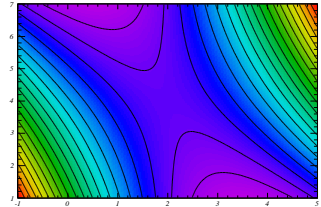




Statistische Methoden der Datenanalyse II

Michael Schmelling – MPI für Kernphysik

- *Einführung*
- *Fehler und Fehlerfortpflanzung*
- *Kleinste Quadrate & Maximum Likelihood*
- *Multivariate Analyse*
- *sWeights*
- *Markov Chain Monte Carlo*
- *Entfaltung und Parametrisierung*
- *Harmonische Analyse*





→ *selected books and papers in alphabetical order. . .*

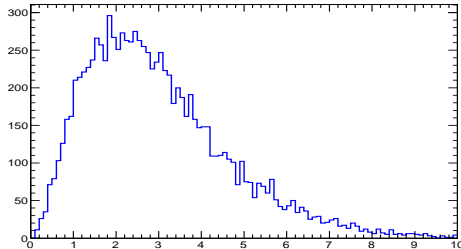
- R.J. Barlow, *Statistics*, Wiley
- S. Brand, *Data Analysis*, Springer
- G.D. Cowan, *Statistical Data Analysis*, Oxford University Press
- H.L. Harney, *Bayesian Inference*, Springer
- A. Hoecker et al., *TMVA 4 Users Guide*, <http://tmva.sourceforge.net>
- F. James, *Statistical Methods in Experimental Physics*, World Scientific
- D.E. Knuth, *The Art of Computer Programming*, Addison Wesley
- M. Pivk, F. R. Le Diberder. *sPlot*, NIM A555(2005)356, physics/0402083
- W.T. Press et al., *Numerical Recipes*, Cambridge University Press
- D.S. Sivia, *Data Analysis - A Bayesian Tutorial*, Oxford University Press



→ *What are statistical methods?*

- recipes for data reduction: large data set → single number e.g. . . .
 - md5sum: fingerprint characterizing the data set
 - particle lifetime from decay time measurements
 - CP violating phase from reconstructed B decays
- statistical methods are **constructed**
 - neither “right” nor “wrong” – characterized by properties
 - properties of a method need to be understood
to judge the applicability and to interpret the results
- example: “central value” and “spread” of a set of measurements
 - different people will associate different things with those terms
 - usually no problem for qualitative discussions
 - quantitative science requires an exact definition

→ how to characterize a data set



■ “central value”

- maximum value (after smoothing the distribution?)
- median value - same number of measurements above and below
- arithmetic average

■ “spread”

- Full-Width-at-Half-Maximum (FWHM) - but how to define the maximum
- central 68% quantile
- rms - average quadratic deviation from the mean

→ start at the beginning



→ “probability” of an event: what does this mean?

■ probability $p = 0$: the event will not happen

■ probability $p = 1$: the event will happen

■ probability $p = 1/3$: suggestions?

→ the event will happen every third try

◆ not consistent: equivalent to a sequence of $p=0, p=0, p=1$

→ the event will happen in 1/3 of infinitely many tries?

◆ OK if the next result cannot be predicted from previous ones

◆ provides a measurement prescription for repeatable tries

◆ only approximate realization possible in practice

→ I should get paid 3 EUR if I invest 1 EUR and the event happens

◆ OK - applicable also for non-repeatable events

◆ basis of the world's financial system



→ *define properties of probabilities - don't care what they are!*

Build probability theory on a mapping of sets → real numbers.

❖ Definitions:

Ω : the entire set

E : partial set of Ω

$p(E)$: probability of E

❖ Axioms:

1. $0 \leq p(E) \leq 1$

2. $p(\Omega) = 1$

3. $p(E_1 \cup E_2) = p(E_1) + p(E_2)$ if $E_1 \cap E_2 = \emptyset$

Math of probabilities follows unambiguously - interpretation is left open.



Consider the probability of an event B occurring together with another one from a set of disjoint events $A_i, i = 1, \dots, n$.

$$P(A_i, B) = p(B|A_i) p(A_i) = p(A_i|B) p(B)$$

It follows:

$$p(A_i|B) = \frac{p(B|A_i) p(A_i)}{p(B)} \quad \text{Bayes' theorem}$$

Having seen B , the prior $p(A_i)$ for A_i is updated to $p(A_i|B)$.

Bayes' theorem is at the heart of statistical inference based on empirical input. If the probabilities of the A_i sum up to unity, then one has

$$p(B) = \sum_i p(B|A_i)p(A_i)$$

and thus:

$$p(A_k|B) = \frac{p(B|A_k)p(A_k)}{\sum_i p(B|A_i)p(A_i)}$$



Applying Bayes' theorem

Consider a test that detects the common cold in the early stages of an infection, where an efficient cure is available. The probability to test positive in case of an infection is $p(+|I) = 0.98$, the probability for a negative result on a healthy subject is $p(-|H) = 0.97$. In summer, the a priori probability for infection is $p(I) = 0.001$.

What's the probability for a person tested positive to be infected?

the probabilities are:	$p(I)$	=	0.001	$p(H)$	=	0.999
	$p(+ I)$	=	0.980	$p(- I)$	=	0.020
	$p(+ H)$	=	0.030	$p(- H)$	=	0.970

The rows sum up to unity. Application of Bayes' theorem yields

$$p(I|+) = \frac{p(+|I)p(I)}{p(+|I)p(I) + p(+|H)p(H)} \approx 0.032$$

Sweets for all patients diagnosed "infected" will yield 97% "healing rate"!



→ Kolmogorov's axiom for discrete sets: discrete probabilities

Enumerate discrete probabilities by $i = 0, 1, 2, \dots$

$$\sum_i p_i = 1 \quad \text{with} \quad p_i = \text{probability to find state } i$$

→ continuous sets: probability density functions (PDFs)

A function $f(x)$ can be interpreted as a PDF if

$$f(x) \geq 0 \quad \forall x \quad \text{and} \quad \int_{-\infty}^{+\infty} dx f(x) = 1 .$$

The PDF gives the probability to observe an event in $[x, x + dx]$:

$$p(x, x + dx) = \int_x^{x+dx} dx f(x) \approx f(x) dx$$



→ the uniform distribution

The probability density inside a range $[a, b]$ is constant.

- (most) fundamental, simple PDF
- convenient starting point to derive more complex PDFs
- core of numerical random generators

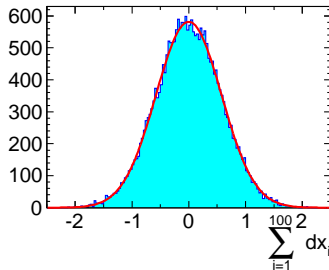
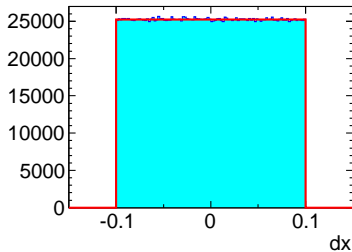
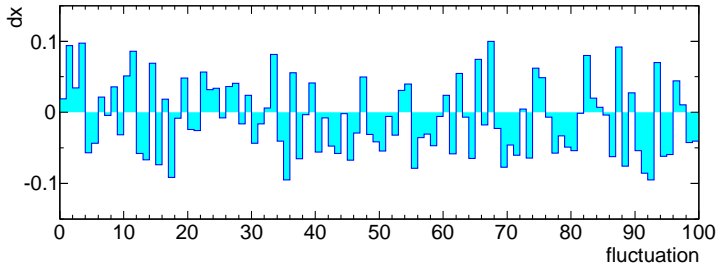
→ modelling measurement errors

- example 1: astronomical observations
 - light rays are scattered at density variations in the atmosphere
- example 2: current over a resistor
 - current variations from thermal motion of many electrons

❖ common feature

Many small variations add up to deviations between measurement und true value. Do a numerical study with uniform PDF for the variations.





→ observation

The sum of many random fluctuations is described by a **Gaussian PDF**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$

- symmetric around zero
- one parameter σ describing the width
- first published in by C.F. 1809 Gauss in
"Theoria motus corporum coelestium in
sectionibus conicis solem ambientium"
(with Least-Squares and Maximum-Likelihood method)
- the exact conditions for convergence to a Gaussian are formally described
by the **central limit theorem**
- due to its fundamental nature also referred to as “normal” distribution





→ statistics of counting experiments

■ examples:

- decays in a radioactive source
- cosmic muons observed at surface level on earth
- number of soldiers in the Prussian army killed accidentally by horse kicks (Ladislaus Bortkiewicz, 1898)

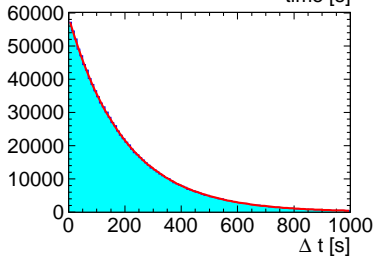
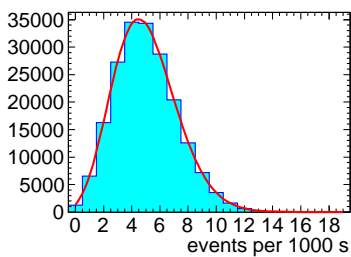
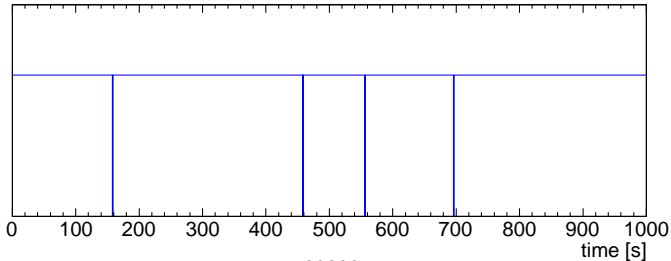
■ quantities of interest

- time differences between subsequent events
- number of events in time interval T

❖ numerical simulation

- split T into (many) subsequent time slices
- assume a probability to observe an event in a time slice $p \ll 1$

see what happens →



→ observation

- results are described by simple functions of a single parameter (consequence of the single probability for an event per time slice)
- event counts per time interval: **Poisson distribution**
 - first published by Siméon Denis Poisson 1837 in “Recherches sur la probabilité des jugements en matière criminelle et en matière civile”

$$p_n = e^{-\mu} \frac{\mu^n}{n!}$$

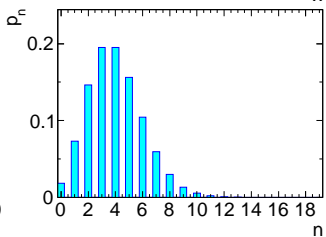
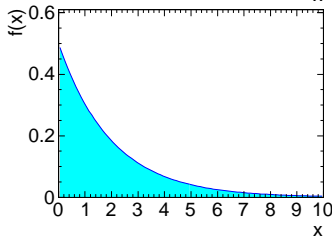
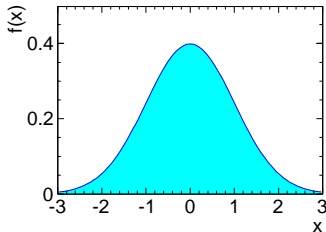
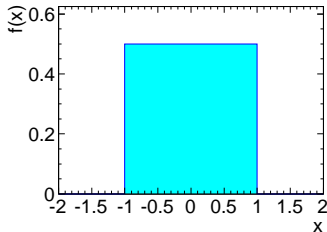
- time difference between events: **exponential distribution**

$$f(t) = \frac{1}{\tau} e^{-t/\tau}$$





→ consider the distributions introduced before



❖ wanted: location and width



→ “typical” x -values

- maximum of the distribution
 - not always well defined
 - can be at one edge of the distribution
- median value m

$$\int_{-\infty}^m dx f(x) = \int_m^{\infty} dx f(x)$$

- not obvious for discrete distributions; insensitive to tails
- center-of-gravity $\langle \cdot \cdot \rangle$ – usually referred to as mean value

$$\langle x \rangle = \int dx x f(x) \quad \text{or} \quad \langle n \rangle = \sum_{n=0}^{\infty} n p_n$$

- well defined for continuous and discrete distributions
- sensitive to asymmetric tails; may even diverge
- median and center-of-gravity coincide for symmetric distributions
- all three “typical values” coincide for symmetric uni-modal distributions



→ “typical” range covered by x -values

- **FWHM**: full width at half the maximum value
 - not obvious for discrete distributions; insensitive to tails
- **central $q\%$ quantile** $[a, b]$, with, e.g., $q=68.3\%$, 90% or 95%

$$\int_{-\infty}^a dx f(x) = \int_b^{\infty} dx f(x) = \frac{1}{2}(1 - q)$$

- not obvious for discrete distributions; insensitive to tails
- **standard deviation** σ

$$\sigma^2 = \int_{-\infty}^{\infty} dx f(x) (x - \langle x \rangle)^2 \quad \text{or} \quad \sigma^2 = \sum_{n=0}^{\infty} (n - \langle n \rangle)^2 p_n$$

- well defined for continuous and discrete distributions
- sensitive to the functional form of the tails - may even diverge
- simple linear operation on the PDF

“unified” approach →

A measure for the scatter s of x with PDF $f(x)$ around a point a is:

$$s^2 = \int dx (x - a)^2 f(x)$$

For s to characterize $f(x)$, a should be chosen to minimize s :

$$\frac{\partial s^2}{\partial a} = -2 \int dx (x - a) f(x) \stackrel{!}{=} 0 \quad \text{i.e.} \quad a = \int dx x f(x) = \langle x \rangle$$

The **mean value** $\langle x \rangle$ is the location parameter that minimizes the scatter s . The minimal scatter s is called **standard deviation**, σ . Its square is called **variance**, σ^2 .

→ *note:*

- for symmetric PDFs $\langle x \rangle$ is the symmetry point
- the scatter around $\langle x \rangle$ is called “**standard deviation**” σ
- σ is also referred to as “**rms**”-width



→ uniform, gaussian, exponential and poisson distributions

$$\frac{1}{2w} \Theta(x+w) \Theta(w-x) \quad , \quad \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \quad , \quad \frac{e^{-x/\tau}}{\tau} \quad , \quad e^{-\mu} \frac{\mu^n}{n!}$$

	median	mean	FWHM	68.3% quant.	stdev
uniform	0	0	$2w$	$1.366 w$	$w/\sqrt{3}$
gaussian	0	0	$\sqrt{8 \ln 2} \sigma$	2σ	σ
exponential	$\tau \ln 2$	τ	$\tau \ln 2$	$-\tau \ln 0.317$	τ
poisson		μ			$\sqrt{\mu}$

- ratios of different width or location estimators are $O(1)$
- analytically most convenient: mean value and standard deviation
 - simple integrals over the entire distributions
 - most commonly used estimators for location and width



→ *generalization of concepts introduced before:*

Given a PDF $f(x)$ and a function $a(x)$, the expectation value $\langle a \rangle$ is:

$$\langle a \rangle = \int_{-\infty}^{\infty} dx \, a(x) f(x)$$

- mapping of functions $f(x)$ to a real numbers – if the integral exists
- important property: **linearity** i.e. $\langle \alpha A + \beta B \rangle = \alpha \langle A \rangle + \beta \langle B \rangle$

❖ **examples:**

$\langle x \rangle$: mean value

$\langle (x - \langle x \rangle)^2 \rangle$: variance

→ *note:*

$$\begin{aligned} \sigma^2 &= \int dx \, (x - \langle x \rangle)^2 f(x) = \int dx \, (x^2 - 2x \langle x \rangle + \langle x \rangle^2) f(x) \\ &= \left(\int dx \, x^2 f(x) \right) - \langle x \rangle^2 = \langle x^2 \rangle - \langle x \rangle^2 \end{aligned}$$

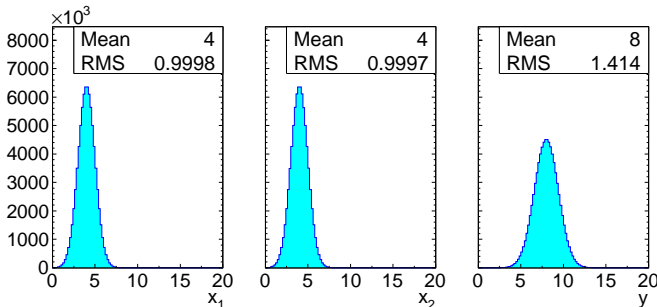


→ *the problem:*

Given PDFs $f_1(x_1)$ and $f_2(x_2)$, determine the PDF $g(y)$ of $y = h(x_1, x_2)$.

→ *solution by Monte Carlo*

- generate x_1 and x_2 according to $f_1(x_1)$ and $f_2(x_2)$
- calculate and histogram $y = h(x_1, x_2)$





→ *the problem:*

Given PDFs $f_1(x_1)$ and $f_2(x_2)$, determine the PDF $g(y)$ of $y = h(x_1, x_2)$.

→ *analytic solution*

For the cumulative distribution $G(Y)$ one has:

$$G(Y) \equiv \int_{-\infty}^Y dy g(y) = \int dx_1 dx_2 f_1(x_1) f_2(x_2) \Theta(Y - h(x_1, x_2))$$

Sum all probability elements $dp_1 dp_2$, with $dp_i = dx_i f_i(x_i)$, which satisfy the constraint $h(x_1, x_2) < Y$. Differentiation with respect to the upper limit Y then yields the solution:

$$g(y) = \left. \frac{d}{dY} G(Y) \right|_{Y=y} = \int dx_1 dx_2 f_1(x_1) f_2(x_2) \delta(y - h(x_1, x_2))$$



→ normalization, mean value and variance of $y = x_1 + x_2$

$$\begin{aligned}\langle y^k \rangle &= \int dy y^k g(y) = \int dy y^k \int dx_1 dx_2 f_1(x_1) f_2(x_2) \delta(y - x_1 - x_2) \\ &= \int dx_1 dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^k\end{aligned}$$

expectation values:

$$\langle y^0 \rangle = \int dx_1 dx_2 f_1(x_1) f_2(x_2) = 1$$

$$\langle y^1 \rangle = \int dx_1 dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2) = \langle x_1 \rangle + \langle x_2 \rangle$$

$$\langle y^2 \rangle = \int dx_1 dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^2 = \langle x_1^2 \rangle + 2 \langle x_1 \rangle \langle x_2 \rangle + \langle x_2^2 \rangle$$

$$\text{and thus} \quad \langle y^2 \rangle - \langle y \rangle^2 = [\langle x_1^2 \rangle - \langle x_1 \rangle^2] + [\langle x_2^2 \rangle - \langle x_2 \rangle^2]$$

❖ *convolutions are normalized, mean values and variances are added!*



→ generalization of 1-dim PDFs

- non-negative, normalizable functions in n dimensions
- discuss the most important concepts with 2-dim PDFs

❖ 2-dim PDF:

$$f(x, y) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y) = 1$$

❖ interpretation:

the Probability for (x, y) in the rectangle $[x, x + dx] \times [y, y + dy]$ is

$$p(x, x + dx; y, y + dy) = \int_x^{x+dx} dx \int_y^{y+dy} dy f(x, y) \approx f(x, y) dx dy$$

❖ independence of variables:

x and y are independent if the PDF factorizes: $f(x, y) = g_1(x) \cdot g_2(y)$



→ look for expectation values that are sensitive to dependencies

0th order $\langle 1 \rangle$

1st order $\langle x \rangle, \langle y \rangle$

2nd order $\langle x^2 \rangle, \langle xy \rangle, \langle y^2 \rangle$

The lowest order term probing dependencies between x and y is $\langle xy \rangle$.

For independent variables with $f(x, y) = g_1(x) g_2(y)$ one finds

$$\begin{aligned}\langle xy \rangle &= \int dx \int dy (x y) g_1(x) g_2(y) \\ &= \left(\int dx x g_1(x) \right) \left(\int dy y g_2(y) \right) = \langle x \rangle \langle y \rangle\end{aligned}$$

❖ measure of correlation: the “covariance” of x and y

$$C_{xy} = \langle x y \rangle - \langle x \rangle \langle y \rangle$$

not the only possibility, but simple and useful . . .



→ *dimensionless measures of correlation between two variables*

$$\rho = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{C_{xy}}{\sqrt{C_{xx} C_{yy}}}$$

❖ properties

■ $-1 \leq \rho \leq 1$

■ $y = a x + b \rightarrow \rho = \text{sign}(a)$ (“100% (anti)correlation”)

■ $\rho = 0$ necessary, but not sufficient for independence of x and y

❖ example: function $y = a x^2 + b x + c$ with gaussian distributed x

$$\rho = \frac{b}{\sqrt{2a^2\sigma_x^2 + b^2}}$$

→ $|\rho| < 1$ if a parabolic term is present

→ $\rho = 0$ if the linear term is absent



→ array of covariances between all variable-pairs of an n -dim PDF:

$$C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

Expressed through standard deviations and correlation coefficients it is

$$C_{ij} = \rho_{ij} \cdot \sigma_i \sigma_j \quad \text{with} \quad \rho_{ii} = 1 .$$

→ note:

- the diagonal terms C_{ii} are the variances of the individual variables
- off-diagonal terms are covariances
- the covariance matrix is symmetric and positive definite
- it can be diagonalized by rotation in the space of the variables
- C also is referred to as “error matrix”
- C describes the extension and orientation of an n -dim PDF



→ exploit the linearity of expectation values

Consider a linear transformation $y_k = \sum_i M_{ki} x_i$. Given the covariance matrix $C_{ij}(x)$, the covariance matrix $C_{kl}(y)$ is

$$\begin{aligned} C_{kl}(y) &= \langle y_k y_l \rangle - \langle y_k \rangle \langle y_l \rangle \\ &= \left\langle \left(\sum_i M_{ki} x_i \right) \left(\sum_j M_{lj} x_j \right) \right\rangle - \left\langle \sum_i M_{ki} x_i \right\rangle \left\langle \sum_j M_{lj} x_j \right\rangle \\ &= \sum_{ij} M_{ki} M_{lj} (\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle) = \sum_{ij} M_{ki} M_{lj} C_{ij}(x) \end{aligned}$$

or in matrix notation:

$$\vec{y} = M \cdot \vec{x} \quad \text{and} \quad C(y) = M \cdot C(x) \cdot M^T$$

- if $C(x)$ is positive definite, so is $C(y)$
- M need not be a square matrix - the number of rows is arbitrary



2. ERRORS AND ERROR PROPAGATION



→ *what are errors?*

- “errors” are uncertainties - not to be confused with “mistakes”
- quantify how well one knows e.g. a constant of nature - but how?
- engineer: tolerance = maximum possible deviation
- physicist: many different conventions. . .
 - standard deviation σ
 - 3- σ uncertainties
 - confidence level intervals containing the true value. . .
 - ◆ in a **certain fraction** of experiments (frequentist)
 - ◆ with a **certain probability** (bayesian)

→ ask the professionals. . .



→ *Expression of experimental uncertainties*

- 1 The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:
 - A those which are evaluated by statistical methods,
 - B those which are evaluated by other means.

There is not always a simple correspondence between the classification into categories A or B and the previously used classification into “random” and “systematic” uncertainties. The term “systematic uncertainty” can be misleading and should be avoided. Any detailed report of the uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical value.



- 2 The components in category A are characterized by the estimated variances s_i^2 (or the estimated “standard deviations” s_i) and the number of degrees of freedom ν_i . Where appropriate, the covariances should be given.
- 3 The components in category B should be characterized by quantities u_j^2 , which may be considered as approximations to the corresponding variances, the existence of which is assumed. The quantities u_j^2 may be treated like variances and the quantities u_j like standard deviations. Where appropriate, the covariances should be treated in a similar way.
- 4 The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined uncertainty and its components should be expressed in the form of “standard deviations”.
- 5 If, for particular applications, it is necessary to multiply the combined uncertainty by a factor to obtain an overall uncertainty, the multiplying factor used must always be stated.

(end of quote)



→ why define uncertainties by variances and standard deviations

- well defined procedures how to handle them
 - when propagating uncertainties into derived variables
 - for the combination of independent measurements
- rigorous limits on probability contents in the tails
- often asymptotically gaussian behaviour (central limit theorem)
- no (little) danger of mis-interpretation
- confidence level intervals . . .
 - not always obvious how they are defined
 - not obvious how to combine them
- warning: many physics papers actually mix concepts, combining “one-sided” variances in quadrature with confidence level intervals . . .

❖ focus first on variances/standard deviations!



→ *probability content in the tails of a distribution*

Take any PDF $f(x)$, function $w(x) \geq 0$ and x -region with $w(x) \geq C$:

$$\langle w \rangle = \int dx f(x) w(x) \geq \int_{w(x) \geq C} dx f(x) w(x) \geq C \int_{w(x) \geq C} dx f(x) = C p(w(x) \geq C)$$

$$\text{it follows} \quad p(w(x) \geq C) \leq \frac{\langle w \rangle}{C}.$$

For the special choice $w(x) = (x - \langle x \rangle)^2$ and $C = k^2 \sigma^2$ one finds:

$$p_k \equiv p((x - \langle x \rangle)^2 > k^2 \sigma^2) \leq \frac{1}{k^2}$$

- the probability beyond $\pm k \sigma$ around $\langle x \rangle$ is at most $1/k^2$
- actual probability contents for most PDFs are much lower
 - e.g. gaussian: $\{p_1, p_2, p_3\} \approx \{0.317, 0.0555, 0.0027\}$



→ definitions:

- \vec{x} : vector of observed quantities
- $\langle \vec{x} \rangle$: expectation values of \vec{x} - assumed to be the true values \vec{x}^t
- $C(\vec{x})$: covariance matrix of \vec{x} - assumed to be known
- $\vec{y} = \vec{g}(\vec{x})$: vector of derived quantities
- $\vec{y}^t = \vec{g}(\langle \vec{x} \rangle)$: true vector of derived quantities
- $C(\vec{y})$: covariance matrix of \vec{y} - to be determined

❖ study properties of the transition $\vec{x} \rightarrow \vec{y}$

- expectation values
- uncertainties



→ the expectation value of \vec{y} is biased: $\langle \vec{y} \rangle \neq \vec{y}^t$

Taylor expansion for a single component around $\langle x \rangle$ shows

$$y_k = g_k(\langle \vec{x} \rangle) + \sum_i \frac{\partial g_k(\langle \vec{x} \rangle)}{\partial x_i} (x_i - \langle x_i \rangle) + \frac{1}{2} \sum_{i,j} \frac{\partial^2 g_k(\langle \vec{x} \rangle)}{\partial x_i \partial x_j} (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) + \dots$$

and taking the expectation value yields:

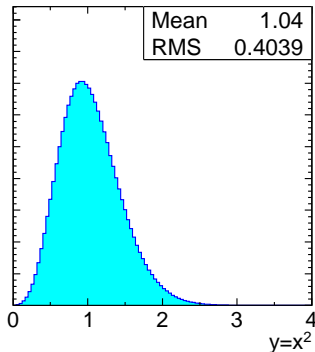
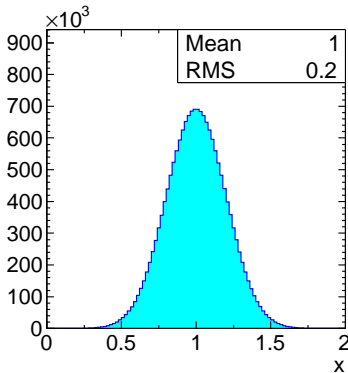
$$\langle y_k \rangle = y_k^t + \frac{1}{2} \sum_{i,j} \frac{\partial^2 g_k(\langle \vec{x} \rangle)}{\partial x_i \partial x_j} C_{ij}(x) + \dots$$

❖ discussion

- ❑ in many cases the bias is small and can be neglected
- ❑ the leading order correction in principle is known
- ❑ don't average biased estimates of \vec{y} - average the unbiased \vec{x}



→ transformation of a gaussian distributed $x \rightarrow y = x^n$



- small non-linearities or small σ are uncritical
- biases are usually small compared to standard deviations
- bias correction is needed when averaging transformed values



→ leading order treatment in n dimensions

$$y_k \approx g_k(\langle \vec{x} \rangle) + \sum_{i=1}^n \frac{\partial g_k(\langle \vec{x} \rangle)}{\partial x_i} (x_i - \langle x_i \rangle) \quad \text{expansion around } \langle \vec{x} \rangle$$

$$\approx g_k(\langle \vec{x} \rangle) + \sum_{i=1}^n \frac{\partial g_k(\vec{x})}{\partial x_i} (x_i - \langle x_i \rangle) \quad \text{derivatives taken at } \vec{x}$$

$$\approx \langle y_k \rangle + \sum_{i=1}^n \frac{\partial g_k(\vec{x})}{\partial x_i} (x_i - \langle x_i \rangle) \quad \text{assume } \vec{y}^t = \langle \vec{y} \rangle$$

then calculate the covariance matrix $C_{kl}(y) = \langle (y_k - \langle y_k \rangle)(y_l - \langle y_l \rangle) \rangle$:

$$C_{kl}(y) \approx \sum_{i,j=1}^n \frac{\partial g_k}{\partial x_i} \frac{\partial g_l}{\partial x_j} \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle = \sum_{i,j=1}^n \frac{\partial g_k}{\partial x_i} \frac{\partial g_l}{\partial x_j} C_{ij}(x)$$

(note: derivatives are taken at the measured \vec{x} .)



→ *matrix notation*

Under a transformation $\vec{y} = \vec{g}(\vec{x})$ the covariance matrix transforms as

$$C(y) = M(x) \cdot C(x) \cdot M^T(x)$$

with jacobian $M(x)$ and matrix elements $M_{ij} = \frac{\partial g_i}{\partial x_j}$.

The argument to M indicates that the derivatives are with respect to \vec{x} .

If $M(x)$ can be inverted then no information is lost in the transformation.

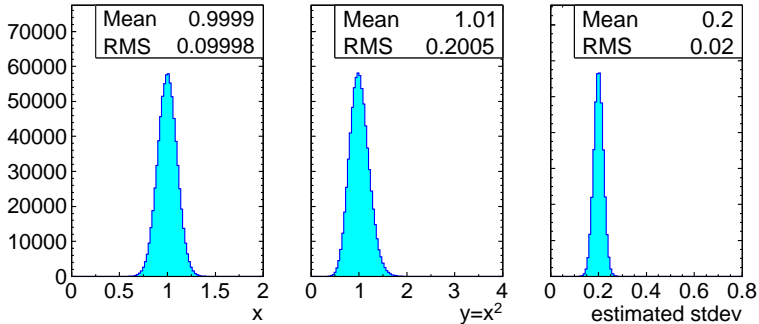
When chaining transformations one has:

$$\vec{y} = \vec{h}(\vec{g}(\vec{x})) \quad \text{and} \quad M_{ij} = \sum_{k=1}^n \frac{\partial h_i}{\partial g_k} \frac{\partial g_k}{\partial x_j} \quad \text{or} \quad M = M(g) \cdot M(x).$$

Gaussian error propagation is consistent. The final covariance matrix is the same, if a transformation is done in one or in several steps.



→ estimated and exact standard deviations for $x \rightarrow y = x^n$

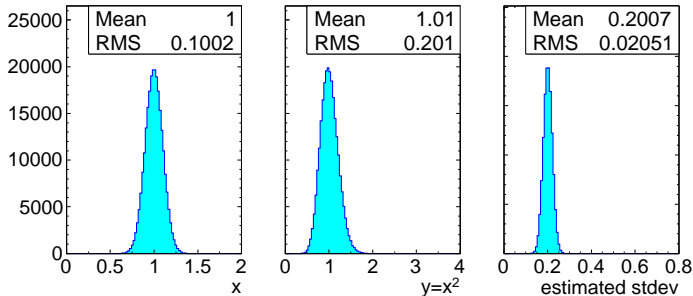


- average error estimates are OK
- actual values scatter proportional to relative errors of x



→ *error MC and exact standard deviations for $x \rightarrow y = x^n$*

Fluctuate every measured value x by its known variance and estimate the standard deviation of y from the transformed x -values.



- similar behaviour as analytical results (slightly larger scatter)
- easy to implement as no derivatives are required
- small sensitivity to PDF of fluctuations



→ *when uncertainties are quantified by the covariance matrix. . .*

- gaussian error propagation is . . .
 - consistent when chaining transformations
 - exact for linear transformations
 - approximate for non-linear transformation
- error propagation via MC is . . .
 - easy to implement
 - approximately the same accuracy as gaussian error propagation
- error estimates for non-linear transformation can have relative uncertainties of the same order of magnitude as the measured quantities - even if the variances of the measurements are known!
- non-linear transformation induces a bias
- leading order bias correction is recommended before averaging



→ alternative ways to quantify uncertainties

- no longer distribution-free – the underlying PDFs need to be known
- propagation of uncertainties usually not possible
 - requires full PDFs or likelihood functions
 - usually only the intervals are provided
- combination of uncertainties not well defined
 - common practice:

$$a = 42 \pm_3^8 \pm_4^6 = 42 \pm_5^{10}$$

- little or no theoretical backing
- implies the concept of asymmetric variance
- implies that confidence level intervals behave like variances
- different concepts in bayesian and frequentist frameworks

a simple case study →



→ setting the scene:

A counting experiment has **observed n events**. The experiment did counted independent random processes with a constant probability per time interval to happen, such as e.g. radiocative decays. It thus is known that n is a **poissonian distributed** random variable, i.e. the **probability P_n to observe n events** is:

$$P_n = P(n; \mu) = e^{-\mu} \frac{\mu^n}{n!}$$

→ question:

What can be inferred about the expectation value μ ?



→ quick check of a few hypotheses ...

→ $P(2; \mu = 0.1) \approx 0.0045$

→ $P(2; \mu = 1.0) \approx 0.1839$

→ $P(2; \mu = 10.) \approx 0.0023$

- in principle any value for μ is possible
- a value $\mu = O(1)$ seems more plausible

❖ try to be quantitative about a certain range of μ

- discuss
 - the Bayesian approach
 - the frequentist approach



→ treat μ as a random variable

- formally possible even if μ has a well defined true physical value
- interpret the PDF of μ as encoding the **knowledge** about μ
- use **Bayes' theorem** to **improve** the knowledge by the measurement:

$$P(\mu|n) P(n) = P(n|\mu) P(\mu)$$

- $P(\mu)$: prior PDF of μ - to be defined
- $P(n|\mu)$: Likelihood function
- $P(n)$: probability for n , unknown **constant**
- $P(\mu|n)$: posterior PDF for μ after the measurement

❖ it follows

$$P(\mu|n) \propto P(n|\mu) P(\mu) \quad \text{and thus} \quad P(\mu|n) = \frac{P(n|\mu) P(\mu)}{\int d\mu P(n|\mu) P(\mu)}$$



→ choice of prior distribution

$$P(\mu) = \mu^k$$

- ad hoc - but allows to test sensitivity to prior, special cases:
- $k = 0$: equal probability for all possible values
- $k = -1$ Jeffries prior: invariance w.r.t scale-transformations $\mu \rightarrow \alpha \mu$

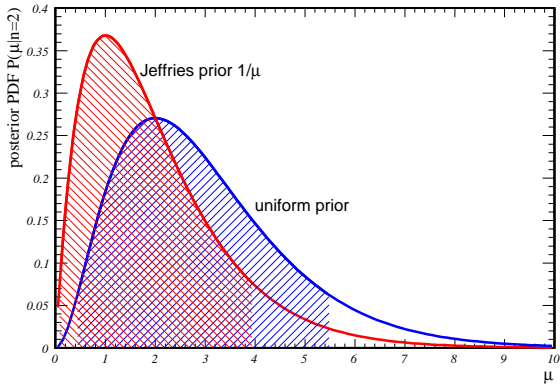
$$P(\mu|n) = \frac{e^{-\mu} \mu^{n+k}}{\int_0^\infty d\mu e^{-\mu} \mu^{n+k}} = e^{-\mu} \frac{\mu^{n+k}}{(n+k)!}$$

→ equal to Poisson-likelihood to observe $n + k$ for given μ

results →



→ posterior distributions and 90% CL intervals



- $X\%$ confidence intervals are regions with $X\%$ probability content
→ many possibilities - usually take the **smallest interval**
- most probable values and confidence intervals depend on the prior



- bayesian approach formalizes gain of knowledge by measurement

→ posterior of first measurement can be prior of second, etc.

$$\begin{aligned}P(\mu|n_2, n_1) &\propto P(n_2|\mu)P(n_1|\mu)P(\mu) \\&= P(n_2, n_1|\mu)P(\mu) \\&= P(n_2|\mu)P_1(\mu) \quad \text{with} \quad P_1(\mu) = P(n_1|\mu)P(\mu)\end{aligned}$$

- consistent if a non-uniform prior (e.g. Jeffries') is used only once

→ avoid non-uniform priors for single measurements

→ if needed, use a non-uniform prior once when combining results

→ possible if likelihood functions are published

- use of uniform priors corresponds to maximum likelihood approach

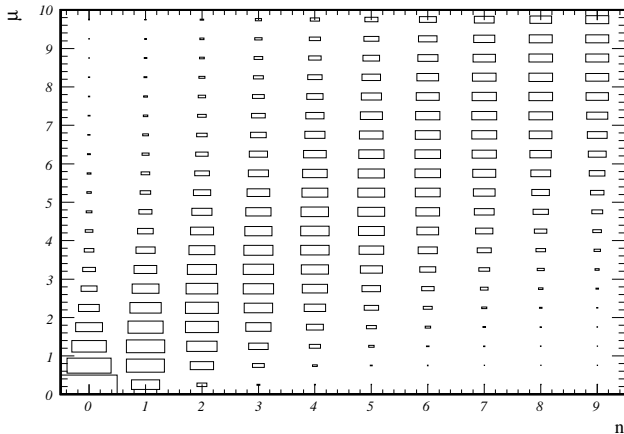
- caveat: uniformity depends on the definition of the parameter

→ example: uniform in μ is non-uniform in $\sqrt{\mu}$



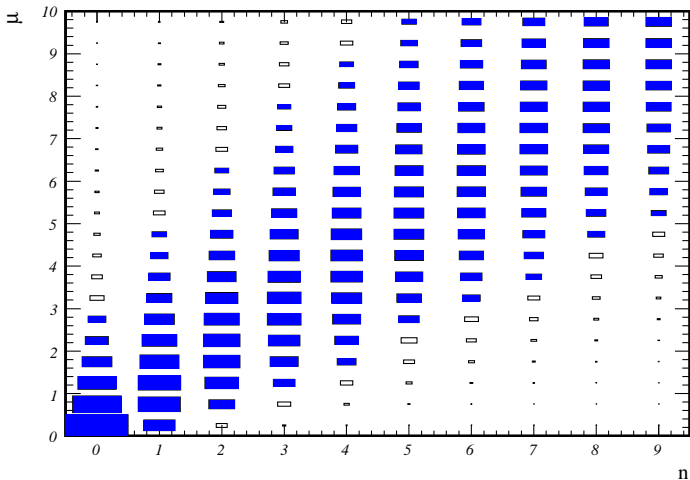
→ Likelihood-function-only based “Neyman construction”

- start from table of probabilities for any observation and any value μ



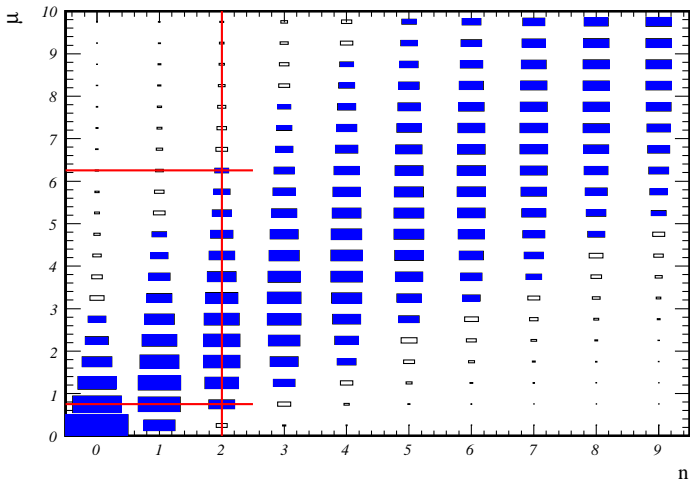


- determine the shortest $\geq 90\%$ horizontal range for each μ





■ given n , take the range of μ with n in the $\geq 90\%$ probability range





- a fixed interval for μ is assigned to every measurement n
- every interval contains the true value with 90% probability
 - false from the frequentist point of view – the true value μ is either inside or outside; it is a fixed value and does not depend on n .
- from an ensemble of measurements (at least) 90% of the confidence level intervals are expected to contain the true value
 - true – for any true μ , different measurements will find different values n and thus will quote different intervals. Take for example $\mu = 4.25$. It is contained in the intervals of $n = 1, \dots, 7$, and by construction, (at least) 90% of the measurements are in that range. Analogous reasoning holds for all μ .
- the interval contains no information about preferred values!



→ some common themes. . .

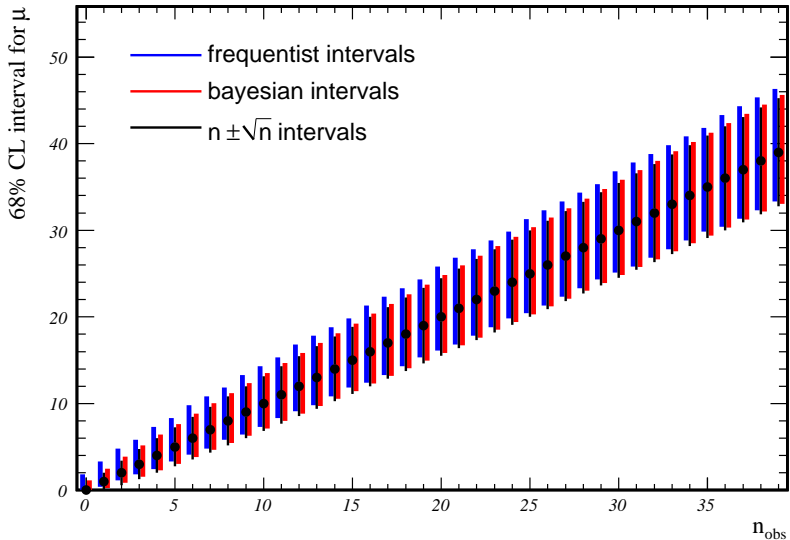
- ❑ bayesian and frequentist methods define **regions** $[\mu_l(n), \mu_h(n)]$
- ❑ for each observation n there is a well defined interval
- ❑ another commonly used interval is $n \pm \sqrt{n}$
 - estimate for the standard deviation of the measurement
 - often taken also as approximate 68.3% confidence level interval

→ further studies. . .

- ❑ compare the intervals defined by the different schemes
- ❑ MC check which fraction of intervals contain the true value
 - do the check as a **function of** the unknown true μ
 - check that the frequentists intervals have **coverage**
 - calculate coverage **also for bayesian** intervals
 - ◆ even if **bayesians do not care** about coverage . . .

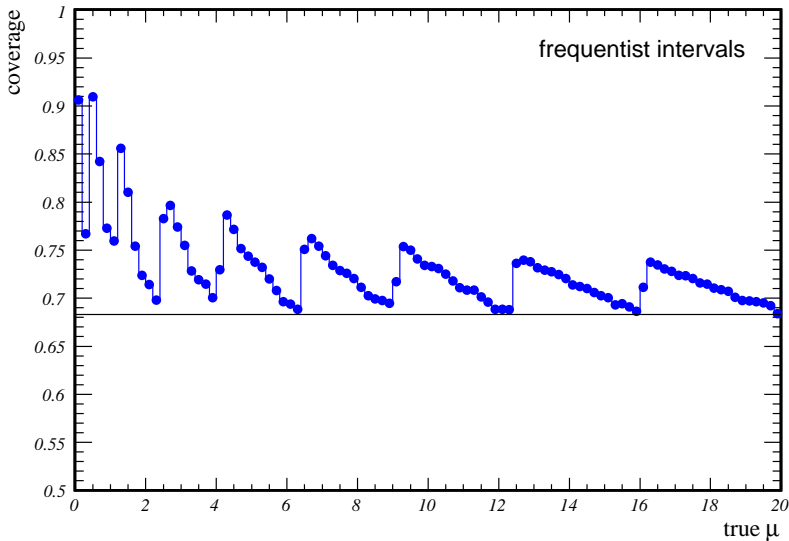


68.3% confidence level intervals



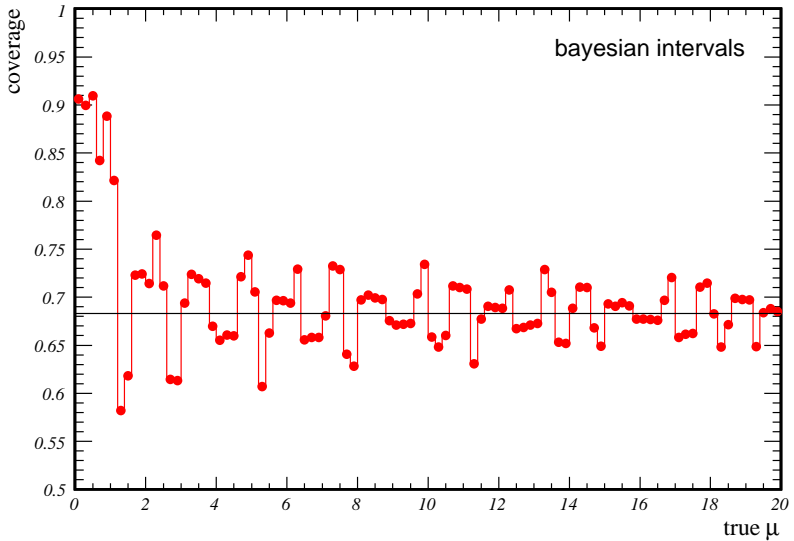


Coverage of 68.3% confidence level intervals



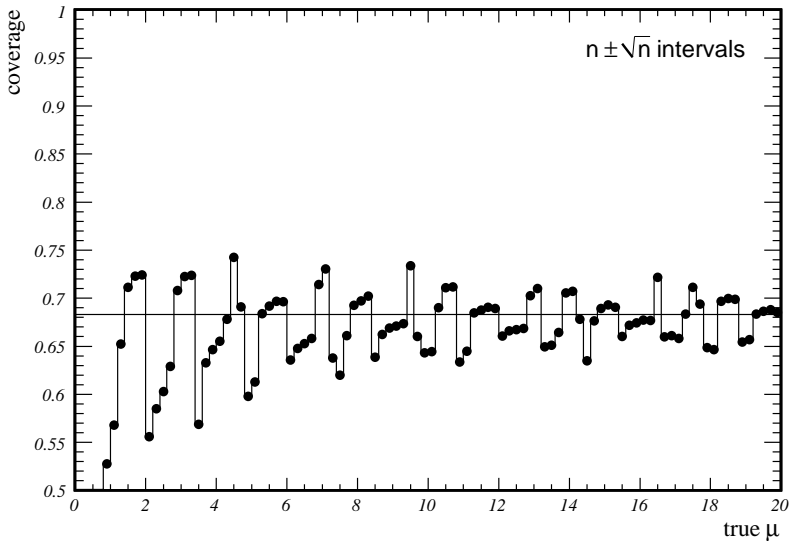


Coverage of 68.3% confidence level intervals





Coverage of 68.3% confidence level intervals





- bayesians makes statements about the theory
 - “The true value μ is with 90% probability inside the 90% confidence level interval”
 - the conclusion depends on the assumed prior
- frequentists makes statements about the data
 - “90% of the 90% confidence level intervals are expected to contain the true value μ ”
 - these confidence level intervals have “coverage”
 - for continuous PDFs exact coverage can be obtained
 - discrete probabilities are chosen to have over-coverage
- bayesians & frequentists base CL-intervals on the likelihood function
- confidence level intervals from maximum likelihood or least squares fits based on $\Delta\chi^2$ or $\Delta \ln L$ are exact only for gaussian PDFs. In most cases they don't have coverage.
- treating confidence level intervals like variances is questionable



3. LEAST SQUARES & MAXIMUM LIKELIHOOD



→ *extract physics parameters from a set of measurements*

❖ properties which are assumed to be satisfied:

- ▣ individual measurements fluctuate with known variance
- ▣ individual measurements are unbiased

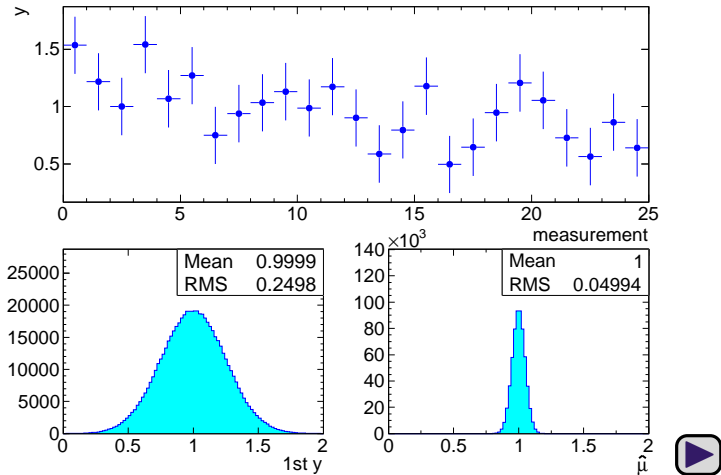
→ *measurements of the same physical quantity*

▣ scenario

- n measurements y_i with $i = 1, 2, \dots, n$
- all measurements fluctuate around an unknown true value μ
- all measurements have the standard deviation σ_i

- ▣ each measurement is an estimate for μ with uncertainty σ_i
- ▣ task: combine the measurements for a better estimate of μ

→ try the arithmetic average $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$



- big improvements if all variances are the same
- less/no improvement w.r.t. best measurement for different variances



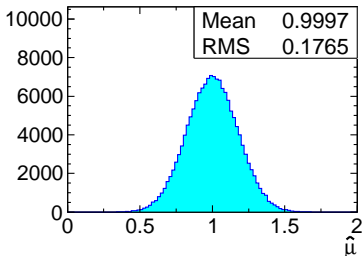
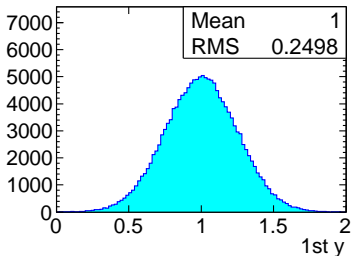
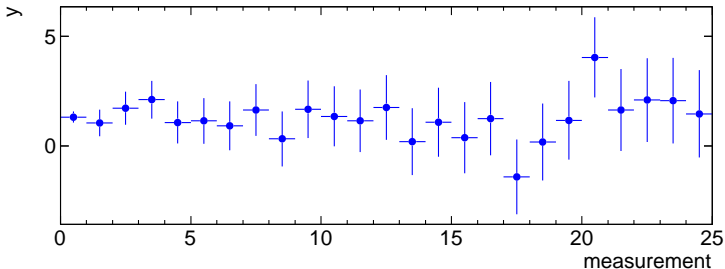
→ *modification of the arithmetic average*

$$\hat{\mu} = \sum_{i=1}^n w_i y_i \quad \text{with} \quad \sum_{i=1}^n w_i = 1$$

- consistent results for arbitrary weights: $\hat{\mu} = \mu$ if $y_i = \mu$
- try to find weights which minimize the variance of $\hat{\mu}$

$$\sigma^2(\hat{\mu}) = \sum_{i=1}^n w_i^2 \sigma_i^2 \stackrel{!}{=} \min$$

- constrained minimization problem
- minimum for $w_i \propto 1/\sigma_i^2$
- recovers unweighted average if all σ_i are the same





→ use case: straight line fit

Consider uncorrelated measurements y_i , $i = 1, \dots, n$ with known variances σ_i^2 , recorded for certain values x_i of a control parameter x . The expectation value of the measurements is $\langle y_i \rangle = a_0 + a_1 x_i$, where the parameters a_0 and a_1 are not known.

→ wanted: a method to find estimates \hat{a}_0 and \hat{a}_1 for a_0 and a_1

❖ discussion

- ❑ control parameters x_i are known
- ❑ the measurements y_i are unbiased
- ❑ variances σ_i^2 are known
- ❑ exact shape of PDFs describing the fluctuations of the y_i is irrelevant
 - any PDF with variance σ_i^2 would do
 - different measurements can fluctuate with different PDFs



→ the case of two measurements

$$\langle y_1 \rangle = a_0 + a_1 x_1 \quad \text{and} \quad y_1 = \langle y_1 \rangle + r_1$$

$$\langle y_2 \rangle = a_0 + a_1 x_2 \quad \text{and} \quad y_2 = \langle y_2 \rangle + r_2$$

- system of linear equations relating $\langle y_i \rangle$ and x_i
- measurements y_i have random deviation r_i from $\langle y_i \rangle$
- unbiasedness of y_i implies $\langle r_i \rangle = 0$
- estimate a_0 and a_1 by assuming $r_i = 0$, i.e. make the ansatz:

$$y_1 = \hat{a}_0 + \hat{a}_1 x_1$$

$$y_2 = \hat{a}_0 + \hat{a}_1 x_2$$

❖ result:

$$\hat{a}_0 = y_1 - \hat{a}_1 x_1 = \frac{x_2}{x_2 - x_1} y_1 - \frac{x_1}{x_2 - x_1} y_2$$

$$\hat{a}_1 = \frac{y_2 - y_1}{x_2 - x_1} = -\frac{1}{x_2 - x_1} y_1 + \frac{1}{x_2 - x_1} y_2$$



→ *does the estimate make sense?*

- parameter estimates are linear combinations of the measurements
- parameter estimates are random variables
- parameter estimates fluctuate with the measurements
- check the expectation values . . .

$$\langle \hat{a}_0 \rangle = \left\langle \frac{1}{x_2 - x_1} (x_2 y_1 - x_1 y_2) \right\rangle = \frac{1}{x_2 - x_1} (x_2 \langle y_1 \rangle - x_1 \langle y_2 \rangle) = a_0$$

$$\langle \hat{a}_1 \rangle = \left\langle \frac{1}{x_2 - x_1} (-y_1 + y_2) \right\rangle = \frac{1}{x_2 - x_1} (-\langle y_1 \rangle + \langle y_2 \rangle) = a_1$$

❖ **conclusion:**

- the estimates for the unknown parameters are unbiased
- the parameter errors can be determined by error propagation

yes, the parameter estimates make sense!



→ the case of $n > 2$ measurements

Take the lessons learnt from the case $n = 2$ and try to estimate the unknown parameters by a linear combination of the measurements.

$$\hat{a}_0 = \sum_{i=1}^n p_i y_i \quad \text{and} \quad \hat{a}_1 = \sum_{i=1}^n q_i y_i$$

- this is a convenient ansatz, not derived from any “first principles”
- it is not the only possible generalization of the case $n = 2$
- nor will it give the best possible estimates for a_0 and a_1
- but it is simple and robust, requiring only minimal input
- and turns out to be surprisingly powerful . . .

→ determine parameters p_i and q_i . . .



→ exploit the freedom of the linear ansatz to . .

- make sure that the estimates are unbiased
- and that the estimates are as accurate as possible

❖ condition for unbiased estimates:

$$\langle \hat{a}_0 \rangle = \sum_{i=1}^n p_i \langle y_i \rangle = \sum_{i=1}^n p_i (a_0 + a_1 x_i) = a_0 \sum_{i=1}^n p_i + a_1 \sum_{i=1}^n p_i x_i \stackrel{!}{=} a_0$$

$$\langle \hat{a}_1 \rangle = \sum_{i=1}^n q_i \langle y_i \rangle = \sum_{i=1}^n q_i (a_0 + a_1 x_i) = a_0 \sum_{i=1}^n q_i + a_1 \sum_{i=1}^n q_i x_i \stackrel{!}{=} a_1$$

one obtains 4 conditions:

$$\sum_{i=1}^n p_i = 1 \quad \sum_{i=1}^n q_i = 0 \quad \sum_{i=1}^n p_i x_i = 0 \quad \sum_{i=1}^n q_i x_i = 1$$



- only 4 constraints for $2n$ parameters
- easy to satisfy both for p_i and q_i
 - start from a set of random numbers e.g. for p_i
 - subtract a constant such that the “0-constraint” is satisfied
 - scale the numbers such that the “1-constraint” is satisfied
- additional criterion needed to fix the coefficients
- require minimal variance for the parameter estimates
 - constrained minimization problem

❖ variance of parameter estimates from error propagation:

$$\sigma^2(\hat{a}_0) = \sum_{i=1}^n \left(\frac{\partial \hat{a}_0}{\partial y_i} \right)^2 \sigma_i^2 = \sum_{i=1}^n p_i^2 \sigma_i^2 \quad \text{and} \quad \sigma^2(\hat{a}_1) = \sum_{i=1}^n q_i^2 \sigma_i^2$$

→ constrained minimization



→ minimization using Lagrange multipliers for the constraints

$$\sum_{i=1}^n p_i^2 \sigma_i^2 + 2\alpha_0 \left(1 - \sum_{i=1}^n p_i \right) + 2\beta_0 \left(- \sum_{i=1}^n p_i x_i \right) \stackrel{!}{=} \min$$

requiring zero derivatives with respect to p_i then yields:

$$2p_i \sigma_i^2 - 2\alpha_0 - 2\beta_0 x_i = 0 \quad \rightarrow \quad p_i = \frac{1}{\sigma_i^2} (\alpha_0 + \beta_0 x_i)$$

α_0 and β_0 follow from the constraint to have unbiased estimates:

$$\sum_{i=1}^n p_i = \alpha_0 \sum_{i=1}^n \frac{1}{\sigma_i^2} + \beta_0 \sum_{i=1}^n \frac{x_i}{\sigma_i^2} = \alpha_0 S_1 + \beta_0 S_x = 1$$

$$\sum_{i=1}^n p_i x_i = \alpha_0 \sum_{i=1}^n \frac{x_i}{\sigma_i^2} + \beta_0 \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} = \alpha_0 S_x + \beta_0 S_{xx} = 0$$



→ minimization using Lagrange multipliers for the constraints

$$\sum_{i=1}^n q_i^2 \sigma_i^2 + 2\alpha_1 \left(-\sum_{i=1}^n q_i \right) + 2\beta_1 \left(1 - \sum_{i=1}^n q_i x_i \right) \stackrel{!}{=} \min$$

requiring zero derivatives with respect to q_i then yields:

$$2q_i \sigma_i^2 - 2\alpha_1 - 2\beta_1 x_i = 0 \quad \rightarrow \quad q_i = \frac{1}{\sigma_i^2} (\alpha_1 + \beta_1 x_i)$$

α_1 and β_1 follow from the constraint to have unbiased estimates:

$$\sum_{i=1}^n q_i = \alpha_1 \sum_{i=1}^n \frac{1}{\sigma_i^2} + \beta_1 \sum_{i=1}^n \frac{x_i}{\sigma_i^2} = \alpha_1 S_1 + \beta_1 S_x = 0$$

$$\sum_{i=1}^n q_i x_i = \alpha_1 \sum_{i=1}^n \frac{x_i}{\sigma_i^2} + \beta_1 \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2} = \alpha_1 S_x + \beta_1 S_{xx} = 1$$



Solving the linear equations for the Lagrange parameters $\alpha_{\{0,1\}}$ and $\beta_{\{0,1\}}$

$$\begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 & \alpha_1 \\ \beta_0 & \beta_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and substituting the results into p_i, q_i , with $D = S_1 S_{xx} - S_x^2$, yields

$$p_i = \frac{1}{\sigma_i^2}(\alpha_0 + \beta_0 x_i) = \frac{1}{D} \left(S_{xx} \frac{1}{\sigma_i^2} - S_x \frac{x_i}{\sigma_i^2} \right)$$
$$q_i = \frac{1}{\sigma_i^2}(\alpha_1 + \beta_1 x_i) = \frac{1}{D} \left(-S_x \frac{1}{\sigma_i^2} + S_1 \frac{x_i}{\sigma_i^2} \right)$$

and thus

$$\hat{a}_0 = \frac{1}{D}(S_{xx} S_y - S_x S_{xy}) \quad \text{and} \quad \hat{a}_1 = \frac{1}{D}(S_1 S_{xy} - S_x S_y)$$

where

$$S_{\{1,x,xx,y,xy\}} = \sum_{i=1}^n \frac{\{1, x_i, x_i x_i, y_i, x_i y_i\}}{\sigma_i^2}.$$

→ linear error propagation

$$C_{kl}(\hat{a}) = \sum_{i=1}^n \frac{\partial \hat{a}_k}{\partial y_i} \frac{\partial \hat{a}_l}{\partial y_i} \sigma_i^2$$

yields

$$C_{00}(\hat{a}) = \sum_{i=1}^n p_i^2 \sigma_i^2 = \frac{S_1}{D^2} (S_{xx} S_1 - S_x^2) = \frac{S_1}{D}$$

$$C_{11}(\hat{a}) = \sum_{i=1}^n q_i^2 \sigma_i^2 = \frac{S_{xx}}{D^2} (S_{xx} S_1 - S_x^2) = \frac{S_{xx}}{D}$$

$$C_{01}(\hat{a}) = \sum_{i=1}^n p_i q_i \sigma_i^2 = \frac{-S_x}{D^2} (S_{xx} S_1 - S_x^2) = \frac{-S_x}{D}$$

... the well known textbook formulae for straight line fits.



→ re-write the solution derived before. . .

$$\hat{a}_0 = \frac{1}{D}(S_{xx}S_y - S_xS_{xy}) \quad \text{and} \quad \hat{a}_1 = \frac{1}{D}(S_1S_{xy} - S_xS_y)$$

to make the structure more evident:

$$\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \frac{1}{D} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S_1 \end{pmatrix} \cdot \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix} \rightarrow \begin{pmatrix} S_1 & S_x \\ S_x & S_{xx} \end{pmatrix} \cdot \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

or

$$S_1 \hat{a}_0 + S_x \hat{a}_1 - S_y = 0$$

$$S_x \hat{a}_0 + S_{xx} \hat{a}_1 - S_{xy} = 0$$

i.e. two equations which define the **best fit parameters** as the zero of a **two-dimensional function**. Now exploit the fact that it's always possible to interpret the zero of a function as a stationary point (e.g. minimum) of its primitive.

p.t.o. →



→ introducing $F(a_0, a_1)$ such that

$$\left. \frac{\partial F}{\partial a_0} \right|_{\{a_0, a_1\}=\{\hat{a}_0, \hat{a}_1\}} = 0 \quad \text{and} \quad \left. \frac{\partial F}{\partial a_1} \right|_{\{a_0, a_1\}=\{\hat{a}_0, \hat{a}_1\}} = 0$$

it follows (from dimensional considerations)

$$\frac{\partial F}{\partial a_0} = S_1 a_0 + S_x a_1 - S_y \quad \text{and} \quad \frac{\partial F}{\partial a_1} = S_x a_0 + S_{xx} a_1 - S_{xy} .$$

Integration of the first equation yields

$$F = \frac{1}{2} S_1 a_0^2 + S_x a_0 a_1 - a_0 S_y + g(a_1)$$

where $g(a_1)$ does not depend on a_0 . Taking the derivative with respect to a_1 and comparing with the known derivative determines $g'(a_1)$:

$$\frac{\partial F}{\partial a_1} = S_x a_0 + g'(a_1) = S_x a_0 + S_{xx} a_1 - S_{xy}$$

p.t.o. →



It follows

$$g'(a_1) = S_{xx} a_1 - S_{xy} \quad \text{and thus} \quad g(a_1) = \frac{1}{2} S_{xx} a_1^2 - S_{xy} a_1 + \frac{C}{2}$$

with an arbitrary constant C . Asking $F_{\min} = 0$ yields $C = \sum y_i^2 / \sigma_i^2$ and

$$\begin{aligned} 2F &= S_1 a_0^2 + S_{xx} a_1^2 + 2S_x a_0 a_1 - 2S_y a_0 - 2S_{xy} a_1 + C \\ &= \sum_{i=1}^n \frac{1}{\sigma_i^2} (a_0^2 + a_1^2 x_i^2 + 2a_0 a_1 x_i - 2a_0 y_i + 2a_1 x_i y_i + y_i^2) \\ &= \sum_{i=1}^n \frac{(y_i - a_0 - a_1 x_i)^2}{\sigma_i^2}, \end{aligned}$$

and setting $2F = \chi^2$, the cost-function becomes

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - f_i(a_0, a_1))^2}{\sigma_i^2} \quad \text{with} \quad f_i(a_0, a_1) = a_0 + a_1 x_i.$$



- the best parameter estimates minimize the **distance** between data and model, measured **in units of standard deviations**
- the derivation was for uncorrelated data points y_i
- general expression, also for correlated data, using $1/\sigma_i^2 = C_{ii}^{-1}$:

$$\chi^2 = \sum_{i,j=1}^n (y_i - f_i(a_0, a_1)) (y_j - f_j(a_0, a_1)) C_{ij}^{-1}$$

or in matrix notation

$$\chi^2 = \vec{r}^T C^{-1} \vec{r} \quad \text{with} \quad \vec{r} = \vec{y} - \vec{f}(a_0, a_1)$$

→ *Invariance under linear transformations M :*

$$\vec{r}' = M \vec{r} \quad , \quad C' = M C M^T \quad , \quad C'^{-1} = (M^T)^{-1} C^{-1} M^{-1}$$

$$\text{and thus} \quad (\chi^2)' = \chi^2$$



→ (average) measurements are linear functions of parameters \vec{a}

$$\begin{aligned}\chi^2 &= (\vec{y} - M\vec{a})^T C^{-1} (\vec{y} - M\vec{a}) \\ &= \vec{y}^T C^{-1} \vec{y} - 2\vec{a}^T [M^T C^{-1} \vec{y}] + \vec{a}^T [M^T C^{-1} M] \vec{a}\end{aligned}$$

minimization:

$$\frac{\partial \chi^2}{\partial \vec{a}} = -2 [M^T C^{-1} \vec{y}] + 2 [M^T C^{-1} M] \vec{a} = 0$$

result: the best fit parameters are linear functions of the measurements

$$\vec{a} = Q \vec{y} \quad \text{with} \quad Q = [M^T C^{-1} M]^{-1} M^T C^{-1}$$

with covariance matrix

$$C(a) = Q C Q^T = [M^T C^{-1} M]^{-1} = \left(\frac{1}{2} \frac{\partial \chi^2}{\partial \vec{a}^2} \right)^{-1}$$



- unbiased parameter estimates (for any constant matrix C^{-1})

$$\langle \vec{y} \rangle = M \vec{a}_{\text{true}} \rightarrow \langle \vec{a} \rangle = [M^T C^{-1} M]^{-1} M^T C^{-1} \langle \vec{y} \rangle = \vec{a}_{\text{true}}$$

- minimum χ^2 value

$$\begin{aligned}\chi_{\min}^2 &= \vec{y}^T C^{-1} \vec{y} - \vec{a}^T [M^T C^{-1} \vec{y}] \\ &= \vec{y}^T C^{-1} \vec{y} - \vec{a}^T [M^T C^{-1} M] \vec{a} \\ &= \text{Tr} (C_y^{-1} \vec{y} \vec{y}^T - C_a^{-1} \vec{a} \vec{a}^T)\end{aligned}$$

- expectation value χ_{\min}^2 , using $C_x = \langle \vec{x} \vec{x}^T \rangle - \langle \vec{x} \rangle \langle \vec{x} \rangle^T$

$$\begin{aligned}\langle \chi_{\min}^2 \rangle &= \text{Tr} \left(C_y^{-1} (C_y + \langle \vec{y} \rangle \langle \vec{y} \rangle^T) - C_a^{-1} (C_a + \langle \vec{a} \rangle \langle \vec{a} \rangle^T) \right) \\ &= n_y - n_a + \text{Tr} \left(C_y^{-1} \langle \vec{y} \rangle \langle \vec{y} \rangle^T - C_a^{-1} \langle \vec{a} \rangle \langle \vec{a} \rangle^T \right) = n_y - n_a\end{aligned}$$

The last step follows from $C_a^{-1} = M^T C_y^{-1} M$ and $M \langle \vec{a} \rangle = \langle \vec{y} \rangle$.



- formulation via the cost function . . .
 - derived for linear models and explains the name “least squares”
 - easily generalizes to multi-dimensional and non-linear problems
- least squares are a distribution-free way for parameter estimates
 - requires only data and covariance matrix of the data
 - weight matrix C^{-1} must be fixed
 - approximately gaussian errors due to the central limit theorem
- for linear models
 - unbiased estimates of the true parameters
 - parameter estimates are linear combinations of the measurements
- when using the inverse of the covariance matrix as weight matrix
 - linear estimates with minimal variance
 - independent of the shape of the PDF of the fluctuations
 - $\langle \chi_{\min}^2 \rangle = N_{\text{data}} - N_{\text{par}} \equiv N_{\text{ndf}}$
 - can be used to judge goodness of fit or estimate size of variances

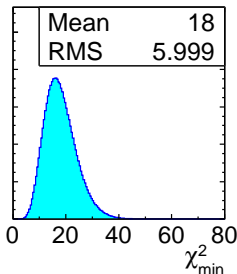
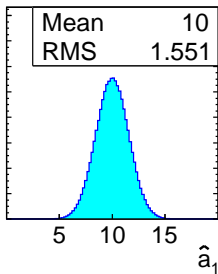
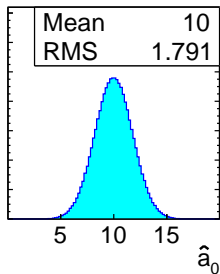
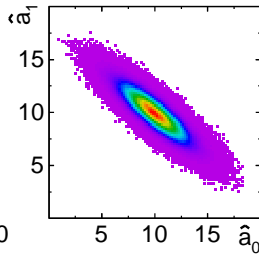
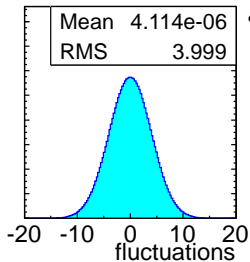
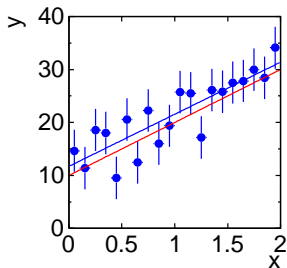


→ *straight line fit*: $y = a_0 + a_1 x$

- expectation values of measurements $y(x)$: $\langle y \rangle = 10 + 10x$
- take 20 equidistant points in the range $0 < x < 2$
- measurements fluctuate with **rms**= 4 around the expectation value
 - gaussian distribution
 - exponential distribution
 - uniform distribution
- same covariance matrix and $\langle \chi^2_{\min} \rangle = 18$ in all cases

$$C(a) \approx \begin{pmatrix} 3.206 & -2.406 \\ -2.406 & 2.406 \end{pmatrix} \quad \begin{array}{l} \sigma(a_0) \approx 1.7905 \\ \sigma(a_1) \approx 1.5511 \end{array} \quad \rho \approx -0.8663$$

- study also poisson distributed measurements. . .
 - fit with **correct** standard deviations: $\sqrt{\langle y \rangle}$
 - fit with **estimated** standard deviations: \sqrt{y}





→ exploring the least squares approach

Given: measurements y_i with known variances σ_i^2 , a parametric model $f_i(a)$, and positive weights $w_i > 0$. Wanted: parameter estimates \hat{a} .

$$\text{Ansatz: } S^2(a) = \sum_i w_i (y_i - f_i(a))^2 \stackrel{!}{=} \min$$

❖ reminder:

- ❑ the best fit \hat{a} makes the model get “as close as possible” to the data
- ❑ the weights allow to (de)emphasize selected points
- ❑ a priori arbitrary weights are allowed
- ❑ for independent measurements the optimal weights are $w_i = 1/\sigma_i^2$

study an analytically solvable problem →



→ problem:

n poisson distributed values y_i , $i = 1, \dots, n$, such as measurements from a counting experiments, which are distributed according to the discrete probability distribution

$$p_n(\mu) = e^{-\mu} \frac{\mu^n}{n!}$$

with, for example, actual values

$$y_i = \{2, 2, 5, 2, 3, 3, 1, 3, 3, 2, 3, 2, 10, 2, 3, 1, 2, 6, 4, 3 \dots\}$$

→ solution:

- ☐ least squares fit of a constant
- ☐ study different terms for the variance in the χ^2 function



→ the ideal χ^2 function

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - c)^2}{\mu} \quad \rightarrow \quad \hat{c} = \frac{1}{n} \sum_{i=1}^n y_i \pm \sqrt{\frac{\mu}{n}}$$

exact properties:

$$\langle \hat{c} \rangle = \mu \quad \text{and} \quad \frac{\langle \chi_{\min}^2 \rangle}{n-1} = 1$$

remarks:

- parameter estimate by arithmetic average
- ansatz questionable since μ is not known, but. . .
- \hat{c} does not depend on μ , only its uncertainty and χ^2
 - determine \hat{c} and use $\mu = \hat{c}$ in the χ^2 function

$$\text{result:} \quad \langle \hat{c} \rangle = \mu \quad \text{and} \quad \frac{\langle \chi_{\min}^2 \rangle}{n-1} \stackrel{n \rightarrow \infty}{=} 1$$



→ the RooFit default

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - c)^2}{c} \quad \rightarrow \quad \hat{c} = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2} \pm \sqrt{\frac{\hat{c}}{n}}$$

asymptotic properties:

$$\langle \hat{c} \rangle \stackrel{n \rightarrow \infty}{=} \sqrt{\mu(\mu + 1)} \quad \text{and} \quad \frac{\langle \chi_{\min}^2 \rangle}{n - 1} \stackrel{n \rightarrow \infty}{=} 2(\sqrt{\mu(\mu + 1)} - \mu)$$

remarks:

- parameter estimate by quadratic average
- non-linear fit model (non-parabolic cost function)
- biased parameter estimate
- biased χ_{\min}^2 values – p -values are of limited use



→ *alternative* RooFit *setting*

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - c)^2}{y_i} \quad \rightarrow \quad \hat{c} = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{y_i} \right)^{-1} \pm \sqrt{\frac{\hat{c}}{n}}$$

asymptotic properties:

$$\langle \hat{c} \rangle \stackrel{n \rightarrow \infty}{=} \frac{1}{\langle 1/y \rangle} \quad \text{and} \quad \frac{\langle \chi_{\min}^2 \rangle}{n-1} \stackrel{n \rightarrow \infty}{=} \frac{\mu}{1 - e^{-\mu}} - \frac{1}{\langle 1/y \rangle},$$

remarks:

- parameter estimate by **harmonic average**
- necessity to discard values $y_i = 0$
- linear model
- biased parameter estimate
- biased χ_{\min}^2 values – p -values are of limited use



→ avoid discarding zero bins $z_i = y_i + 1$

$$\chi^2 = \sum_{i=1}^n \frac{(z_i - c)^2}{z_i} \quad \rightarrow \quad \hat{c} = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{y_i + 1} \right)^{-1} \pm \sqrt{\frac{\hat{c}}{n}}$$

asymptotic properties:

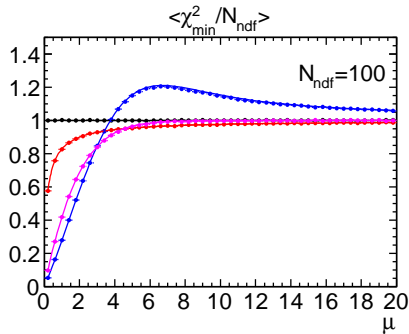
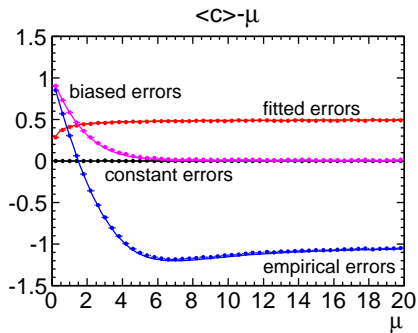
$$\langle c \rangle \stackrel{n \rightarrow \infty}{=} \frac{\mu}{1 - e^{-\mu}} \quad \text{and} \quad \frac{\langle \chi_{\min}^2 \rangle}{n - 1} \stackrel{n \rightarrow \infty}{=} 1 - \frac{\mu}{e^{\mu} - 1}.$$

remarks:

- parameter estimate by **harmonic average**
- allows to include also values $y_i = 0$
- linear model
- asymptotically unbiased parameter estimate
- asymptotically unbiased χ_{\min}^2 values



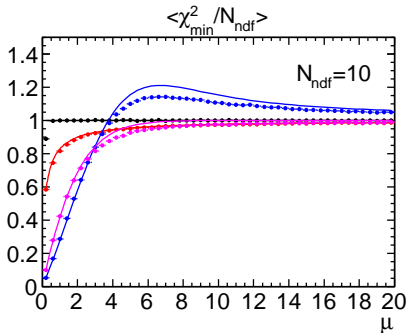
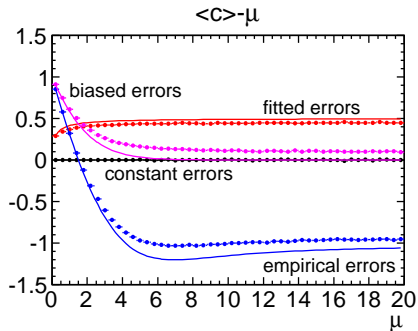
→ expectation values vs μ for $n = 101$



- data points: simulations for $n = 101$ data points
- curves: asymptotic expectations



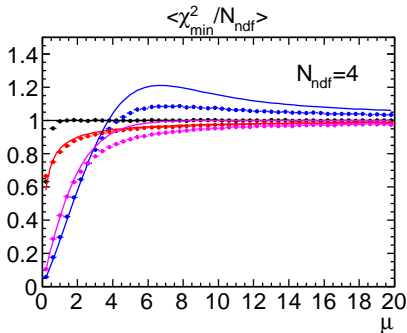
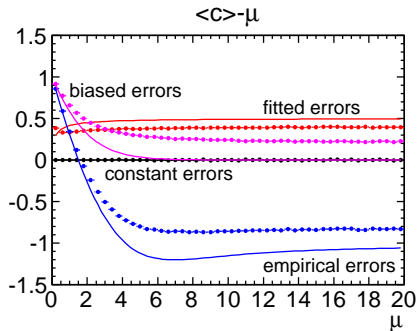
→ expectation values vs μ for $n = 11$



- data points: simulations for $n = 11$ data points
- curves: asymptotic expectations



→ expectation values vs μ for $n = 5$



- data points: simulations for $n = 5$ data points
- curves: asymptotic expectations



→ introductory remarks

- common wisdom: least squares fits need. . .
 - gaussian fluctuations
 - sufficiently large event counts for poisson distributed data
- in the derivation of the method none of the above entered
 - only proper variance estimates were assumed
 - the variances are treated as constants in the χ^2 minimization
 - the variance estimates should not be correlated to the data

case study, keeping an eye on those points when doing fits →



→ determination of the lifetime of an unstable particle

- lifetime distribution

$$\frac{dn}{dt} = \frac{1}{\mu} e^{-t/\mu} \quad \text{with} \quad \mu = 1 \text{ ns}$$

- MC study of test experiments with fixed number N of decays

- histogram representation of the measurement

- 100 bins for $0 < t < 10 \text{ ns}$

- optimal parameter estimate:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N t_i \quad \text{for } \mu = 1: \quad \hat{\mu} = 1 \pm \frac{1}{\sqrt{N}}$$

- parametric model for bin contents n_i in Least Squares fit

$$f_i(\mu) = N \int_{\text{bin } i} dt \frac{dn}{dt}$$

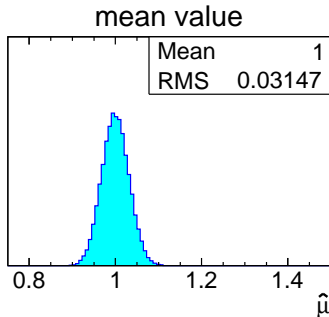
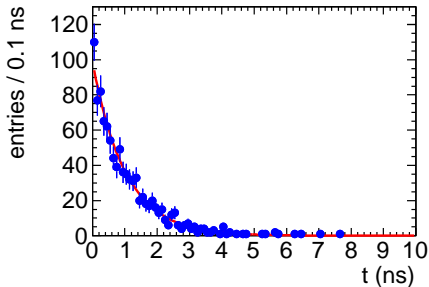


→ test different weight-assignments

- $w_i = 1$ for all bins
 - unsophisticated but hopefully robust unweighted fit
- $w_i = 1$ for all bins with non-zero entries
 - pretend that empty bins don't have informations
- $w_i = 1/n_i$ for all bins with non-zero entries
 - use empirical variance estimates
- $w_i = 1/f_i$ for all bins
 - naive way to use the theoretical variances
- iterative fit with $w(0) = 1$ and $w_i(m) = 1/f_i(\hat{\mu}_{m-1})$ for all bins
 - proper way to use the theoretical variances
 - implements that variances must be fixed in minimization
 - weak correlation between variance estimates and data
- for comparison: simple arithmetic mean of all entries



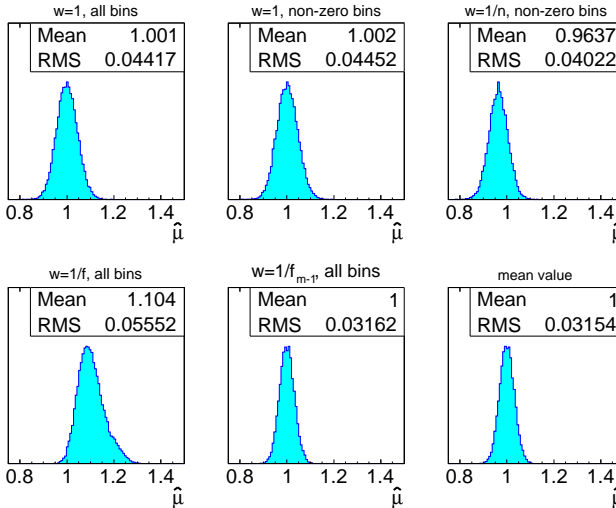
→ best fit performance for $N = 1000$ events



- check standard deviation and bias of fitted $\hat{\mu}$
 - as a function of available statistics
 - for the different choices of the weight function

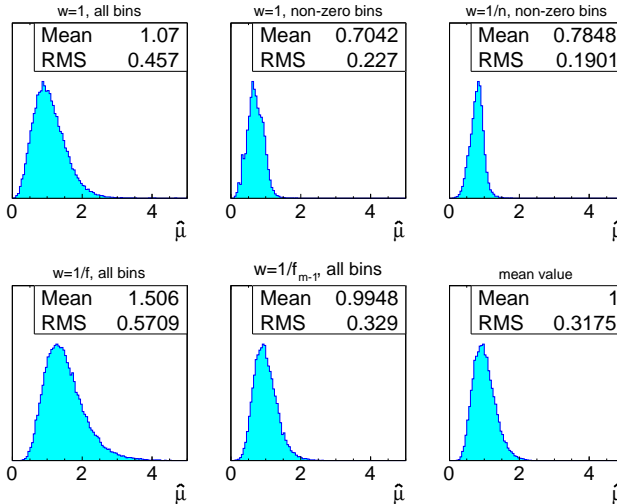


→ parameter estimates for $N = 1000$ events





→ parameter estimates for $N = 10$ events





→ *properties of different weight-assignments*

- $w_i = 1$ for all bins
 - OK, generally unbiased, but not with optimal precision
 - do not use Hessian of χ^2 function for error estimates
- $w_i = 1$ for non-zero bins
 - needless loss of information and bias at low statistics
- $w_i = 1/n_i$ for all bins with non-zero entries
 - biased – violates the least squares ansatz
- $w_i = 1/f_i$ for all bins
 - biased – violates the least squares ansatz
- iterative fit with $w(0) = 1$ and $w_i(m) = 1/f_i(\hat{\mu}_{m-1})$ for all bins
 - close to optimum (maximum likelihood fit)
 - works also at low statistics