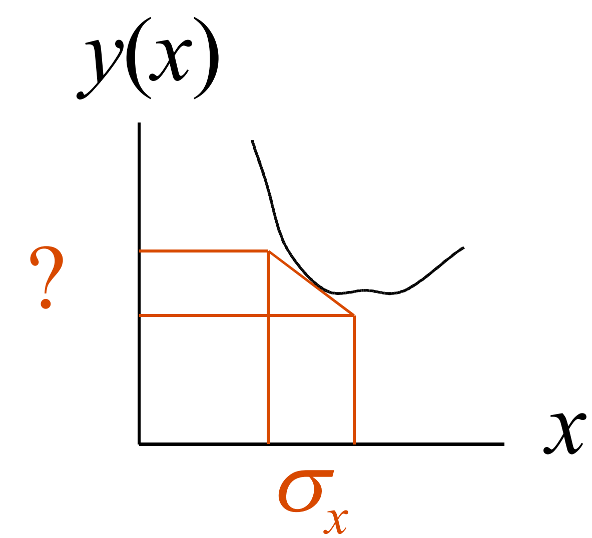
$$A_{ij} = \left[\frac{\partial y_i}{\partial x_j} \right]_{\vec{x}=\vec{\mu}}$$

Error propagation (5)

The 'error propagation' formulae tell us the covariances of a set of functions $\vec{y}(\vec{x}) = (y_1(\vec{x}), \ldots, y_m(\vec{x}))$ in terms of the covariances of the original variables.



Limitations: exact only if $\vec{y}(\vec{x})$ linear. Approximation breaks down if function nonlinear over a region comparable in size to the σ_i.



N.B. We have said nothing about the exact pdf of the x_i, e.g., it doesn't have to be Gaussian.

Error propagation − special cases

$$y = x_1 + x_2 \quad \rightarrow \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\mathrm{cov}[x_1, x_2]$$

$$y = x_1 x_2 \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2}$$

That is, if the x_i are uncorrelated:

add errors quadratically for the sum (or difference),

add relative errors quadratically for product (or ratio).

But correlations can change this completely...

Error propagation − special cases (2)

Consider $y = x_1 - x_2$ with

$$\mu_1 = \mu_2 = 10, \quad \sigma_1 = \sigma_2 = 1, \quad \rho = \frac{\mathrm{cov}[x_1, x_2]}{\sigma_1 \sigma_2} = 0 \, .$$

$$V[y] = 1^2 + 1^2 = 2, \rightarrow \sigma_y = 1.4$$

Now suppose $\rho = 1$. Then

$$V[y] = 1^2 + 1^2 - 2 = 0, \rightarrow \sigma_y = 0$$

i.e. for 100% correlation, error in difference → 0.

2. Parameter Estimation

Parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v. parameter

Suppose we have a sample of observed values: $\vec{x} = (x_1, \dots, x_n)$

We want to find some function of the data to estimate the parameter(s):

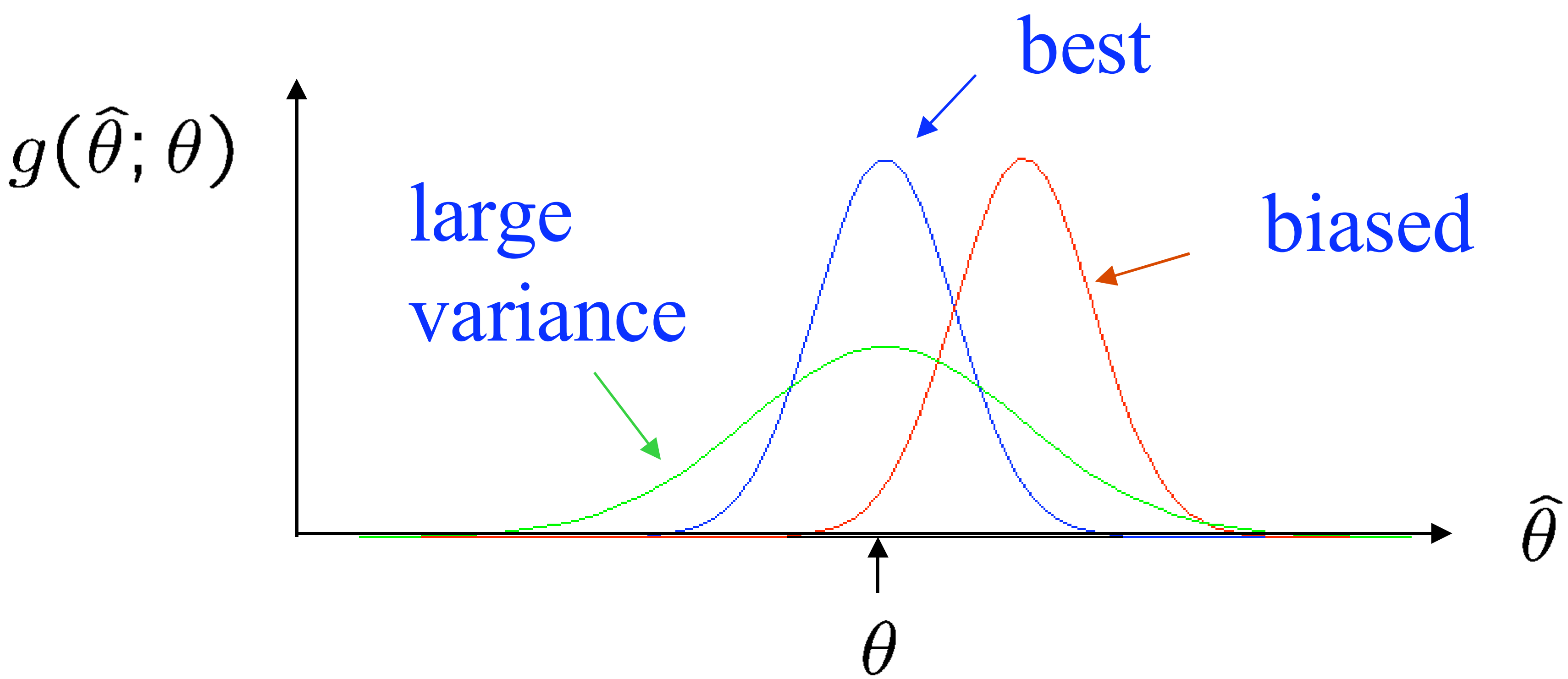
$$\widehat{\theta}(\vec{x})$$

← estimator written with a hat

Sometimes we say 'estimator' for the function of $x_1, \dots, x_n$;
'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

An estimator for the mean (expectation value)

Parameter: $\mu = E[x]$

Estimator: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \equiv \overline{x}$ ('sample mean')

We find: $b = E[\hat{\mu}] - \mu = 0$

$$V[\hat{\mu}] = \frac{\sigma^2}{n} \qquad \left(\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$

An estimator for the variance

Parameter: $\sigma^2 = V[x]$

Estimator: $\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \equiv s^2$ ('sample variance')

We find:

$$b = E[\widehat{\sigma^2}] - \sigma^2 = 0$$ (factor of n–1 makes this so)

$$V[\widehat{\sigma^2}] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2\right) \, ,$$ where

$$\mu_k = \int (x-\mu)^k f(x)\, dx$$

The likelihood function

Suppose the outcome of an experiment is: $x_1, ..., x_n$, which is modeled as a sample from a joint pdf with parameter(s) θ:

$$f(x_1, \ldots, x_n; \theta)$$

Now evaluate this with the data sample obtained and regard it as a function of the parameter(s). This is the likelihood function:

$$L(\theta) = f(x_1, \ldots, x_n; \theta)$$

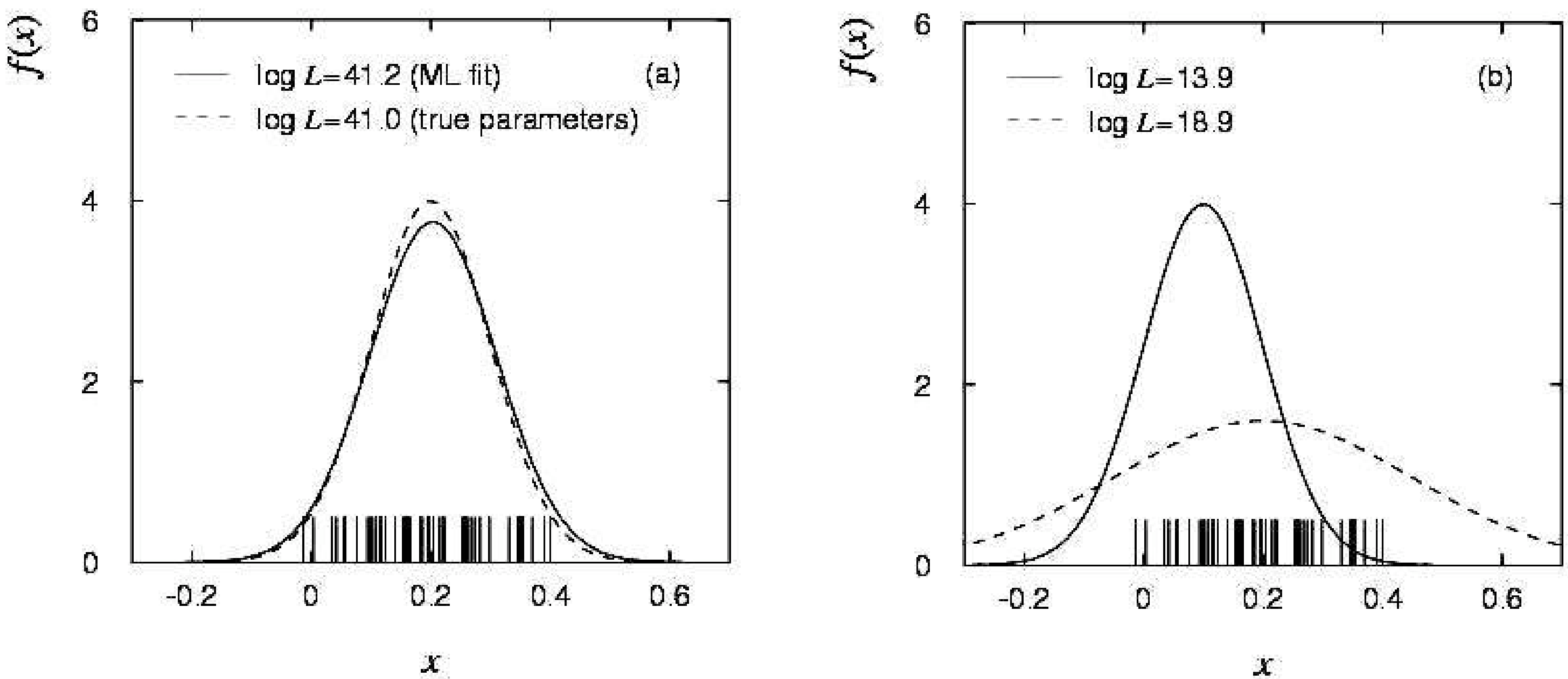
(x_i constant)

If the x_i are independent observations of $x \sim f(x;\theta)$, then,

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

Maximum likelihood estimators

If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

ML estimators not guaranteed to have any 'optimal' properties, (but in practice they're very good).

ML example: parameter of exponential pdf

Consider exponential pdf,
$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

and suppose we have data,
$$t_1, \ldots, t_n$$

The likelihood function is
$$L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

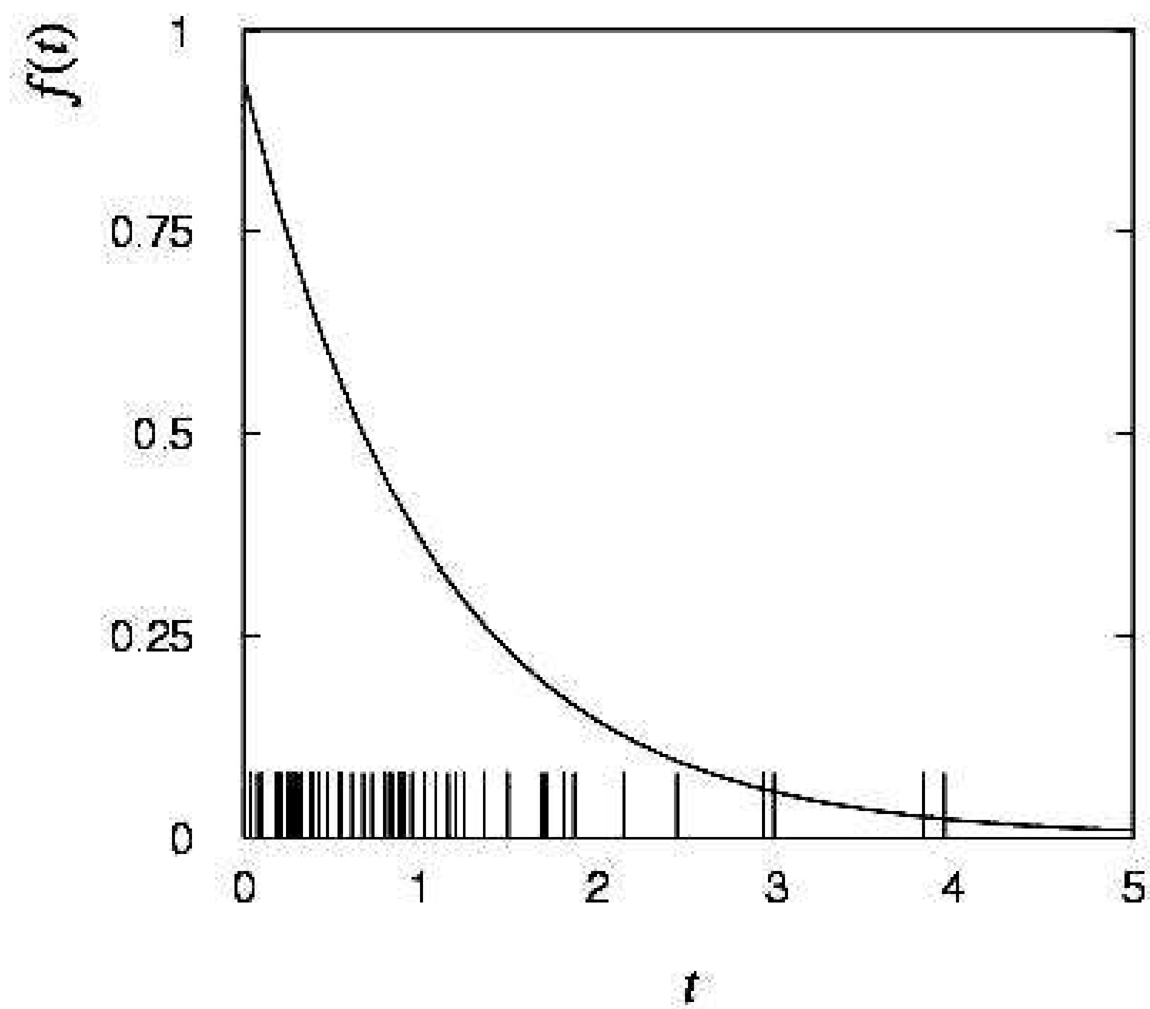
ML example: parameter of exponential pdf (2)

Find its maximum by setting $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Monte Carlo test:

generate 50 values using $\tau = 1$:



We find the ML estimate:

$$\hat{\tau} = 1.062$$

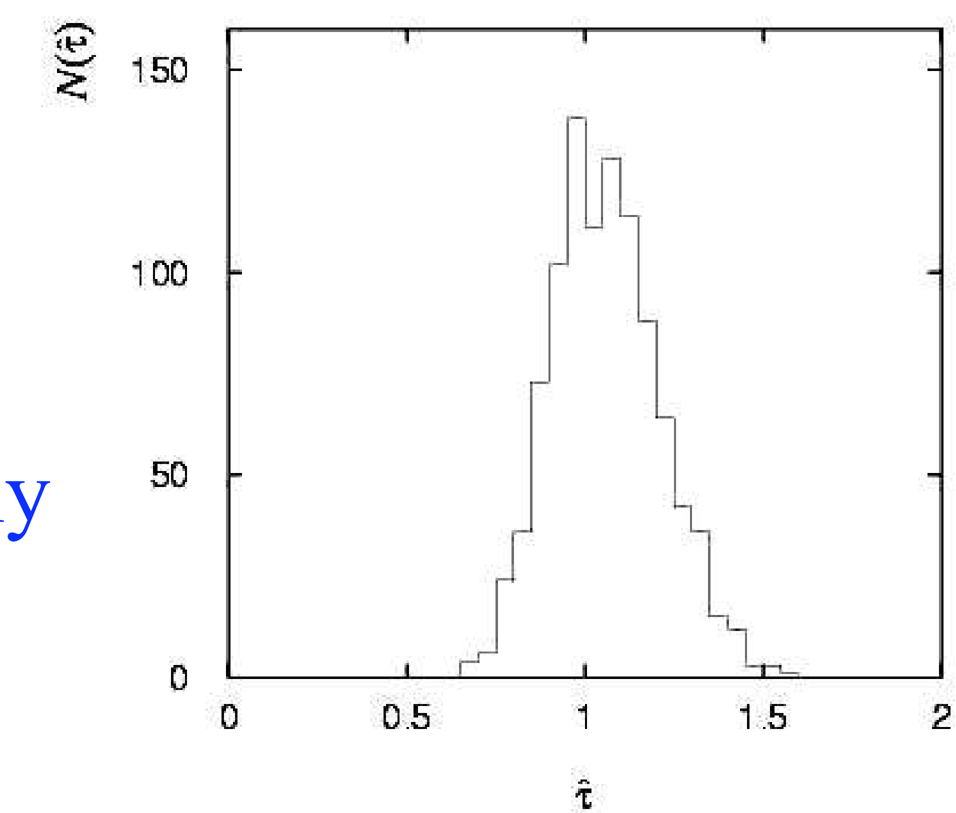
Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$



Note distribution of estimates is roughly Gaussian − (almost) always true for ML in large sample limit.

Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

$$V[\widehat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] \qquad (b = E[\widehat{\theta}] - \theta)$$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\widehat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

$$\widehat{V}[\widehat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\widehat{\theta}}$$

Example of variance by graphical method

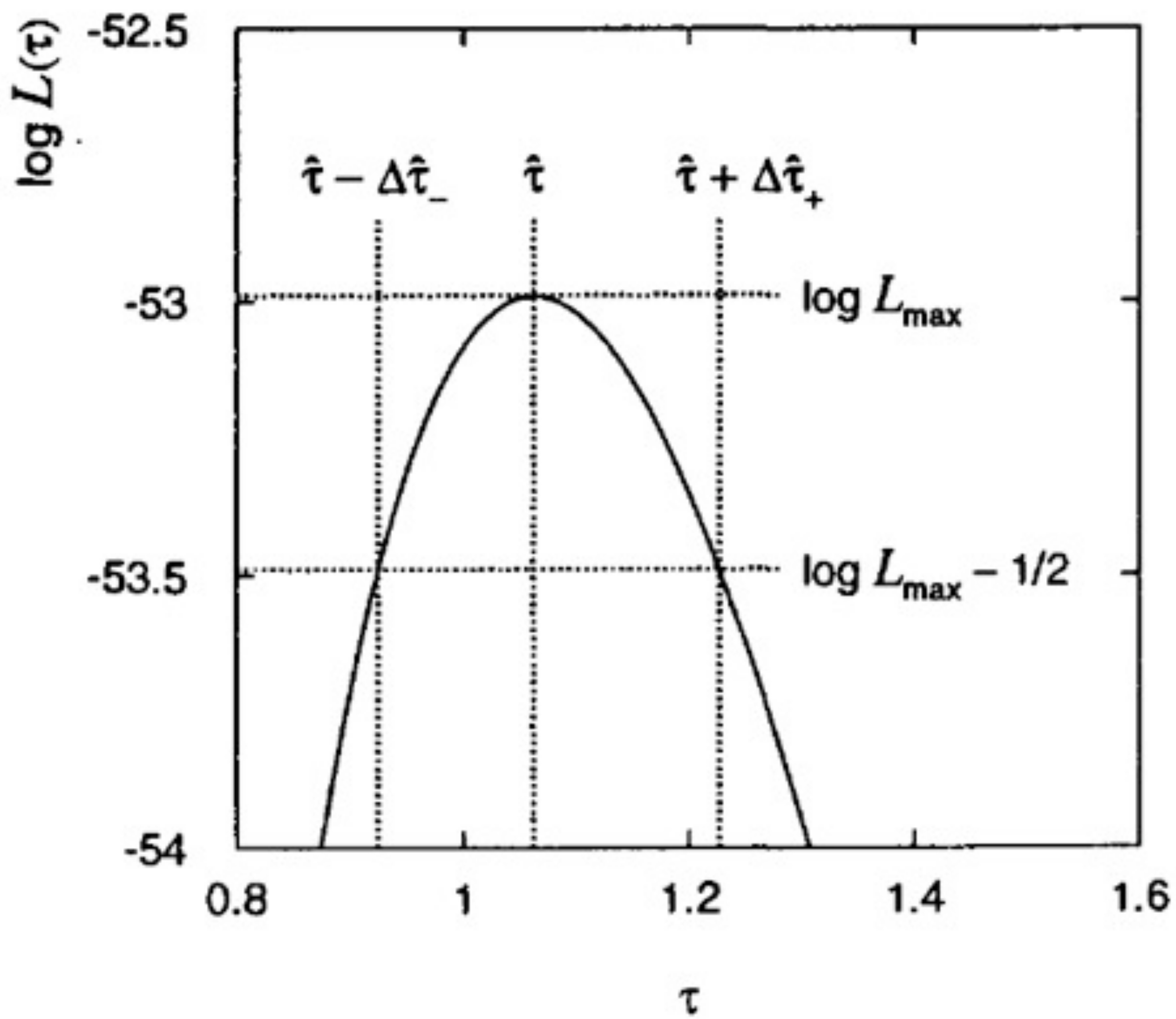
ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$

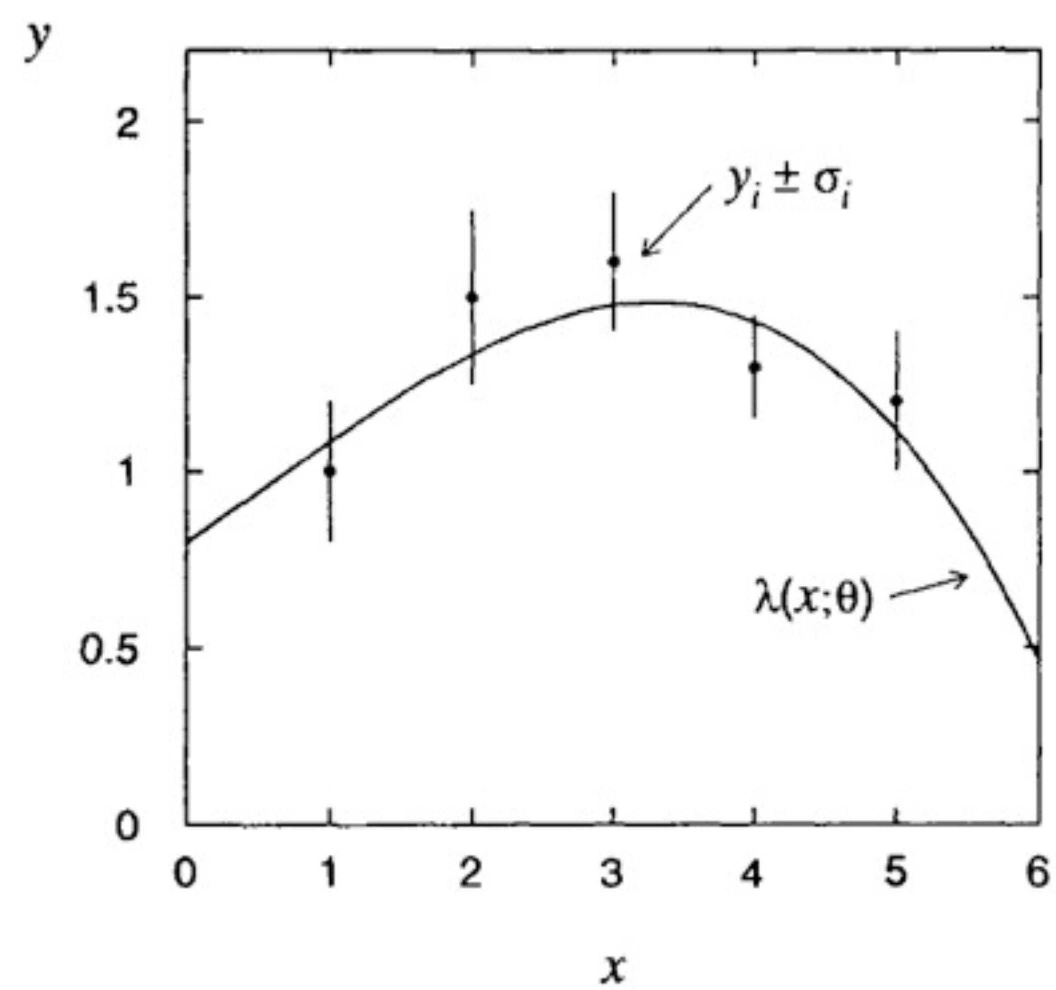


Not quite parabolic ln L since finite sample size (n = 50).

The method of least squares

Suppose we measure N values, $y_1, \ldots, y_N$, assumed to be independent Gaussian r.v.s with

$$E[y_i] = \lambda(x_i;\theta) \; .$$



Assume known values of the control variable $x_1, \ldots, x_N$ and known variances

$$V[y_i] = \sigma_i^2 \; .$$

We want to estimate θ, i.e., fit the curve to the data points.

The likelihood function is

$$L(\theta) = \prod_{i=1}^{N} f(y_i;\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - \lambda(x_i;\theta))^2}{2\sigma_i^2} \right]$$

The method of least squares (2)

The log-likelihood function is therefore

$$\ln L(\theta) = -\frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2} + \text{terms not depending on } \theta$$

So maximizing the likelihood is equivalent to minimizing

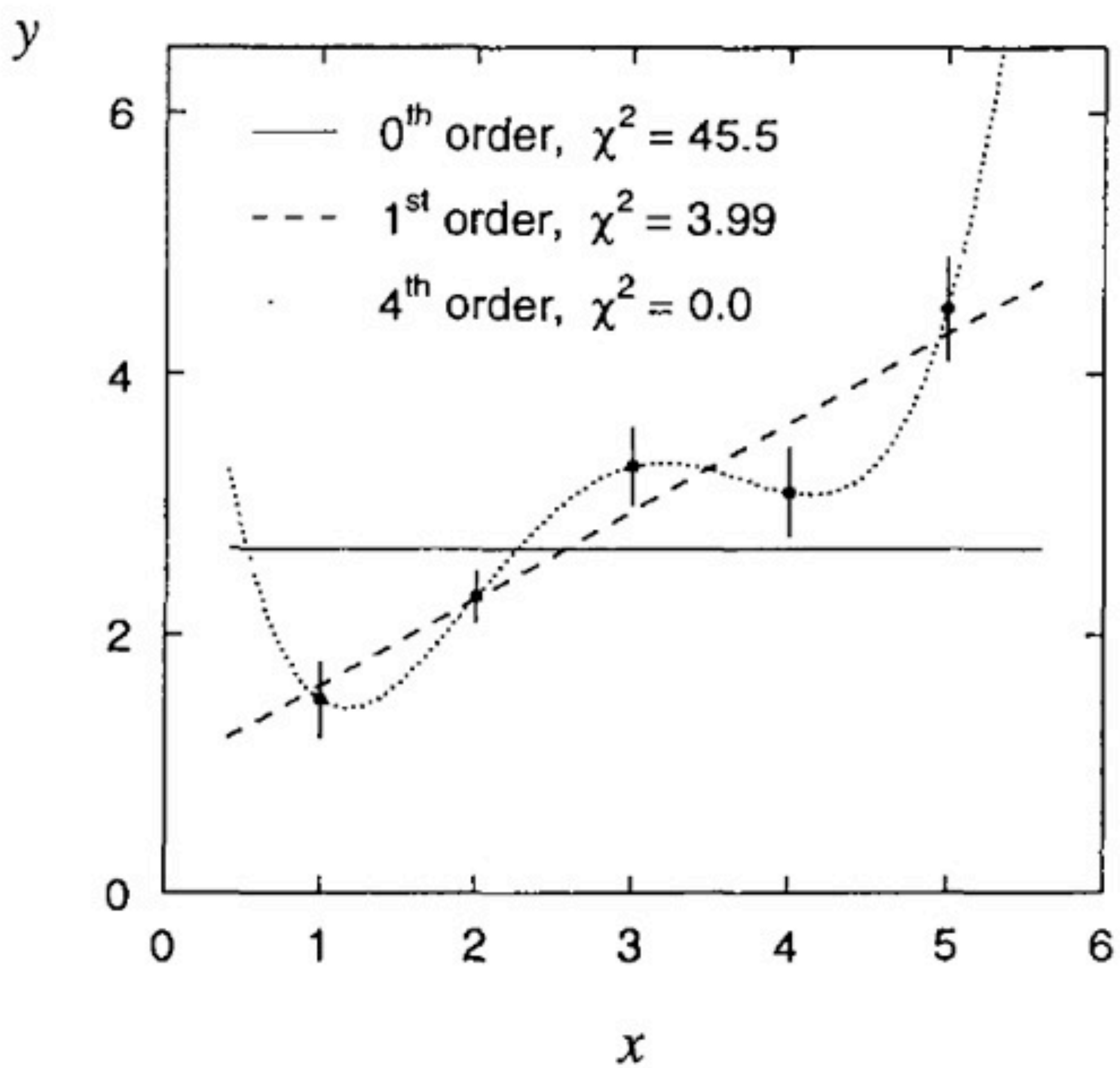
$$\chi^2(\theta) = \sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

Minimum of this quantity defines the least squares estimator $\hat{\theta}$.

Often minimize χ^2 numerically (e.g. program `MINUIT`).

Example of least squares fit

Fit a polynomial of order p: $\lambda(x;\theta_0,\ldots,\theta_p)=\sum_{n=0}^{p}\theta_n x^n$



Variance of LS estimators

In most cases of interest we obtain the variance in a manner similar to ML. E.g. for data ~ Gaussian we have

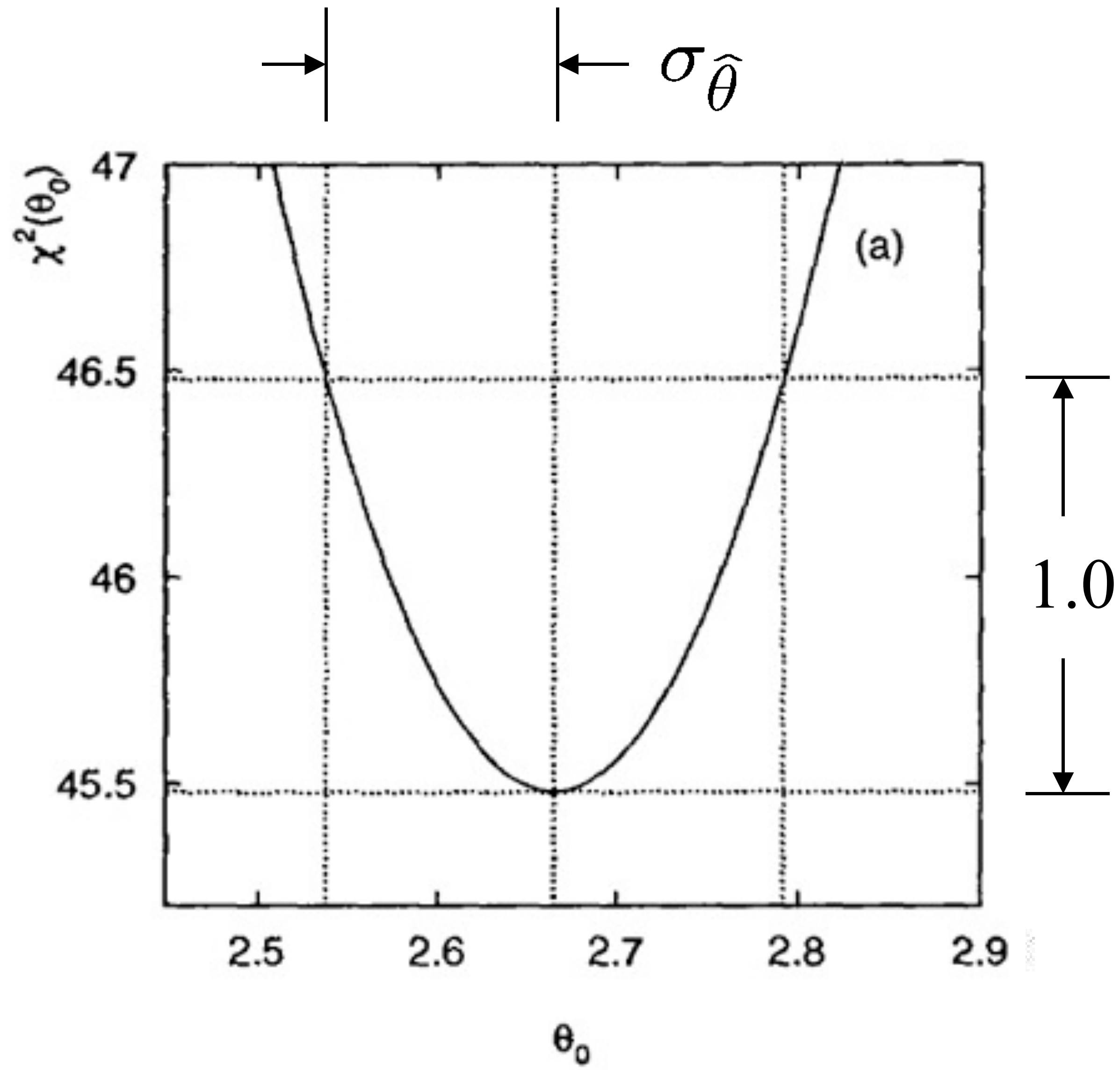
$$\chi^2(\theta) = -2 \ln L(\theta)$$

and so

$$\widehat{\sigma^2}_{\hat{\theta}} \approx 2 \left[\frac{\partial^2 \chi^2}{\partial \theta^2} \right]^{-1}_{\theta=\hat{\theta}}$$

or for the graphical method we take the values of θ where

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



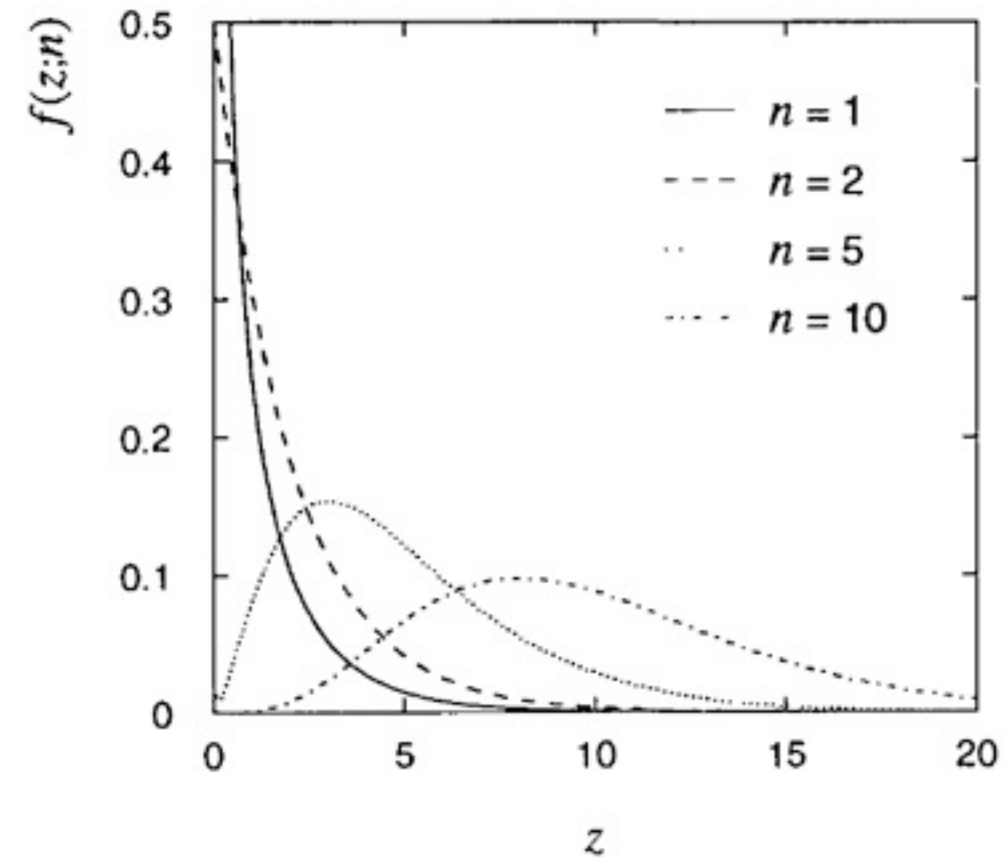
Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z ($z \geq 0$) is defined by

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

$n = 1, 2, ... =$ number of 'degrees of freedom' (dof)

$$E[z] = n \,, \qquad V[z] = 2n \,.$$



For independent Gaussian x_i, $i = 1, ..., n$, means μ_i, variances σ_i^2,

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$ follows χ^2 pdf with n dof.

Example: goodness-of-fit test variable especially in conjunction with method of least squares.

Goodness-of-fit with least squares

The value of the χ^2 at its minimum is a measure of the level of agreement between the data and fitted curve:

$$\chi^2_{\min} = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \hat{\theta}))^2}{\sigma_i^2}$$

It can therefore be employed as a goodness-of-fit statistic to test the hypothesized functional form $\lambda(x; \theta)$.

We can show that if the hypothesis is correct, then the statistic $t = \chi^2_{\min}$ follows the chi-square pdf,

$$f(t; n_{\mathrm{d}}) = \frac{1}{2^{n_{\mathrm{d}}/2}\Gamma(n_{\mathrm{d}}/2)} t^{n_{\mathrm{d}}/2-1} e^{-t/2}$$

where the number of degrees of freedom is

n_{d} = number of data points – number of fitted parameters

Goodness-of-fit with least squares (2)

The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if $\chi^2_{\min} \approx n_d$ the fit is 'good'.

More generally, find the p-value: $$p = \int_{\chi^2_{\min}}^{\infty} f(t; n_{\mathrm{d}})\, dt$$

This is the probability of obtaining a $\chi^2_{\min}$ as high as the one we got, or higher, if the hypothesis is correct.

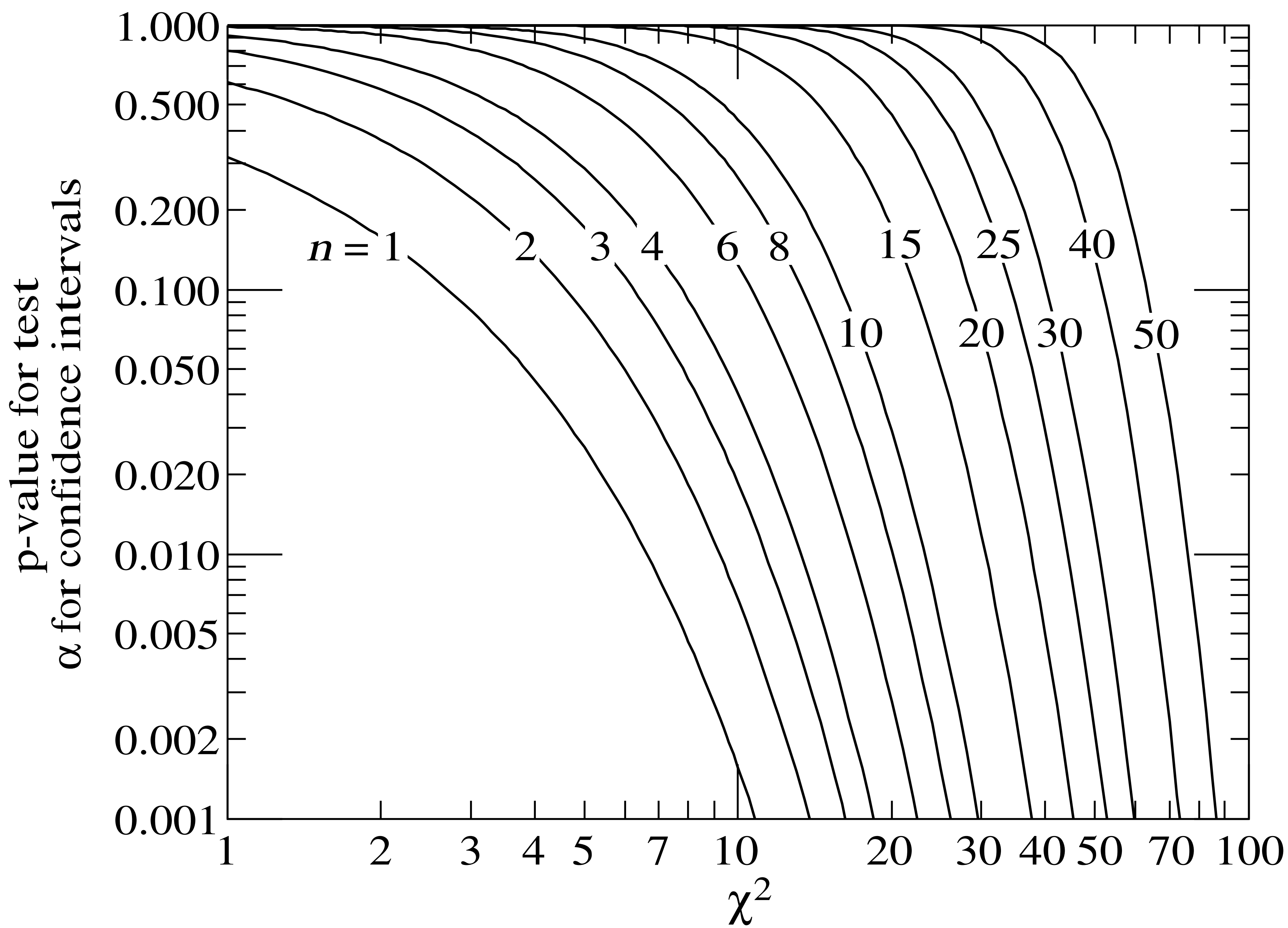
E.g. for the previous example with 1st order polynomial (line),

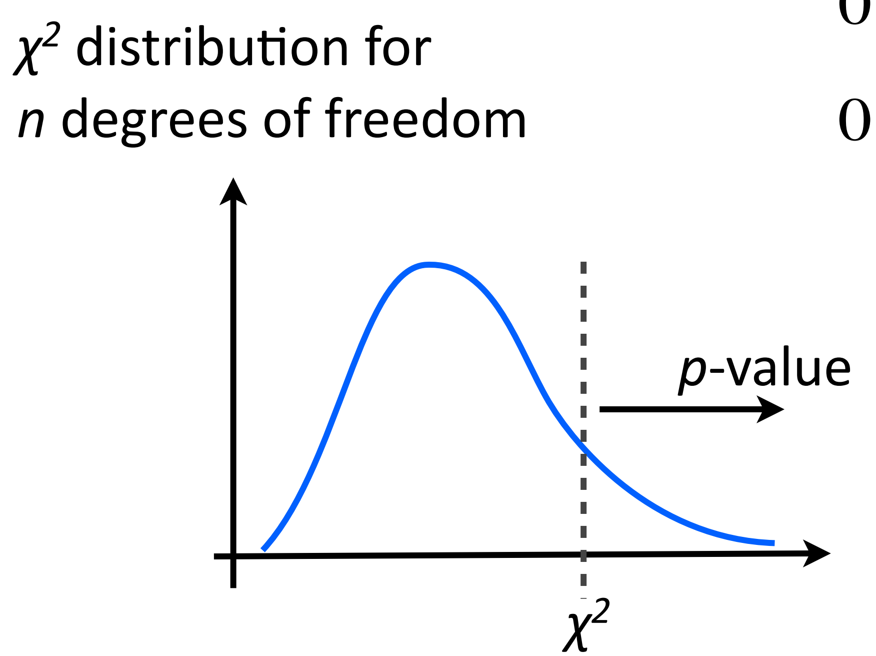
$$\chi^2_{\min} = 3.99\,, \qquad n_{\mathrm{d}} = 5-2 = 3\,, \qquad p = 0.263$$

whereas for the 0th order polynomial (horizontal line),

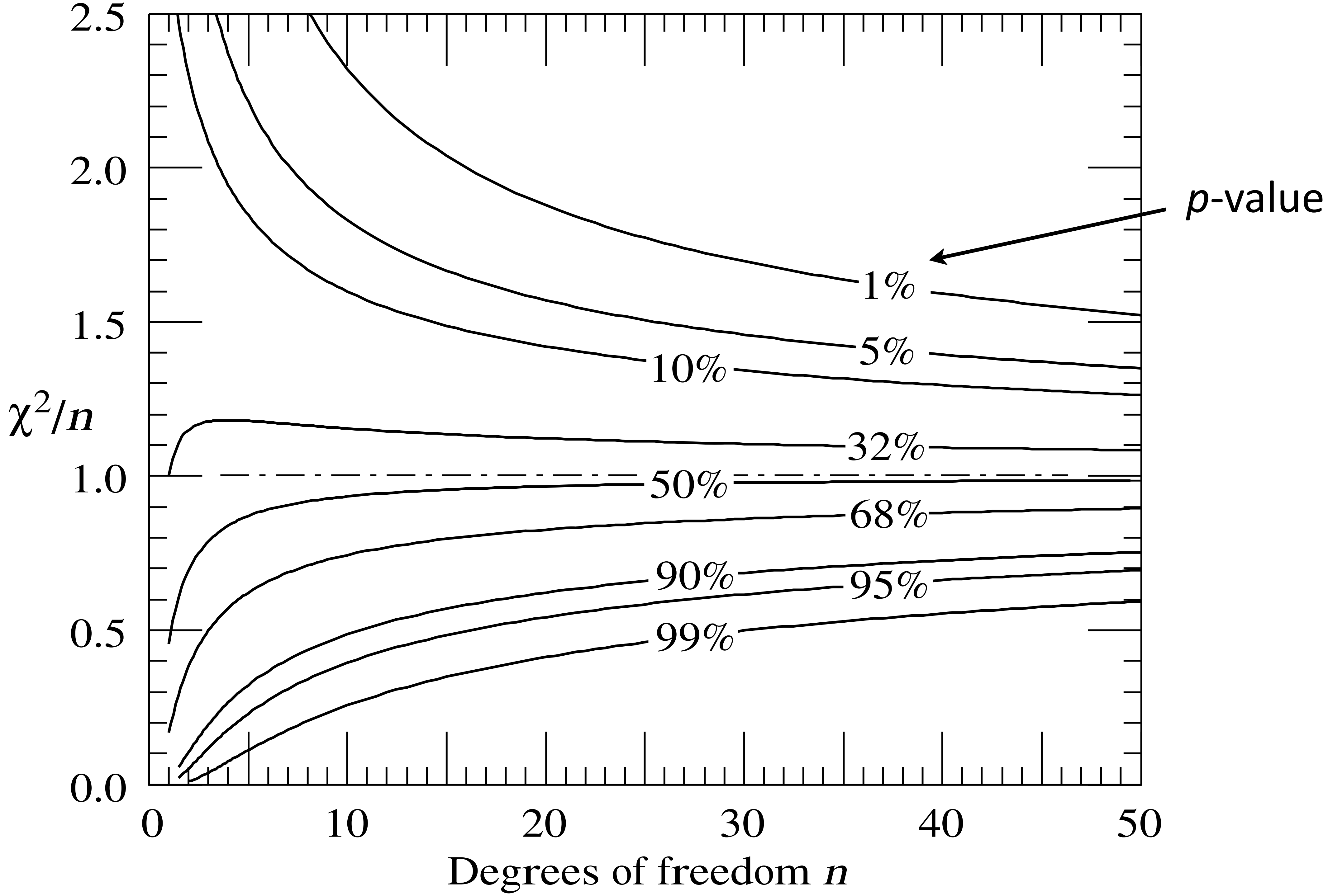
$$\chi^2_{\min} = 45.5\,, \qquad n_{\mathrm{d}} = 5-1 = 4\,, \qquad p = 3.1\times 10^{-9}$$

χ^2 Cumulative Distribution for *n* Degrees of Freedom

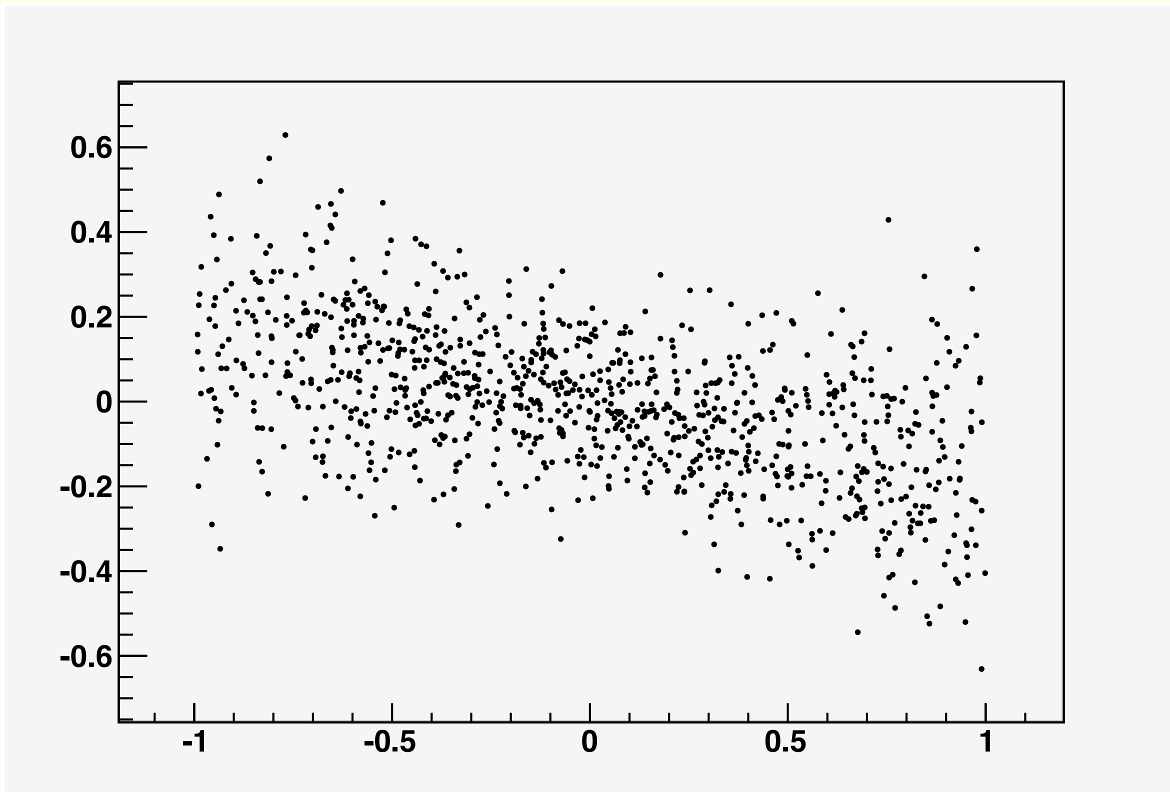


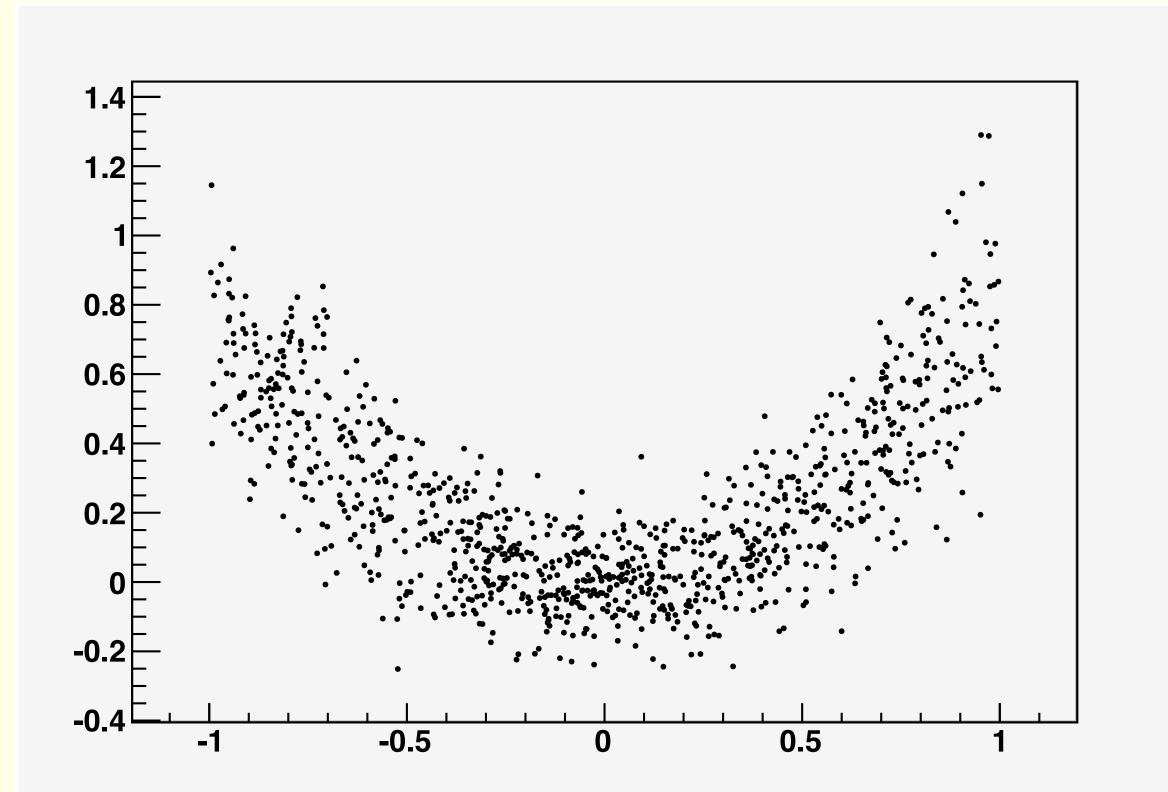


‚Reduced' χ^2 (χ^2/n)



Hands-On Exercise 1: Correlation Coefficients





Starting point: hands-on/statistics/01_corr_coeff/corr_coeff_v0.C

Calculate the correlation coefficient for the shown data sets.

Hands-On Exercise 2: Weighted Average (I)

In case of independent measurements x_i of the same quantity with uncertainties σ_i the weighted average is given by

$$\bar{x} = \frac{\sum x_i / \sigma_i^2}{\sum 1 / \sigma_i^2} \qquad V(\bar{x}) = \frac{1}{\sum 1 / \sigma_i^2}$$

a) Find the best combined result and error from the three measurements of *c*:

299 798 000 ± 5000 m/s
299 789 000 ± 4000 m/s
299 797 000 ± 8000 m/s

b) A long-lived source gives 389 counts in the first minute and 423 in the second minute. What is the best combined result? (Note: it is not 405.3)

Hands-On Exercise 2: Weighted Average (II)

5.2.2. *Unconstrained averaging:* To average data, we use a standard weighted least-squares procedure and in some cases, discussed below, increase the errors with a "scale factor." We begin by assuming that measurements of a given quantity are uncorrelated, and calculate a weighted average and error as

$$\overline{x} \pm \delta\overline{x} = \frac{\sum_i w_i\, x_i}{\sum_i w_i} \pm \left(\textstyle\sum_i w_i\right)^{-1/2} , \qquad (1)$$

where

$$w_i = 1/(\delta x_i)^2 \,.$$

Here x_i and δx_i are the value and error reported by the ith experiment, and the sums run over the N experiments. We then calculate $\chi^2 = \sum w_i(\overline{x} - x_i)^2$ and compare it with $N-1$, which is the expectation value of χ^2 if the measurements are from a Gaussian distribution.

If $\chi^2/(N-1)$ is less than or equal to 1, and there are no known problems with the data, we accept the results.

If $\chi^2/(N-1)$ is very large, we may choose not to use the average at all. Alternatively, we may quote the calculated average, but then make an educated guess of the error, a conservative estimate designed to take into account known problems with the data.

Finally, if $\chi^2/(N-1)$ is greater than 1, but not greatly so, we still average the data, but then also do the following:

(a) We increase our quoted error, $\delta\overline{x}$ in Eq. (1), by a scale factor S defined as

$$S = \left[\chi^2/(N-1)\right]^{1/2} . \qquad (2)$$

Our reasoning is as follows. The large value of the χ^2 is likely to be due to underestimation of errors in at least one of the experiments. Not knowing which of the errors are underestimated, we assume they are all underestimated by the same factor S. If we scale up all the input errors by this factor, the χ^2 becomes $N-1$, and of course the output error $\delta\overline{x}$ scales up by the same factor. See Ref. 3.

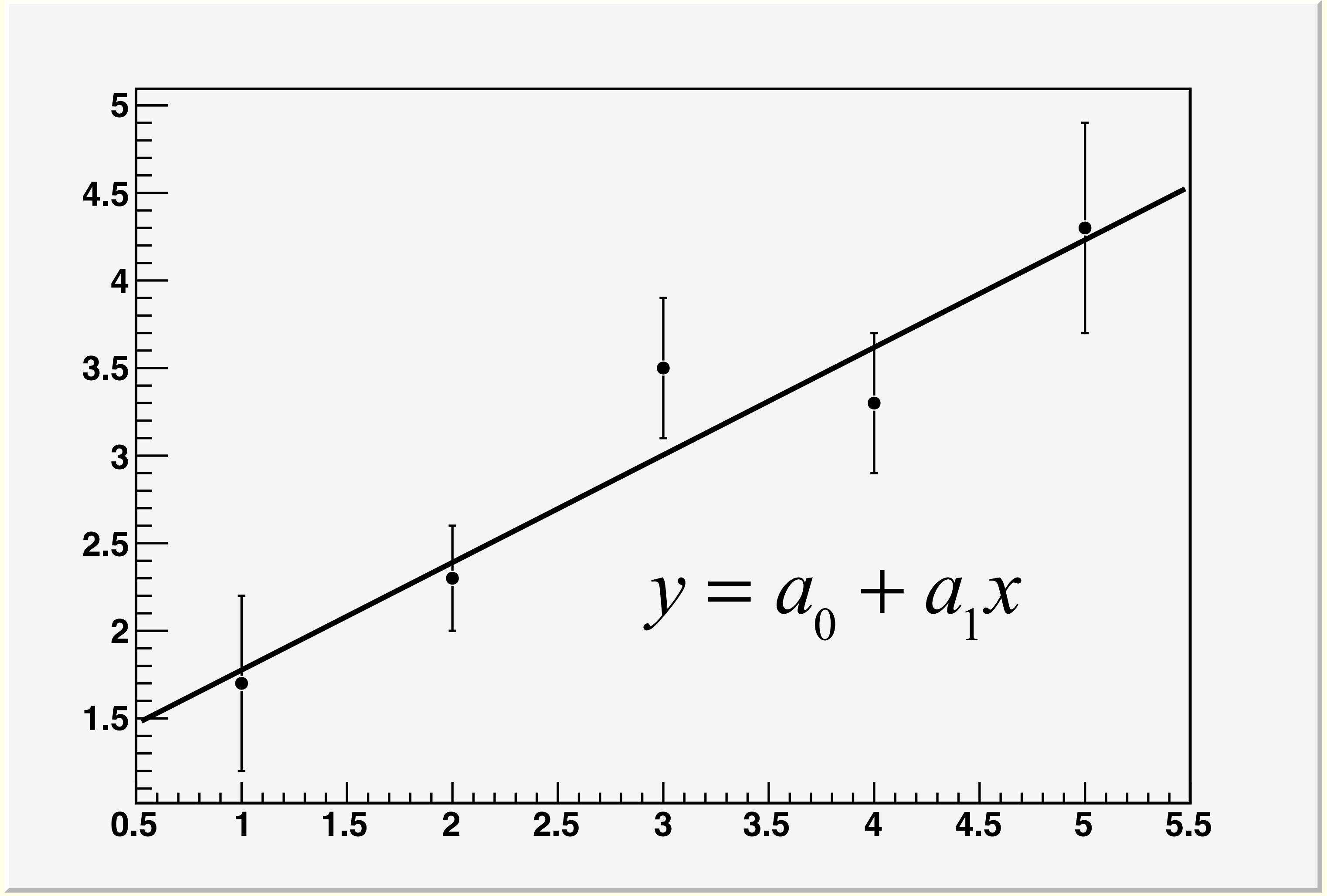
The PDG increases the error of the weighted average if the X^2 is large.

Find the best combined result and error for the three measurements of the π^0 mass following the PDG recipe:

(135 ± 1.5) MeV,
(139 ± 2.0) MeV,
(134 ± 1.0) MeV

Review of Particle Physics (2010)

Hands-On Exercise 3: Error Propagation



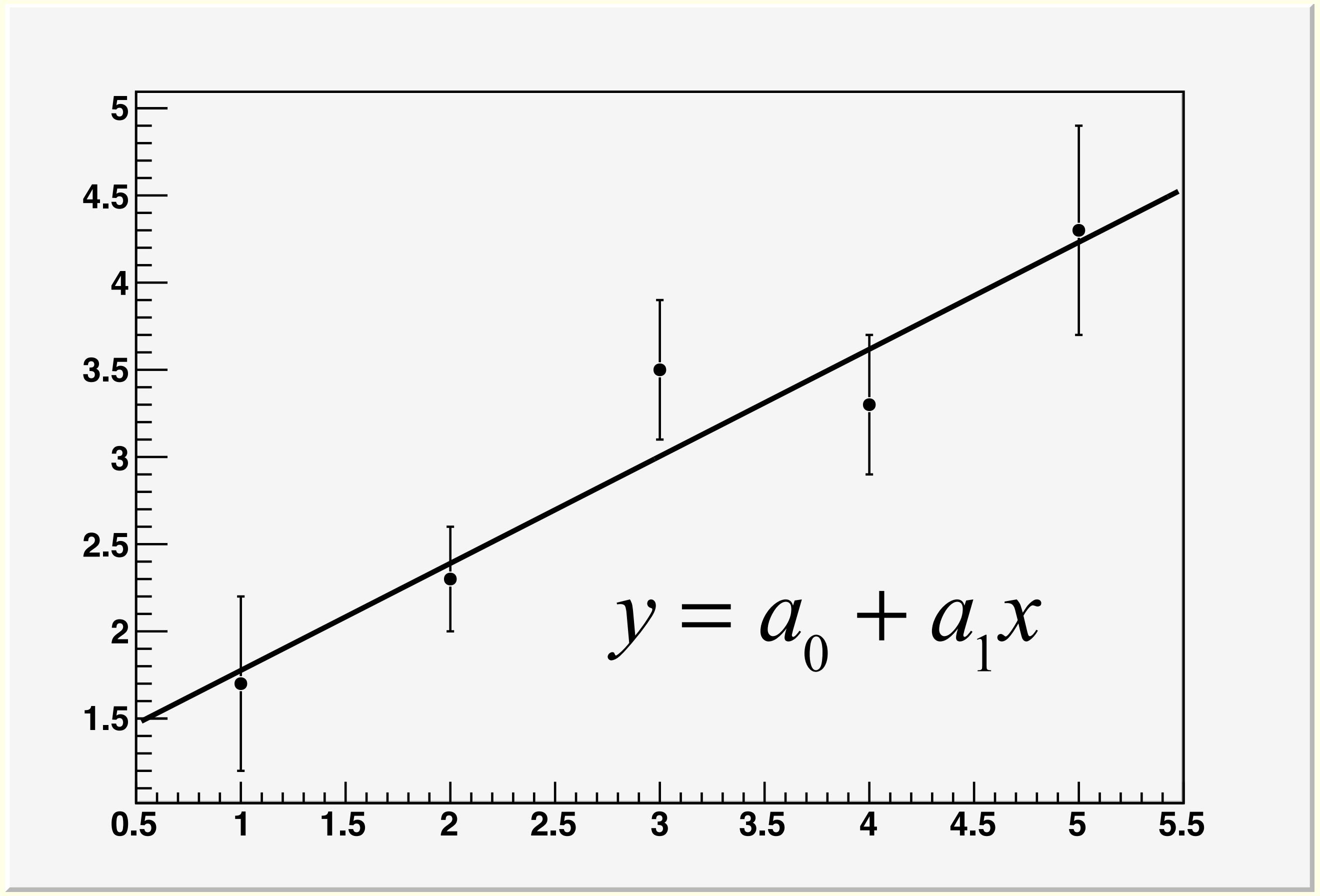
Hint on plotting error bands:

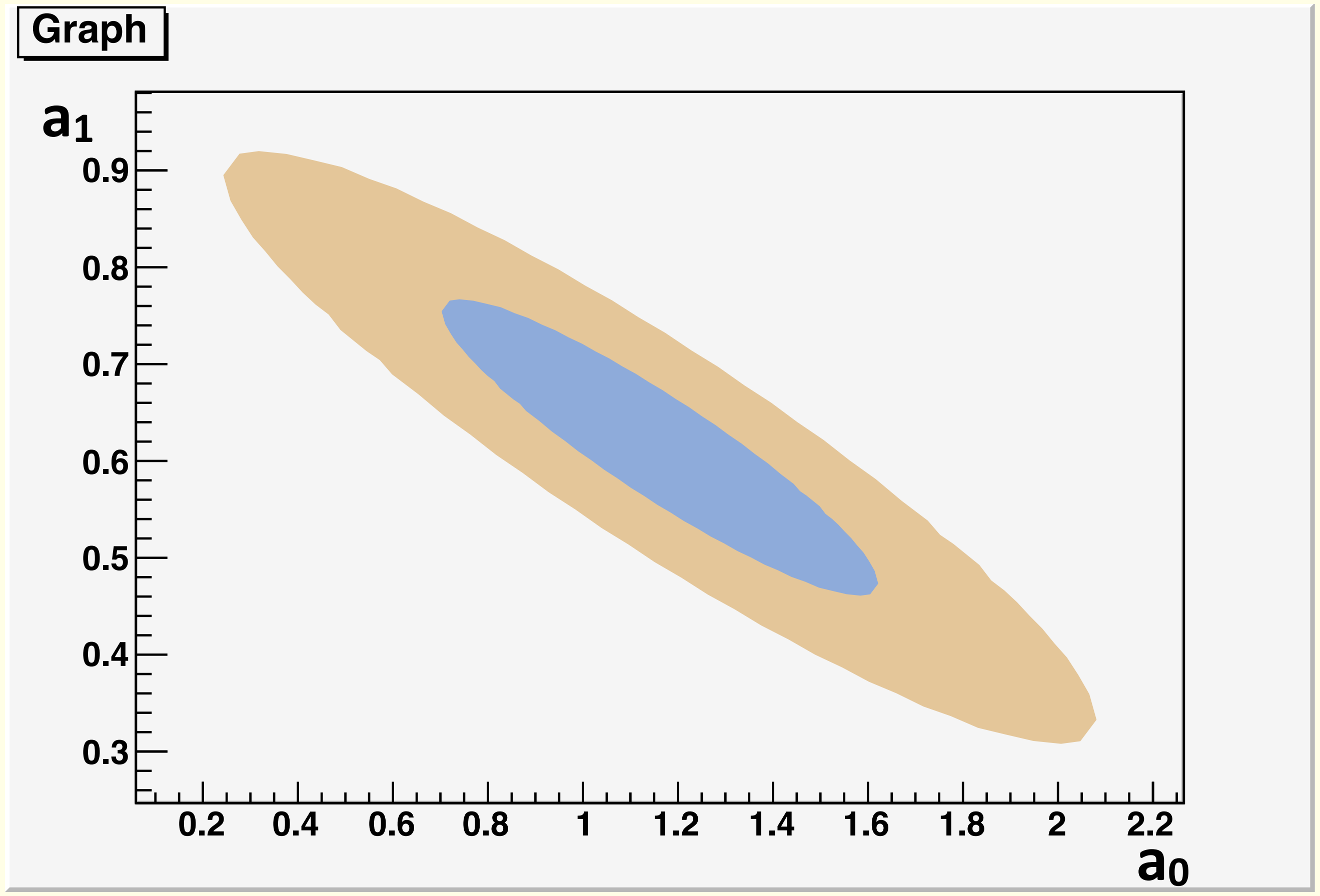
```
// draw errors of histo as bands
hErr->SetFillColor(kRed);
hErr->Draw("e3 same");
```

Starting point: hands-on/statistics/03_error_propagation/error_prop_v0.C

The data-points are fitted with a linear function. Use the covariance matrix of the fit to calculate the uncertainty of y. How big is σ_y a at $x = 1.5$? Plot the uncertainty σ_y as a band around $y(x)$. Compare your result with the one obtained from the root method used in the example $ROOTSYS/tutorials/fit/ConfidenceIntervals.C.

Hands-On Exercise 4: Error Ellipses





Starting point: hands-on/statistics/04_error_ellipses/fit1_v0.C

Look at fit1_v0.C to learn how to draw error ellipses of fit parameters.
Fit the p_T spectrum in fit2_v0.C with a Hagedorn function and plot error ellipses for all 3 pair combination of the 3 parameters.

Hands-On Exercise 5*: Minuit

Implement the fit of exercise 3 using Minuit directly. Follow the example in $ROOTSYS/tutorials/fit/Ifit.C.

Reference: http://root.cern.ch/root/html526/TMinuit.html