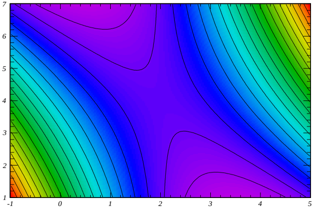
Statistische Methoden der Datenanalyse II

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*



Literature

➜ *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

1. Introduction

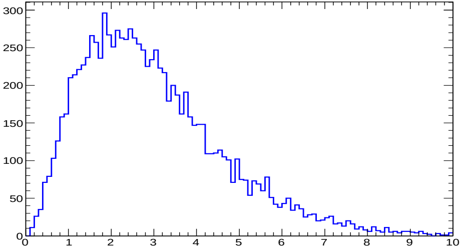
➜ *What are statistical methods?*

- 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

Example for a distribution of measurements



- “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

➜ *“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

Kolmogorov´s axioms on probability

➜ *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 $E_1 \cap E_2 = 0$

Math of probabilities follows unambiguously - interpretation is left open.

Bayes´ theorem

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

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

It follows:

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

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:

$$\begin{array}{lclclcl} p(I) & = & 0.001 & p(H) & = & 0.999 \\ p(+|I) & = & 0.980 & p(-|I) & = & 0.020 \\ p(+|H) & = & 0.030 & p(-|H) & = & 0.970 \end{array}$$

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"!

Probabilities and probability density functions

➜ *Kolmogorov's axiom for discrete sets: discrete probabilities*

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

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

➜ *continuous sets: probability density functions (PDFs)*

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

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

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

$$p(x, x + dx) = \int\limits_{x}^{x+dx} dx \; f(x) \approx f(x) dx$$

Probability distributions (i)

➜ *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*

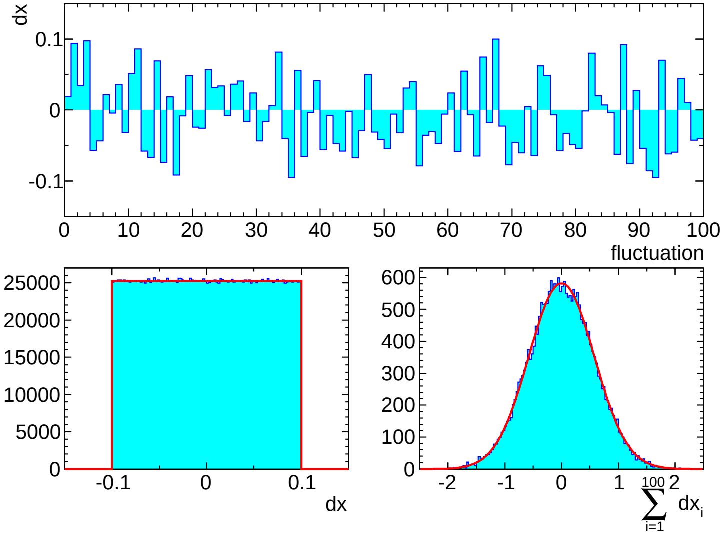
- example 1: astronomical observations
  - ➜ light rays are scattered at density variations in the atmosphere
- example 2: current over a resistor
  - ➜ current variations from thermal motion of many electrons

❖ common feature

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

➜

Simulation study



Discussion

➜ *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

Probability distributions (ii)

➜ *statistics of counting experiments*

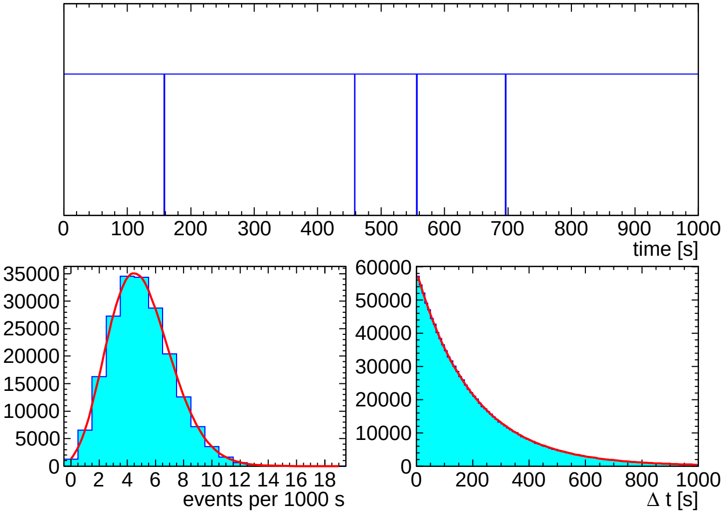
- examples:
  - ➜ decays in a radiactive source
  - ➜ cosmic muons observed at surface level on earth
  - ➜ number of soldiers in the Prussian army killed accidentally by horse kicks (Ladislaus Bortkiewicz, 1898)
- quantities of interest
  - ➜ time differences between subsequent events
  - ➜ number of 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 ➜

Simulation study

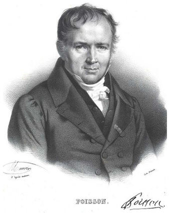


Discussion

➜ *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 Simèon Denis Poisson 1837 in “Recherches sur la probabilité des jugements en matière criminelle et en matière civile”

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

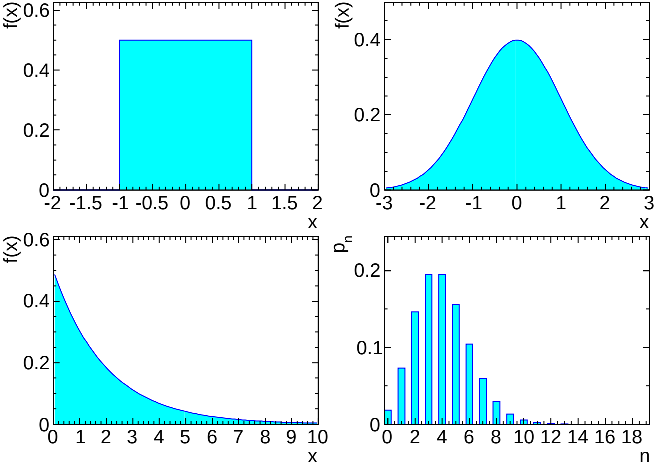


- time difference between events: exponential distribution

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

Characterizing probability distributions

➜ *consider the distributions introduced before*



❖ wanted: location and width

Characterizing the location of a distribution

➜ *“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 $\langle \cdots \rangle$ – 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

Characterizing the width of a distribution

➜ *“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 ➜

Mean value, standard deviation and variance

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:

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

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

Comparisons

➜ *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	$\sqrt{8\ln 2}\,\sigma$	$2\,\sigma$	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

Expectation values

➜ *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:

$$\begin{array}{rll} \langle x \rangle & : & \text{mean value} \\ \left\langle (x - \langle x \rangle)^2 \right\rangle & : & \text{variance} \end{array}$$

➜ *note:*

$$\sigma^2 = \int dx \; (x - \langle x \rangle)^2 \, f(x) = \int dx \; (x^2 - 2x \langle x \rangle + \langle x \rangle^2) \, f(x)$$

$$= \left(\int dx \; x^2 \, f(x) \right) - \langle x \rangle^2 = \left\langle x^2 \right\rangle - \langle x \rangle^2$$

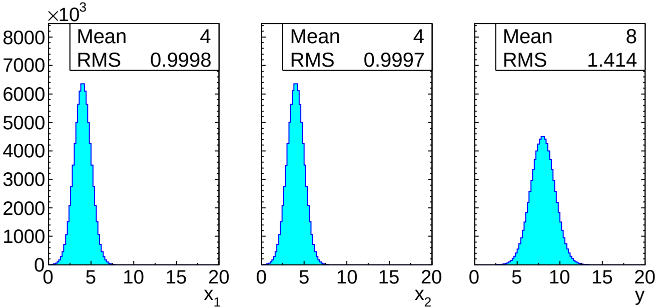
Transformation of random variables

→ *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 $y = h(x_1, x_2)$



Transformation of random variables

➜ *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:

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

Example: convolution of random variables

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

$$\langle y^k \rangle = \int dy\, y^k g(y) = \int dy\, y^k \int dx_1\, dx_2 f_1(x_1) f_2(x_2) \delta(y - x_1 - x_2)$$

$$= \int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^k$$

expectation values:

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

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

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

and thus $\langle y^2 \rangle - \langle y \rangle^2 = \left[\langle x_1^2 \rangle - \langle x_1 \rangle^2 \right] + \left[\langle x_2^2 \rangle - \langle x_2 \rangle^2 \right]$

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

Multidimensional PDFs

➜ *generalization of 1-dim PDFs*

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

❖ 2-dim PDF:

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

❖ interpretation:

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

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

❖ independence of variables:

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

The covariance between two variables

➜ *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

$$\langle xy \rangle = \int dx \int dy \, (x\, y)\, g_1(x)\, g_2(y)$$
$$= \left(\int dx\, x\, g_1(x) \right) \left(\int dy\, y\, g_2(y) \right) = \langle x \rangle \langle y \rangle$$

❖ 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 . . .

The correlation coefficient

➜ *dimensionless measures of correlation between two variables*

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

❖ properties

- $-1 \leq \rho \leq 1$
- $y = a\,x + b$ ➜ $\rho = \mathrm{sign}(a)$ ("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

The covariance matrix

➜ *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

Linear transformation of covariance matrices

➜ *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:

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

- ➜ 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. . .

Joint Committee for Guides in Metrology

WG 1 (JCGM 100:2008, Recommendation INC-1 (1980)

➜ *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)

Discussion

➜ *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!

The Bienaymé-Chebycheff-inequality

➜ *probability content in the tails of a distribution*

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

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

it follows $\quad p(w(x) \geq C) \;\leq\; \dfrac{\langle w \rangle}{C}\,.$

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

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

- 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: $\{p_1, p_2, p_3\} \approx \{0.317, 0.0555, 0.0027\}$

Error propagation - setting the stage

➜ *definitions:*

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

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

- ➜ expectation values
- ➜ uncertainties

Expectation values of transformed variables

➜ *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

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

and taking the expectation value yields:

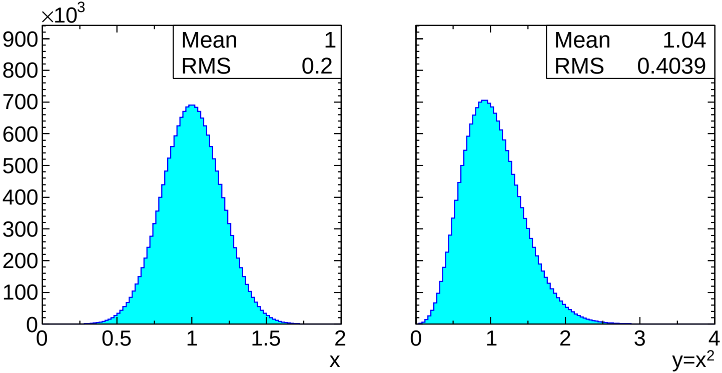
$$\langle y_k\rangle = y_k^t + \frac{1}{2}\sum_{i,j}\frac{\partial^2 g_k(\langle\vec{x}\rangle)}{\partial x_i \partial x_j} C_{ij}(x) + \ldots$$

❖ 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 $\vec{x}$

Numerical example

➜ *transformation of a gaussian distributed* $x \to y = x^n$



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

Gaussian error propagation

➜ *leading order treatment in n dimensions*

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

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

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

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 $\vec{x}$.)

Discussion

➜ *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.

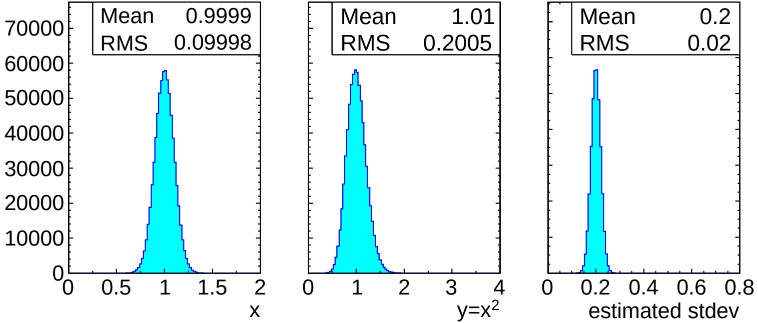
When chaining transformations one has:

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

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

Numerical studies (i)

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

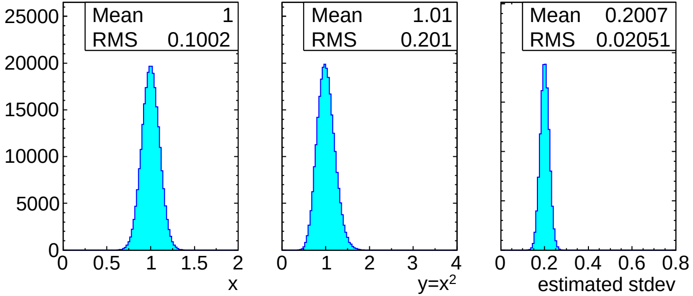


- average error estimates are OK
- actual values scatter proportional to relative errors of x

Numerical studies (ii)

➜ *error MC and exact standard deviations for* $x \to y = x^n$

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

Summary up to this point

➜ *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

Confidence level intervals

➜ *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:

$$a = 42 \pm^{8}_{3} \pm^{6}_{4} = 42 \pm^{10}_{5}$$

  - ➜ 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 ➜

Poisson measurements without background

➜ *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?

Example: $n = 2$

➜ *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

The Bayesian 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

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

Application of the Bayesian approach

➜ *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 $\mu \to \alpha\,\mu$

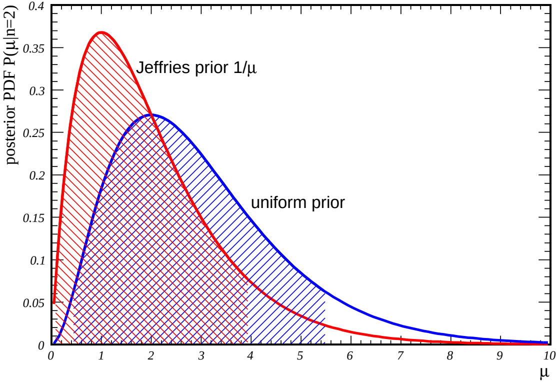
$$P(\mu|n) = \frac{e^{-\mu}\,\mu^{n+k}}{\int_0^\infty d\mu\; e^{-\mu}\,\mu^{n+k}} = e^{-\mu}\,\frac{\mu^{n+k}}{(n+k)!}$$

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

results ➜

Bayesian result

➜ *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

Discussion

- bayesian approach formalizes gain of knowledge by measurement
  - ➜ posterior of first measurement can be prior of second, etc.

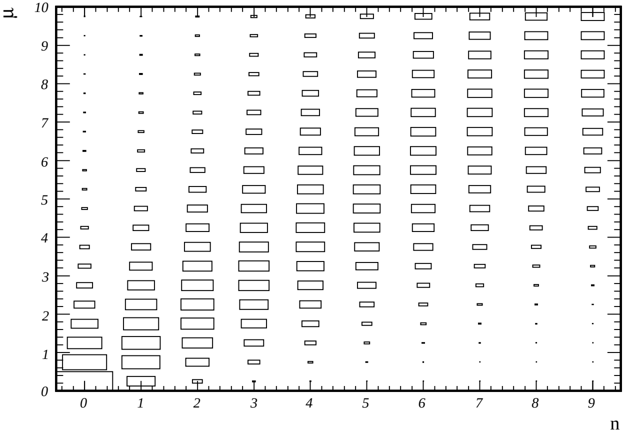
$$\begin{aligned} P(\mu|n_2,n_1) &\propto P(n_2|\mu)P(n_1|\mu)P(\mu) \\ &= P(n_2,n_1|\mu)P(\mu) \\ &= P(n_2|\mu)P_1(\mu) \quad \text{with} \quad P_1(\mu) = P(n_1|\mu)P(\mu) \end{aligned}$$

- 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 $\sqrt{\mu}$

The frequentist approach

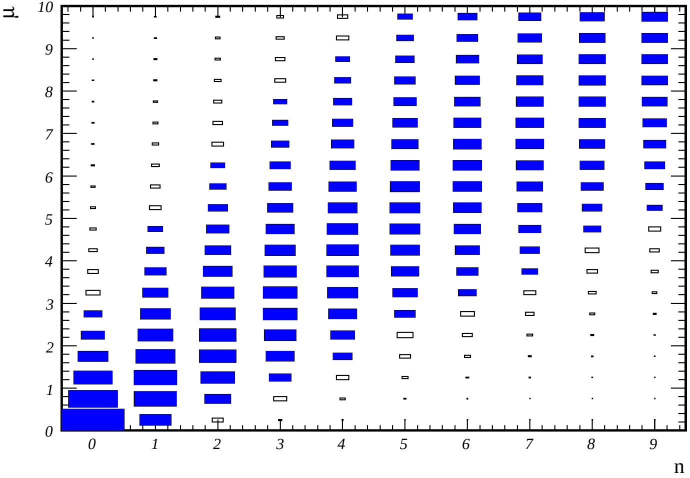
➜ *Likelihood-function-only based “Neyman construction”*

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



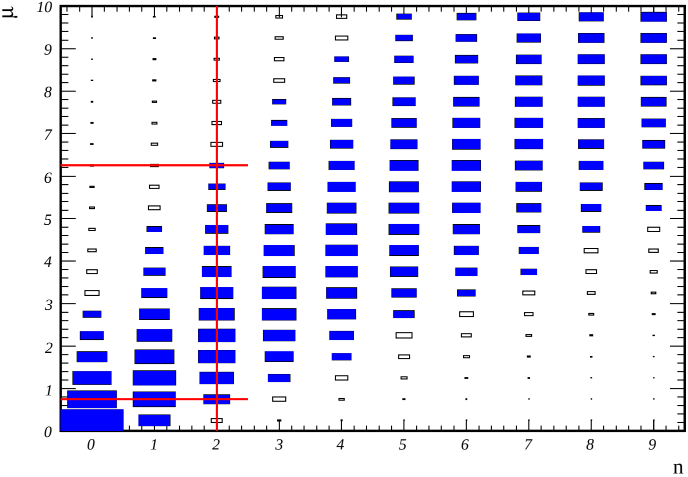
The Neyman construction (i)

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



The Neyman construction (ii)

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



Properties of the Neyman construction

- 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 $n = 1, \ldots, 7$, 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!

Discussion

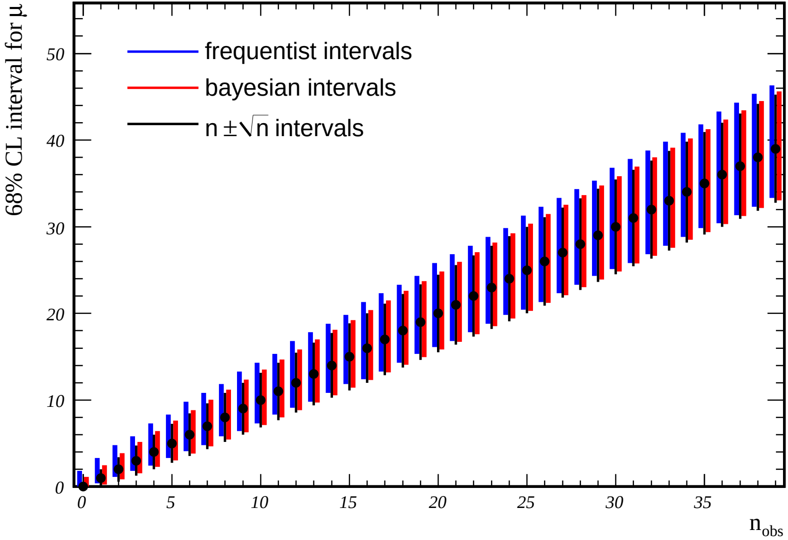
➜ *some common themes. . .*

- 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

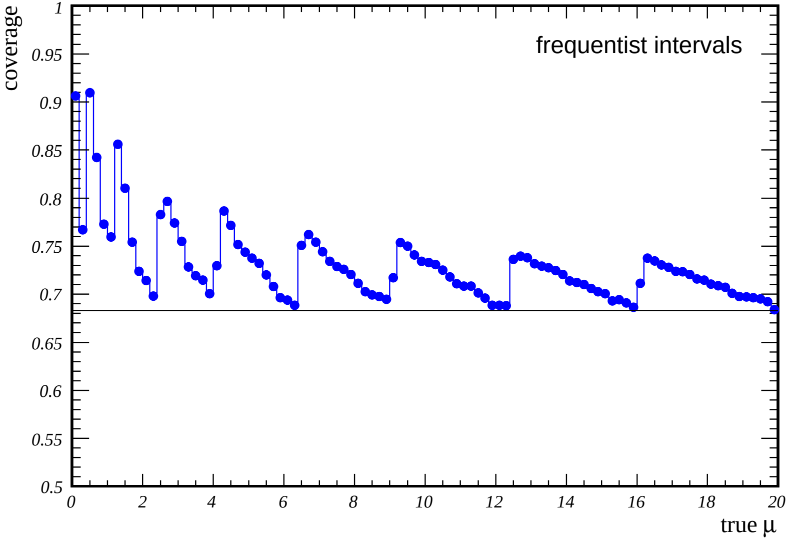
➜ *further studies. . .*

- 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 . . .

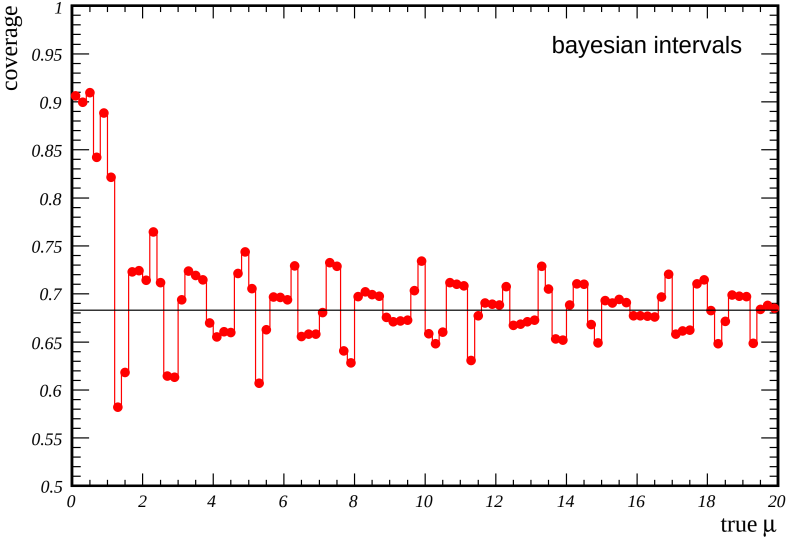
68.3% confidence level intervals



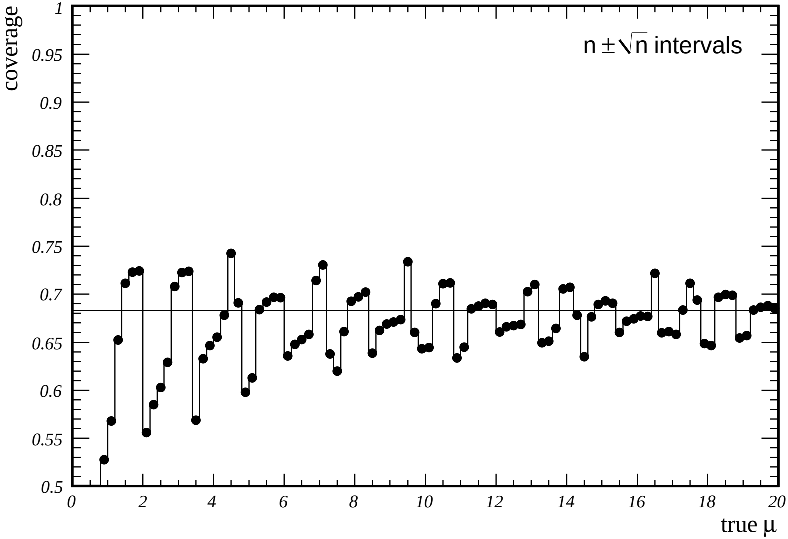
Coverage of 68.3% confidence level intervals



Coverage of 68.3% confidence level intervals



Coverage of 68.3% confidence level intervals



Concluding remarks

- 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:

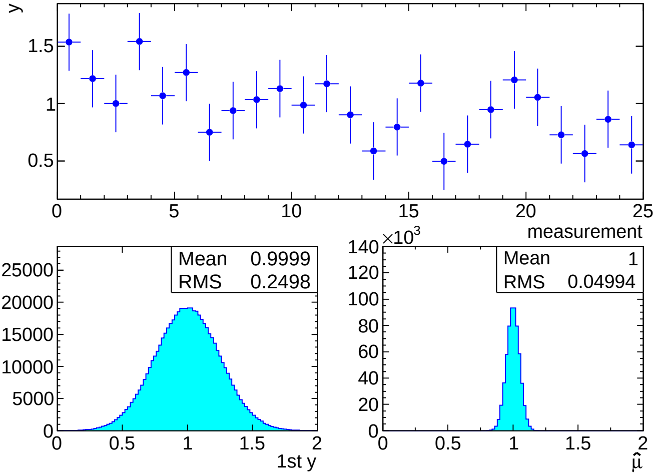
- individual measurements fluctuate with known variance
- individual measurements are unbiased

➜ *measurements of the same physical quantity*

- scenario
  - ➜ n measurements y_i with $i = 1, 2, \ldots, n$
  - ➜ 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 $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Numerical simulation



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

The weighted average

➜ *modification of the arithmetic average*

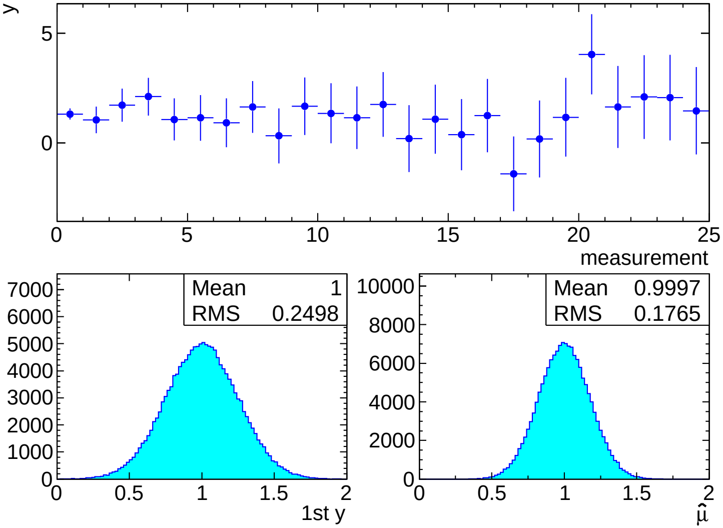
$$\hat{\mu} = \sum_{i=1}^{n} w_i y_i \qquad \text{with} \qquad \sum_{i=1}^{n} w_i = 1$$

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

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

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

Numerical simulation



The road to Least Squares

→ *use case: straight line fit*

Consider uncorrelated measurements $y_i, i = 1, \ldots, n$ with known variances σ_i^2, recorded for certain values x_i of a control paramater 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

Constructing parameter estimates

➜ *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:

$$\hat{a}_0 = y_1 - \hat{a}_1 x_1 = \frac{x_2}{x_2 - x_1}\, y_1 - \frac{x_1}{x_2 - x_1}\, y_2$$
$$\hat{a}_1 = \frac{y_2 - y_1}{x_2 - x_1} = -\frac{1}{x_2 - x_1}\, y_1 + \frac{1}{x_2 - x_1}\, y_2$$

Discussion

➜ *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!

BLUE (Best Linear Unbiased Estimator)

➜ *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 . . .

Optimizing the parameter estimates

➜ *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:

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

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

one obtains 4 conditions:

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

Discussion

- 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

Fixing the coefficients p_i

➜ *minimization using Lagrange multipliers for the constraints*

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

requiring zero derivatives with respect to p_i then yields:

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

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

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

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

Fixing the coefficients q_i

➜ *minimization using Lagrange multipliers for the constraints*

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

requiring zero derivatives with respect to q_i then yields:

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

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

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

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

Putting it all together

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

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

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

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

Covariance matrix of the parameter estimates

→ *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.

The cost function

➜ *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:

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

or

$$\begin{aligned} S_1\ \hat{a}_0 + S_x\ \ \hat{a}_1 - S_y\ \ &= 0 \\ S_x\ \hat{a}_0 + S_{xx}\ \hat{a}_1 - S_{xy} &= 0 \end{aligned}$$

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. ➜

Constructing the cost function

➜ *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. ➜

Constructing the cost function (continued)

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

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

Discussion

- 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

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

➜ *Invariance under linear transformations M :*

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

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

Linear models

➜ *(average) measurements are linear functions of parameters $\vec{a}$*

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

minimization:

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

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

$$\vec{a} = Q\;\vec{y} \qquad \text{with} \qquad Q = \left[M^T C^{-1} M\right]^{-1} M^T C^{-1}$$

with covariance matrix

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

Properties

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

$$\langle \vec{y} \rangle = M\,\vec{a}_{\mathrm{true}} \quad \rightarrow \quad \langle \vec{a} \rangle = \left[M^T C^{-1} M \right]^{-1} M^T C^{-1} \langle \vec{y} \rangle = \vec{a}_{\mathrm{true}}$$

- minimum χ^2 value

$$\begin{aligned} \chi^2_{\mathrm{min}} &= \vec{y}^{\,T} C^{-1} \vec{y} - \vec{a}^{\,T} \left[M^T C^{-1} \vec{y} \right] \\ &= \vec{y}^{\,T} C^{-1} \vec{y} - \vec{a}^{\,T} \left[M^T C^{-1} M \right] \vec{a} \\ &= \mathrm{Tr}\left(C_y^{-1} \vec{y}\vec{y}^{\,T} - C_a^{-1} \vec{a}\vec{a}^{\,T} \right) \end{aligned}$$

- expectation value χ^2_{min}, using $C_x = \left\langle \vec{x}\vec{x}^{\,T} \right\rangle - \langle \vec{x} \rangle \langle \vec{x} \rangle^T$

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

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

Property list of the least squares method

- 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
  - ➜ $\langle \chi^2_{\mathrm{min}} \rangle = N_{\mathrm{data}} - N_{\mathrm{par}} \equiv N_{\mathrm{ndf}}$
    - ➢ can be used to judge goodness of fit or estimate size of variances

Numerical example

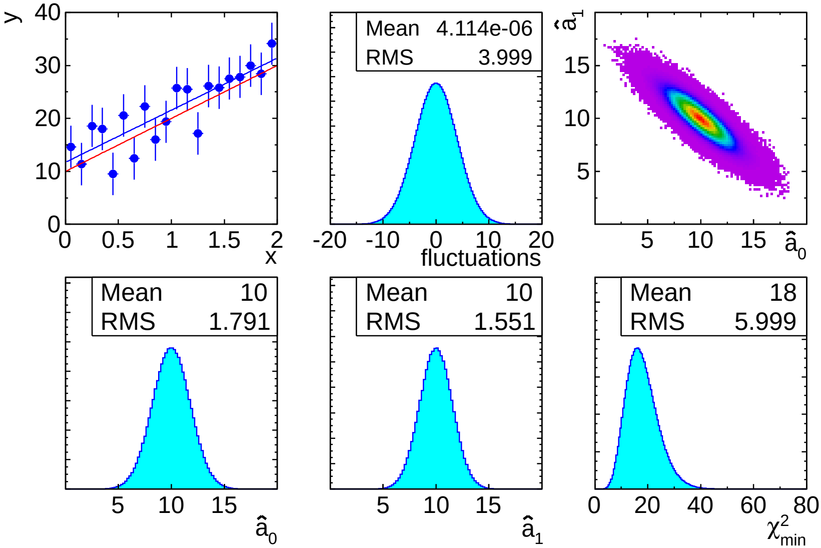
➜ *straight line fit: $y = a_0 + a_1 x$*

- expectation values of measurements $y(x)$: $\langle y \rangle = 10 + 10\,x$
- 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

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

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

Illustration



Dos and dont´s in least squares fitting

➜ *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 $\hat{a}$.

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

❖ 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 ➜

Estimating a poisson average

➜ *problem:*

n poisson distributed values y_i, $i = 1, \ldots, n$, 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

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

➜ *solution:*

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

Exact variance

➜ *the ideal χ^2 function*

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

exact properties:

$$\langle \hat{c} \rangle = \mu \qquad \text{and} \qquad \frac{\langle \chi^2_{\mathrm{min}} \rangle}{n-1} = 1$$

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

result: $\langle \hat{c} \rangle = \mu$ and $\frac{\langle \chi^2_{\mathrm{min}} \rangle}{n-1} \overset{n\to\infty}{=} 1$

Variance estimate from the fit

➜ *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:

$$\langle \hat{c} \rangle \overset{n\to\infty}{=} \sqrt{\mu(\mu+1)} \qquad \text{and} \qquad \frac{\langle \chi^2_{\min} \rangle}{n-1} \overset{n\to\infty}{=} 2\left(\sqrt{\mu(\mu+1)} - \mu\right)$$

remarks:

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

Empirical variances

➜ *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:

$$\langle \hat{c} \rangle \overset{n\to\infty}{=} \frac{1}{\langle 1/y \rangle} \quad \text{and} \quad \frac{\langle \chi^2_{\min} \rangle}{n-1} \overset{n\to\infty}{=} \frac{\mu}{1-e^{-\mu}} - \frac{1}{\langle 1/y \rangle} ,$$

remarks:

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

Biased variances

➜ *avoid discarding zero bins* $z_i = y_i + 1$

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

asymptotic properties:

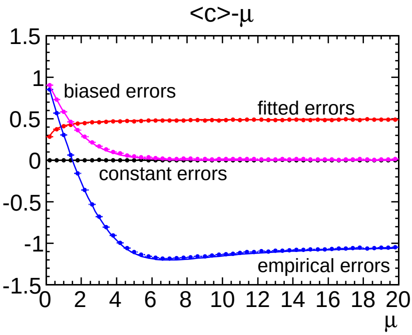
$$\langle c \rangle \overset{n\to\infty}{=} \frac{\mu}{1 - e^{-\mu}} \quad \text{and} \quad \frac{\langle \chi^2_{\min} \rangle}{n-1} \overset{n\to\infty}{=} 1 - \frac{\mu}{e^{\mu} - 1} \, .$$

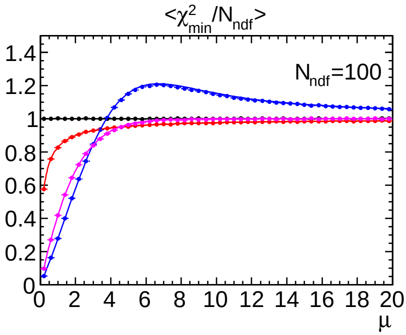
remarks:

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

Numerical studies (i)

➜ *expectation values vs* μ *for* $n = 101$

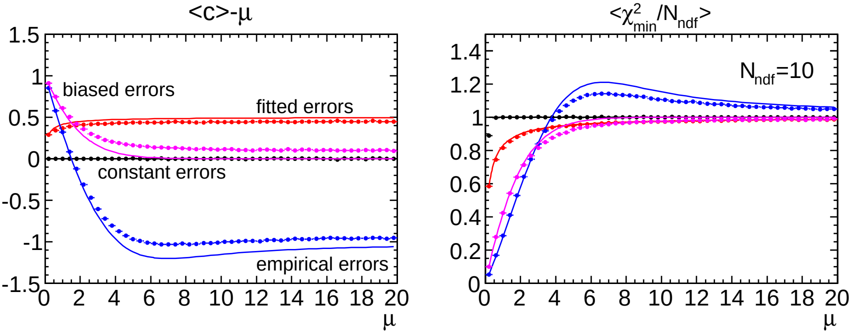




- data points: simulations for $n = 101$ data points
- curves: asymptotic expectations

Numerical studies (ii)

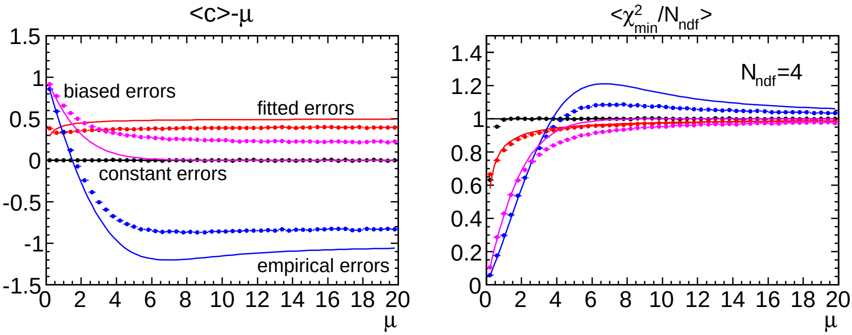
➜ *expectation values vs μ for $n = 11$*



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

Numerical studies (iii)

➜ *expectation values vs μ for $n = 5$*



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

Least squares for low statistics

➜ *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 ➜

Numerical example

➜ *determination of the lifetime of an unstable particle*

- lifetime distribution

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

- 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:

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

- parametric model for bin contents n_i in Least Squares fit

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

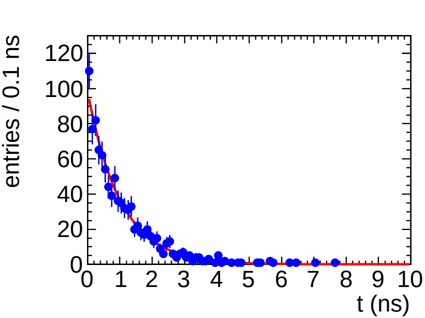
Fit scenarios

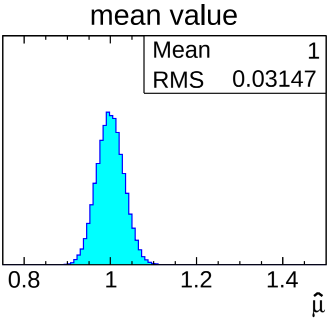
➜ *test different weight-assignments*

- $w_i = 1$ 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

Monte Carlo simulation

➜ *best fit performance for* $N = 1000$ *events*

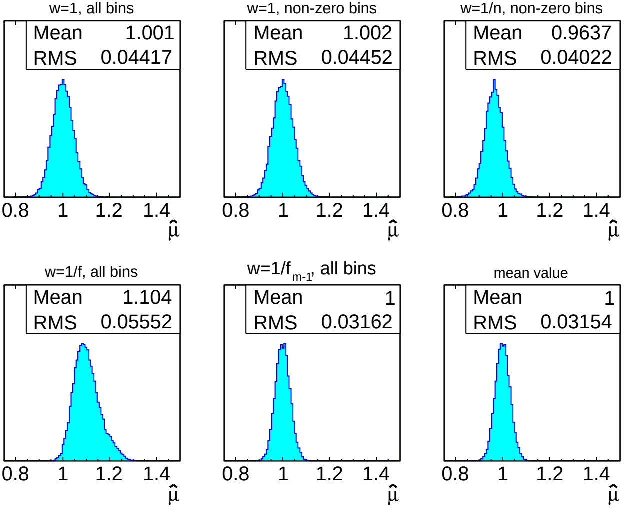




- check standard deviation and bias of fitted $\hat{\mu}$
  - ➜ as a function of available statistics
  - ➜ for the different choices of the weight function

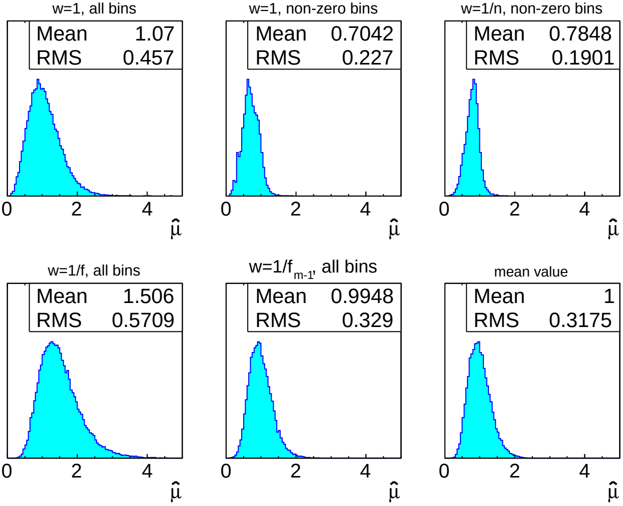
Performance comparison

➜ *parameter estimates for $N = 1000$ events*



Performance comparison

➜ *parameter estimates for $N = 10$ events*


w=1, all bins

w=1, non-zero bins

w=1/n, non-zero bins

w=1/f, all bins

w=1/f$_{m-1}$, all bins

mean value

Conclusions

➜ *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

The Maximum Likelihood Method

➜ *a limiting case of the Least Squares Method*

- uncorrelated single measurements
- counting statistics
- infinitesimal bin widths - i.e. zero or one entry per bin

❖ least-squares fitting of a single parameter with a fixed number of events N:

- ➜ estimate the parameter a of the PDF $f(x;a)$ of the measurements
- ➜ iterative minimization with $\hat{a}$ the estimate from the previous step
- ➜ bin contents $y_i \in 0, 1$

$$\chi^2 = \sum_i \frac{(y_i - N\,p_i)^2}{N\,\hat{p}_i} \quad \text{with} \quad p_i = f(x_i;a)\Delta x \quad \text{and} \quad \hat{p}_i = f(x_i;\hat{a})\Delta x$$

expanding the numerator:

$$\chi^2 = \sum_i \frac{y_i^2}{N\hat{p}_i} - 2\sum_i \frac{y_i p_i}{\hat{p}_i} + N\sum_i \frac{p_i^2}{\hat{p}_i}$$

- ➜ the 1st term is arbitrary ($\propto 1/\Delta x$) and independent of a
- ➜ the 2nd and 3rd terms are functions of a

For infinitesimal bin widths one obtains

$$-2\sum_{\mathrm{bins},i}\frac{y_i p_i}{\hat{p}_i}\rightarrow -2\sum_{\mathrm{events},i}\frac{p_i}{\hat{p}_i} = -2\sum_{\mathrm{events},i}\frac{f(x_i;\,a)}{f(x_i;\,\hat{a})}$$

and

$$N\sum_{\mathrm{bins},i}\frac{p_i^2}{\hat{p}_i}\rightarrow N\int dx\,\frac{f^2(x;\,a)}{f(x;\,\hat{a})}$$

and minimization of χ^2 with convergence $\hat{a}\rightarrow a$ leads to:

$$\frac{\partial}{\partial a}\chi^2 = -2\sum_{\mathrm{events},i}\frac{f'(x_i;\,a)}{f(x_i;\,\hat{a})} + N\int dx\,\frac{2\,f(x;\,a)\,f'(x;\,a)}{f(x;\,\hat{a})}$$

$$\overset{\hat{a}\rightarrow a}{=} -2\sum_{\mathrm{events},i}\frac{f'(x_i;\,a)}{f(x_i;\,a)} + 2N\int dx\,f'(x;\,a)$$

$$= 2\frac{\partial}{\partial a}\left(-\sum_{\mathrm{events},i}\ln f(x_i;\,a) + N\int dx\,f(x;\,a)\right)$$

Result

➜ *since $f(x;a)$ is normalized when integrating over x:*

$$\frac{\partial}{\partial a}\left(\frac{1}{2}\chi^2\right) = \frac{\partial}{\partial a}\left(-\ln L(\vec{x};a)\right) \overset{!}{=} 0 \quad \text{with} \quad L(\vec{x};a) = \prod_{\text{events},i} f(x_i;a)$$

❖ discussion:

- the best fit parameter is obtained by maximising the likelihood of the data
- for uncorrelated measurements it is the estimate with the smallest variance
- in presence of correlations the least-squares approach with the full covariance matrix is more powerful
- going to infinitesimal bin sizes, the χ^2-minimum becomes arbitrary, i.e. the maximum of the likelihood contains no information about the quality of the fit
- maximum likelihood and least squares fits have very similar properties

$$\Delta(-\ln L) = \frac{1}{2}\Delta\chi^2$$

Extended Maximum Likelihood

➜ *ansatz to estimate also the normalisation when n events were seen:*

$$\chi^2 = \sum_i \frac{(y_i - N\,p_i)^2}{\hat{N}\,\hat{p}_i} \quad \text{with} \quad p_i = f(x_i;a)\Delta x \quad \text{and} \quad \hat{p}_i = f(x_i;\hat{a})\Delta x$$

Expanding the numerator yields:

$$\chi^2 = \sum_i \frac{y_i^2}{\hat{N}\,\hat{p}_i} - 2\frac{N}{\hat{N}}\sum_i \frac{y_i p_i}{\hat{p}_i} + \frac{N^2}{\hat{N}}\sum_i \frac{p_i^2}{\hat{p}_i}$$

- ➜ the 1st term is an arbitrary offset C
- ➜ the remaining terms depend in N and a

χ^2 function in the limit of infinitesimal bin widths:

$$\chi^2 = C - 2\frac{N}{\hat{N}}\sum_{\text{events},i}^{n} \frac{f(x_i;a)}{f(x_i;\hat{a})} + \frac{N^2}{\hat{N}}\int dx\,\frac{f^2(x;a)}{f(x;\hat{a})}$$

Derivatives w.r.t. N and a must vanish; consider $\hat{N} \to N$ and $\hat{a} \to a$.

Result

Taking first the partial derivatives and then the limit $\hat{N} \to N$ and $\hat{a} \to a$ yields

$$\frac{\partial}{\partial N}\chi^2 = -2\frac{n}{N} + 2 = 0$$
$$\frac{\partial}{\partial a}\chi^2 = -2\sum_{\text{events},i}^{n} \frac{f'(x_i;a)}{f(x_i;a)} = 0$$

which corresponds to

$$\frac{\partial}{\partial N}(-\ln L) = \frac{\partial}{\partial a}(-\ln L) = 0$$

with

$$-\ln L = N - n\ln N - \sum_{\text{events},i}^{n} \ln f(x_i;a) = N - \sum_{\text{events},i}^{n} \ln[N\,f(x_i;a)]$$

➜ *standard and extended maximum likelihood method follow from least squares*

Significance of fit parameters

➜ *S.S. Wilks, March 26, 1937*

If a population with a variate x is distributed according to the probability distribution $f(x,\theta_1,\theta_2,\ldots,\theta_h)$, such that optimum estimates $\hat\theta_i$ of θ_i exist which are distributed in large samples according to (1), then when the hypothesis H is true that $\theta_i=\theta_{0i}, i=m+1,m+2,\ldots h$, the distribution of $-2\ln\lambda$, where λ is given by (2) is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h-m$ degrees of freedom.

(1) a PDF deviating from a d-dim Gaussian only by terms of order $1/\sqrt{n}$

(2) the ratio of the best fit likelihoods fitting all or only m parameters, fixing the others to the true values

$$\lambda=\frac{P(\hat\theta_1,\ldots,\hat\theta_m,\hat\theta_{0m+1},\ldots,\hat\theta_{0h})}{P(\hat\theta_1,\ldots,\hat\theta_m,\hat\theta_{m+1},\ldots,\hat\theta_h)}$$

❖ likelihoods are meaningless, likelihood ratios are significant

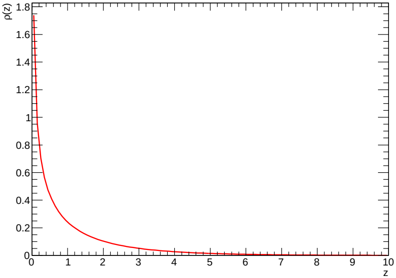
Application of Wilks´ theorem

➜ *test for the existence of a signal s component in data*

- fit with free parameter s: $F_s = -\ln L_{\text{best}}(s)$
- fit with parameter $s = 0$: $F_0 = -\ln L_{\text{best}}(s = 0)$

➜ one has $F_s < F_0$ and $z = 2(F_0 - F_s) > 0$

PDF of z if $s = 0$ is true: $\rho(z) = \dfrac{1}{\sqrt{2\pi z}}\, e^{-z/2}$



➜ *p-value for observed z_{obs}*

$$p = \int_{z=z_{\text{obs}}}^{\infty} dz\ \rho(z)$$

discovery $s \neq 0$ if e.g. $p < 5.7 \cdot 10^{-7}$

4. Multivariate Analysis

➜ *objective: decide between hypotheses*

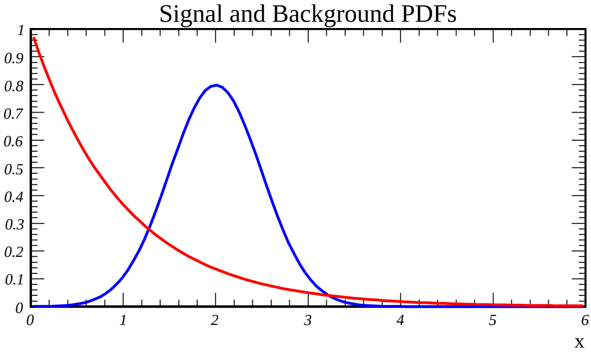
- e.g. classification of events or candidates
  - ➜ H_0: signal
  - ➜ H_1: background
- error of 1. kind : H_0 is wrongly rejected with probability α
- error of 2. kind: H_1 is wrongly rejected with probability β

	truth	
classification	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

- in the following: PDFs of signal and background are known
- try optimal separation of both components

Example

➜ *gaussian signal on exponential background*

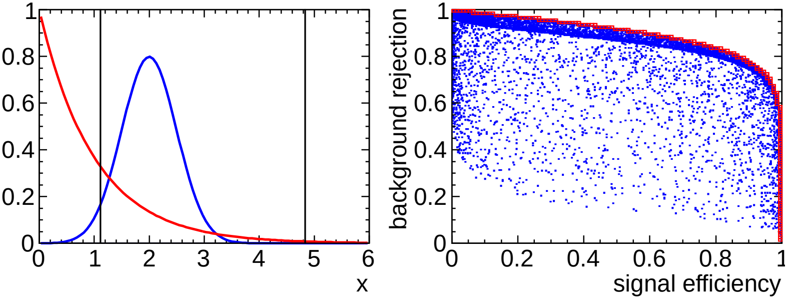


❖ study signal selection

- try different signal windows
- gauge performance by background rejection vs signal efficiency

Best selection by likelihood ratio

selection: if $\dfrac{f(x|H_0)}{f(x|H_1)} \le c$ then reject x



- best “Receiver Operation Characteristic” (ROC-curve)
  - ➜ largest background rejection for fixed signal efficiency
  - ➜ smallest errors of 2nd kind for fixed errors of 1st kind
  - ➜ parameter c determines signal efficiency

Neyman-Pearson´s lemma

➜ *definitions and conditions*

- $\vec{x}$: point in configuration space
- $f(\vec{x}|H_k)$: PDF for $\vec{x}$ in case of H_k
- critical region S: configuration space volume with probability α for H_0

$$P(\vec{x} \in S|H_0) = \int_S d^n x\, f(\vec{x}|H_0) = \alpha$$

- S_c: critical region satisfying

$$\frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} \leq c$$

➜ *conjecture:*

The critical region S_c is optimal in the sense, that it minimizes errors of the second kind (minimal probability to accept background).

➜ proof

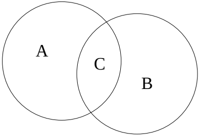
Proof of Neyman-Pearson´s lemma (i)

Take two critical regions S_c and S with equal probability content for H_0

$$\int_{S_c} d^n x\, f(\vec{x}|H_0) = \int_S d^n x\, f(\vec{x}|H_0) = \alpha$$

In general the regions will overlap and one can write:

$$S_c = A \cup C \quad \text{and} \quad S = B \cup C$$



Thus C contributes equally to S_c and S and one has

$$\int_A d^n x\, f(\vec{x}|H_0) = \int_B d^n x\, f(\vec{x}|H_0)\ .$$

Region A is inside S_c, B is outside, i.e. by construction

$$\frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} \leq c \quad \text{if } \vec{x} \in A \qquad \text{and} \qquad \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} > c \quad \text{if} \quad \vec{x} \in B$$

Proof of Neyman-Pearson´s lemma (ii)

It follows:

$$\int_A d^n x\, f(\vec{x}|H_0) \leq c \int_A d^n x\, f(\vec{x}|H_1)$$
$$\int_B d^n x\, f(\vec{x}|H_0) \geq c \int_B d^n x\, f(\vec{x}|H_1)$$

Compare now the H_1 (background) probabilities in S_c and S:

$$\begin{aligned} P(\vec{x} \in S_c|H_1) &= \int_A d^n x\, f(\vec{x}|H_1) + \int_C d^n x\, f(\vec{x}|H_1) \\ &\geq \frac{1}{c} \int_A d^n x\, f(\vec{x}|H_0) + \int_C d^n x\, f(\vec{x}|H_1) \\ &= \frac{1}{c} \int_B d^n x\, f(\vec{x}|H_0) + \int_C d^n x\, f(\vec{x}|H_1) \\ &\geq \int_B d^n x\, f(\vec{x}|H_1) + \int_C d^n x\, f(\vec{x}|H_1) = P(\vec{x} \in S|H_1) \end{aligned}$$

Result

➜ *comparison of the background probability shows:*

$$P(\vec{x} \in S_c | H_1) \geq P(\vec{x} \in S | H_1) \; .$$

- critical regions are rejected for signal selections
- by construction all critical regions have the same α
  - ➜ the same signal efficiency $1 - \alpha$
- the region S_c has the largest background probability
  - ➜ largest possible rejection for given signal efficiency
  - ➜ smallest errors of 2nd kind for given errors of 1st kind
- in S_c one has

$$\frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} \leq c$$

- optimal solution of the selection problem if all PDFs are known

Multivariate analysis in practice

➜ *problem*

- PDFs are not known
- only finite samples exists to estimates the PDFs of H_0 and H_1
- multi-dimensional PDFs hard to determine ("curse of dimensionality")

➜ *general strategy*

- construct test variables or functions (classifier) in configuration space
- start with training
  - ➜ estimate PDFs L_S and L_B for signal and background
  - ➜ avoid "overtraining" (learning fluctuations in the training sample)
- performance test with independent signal and background samples

➜ *z.B. open source implementation: TMVA*

Toolkit for MultiVariate Analysis with ROOT

`arXiv:physics/0703039, CERN-OPEN-2007-007`

`http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf`

Introductory remarks

- two types of classifiers
  - ➜ optimized classifiers for predefined signal efficiency $1-\alpha$
  - ➜ continuous probability-like classifiers t provided by TMVA
- raw ranges $t_{\mathrm{min}} \leq t \leq t_{\mathrm{max}}$, possibly peaking towards limit
  - ➜ transform to normalized classifiers to $-1 \leq t' \leq +1$

$$t' = \frac{1}{N} \ln \frac{t - t_{\mathrm{min}} + \delta}{t_{\mathrm{max}} - t + \delta} \quad \text{with} \quad \delta = \frac{t_{\mathrm{max}} - t_{\mathrm{min}}}{\exp(N) - 1}$$

  - ➜ $N \to 0$: linear rescaling to $[-1, +1]$
  - ➜ $N > 0$: remove singularities at the end points
- classifiers are not invariant under transformations of variables
  - ➜ human understanding of the problem still vital
- in most cases the theoretical optimum is not reached
- note: biased training samples lead to biased efficiency estimates
  - ➜ ongoing work to understand and control such systematics

Overview of methods

- discussed below (and available in TMVA)
  - ➜ projected 1-dim likelihood ratios
  - ➜ KNN
  - ➜ PDEFoam
  - ➜ Fisher discriminant
  - ➜ multilayer-perceptron neural networks
  - ➜ Boosted Decision Trees
- common preprocessing steps
  - ➜ decorrelation and gaussianization
  - ➜ combinations and iterations of the above
- use TMVA methods with default settings
  - ➜ no tuning of internal parameters and options
  - ➜ no preprocessing of variables
- performance classification by ROC-curves

Projected likelihoods ratios

➜ *attempt to apply the Neyman-Pearson lemma*

$$f(x_1, x_2, x_3 \ldots) \to L = f_1(x_1)\, f_2(x_2)\, f_3(x_3) \ldots$$

$$\text{with} \quad f_1(x_1) = \int dx_2\, dx_3 \ldots f(x_1, x_2, x_3, \ldots)$$

$$f_2(x_2) = \int dx_1\, dx_3 \ldots f(x_1, x_2, x_3, \ldots) \quad \text{etc.}$$

- parametrize the projected PDFs
- classifier $c(i)$ for each event i

$$c(i) = \frac{L_{\mathrm{sig}}(i)}{L_{\mathrm{sig}}(i) + L_{\mathrm{bkg}}(i)} = \frac{1}{1 + L_{\mathrm{bkg}}(i)/L_{\mathrm{sig}}(i)}$$

- projections avoid “curse of dimensionality”
- loss of performance if true PDFs do not factorize

PDEFoam

➜ *attempt proper n-dim density estimates*

- subdivide the phase space into a given number of hyper-rectangles with (about) equal numbers of entries per cell
- search subdivision which minimizes the density variance in the cells
- assume constant density per cell
- do this separately for signal and background
- construct classifier based on likelihood ratios
- properties:
  - ➜ non-parametric description of PDFs
  - ➜ correlations are taken into account
  - ➜ very few entries per cell in high-dimensional spaces

k-Nearest Neighbor classifier (KNN)

➜ *compare a candidate event to training sample densities*

- non-parametric density estimates
- k training events (signal plus background) closest to the candidate
- classifier:

relative signal-probability $$c_{\mathrm{KNN}} = \frac{k_{\mathrm{sig}}}{k_{\mathrm{sig}} + k_{\mathrm{bkg}}} = \frac{k_{\mathrm{sig}}}{k}$$

- empirical finding: $10 < k < 100$ shows good performance
  - ➜ too large value: local density variations are not seen
  - ➜ too small value: estimates suffer from large fluctuations
- performance depends on the metric

$$R^2 = \sum_{i=1}^{n_{\mathrm{dim}}} \frac{1}{w_i^2} (x_i - y_i)^2$$

  - ➜ w_i allows adaption to spread of input variables
  - ➜ intrisically adaptive – no problem with large number of dimensions

The Fisher discriminant (i)

➜ *test variable from a linear combination of the measurements x_i*

$$t(\vec{x}) = a_0 + \sum_{i=1}^{n} a_i x_i = a_0 + \vec{a}^{\,T}\vec{x}$$

❖ geometrical interpretation

- $\vec{a}$ and a_0 define a hyperplane in n dimensions
  - ➜ $\vec{a}$ is a vector normal to the plane
  - ➜ a_0 is the distance of the plane from the origin
- constant values $t(\vec{x})$ for points $\vec{x}$ on a plane parallel to the hyperplane defined by $\vec{a}$ and a_0
- adjust the orientation of $\vec{a}$ and the offset a_0 to get optimal separation between H_0 (signal) and H_1 (background)

The Fisher discriminant (ii)

➜ *realization:*

expectation values and covariance matrix of $\vec{x}$ for hypotheses H_k are

$$\langle\vec{x}\rangle\big|_{H_k} = \vec{\mu}_k \quad \text{and} \quad \left.\left\langle\vec{x}\cdot\vec{x}^{\,T}\right\rangle - \langle\vec{x}\rangle\cdot\langle\vec{x}\rangle^T\right|_{H_k} = V_k$$

for mean and variance of t under hypothesis H_k follows

$$\langle t_k\rangle = a_0 + \vec{a}^{\,T}\vec{\mu}_k \quad \text{and} \quad V_k(t) = \vec{a}^{\,T}\cdot V_k\cdot\vec{a}$$

and a measure $J(\vec{a})$ for the separation between the hypotheses is

$$J(\vec{a}) = \frac{(\langle t_0\rangle - \langle t_1\rangle)^2}{V_0(t) + V_1(t)} = \frac{\left(\vec{a}^{\,T}(\vec{\mu}_0 - \vec{\mu}_1)\right)^2}{\vec{a}^{\,T}\cdot(V_0 + V_1)\cdot\vec{a}}$$

or, introducing $V = V_0 + V_1$ and $\vec{\mu} = \vec{\mu_0} - \vec{\mu_1}$,

$$J(\vec{a}) = \frac{\vec{a}^{\,T}\vec{\mu}\vec{\mu}^{\,T}\vec{a}}{\vec{a}^{\,T}V\vec{a}} \overset{!}{=} \max\ .$$

The Fisher discriminant (iii)

➜ *construction of the solution*

- $J(\vec{a})$ does not depend on the normalization of $\vec{a}$ or a_0
- boundary condition $\vec{a}^T\vec{\mu} = c$ defines unique solution
- constrained maximum:

$$\frac{\partial}{\partial \vec{a}^T}\left[\frac{\vec{a}^T\vec{\mu}\vec{\mu}^T\vec{a}}{\vec{a}^T V\vec{a}} - \lambda(c - \vec{a}^T\vec{\mu})\right] = 0$$

and thus
$$\frac{\vec{\mu}\vec{\mu}^T\vec{a}}{\vec{a}^T V\vec{a}} - \frac{\vec{a}^T\vec{\mu}\vec{\mu}^T\vec{a}}{(\vec{a}^T V\vec{a})^2}(V\vec{a}) + \lambda\vec{\mu} = 0$$

➜ a solution exists if $V\vec{a} \propto \vec{\mu}$, i.e.

$$\vec{a} \propto V^{-1}\vec{\mu}$$

➜ and adjustment of c and a_0 allows to have

$$\langle t_0\rangle = 1 \qquad \text{and} \qquad \langle t_1\rangle = 0$$

The Fisher discriminant (iv)

➜ *result:*

$$t(\vec{x}) = a_0 + c(\vec{\mu_0} - \vec{\mu_1})^T (V_0 + V_1)^{-1} \vec{x}$$

with freely choosable parameters a_0 and c.

➜ *construction of $t(\vec{x})$ requires of each hypothesis*

- n expectation values
- $n(n+1)/2$ variances and covariances
- in total $n(n+1) + 2n = n^2 + 3n$ parameters
- numerically stable determination
- very small danger of overtraining if $N \gg n^2$

Fisher discriminant and likelihood ratios

➜ *consider 1-dim gaussians*

$$f(x|H_0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_0)^2/2\sigma^2} \quad \text{and} \quad f(x|H_1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_1)^2/2\sigma^2}$$

- different mean values μ_0 and μ_1 but equal standard deviations σ
- then the logarithm of the likelihood ratio

$$\ln \frac{f(x|H_0)}{f(x|H_1)} = -\frac{1}{2\sigma^2}(x^2 - 2x\mu_0 + \mu_0^2 - x^2 + 2x\mu_1 - \mu_1^2)$$

$$= \frac{\mu_1^2 - \mu_0^2}{2\sigma^2} + \frac{\mu_0 - \mu_1}{\sigma^2} x = a_0 + a_1 x$$

has the structure of a Fisher discriminant, i.e.

$$\frac{f(x|H_0)}{f(x|H_1)} \equiv r \propto e^{t(x)}$$

➜ Fisher: optimal for gaussian H_0 and H_1 with equal variance

Neural networks

➜ heuristic approach

Bayes' theorem allows to formulate a relation between Fisher discriminant and bayesian signal probability. Taking equal prior probabilities $p(H_0) = p(H_1)$ for signal and background one finds:

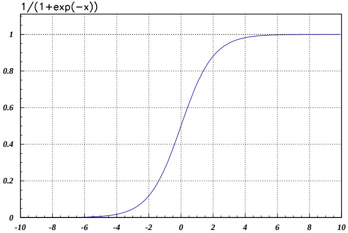
$$P(H_0|\vec{x}) = \frac{f(\vec{x}|H_0)p(H_0)}{f(\vec{x}|H_0)p(H_0) + f(\vec{x}|H_1)p(H_1)} = \frac{1}{1 + 1/r}$$

For equal-width gaussians one had $r = e^t$, which leads to

$$P(H_0|\vec{x}) = \frac{1}{1 + e^{-t}} \equiv s(t) \quad \in \quad [0, 1]$$

➜ "logistic sigmoid function"

useful to describe decisions between hypotheses. . .



The single layer perceptron

➜ *interpretation of the Fisher discriminant as “neural network”*

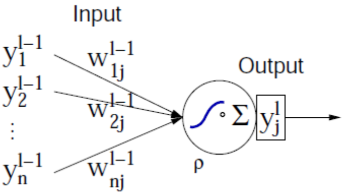
- Fisher discriminant as a weighted sum of the input signals

$$t(\vec{x}) = a_0 + \sum_{i=1}^{n} a_i x_i$$

- interpretation by means of the logistic sigmoid function

$$s(t) = \frac{1}{1 + e^{-t}}$$

single layer perceptron



equivalent to Fisher discriminant

❖ compare:

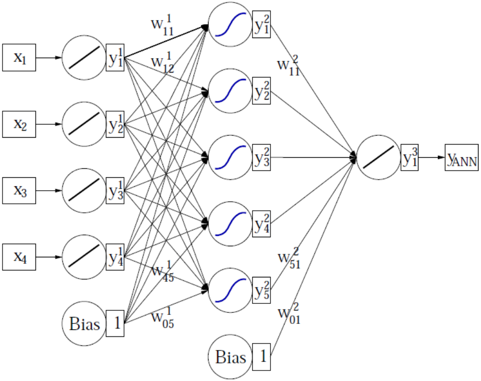
- signal processing in nerve cells
  - ➜ several inputs x_i
  - ➜ weighted summation $\sum_i a_i x_i$
  - ➜ switching according to activation function

Multilayer networks

➜ *cascading of neurons*

- for example: double layer perceptron
- output signal is a function of an inner (hidden) layer of neurons

Input Layer Hidden Layer Output Layer



$$t(\vec{x}) = s\left(a_0 + \sum_{i=1}^{m} a_i h_i(\vec{x})\right)$$

with

$$h_i(\vec{x}) = s\left(w_{i0} + \sum_{k=1}^{n} w_{ik} x_k\right)$$

Discussion

- n neurons in principle allow n^2 directional connections
- reduction of complexity by
  - ➜ arrangement in layers
  - ➜ restriction to feed-forward networks
- neural networks are very efficient universal approximators
  - ➜ even the most general case can be realized with a single hidden layer – but may require a very large number of neurons
  - ➜ alternatively use several hidden layers and fewer neurons
- the optimal network topology for a given application is not known
- there are many possible choices for the activation function, e.g.
  - ➜ logistic sigmoid $s(t)$, $\mathtt{tanh}(t)$, $\ldots$
- determination of the weight usually done numerically

Strategy for the determination of the weights

- generate training samples for each hypothesis
- define decisions as a function of the output signal
- define a cost functions for the quality of the decision
- adjust the weight by minimizing the cost function

example: $$F = \sum_{\vec{x}\in H_1} t^2(\vec{x}) + \sum_{\vec{x}\in H_0} (1 - t(\vec{x}))^2$$

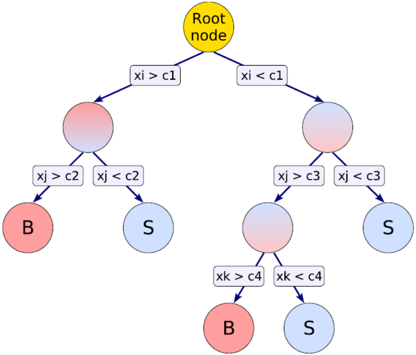
goal: $t(\vec{x}) = 1$ for $\vec{x} \in H_0$ (sig) and $t(\vec{x}) = 0$ für $\vec{x} \in H_1$ (bkg)

Determination of the weights is a hard non linear minimization problem with usually many local minima. It is normally sufficient to find a good minimum instead of the global one. Possible algorithm:

- ➜ take random initial values for the weights
- ➜ get the gradient of the cost function with respect to the weights
- ➜ do a (small) downhill step and iterate until a minimum is reached
- ➜ try other initial values

Boosted Decision Trees

➜ *basic topology of a decision tree*



- sequence of binary decisions
- generalization of cut-sequence for signal selection
- each instance $\vec{x}$ is classified as signal or background

  $h(\vec{x}) = +1$ signal

  $h(\vec{x}) = -1$ background
- classification of a node as signal or background according to the majority of its instances

Construction of a decision tree

➜ *iteratively splitting of each node*

- basic idea

  scan all variables and determine a cut which gives the best improvement in the separation of signal and background
- implementation requires a measure for separation, as e.g. the “Gini Index” S, based on the signal purity p in a node
  - ➜ separation in the mother node:

$$S_M = p(1-p)$$

  - ➜ separation in the daughter nodes:

$$S_T = \frac{n_1}{n_1+n_2} p_1(1-p_1) + \frac{n_2}{n_1+n_2} p_2(1-p_2)$$

- use that variable and the cut which maximizes $S_M - S_T$
- stop splitting if 100% purity or too few events in a node
- problem: small fluctuations can give radically different decision trees
- remedy: boosting

Boosting

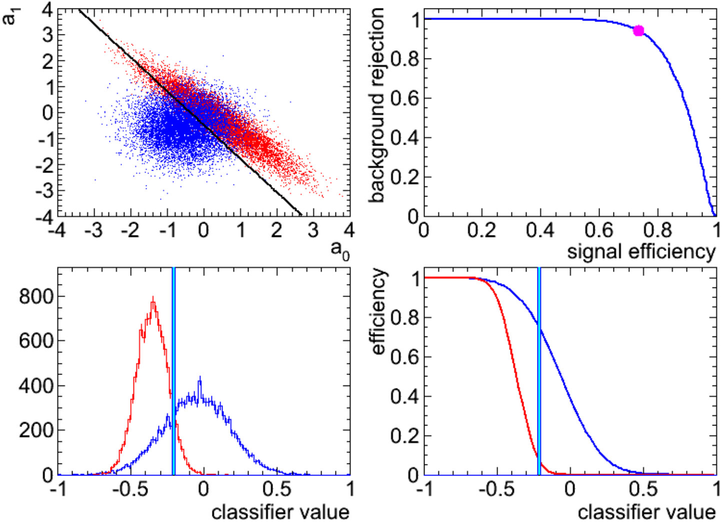
➜ *construction an average decision tree with improved performance*

- iterative generation of decision trees by constructing new training samples based on mis-classification rate ε of the current tree
  - ➜ weight all wrongly classified instances by $(1-\varepsilon)/\varepsilon$
  - ➜ renormalize the sum of all instances to the original value
  - ➜ determine a new decision tree
- decision tree ➜ decision forest
- further improvement:
  - ➜ pruning of the trees by eliminating branches with only a negligible improvement in the separation between signal and background
- BDT-Classifier: weighted average of all classifications in the forest

$$y(\vec{x}) = \sum_{i \in \text{forest}} \ln \frac{1-\varepsilon_i}{\varepsilon_i} \cdot h_i(\vec{x})$$

  i.e. larger weight for trees with smaller mis-identification rate
- often best: weighted mean over many weakly optimized trees

Illustration



Summary on multivariate analysis

- optimal separation of signal and background by likelihood ratios
  - ➜ in practice useful only for few dimensional problems
  - ➜ many different methods for higher dimensional problems
- projected likelihoods: ideal for uncorrelated variables
- PDEFoam and KNN: good start for n-dim likelihood ratios
- Fisher discriminant: simple and robust, optimal for gaussians
- neural networks: probably best, but hard to train
- (Boosted) Decision Trees: very good “out-of-the-box”-method
- performance of many methods depends on choice of variables
  - ➜ pre-processing can result in significant gain
    - ◆ de-correlation
    - ◆ avoid singularities ($x \to \ln x$)
    - ◆ discard insensitive variables to avoid “curse of dimensionality”
- MVA: very active field of research (data mining . . .)

➜ *separating signal and background*

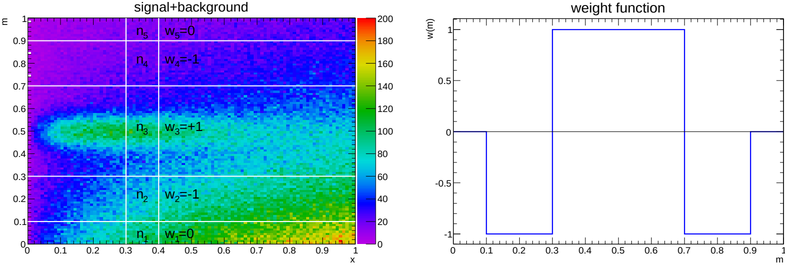
$$f(x,m) = N_s\, s(x,m) + N_b\, b(x,m)$$

- normalized PDFs $s(x,m)$ and $b(x,m)$ for signal and background
- normalizations N_s and N_b for signal and background
- m: "discriminant" variable to tell signal from background
  - ➜ will be treated as a scalar in the following
  - ➜ can equally well be vector
- x: "control" variable to be studied
  - ➜ will be treated as a scalar in the following
  - ➜ can equally well be vector
- try to extract the signal density as function of x

get rid of background by sideband subtraction ➜

Starting point: sideband subtraction

- determine the signal density $N_s\, s(x)$ for a given x



$$N_s\, s(x) \approx \frac{\mathrm{signal}(x-\Delta x/2, x+\Delta x/2)}{\Delta x} = \frac{1}{\Delta x} \sum_{i, x_i \in x \pm \Delta x/2} w(m_i)$$

$$= \frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} dx \int dm\, w(m)\, f(x,m) = \int dm\, w(m)\, f(x,m)$$

Generalization

➜ *sideband subtraction is a special case of an integral transform*

$$N_s\, s(x) = \int dm\, w(m)\, f(x,m)$$

- the weight function $w(m)$ projects out the signal density $s(x)$
- the above equation was derived for one fixed x
- now require that the same $w(m)$ works for all x
- possible if $s(x,m)$ and $b(x,m)$ factorize as a function of x and m

$$f(x,m) = N_s\, s(x)\, s(m) + N_b\, b(x)\, b(m)$$

- assume that $s(m)$ and $b(m)$ are known

❖ Find the optimal weight function $w(m)$ for this case!

Determining the weight function

➜ *necessary condition:*

$$\int dm\, w(m)\, [N_s\, s(x)\, s(m) + N_b\, b(x)\, b(m)] = N_s\, s(x)$$

which implies

$$\int dm\, w(m)\ s(m) = 1 \qquad \text{and} \qquad \int dm\, w(m)\ b(m) = 0$$

- any $w(m)$ orthogonal to $b(m)$ can be normalized to satisfy this
- for $s(m) \propto b(m)$ signal and background cannot be separated
- select $w(m)$ which gives the most precise result for $s(x)$

$$\sum_{events} w^2 \stackrel{!}{=} \min \qquad ➜ \qquad \int dx\, dm\, w^2(m)\, f(x,m) \stackrel{!}{=} \min$$

➜ constrained minimization problem

➜ solved by variational calculus

Optimizing statistical precision

➜ *constrained minimization with Lagrange parameters α and β*

$$\delta\left\{\int dx\,dm\,w^2(m)\,[N_s\,s(x)\,s(m)+N_b\,b(x)\,b(m)]\right.$$

$$\left.+\,2\alpha\left(1-\int dm\,w(m)\,s(m)\right)-2\beta\int dm\,w(m)\,b(m)\right\}=0$$

- ➜ the variation is performed on $w(m)$
- ➜ the constant term 2α is irrelevant for the minimization
- ➜ integration over x gives two factors of unity ➜ single integral over m

$$\delta\left\{\int dm\,w^2(m)\,[N_s\,s(m)+N_b\,b(m)]-2w(m)[\alpha\,s(m)+\beta\,b(m)]\right\}=0$$

- ➜ substituting $\delta w^2(m)=2\,w(m)\delta w(m)$ yields

$$\int dm\,\delta w(m)\,\{w(m)\,[N_s\,s(m)+N_b\,b(m)]-[\alpha\,s(m)+\beta\,b(m)]\}=0$$

- ➜ zero integral for arbitrary $\delta w(m)$ requires $\{\ldots\}=0$

Result

➜ *optimal weight function*

$$w(m) = \frac{\alpha\, s(m) + \beta\, b(m)}{N_s\, s(m) + N_b\, b(m)}$$

- $w(m)$ is a linear combination of signal purity and background purity
- numerical values are for α and β follow from the constraints

$$\int dm\, w(m)\, s(m) = 1 \qquad \text{and} \qquad \int dm\, w(m)\, b(m) = 0$$

- substituting m yields the system of equations:

$$\begin{pmatrix} W_{ss} & W_{sb} \\ W_{sb} & W_{bb} \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{with} \quad W_{uv} = \int dm\, \frac{u(m)\, v(m)}{N_s\, s(m) + N_b\, b(m)}$$

- with solutions

$$\alpha = \frac{W_{bb}}{W_{ss}\, W_{bb} - W_{sb}^2} \quad \text{and} \quad \beta = \frac{-W_{sb}}{W_{ss}\, W_{bb} - W_{sb}^2}$$

Numerical determination of W_{uv}

➜ *consider the binned m-distribution*

- with $i = 1, \ldots, n$ bins with widths Δm and bin centers m_i
- indices u, v referring to signal or background PDF, i.e. $u, v \in \{s, b\}$

$$W_{uv} = \int dm \, \frac{u(m)\, v(m)}{N_s\, s(m) + N_b\, b(m)} \approx \sum_i \Delta m \frac{u(m_i)\, v(m_i)}{N_s\, s(m_i) + N_b\, b(m_i)}$$

$$= \sum \Delta m \frac{u(m_i)\, v(m_i)\, \Delta m}{N_s\, s(m_i)\, \Delta m + N_b\, b(m_i)\, \Delta m} = \sum_i \frac{p_u(m_i)\, p_v(m_i)}{n(m_i)}$$

➜ with $p_{u,v}(m_i)$ the signal or background probability in bin i

$$p_s(m_i) = s(m_i)\, \Delta m \quad \text{and} \quad p_b(m_i) = b(m_i)\, \Delta m$$

➜ and $n(m_i)$ the observed number of events in the bin around m_i

Numerical example

- phase space

$$0 < x < 1 \quad \text{and} \quad 0 < m < 1$$

- signal $s(x,m) = s(x)\,s(m)$

$$s(x) = \frac{25}{1-6e^{-5}}\, x\, e^{-5x} \quad \text{and} \quad s(m) = \frac{20}{\sqrt{2\pi}} e^{-200(m-0.5)^2}$$

- background $b(x,m) = b(x)\,b(m)$

$$b(x) = 1.5\sqrt{x} \quad \text{and} \quad b(m) = \frac{2}{1-e^{-2}}\, e^{-2m}$$

- event statistics

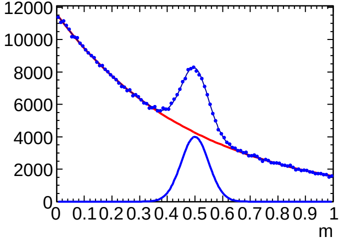
$$N_s = 50\,000 \quad \text{with} \quad N_b = 500\,000$$

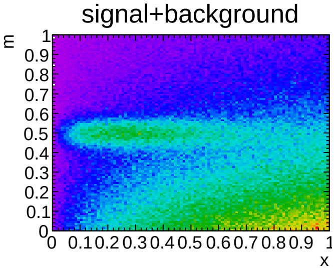
- efficiencies

$$\varepsilon_1(x,m) = 1 \;\;,\;\; \varepsilon_2(x,m) = \frac{m+0.5}{1.5}\,\frac{1.5-x}{1.5} \quad \text{and} \quad \varepsilon_3(x,m) = \frac{x+m}{2}$$

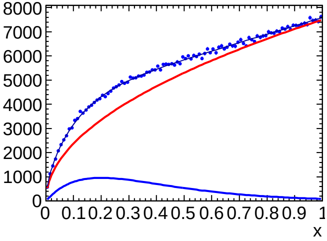
Generated distributions

➜ *sum of signal and background:* $\varepsilon(x, m) = 1$

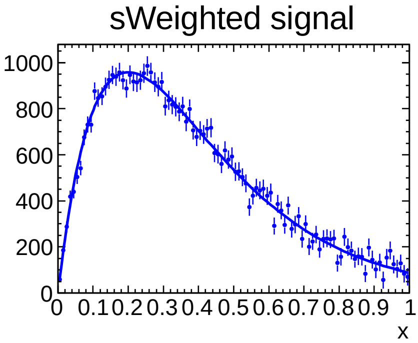


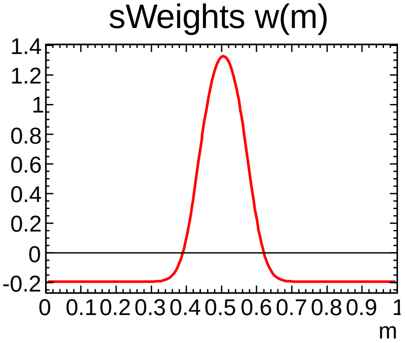


generated distributions



Extracted signal density





➜ histogram all events (x, m) with weights $w(m)$: `hx->Fill(x,w(m))`

sWeights and efficiencies

→ *parameterization of the observations*

$$f(x,m) = \varepsilon(x,m)\,[N_s\, s(x)\, s(m) + N_b\, b(x)\, b(m)]$$

- physics may be expected to factorize in x and m
- detector properties and the observed density often will not factorize

❖ new ansatz for finding a weight function:

$$\int dm\, w(m)\, \frac{f(x,m)}{\varepsilon(x,m)} = N_s\, s(x)$$

with

$$\int dx\, dm\, \left(\frac{w(m)}{\varepsilon(x,m)}\right)^2 f(x,m) \overset{!}{=} \min \quad \text{and constraints}$$

$$\int dm\, w(m)\, s(m) = 1 \qquad \text{and} \qquad \int dm\, w(m)\, b(m) = 0$$

Result

➜ *optimal weight function to extract the efficiency corrected signal*

$$w(m) = \frac{\alpha\, s(m) + \beta\, b(m)}{q(m)} \quad \text{with} \quad q(m) = \int dx\, \frac{f(x,m)}{\varepsilon^2(x,m)}$$

with coefficients α and β

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \frac{1}{W_{ss}\, W_{bb} - W_{sb}^2} \begin{pmatrix} W_{bb} \\ -W_{sb} \end{pmatrix} \quad \text{with} \quad W_{uv} = \int dm\, \frac{u(m)\, v(m)}{q(m)}$$

❖ discussion

- $q(m)$ is the m-spectrum with events weighted by $1/\varepsilon^2(x,m)$
- $s(m)$ and $b(m)$ are the efficiency corrected m-spectra
- the event-by-event weights to extract the signal are $w(m)/\varepsilon(x,m)$
- only if the efficiency factorizes, $\varepsilon(x,m) = \varepsilon(x)\,\varepsilon(m)$, one can also determine sWeights from the uncorrected m-spectra

➜

Efficiencies that factorize (i)

The parameterization of the observed density $f(x, m)$ is given by

$$f(x,m) = N_s\,\varepsilon(x)\,\varepsilon(m)\,s(x)\,s(m) + N_b\,\varepsilon(x)\,\varepsilon(m)\,b(x)\,b(m)$$

which can be re-written by introducing observable quantities

$$s'(x) = \frac{\varepsilon(x)\,s(x)}{\int dx\,\varepsilon(x)\,s(x)} \quad \text{and} \quad b'(x) = \frac{\varepsilon(x)\,b(x)}{\int dx\,\varepsilon(x)\,b(x)}$$

$$s'(m) = \frac{\varepsilon(m)\,s(m)}{\int dm\,\varepsilon(m)\,s(m)} \quad \text{and} \quad b'(m) = \frac{\varepsilon(m)\,b(m)}{\int dm\,\varepsilon(m)\,b(m)}$$

$$N_s' = N_s \int dx\,\varepsilon(x)\,s(x) \int dm\,\varepsilon(m)\,s(m) \quad \text{and}$$

$$N_b' = N_b \int dx\,\varepsilon(x)\,b(x) \int dm\,\varepsilon(m)\,b(m)$$

to the same form discussed before for $\varepsilon = 1$:

$$f(x,m) = N_s'\,s'(x)\,s'(m) + N_b'\,b'(x)\,b'(m)$$

Efficiencies that factorize (ii)

➜ *signal extraction weights $w(m)$ from the observed distributions*

$$w(m) = \frac{\alpha\, s'(m) + \beta\, b'(m)}{q(m)} \quad \text{with} \quad q(m) = \int dx\, f(x,m)$$

application to extract an efficiency corrected x-spectrum:

$$\int dm\, w(m)\, f(x,m) = N_s'\, s'(x) = N_s \varepsilon(x)\, s(x) \int dm\, \varepsilon(m)\, s(m)$$

Introducing the global factor F

$$F = \frac{1}{\int dm\, \varepsilon(m)\, s(m)} = \int dm \frac{1}{\varepsilon(m)} \frac{s(m)\, \varepsilon(m)}{\int dm\, \varepsilon(m)\, s(m)} = \int dm \frac{s'(m)}{\varepsilon(m)}$$

and pulling the m-independent factors $\varepsilon(x)$ and F to the LHS yields:

$$\int dm\, W(x,m)\, f(x,m) = N_s\, s(x) \quad \text{with} \quad W(x,m) = F\, \frac{w(m)}{\varepsilon(x)}$$

Discussion

- use of $w(m)$ in efficiency correction requires factorization
  - ➜ physics must factorize into $s(x) \cdot s(m)$ and $b(x) \cdot b(m)$
  - ➜ efficiencies must factorize into $\varepsilon(x) \cdot \varepsilon(m)$
- correct weight: $W(x, m) = F\, w(m)/\varepsilon(x) \neq w(m)/(\varepsilon(x)\varepsilon(m))$
- per event only the efficiency $\varepsilon(x)$ is used
- $\varepsilon(m)$ enters only via the global factor F – not per event
- $\varepsilon(m)$ per event destroys normalization and orthogonality of $w(x)$

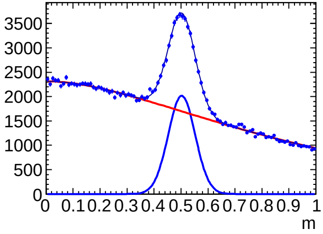
$$\int dm\, w(m)\, b'(m) = 0 \quad ➜ \quad \int dm\, \frac{w(m)}{\varepsilon(m)}\, b'(m) \neq 0$$

$$\int dm\, w(m)\, s'(m) = 1 \quad ➜ \quad \int dm\, \frac{w(m)}{\varepsilon(m)}\, s'(m) \neq 1$$

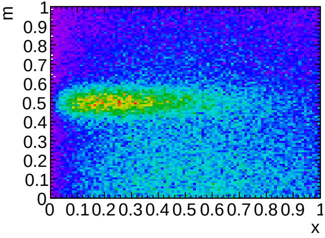
➜ incomplete background subtraction and wrong normalization!

Example - factorizing efficiencies

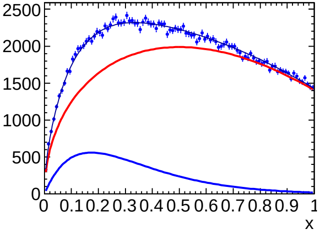
➜ *efficiency:* $\varepsilon(x, m) = ((m + 0.5)/1.5)((1.5 - x)/1.5)$



signal+background

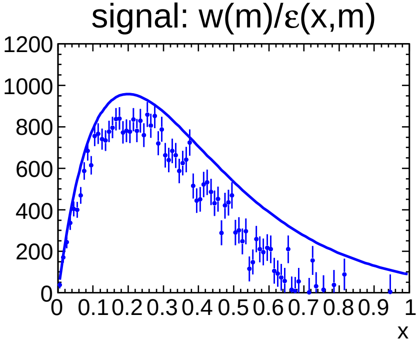


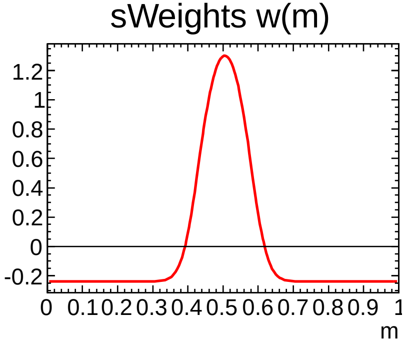
generated distributions



Naive event-by-event efficiency correction

➜ *sWeights determined for observed densities*



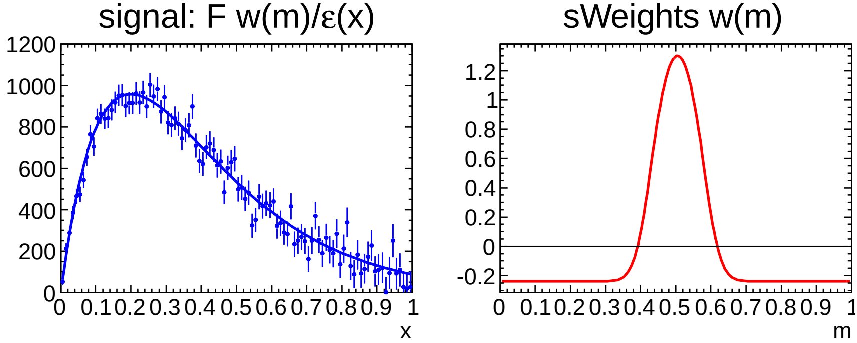


➜ events (x, m) with weights $W(m, x)$: `hx->Fill(x,W(m,x))` where

$$W(m, x) = \frac{w(x)}{\varepsilon(x)\varepsilon(m)}$$

Properly extracted result - version 1

➜ *sWeights determined for observed densities*

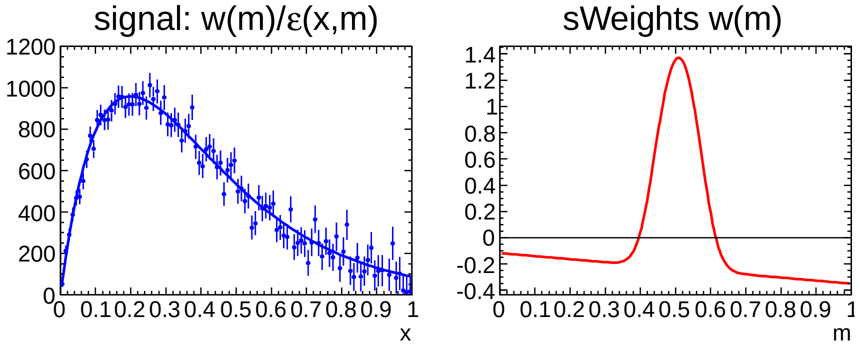


➜ events (x, m) with weights $W(m, x)$: `hx->Fill(x,W(m,x))` where

$$W(m,x) = \frac{w(m)}{\varepsilon(x)} \cdot F \quad \text{with} \quad F = \int dm\, \frac{s'(m)}{\varepsilon(m)}$$

Properly extracted result - version 2

➜ *better precision sWeights determined for true densities*

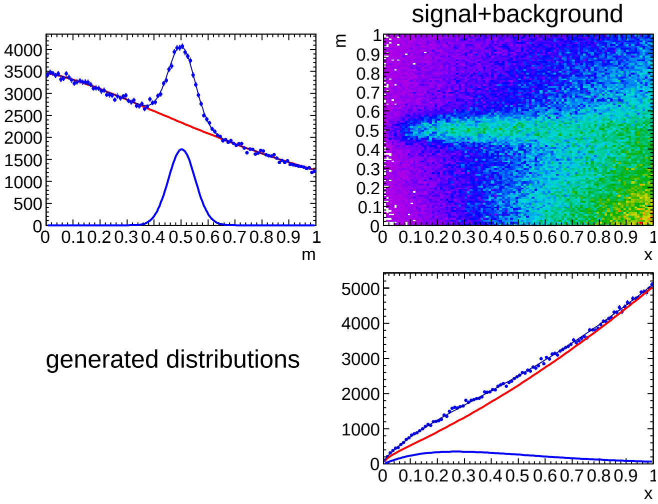


- determined by weighting measurements by $1/\varepsilon(x,m)^2$
- event weights:

$$W(m,x) = \frac{w(m)}{\varepsilon(x,m)}$$

Example - non-factorizing efficiencies

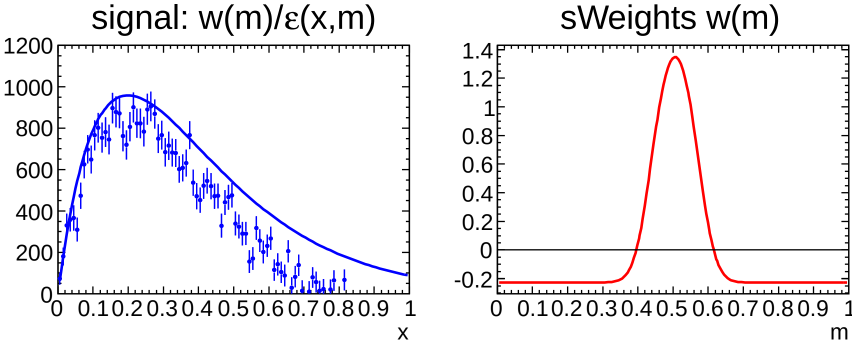
- *efficiency:* $\varepsilon(x, m) = (x + m)/2$



generated distributions

Naive event-by-event efficiency correction

➜ *sWeights determined for observed densities*

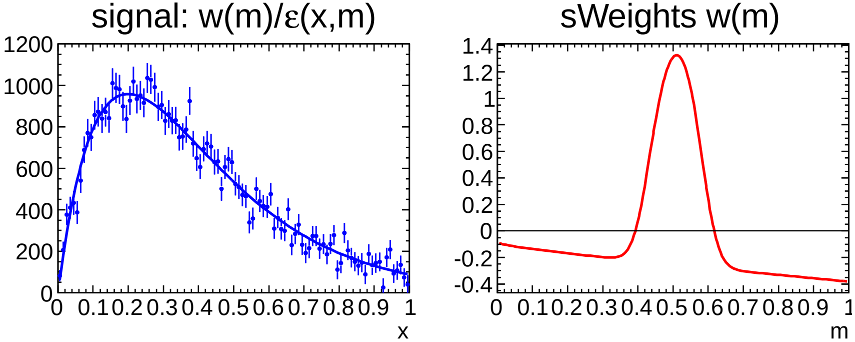


➜ events (x, m) with weights $W(m, x)$: `hx->Fill(x,W(m,x))` where

$$W(m, x) = \frac{w(x)}{\varepsilon(x, m)}$$

Properly extracted result - version 2

➜ *sWeights determined for true densities*



- determined by weighting measurements by $1/\varepsilon(x,m)^2$
- event weights:

$$W(m,x) = \frac{w(m)}{\varepsilon(x,m)}$$

further checks (for general efficiencies) ➜

Relation to arXiv:physics/0402083 (i)

➜ *extract normalizations by a fit to the data*

- signal and background shapes $s(m)$ and $b(m)$ are known
- extract normalization from a least squares fit
  - ➜ assume narrow bins m and x
  - ➜ bin content and variances are n_i and σ_i^2

$$\chi^2 = \sum_i \frac{(n_i - N_s\, s(m_i)\, \Delta m - N_b\, b(m_i)\, \Delta m)^2}{\sigma_i^2} \; .$$

When events are weighted with the inverse of the efficiency one has:

$$n_i = \sum_j \Delta x \Delta m \frac{f(x_j, m_i)}{\varepsilon(x_j, m_i)} \rightarrow n_i = \Delta m \int dx \, \frac{f(x, m_i)}{\varepsilon(x, m_i)} = \Delta m \; p(m_i)$$

$$\sigma_i^2 = \sum_j \Delta x \Delta m \frac{f(x_j, m_i)}{\varepsilon^2(x_j, m_i)} \rightarrow \sigma_i^2 = \Delta m \int dx \, \frac{f(x, m_i)}{\varepsilon^2(x, m_i)} = \Delta m \; q(m_i)$$

Relation to arXiv:physics/0402083 (ii)

function to be minimized:

$$\chi^2 = \sum_i \frac{(n_i - N_s\, s(m_i)\,\Delta m - N_b\, b(m_i)\,\Delta m)^2}{\sigma_i^2}$$

$$= \int dm\, \frac{(p(m) - N_s\, s(m) - N_b\, b(m))^2}{q(m)}$$

covariance matrix of normalizations N_s and N_b:

$$C_{uv}^{-1} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial N_u \partial N_v} = \int dm\, \frac{u(m)\, v(m)}{q(m)}$$ and thus

$$C = \begin{pmatrix} W_{ss} & W_{sb} \\ W_{sb} & W_{bb} \end{pmatrix}^{-1} = \frac{1}{W_{ss}\,W_{bb} - W_{sb}^2} \begin{pmatrix} W_{bb} & -W_{sb} \\ -W_{sb} & W_{ss} \end{pmatrix}$$

❖ sWeights are related to the covariance matrix of the normalization fit

Sanity checks

1. normalization of the signal distribution:

$$\sum_{\text{all events}} \frac{w(m)}{\varepsilon(x,m)} = \int dx\, dm\, \frac{w(m)}{\varepsilon(x,m)} f(x,m)$$

$$= \int dm\, w(m) \left[N_s\, s(m) + N_n\, b(m) \right] = N_s$$

2. variance of the normalization

$$\sum_{\text{all events}} \left(\frac{w(m)}{\varepsilon(x,m)} \right)^2 = \int dx\, dm\, \left(\frac{w(m)}{\varepsilon(x,m)} \right)^2 f(x,m)$$

$$= \int dm\, w^2(m)\, q(m) = \int dm\, w(m) \left[\alpha\, s(m) + \beta\, b(m) \right] = \alpha = C_{ss}$$

- normalization and variance of the signal spectrum are the same as obtained in the fit of the normalizations to the discriminant variable
- if the normalization fit had optimal precision, then sWeights are optimal to extract the signal as a function of the control variable

Summary on sWeights

- sWeights are functions orthogonal to the background density
  - ➜ signal & background are separable in a discriminant variable m
  - ➜ sWeights project out the signal component in a control variable x
  - ➜ sWeights do not quantify “signalness” ($w(m) < 0$ is allowed)
  - ➜ discriminant and control variables have to be independent
- for $\varepsilon(x, m) = \varepsilon(x) \cdot \varepsilon(m)$ sWeights for efficiency corrections can be determined from the observed densities $s'(m)$ and $b'(m)$

$$W(m, x) = F \frac{w(m)}{\varepsilon(x)} \neq \frac{w(m)}{\varepsilon(x, m)} \quad \text{with} \quad F = \int dm \, \frac{s'(m)}{\varepsilon(m)}$$

- for $\varepsilon(x, m) \neq \varepsilon(x) \cdot \varepsilon(m)$ sWeights for efficiency corrections must be determined from the corrected densities $s(m)$ and $b(m)$
  - ➜ to extract the signal use event-by-event weights $w(m)/\varepsilon(x, m)$
- everything also holds if x and m are multi-dimensional