Statistical and numerical methods of data analysis I

V. Blobel
Universität Hamburg

Desy, **June** 14 (+21, 28), **2001**

"The theory of probability and statistics is concerned with situtations in which an observation is made and which contains an element of inherent variability outwith the observer's control. He knows that if he repeated the observation under conditions which were identical in so far as he could control them, the second observation would not necessarily agree exactly with the first." (S.D. Silvey, Statistical inference)

Statistical and numerical methods of data analysis

- Current physics experiments take data at a high rate: many Mbytes per second over a long period of time
- The essential result may be a single number, or a limit
- Statistical and numerical methods are needed in order to extract meaningful information from experimental data.

Examples for new data from HEP

Amplitude of the CP-violating asymmetry:

$\sin 2\beta$	$=$	$0.34 \pm 0.20(\text{stat}) \pm 0.05(\text{syst})$	BABAR
$\sin 2\beta$	$=$	$0.58^{+0.32}_{-0.34}{}^{+0.09}_{-0.10}$	BELLE

Errors, uncertainties and ...

The unavoidable uncertainty is unfortunately called "error".

Guidelines*) for evaluating and expressing the uncertainty of NIST measurement results:

The uncertainty of the result of a measurement generally consists out of several components which ... may be grouped into two categories according to the method used to estimate their numerical values:

A. those which are evaluated by statistical methods,

B. those which are evaluated by other means.

Each component of uncertainty that contributes to the uncertainty of a measurement result is represented by an estimated standard deviation, termed **standard uncertainty** ...

... the term "error" is avoided.

---

*) National Institute of Standards and Technology (NIST), USA, Technical note 1297, by B. N. Taylor

Errors, uncertainties and bugs

Note: the probability content of a certain standard deviation interval depends on the type of the distribution.

Normal distribution: probability for single value outside ± 3 standard deviations is 0.27 %.

But: basic measurement result in HIGH ENERGY PHYSICS follow the normal distribution only up to one standard deviation*): the probability for single value outside ± 3 standard deviations is 5.4 %.

---

*) A. P. Bukhvostov, On the probability distribution of the experimental results, `hep-ph/9705387`

Example: neutrino mass

Data (1996) on the electron-neutrino mass m_ν from tritium β decay

Experiment	m_ν^2	(eV^2)	m_ν 95 % C.L. limit
Tokyo	-65	$\pm 85 \pm 65$	< 13.0 eV
Los Alamos	-147	$\pm 68 \pm 41$	< 9.3 eV
Zürich	-24	$\pm 48 \pm 61$	< 11.7 eV
Mainz	-39	$\pm 34 \pm 15$	< 7.2 eV
Livermore (LLNL)	-130	$\pm 20 \pm 15$	< 7.0 eV
Moskau (INR)	-22	± 4.8	< 4.4 eV

First error of m_ν^2 is statistical, second systematic. Upper limits for the neutrino mass are calculated using a certain method for 95 % confidence level.

PDG 2000: $m_\nu < 3\mathrm{eV}$ Interpretation of tritium β decay is complicated by anomalies near the endpoint ("unknown effects cause an accumulation of events near the endpoint of the electron spectrum"), and the limits are not without ambiguity.

Interpretation of experimental results

First step: check of internal consistency – can the result be accepted?

- measured values outside the physical region?
- if yes, how far?
- consistent with general rules?
- consistent with other experiments?

Second step: derivation of two-sided or one-sided limits, if experimental result acceptable

Both steps are based on the same data – no unique statistical method in difficult cases.

Example: Number of light neutrinos

Mark II collaboration at SLAC (1989): analysis which assumed Standard Model couplings but which fit for the Z mass and the number of neutrinos N_ν with result:

$$\begin{array}{lrcll} & N_\nu & = & 2.8 \pm 0.6 & \\ \text{95 \% C.L. limit} & N_\nu & < & 3.9 & \text{excludes fourth massless neutrino} \end{array}$$

- Existence of at least three types on neutrinos was well established
- Ability to exclude $N_\nu = 4$ in this experiment due to downward fluctuation of measured N_ν from 3 to 2.8

PDG 2000: $N_\nu = 2.994 \pm 0.012$ from Standard Model fits to LEP data.

Literature

R. Barlow Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences, Wiley, 1989

V. Blobel and E. Lohrmann Statistische und numerische Methoden der Datenanalyse, Teubner, 1998

S. Brandt Datenanalyse, Spektrum, 1999

G. Cowan Statistical Data Analysis, Oxford, 1998

L. Lyons Statistics for nuclear and particle physicists, Cambridge, 1989

W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in [. . . C, Fortran] (various progr. lang.), The Art of Scientific Computing

Literature on algorithms

Robert Sedgewick Algorithms, Addison-Wesley (various editions)

Algorithms: complex problems are divided into smaller problems, which hopefully can be solved by standard algorithms.

Important standard problems in data organisation

- sorting of n items according to a key (ascending, descending)
- search in a list of n items

The dependence $f(n)$ of the computer time (or the number of operations) on n for larger values of n (leading dependence) is often an essential property of an algorithm.

Typical leading dependences

Sometimes important:

$A(n) =$ average behaviour $\quad W(n) =$ worst case behaviour

Typical leading dependences on the number n of objects in algorithms with examples of algorithms for sorting and searching.

function	sorting	searching
$f(n) = \text{const.}$		hash search
$f(n) = \text{const.} \cdot \log n$		binary search
$f(n) = \text{const.} \cdot n$		sequential search
$f(n) = \text{const.} \cdot n \log n$	quicksort, heapsort	
$f(n) = \text{const.} \cdot n^2$	quicksort with sorted input	
$f(n) = \text{const.} \cdot 2^n$		

Topics for the three lectures

- Probability and statistics

- Parameter estimation and information
- Least squares methods

- Maximum likelihood method and function minimization
- Confidence intervals and limits

Probability and statistics

- Probability
- Frequentists and Bayesian view
- Bayes theorem
- Application of Bayes theorem

- Expectation values
- Standard distributions
- Transformations and error propagation
- Monte Carlo

- Fitting correlated data
- The unfolding problem and negative correlations
- Correlations beween data points

Relative frequency – the Frequentists view

A, B, ... are outcomes of a repeatable experiment.

Probability $P(A)$ is limit of frequency:

$$P(A) = \lim_{n\to\infty} \frac{\text{outcome is } A}{n}$$

This frequency interpretation of probability is usually most natural for data analysis in physics.

Examples for frequencies:

decay probability of $J/\Psi \to \mu^+\mu^-$ ($\Gamma_i/\Gamma = 6.01\%$)
detector acceptance (determined by MC)
upper limit of electron neutrino mass of 3 eV (95 % C.L.)

Subjective probability – the Bayesian view

A, B, ... are hypotheses (i.e. statements, that are either true or false).

Probability $P(A) =$ degree of belief that A is true.

Subjective probability allows a more natural treatment of phenomena that are not repeatable.

Examples for hypotheses:

the particle is a K meson
nature is supersymmetric
the Higgs mass is 115 GeV
it will rain tomorrow

Sometimes one has to take decisions on the basis of degree of belief (will it rain tomorrow?).

Bayes' theorem ...

... relates the two conditional probabilities $P(\mathcal{A}|\mathcal{B})$ and $P(\mathcal{B}|\mathcal{A})$:

$$\boxed{P(\mathcal{A}|\mathcal{B}) = P(\mathcal{B}|\mathcal{A}) \cdot \frac{P(\mathcal{A})}{P(\mathcal{B})}}$$

which follows directly from

$$P(\mathcal{A} \text{ and } \mathcal{B}) = P(\mathcal{A}) \cdot \; P(\mathcal{B}|\mathcal{A}) = P(\mathcal{B}) \cdot \; P(\mathcal{A}|\mathcal{B})$$

Generalized to n classes of events with properties $\mathcal{A}_i$, $(i = 1, 2, \ldots n)$ and $P(\mathcal{B}) = \sum_i P(\mathcal{B}|\mathcal{A}_i) \cdot P(\mathcal{A}_i)$

$$P(\mathcal{A}_i|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A}_i) P(\mathcal{A}_i)}{\sum_{i=1}^{n} P(\mathcal{B}|\mathcal{A}_i) \cdot P(\mathcal{A}_i)}$$

There is a special interpretation of the Bayes' theorem with controversial discussions → subjective probability

Reverend Thomas Bayes (1702-1761), "An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Society **53** (1763) 370, reprinted in Biometrika **45** (1958) 293

Application of Bayes theorem

A test for a rare, very dangerous illnes K is made.
The result is positive! What does it mean?

The illness K is rare: a-priori $P(K) = 0.001$ $\quad P(\bar{K}) = 0.999$

The test result is not safe, there are known error probabilities:

in case of illness	$P(+|K) = 0.98$	$P(-|K) = 0.02$
no illness	$P(+|\bar{K}) = 0.03$	$P(-|\bar{K}) = 0.97$

Application of Bayes theorem to calculate $P(K|+)$:

$$P(K|+) = \frac{P(+|K) \cdot P(K)}{P(+|K) \cdot P(K) + P(+|\bar{K}) \cdot P(\bar{K})} = \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.03 \cdot 0.999} = 0.033$$

Thus the probability to have illness K after a positive test is 3.3 %.

Expectation values

The *expectation value* of the function $h(x)$, where x is a random variable from a density $f(x)$, is defined by

$$E[h] = \int_{-\infty}^{+\infty} h(x)\, f(x)\, \mathrm{d}x \qquad E[h] = \sum_{i=a}^{b} h(r_i)\, P(r_i)\,.$$

Important special cases are the expectation value of x or k, or mean value, and the *variance* of x or k:

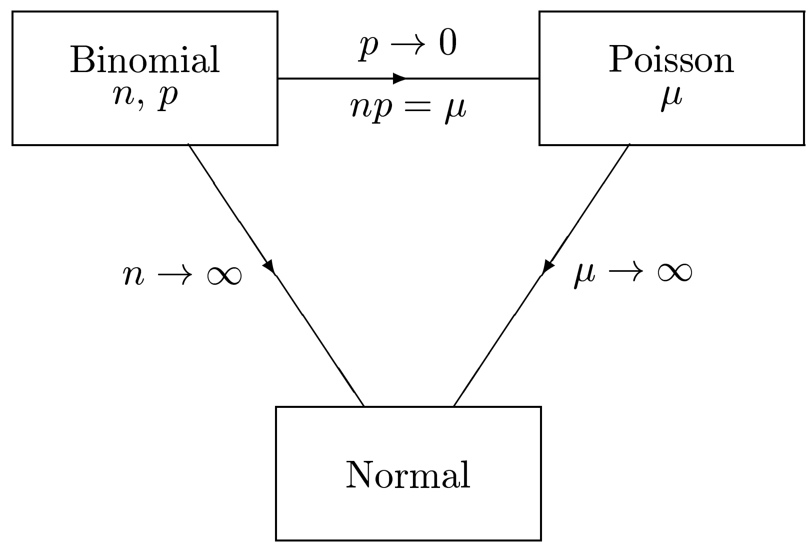
$$E[x] = \int_{-\infty}^{+\infty} x\, f(x)\, \mathrm{d}x = \langle x \rangle \qquad E[k] = \sum_{i=a}^{b} k_i P(k_i) = \langle k \rangle$$

$$V[x] = \int_{-\infty}^{+\infty} (x - \langle x \rangle)^2\, f(x)\, \mathrm{d}x \qquad V[k] = \sum_{i=a}^{b} (k_i - \langle k \rangle)^2\, P(k_i)$$

standard deviation $\qquad \sigma = \sqrt{V}$

Relation between standard distributions

The Binomial distribution follows from combinatorics.



Poisson and normal distribution follow from the Binomial distribution.

Binomial distribution

Single event with two possible outcomes:	**"success"** with probability p, and **"failure"** with probability $q = 1 - p$.

Probability to observe in n experiments exactly k times "success" is given by the Binomial distribution:

$$\boxed{P_n(k) = \binom{n}{k} p^k \, (1-p)^{n-k} \qquad k = 0,\, 1,\, \ldots n}$$

The factor $\binom{n}{k}$ ist the number of possibilities to select k elements out of total n different elements.

$$\text{Expectation value } = E[k] = n\,p \qquad \text{Variance } = V[k] = n\,p(1-p)$$

Law of large numbers

The accuracy of an estimate $h (= k/n)$ of the probability in a Binomial process increases with n:

$$V[h] = V[k/n] = \frac{1}{n^2} V[k] = \frac{1}{n^2} n\,p\,(1-p)$$

Because of $p\,(1-p) \leq 1/4$ for any value of p, one can estimate

$$V[h] \leq \frac{1}{4}\frac{1}{n} \qquad \sigma(h) \propto \frac{1}{\sqrt{n}}$$

The error of the estimate h can be made as small as required by choosing a sufficiently high value of n.

Poisson distribution

Binomial distribution with $n \to \infty$, $p \to 0$, but $n \cdot p \to \mu$ (fixed):

$$\boxed{P_\mu(k) = \frac{\mu^k}{k!}\, e^{-\mu} \qquad k = 0, 1, \ldots}$$

Expectation value $E[k] = \mu$ Variance $V[k] = \mu$

Case of a process with constant event rate:

- number of events k in time interval follows a Poisson distribution, and
- time difference t between events follows exponential distribution:

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

Normal distribution - the most important continous distribution

Theoretically almost anything (which is a sum of many random contributions) follows (approximately) the normal distribution ($\to$ central limit theorem).

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}} \qquad x \in (-\infty, \infty)}$$

Expectation value $E[x] = \mu$ Variance $V[x] = \sigma^2$

$$\begin{aligned} P(|x-\mu| < 1 \cdot \sigma) &= 68.26\% \\ P(|x-\mu| < 2 \cdot \sigma) &= 95.45\% \\ P(|x-\mu| < 3 \cdot \sigma) &= 99.73\% \end{aligned}$$

This relation between standard deviations and probability in % is valid only for the exact normal distribution!

χ^2 distribution, for statistical tests ("goodness of fit")

If $x_i \in N(0,1)$ for $i = 1, 2, \ldots n$, then the sum

$$u = \sum_{i=1}^{n} x_i^2$$

follows a χ^2 distribution $f_n(u)$ with n degrees of freedom:

$$\boxed{f_n(u) = \frac{\frac{1}{2}\left(\frac{u}{2}\right)^{n/2-1} e^{-u/2}}{\Gamma(n/2)}} \qquad \text{density}$$

Example: weighted sum of squares of residuals, minimized in the least squares method, follows the χ^2 distribution.

Transformation of a single variable

A function $y = y(x)$ of a random variable x is itself a random variable.

Consider intervals $(x, x + \mathrm{d}x)$ and $(y, y + \mathrm{d}y)$:

$$f_x(x)\,|\,\mathrm{d}x| = f_y(y)\,|\,\mathrm{d}y|$$

If the inverse of $y(x)$ is not unique, include all branches.

1. Example: $y = x^2$ $\qquad x = \pm\sqrt{y} \qquad \mathrm{d}x = \pm\frac{\mathrm{d}y}{2\sqrt{y}}$

   transf. distribution $\qquad f_y(y) = \dfrac{1}{2\sqrt{y}}\left(f_x(+\sqrt{y}) + f_x(-\sqrt{y})\right)$

2. Example: Linear transformation $y = ax + b$ $\qquad x = (y-b)/a$

transf. distr. $f_y(y) = \dfrac{1}{a} f_x\left(\dfrac{y-b}{a}\right) \qquad E[y] = a{\cdot}E[x]{+}b \qquad V[y] = a^2 V[x]$

Error propagation

Nonlinear transformation $y = y(x)$ may be complicated.

Simpler: transform mean and variance based on expansion

$$y(x) = y(\langle x \rangle) + (x - \langle x \rangle) \left. \frac{\mathrm{d}y}{\mathrm{d}x} \right|_{x=\langle x \rangle} + \frac{1}{2}(x - \langle x \rangle)^2 \left. \frac{\mathrm{d}^2y}{\mathrm{d}x^2} \right|_{x=\langle x \rangle} + \ldots$$

Taking terms up to second order with correction term

$$\text{mean} \quad \langle y \rangle \approx y(\langle x \rangle) \qquad + \frac{1}{2}\sigma_x^2 \left. \frac{\mathrm{d}^2y}{\mathrm{d}x^2} \right|_{x=\langle x \rangle}$$

$$\text{Variance} \qquad V[y] \approx \left(\frac{\mathrm{d}y}{\mathrm{d}x} \right)^2 \cdot V[x] \qquad \text{error propagation}$$

n-dimensional random variable

Consider random vector $\boldsymbol{x}$ with components x_i, $(i = 1, 2, \ldots n)$.

Probability density function:

$$f(x_1, x_2, \ldots x_n) = f(\boldsymbol{x})$$

Expectation values with $\langle \boldsymbol{x} \rangle = E[\boldsymbol{x}]$:

$$E[x_i] = \langle x_i \rangle = \int x_i f(x_1, x_2, \ldots, x_n) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_n$$

$$= \int x_i \cdot f(\boldsymbol{x}) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_n \, .$$

Variance

Generalization of the variance:

covariance matrix $\boldsymbol{V} = \boldsymbol{V}[\boldsymbol{x}] = E\left[(\boldsymbol{x} - E[\boldsymbol{x}])(\boldsymbol{x} - E[\boldsymbol{x}])^T\right]$

with diagonal and not-diagonal elements:

$$\begin{aligned} V_{ii} &= \sigma_i^2 = \text{var}(x_i) = \int (x_i - \langle x_i \rangle)^2 \cdot f(\boldsymbol{x}) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_n \\ V_{ij} &= \sigma_{ij} = \text{cov}(x_i, x_j) = \int (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \cdot f(\boldsymbol{x}) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \ldots \mathrm{d}x_n \, . \end{aligned}$$

correlation coefficient ρ_{ij} $\qquad \sigma_{ij} = \rho_{ij} \, \sigma_i \sigma_j \qquad -1 \le \rho_{ij} \le +1$

x_i and x_j independent $\quad \rightarrow \quad \rho = 0 \quad$ reverse not true!

$\rho = +1$ or $-1 \quad \rightarrow \quad x_i$ and x_j completely correlated

Multivariate normal distribution

$$\boxed{f(\boldsymbol{x}) = \frac{|\boldsymbol{W}|^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})^T \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})\right)}$$

with $\boldsymbol{W} = \boldsymbol{V}^{-1}$ and $|\boldsymbol{W}| = \det \boldsymbol{W} = |\boldsymbol{V}^{-1}|$.

Parameters of the n-dimensional normal distribution are just

- the first moment (n-vector of means $\boldsymbol{\mu}$), and
- the second central moment ($n \times n$ covariance matrix $\boldsymbol{V}$).

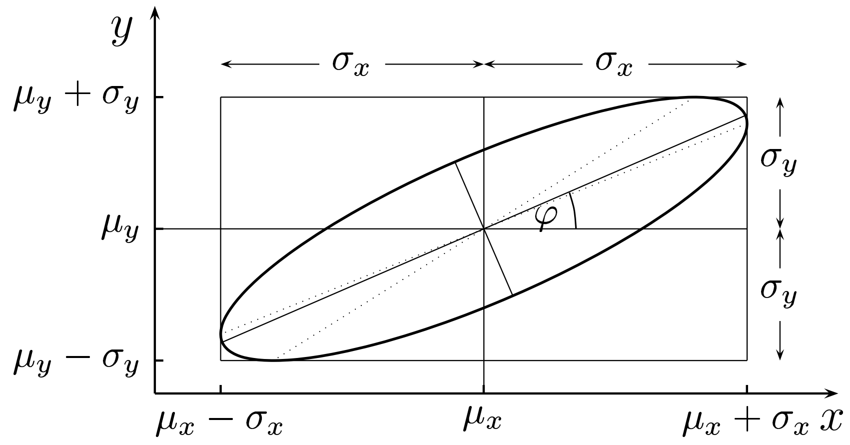
"Distance" $\boldsymbol{\Delta}$ from mean measured by

Mahalanobis distance $\quad S = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta}$

S follows a χ^2 distribution with n degrees of freedom.

Two-dimensional normal distribution - contour

Elliptical 1σ contour (38 % probability content) with mean values μ_x, μ_y, standard deviations σ_x, σ_y and $\rho = +0.8$.



Main axes of the ellipse with the angles φ und $\varphi + \pi/2$ given by $\tan 2\varphi = \pm(2\rho\sigma_x\sigma_y)/(\sigma_x^2 - \sigma_y^2)$.
Dotted lines = conditional means $x - \mu_x = (y - \mu_y)\rho\sigma_x/\sigma_y$ and $y - \mu_y = (x - \mu_x)\rho\sigma_y/\sigma_x$.

Linear transformation $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x}$

n-by-m matrix $\boldsymbol{B}$ with elements $B_{ik} = \partial y_i/\partial x_k$ transforms n-vector $\boldsymbol{x}$ with mean $\boldsymbol{\mu_x}$ and covariance matrix $\boldsymbol{V}(\boldsymbol{x})$ to m-vector $\boldsymbol{y}$

$$\boldsymbol{\mu_y} = E\left[\boldsymbol{y}\right] = E\left[\boldsymbol{B}\boldsymbol{x}\right] = \boldsymbol{B}E\left[\boldsymbol{x}\right] = \boldsymbol{B}\boldsymbol{\mu_x}$$

(expectation is a linear operator)

Covariance matrix: "error" propagation

$$\begin{aligned}\boldsymbol{V}\left[\boldsymbol{y}\right] = E\left[(\boldsymbol{y} - \boldsymbol{\mu_y})(\boldsymbol{y} - \boldsymbol{\mu_y})^T)\right] &= E\left[(\boldsymbol{B}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu_x})(\boldsymbol{B}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\mu_x})^T\right] \\ &= E\left[\boldsymbol{B}(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^T\boldsymbol{B}^T\right] \\ &= \boldsymbol{B} \cdot E\left[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^T\right]\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{V}\left[\boldsymbol{x}\right]\boldsymbol{B}^T\end{aligned}$$

$$\boxed{\boldsymbol{V}\left[\boldsymbol{y}\right] = \boldsymbol{B}\boldsymbol{V}\left[\boldsymbol{x}\right]\boldsymbol{B}^T}$$

Random number generation

Basis for Monte Carlo simulations are uniformly distributed random numbers in $[0,1]$.

$$u_j \in U(0,1) \quad \text{uniform on } (0,1) \qquad\qquad \texttt{u = ran()}$$

Required: computer algorithm to generate sequence of truely uncorrelated random numbers with large period.

Example: linear congruential generator, based on the recurrence relation

$$I_j = (a \cdot I_{j-1} + c) \mod m \qquad\qquad j = 1, 2, \ldots$$

Integer random numbers I_j between 0 and $m-1$ are generated, and converted to a real number $u = I_j/m$.

Test for generators

Basic test is for uniform distribution (histogram).

Up-down- or sequence test: generate $n+1$ random numbers and determine number of sequences $N(k)$ of increasing or decreasing random numbers, satisfying

$$\sum_{k=1}^{n} k \cdot N(k) = n \qquad\qquad \sum_{k=1}^{n} N(k) = \frac{(2n-1)}{3}\,.$$

For truely uncorrelated random numbers u_j the following values are expected:

$$N(1) = \frac{5n+1}{12} \qquad\qquad N(2) = \frac{11n-14}{60}$$

$$N(k) = \frac{2\left[\left(k^2+3k+1\right)n - \left(k^3+3k^2-k-4\right)\right]}{(k+3)!}$$

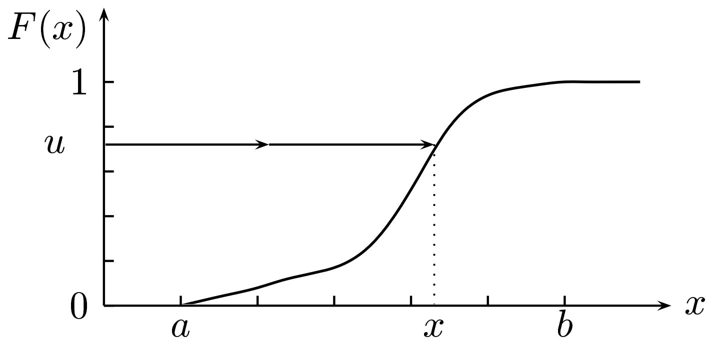
Especially the numbers $N(k)$ from larger values of k allow sensitive tests of the random number generator.

Inverse transfrom for continuous variable

If x is chosen from density $f(x)$, then the cumulative probability $F(x)$ is itself a rn which will occur with uniform density $g(u)$ on $[0,1]$:

$$f(x)\mathrm{d}x = g(u)\,\mathrm{d}u \qquad \int_{-\infty}^{x} f(x')\,\mathrm{d}x' = F(x) = u.$$

The equation $u = F(x)$ is solved for the variable x by $x = F^{(-1)}(u)$ with the inverse function $F^{(-1)}(u)$, if possible.



Example: $f(x)\,\mathrm{d}x = \lambda e^{-\lambda x} \quad \rightarrow \quad x = -\frac{1}{\lambda}\ln u$

Fitting correlated data

Problem of fitting correlated data discussed by G. D'Agostini:
Paper "On the Use of the Covariance Matrix to Fit Correlated Data" (DESY 93-175), NIM A 346 (1994) 306 and `http://www-zeus.roma1.infn.it/~agostini`

Conclusion of the author: *Systematic uncertainties can produce* (in the context of the Least Squares Method) *results that "contradict expectations."*

The problem: fit a theoretical curve through experimental data by minimizing the χ^2

$$\chi^2 = \boldsymbol{\Delta}^T \boldsymbol{V}^{-1} \boldsymbol{\Delta}$$

where $\boldsymbol{V}$ is the error matrix and $\boldsymbol{\Delta}$ is the vector of the differences between the theoretical and the experimental values (standard method of least squares).

Numerical Example with discrepancy

Numerical example, discussed in the paper: " ...the result of two measurements, $8.0 \pm 2\%$ and $8.5 \pm 2\%$, have a 10 % common normalization error. Assuming that the two measurements refer to the same physical quantity, the best estimate of its true value can be obtained by fitting the points to a constant function. Minimizing χ^2 as defined above, with $\boldsymbol{V}$ estimated empirically by the data, one obtains a value of 7.87 ± 0.81, which is at least surprising[a], since the most probable result is outside the interval determined by the two measured values" ("Peelles pertinent Puzzle").

Origin of apparent discrepancy?

(D'A Paper) *"...shows that the effect is a direct consequence of the hypothesis used to estimate the empirical covariance matrix, namely the linearization on which the usual error propagation rely."*

---

[a] The weigthed mean ignoring common error is $8.235 \pm 1.4\%$.

Correct treatment

Is something wrong with the principle of least squares?

The model:

$$E[x_1] = \mu \qquad E[x_2] = \mu$$

(both single measurements are unbiased estimates of the same physical value)

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 + (\varepsilon\mu)^2 & (\varepsilon\mu)^2 \\ (\varepsilon\mu)^2 & \sigma_2^2 + (\varepsilon\mu)^2 \end{pmatrix}$$

(two contribution: a variance σ_1^2 and σ_2^2 for each single value, and a common normalization error, given as relative error).

Result from minimization:

$$8.235 \pm 1.4\% \pm 10\%$$

exactly identical to the expected result.

Approximate treatment

Replace $(\varepsilon\mu)$ by (εx_1) and (εx_2), respectively:

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 + (\varepsilon x_1)^2 & \varepsilon^2 x_1 x_2 \\ \varepsilon^2 x_1 x_2 & \sigma_2^2 + (\varepsilon x_2)^2 \end{pmatrix}$$

Result with "empirical" covariance matrix:

$$7.87 \pm 0.81 \qquad \rightarrow \text{ discrepancy}$$

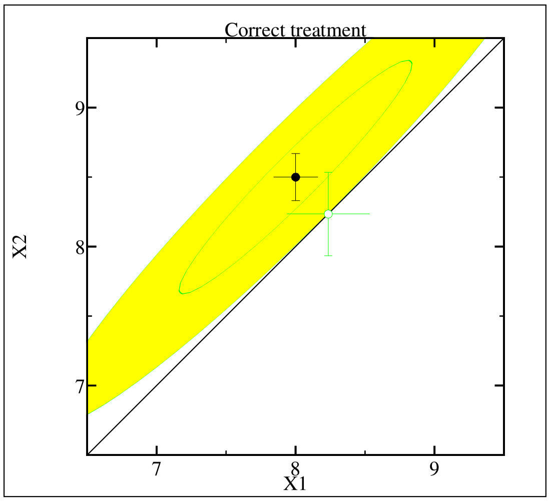
Geometrical explanation: axis of covariance ellipse (= direction of the eigenvector of the "empirical" covariance matrix) is tilted!

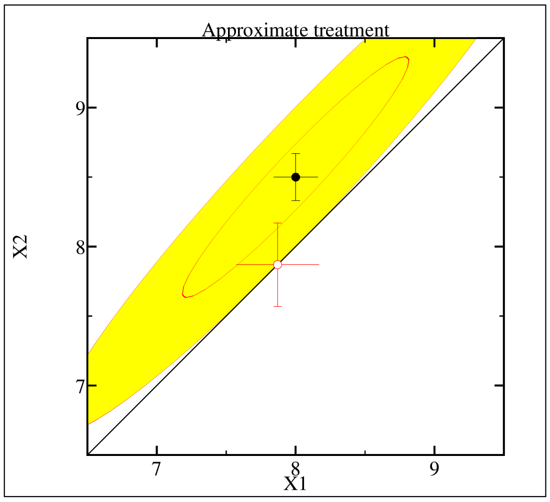
Note: ansatz must be wrong, because normalization changed.

Covariance ellipses displayed . . .

Correct treatment: axis of covariance ellipse under 45°

Approximate treatment: axis of covariance ellipse slightly tilted





The Poisson model

Is the model correct?

Data x_i are cross sections from event numbers n_i:

$$x_i = f \cdot n_i \qquad i = 1, 2$$

where the factor f may have relative uncertainty ε.

Optimal method: Maximum Likelihood based on Poisson statistic[a] with common mean value μ, with result

$$\hat{\mu} = \frac{1}{2}(n_1 + n_2)$$

(The sum of two Poisson distributed numbers follows the Poisson distribution).

Average cross section: $\hat{x} = 1/2\, f \cdot (n_1 + n_2)$.

[a] Example: the case $n_1 = 1$, $n_2 = 4$. The average based on the correct Poisson distribution is simply $1/2(n_1 + n_2) = 2.5$, while the weighted average from the two separate values $n_1 = 1 \pm 1$, $n_2 = 4 \pm 2$ would be 1.6.

The unfolding problem and negative correlations

Example: classification of events into NC (neutral current) and CC (charged current).

$$\begin{pmatrix} NC \\ CC \end{pmatrix}_{\text{obs}} = \begin{pmatrix} NC \to NC & CC \to NC \\ NC \to CC & CC \to CC \end{pmatrix} \cdot \begin{pmatrix} NC \\ CC \end{pmatrix}_{\text{true}}$$

or short $\boldsymbol{g} = \boldsymbol{A f}$ with matrix $\boldsymbol{A}$ containing *positive* probabilities.

Case of two possible true values and two possible observed values.

Unfolding by matrix inversion (*what else?*):

$$\boxed{\begin{pmatrix} NC \\ CC \end{pmatrix}_{\text{true}} = \boldsymbol{f} \;=\; \boldsymbol{A}^{-1}\boldsymbol{g} = \boldsymbol{A}^{-1} \begin{pmatrix} NC \\ CC \end{pmatrix}_{\text{obs}}}$$

i.e. a linear transformation from the observed values to the true values ($\to$ straightforward error propagation).

Unfolding by inversion: properties of the solution

Bias of the estimate $\hat{\boldsymbol{f}} = \boldsymbol{A}^{-1}\boldsymbol{g}_{\text{obs}}$:

$$E[\boldsymbol{g}_{\text{obs}}] = E[\boldsymbol{A}\boldsymbol{f}] = \boldsymbol{A}E[\boldsymbol{f}] = \boldsymbol{A}\cdot\boldsymbol{f}_{\text{true}}$$

$$E[\hat{\boldsymbol{f}}] = E[\boldsymbol{A}^{-1}\boldsymbol{g}_{\text{obs}}] = \boldsymbol{A}^{-1}E[\boldsymbol{g}_{\text{obs}}] = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{f}_{\text{true}} = \boldsymbol{f}_{\text{true}}$$

Thus the bias of the method is zero!

Error propagation from linear mapping $\hat{\boldsymbol{f}} = \boldsymbol{A}^{-1}\boldsymbol{g}_{\text{obs}}$:

$$\boldsymbol{V}_f = \boldsymbol{A}^{-1}\boldsymbol{V}_g\left(\boldsymbol{A}^{-1}\right)^T$$

Method has excellent statistical properties!

Unfolding by inversion

Quantitative example: $\varepsilon =$ probability for migration from original class into other class.

$$\begin{pmatrix} g_1 \\ g_2 \end{pmatrix}_{\text{obs}} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}_{\text{true}}$$

or short: $\boldsymbol{g} = \boldsymbol{A}\boldsymbol{f}$.

Unfolding with error propagation:

$$\boxed{\begin{array}{rcl} \hat{\boldsymbol{f}} &=& \boldsymbol{A}^{-1}\boldsymbol{g} \\ \boldsymbol{V}_f &=& \boldsymbol{A}^{-1}\boldsymbol{V}_g\boldsymbol{A}^{-1} \end{array}}$$

Inverse matrix (with *negative* off-diagonal elements):
(... becomes singular for $\varepsilon = 0.5$)

$$\boldsymbol{A}^{-1} = \frac{1}{1-2\varepsilon}\begin{pmatrix} 1-\varepsilon & -\varepsilon \\ -\varepsilon & 1-\varepsilon \end{pmatrix}$$

Resulting covariance matrix:

$$\boldsymbol{V}_f = \frac{1}{(1-2\varepsilon)^2}\begin{pmatrix} (1-\varepsilon)^2 g_1 + \varepsilon^2 g_2 & -\varepsilon(1-\varepsilon)(g_1+g_2) \\ -\varepsilon(1-\varepsilon)(g_1+g_2) & (1-\varepsilon)^2 g_2 + \varepsilon^2 g_1 \end{pmatrix}$$

A numerical example for a measurement

$$g_1 = 100 \pm 10 \qquad g_2 = 81 \pm 9 \qquad V_g = \begin{pmatrix} 100 & 0 \\ 0 & 81 \end{pmatrix}$$

Result of unfolding by inversion;

ε	$\hat{f}_1$	$\hat{f}_2$	ρ
0	100 ± 10	81 ± 9	0
0.1	102.4 ± 11.3	78.6 ± 10.2	-0.22
0.4	138 ± 35	43 ± 33	-0.92
0.5	$\pm\infty$	$\pm\infty$	singular

Effect in detail for $\varepsilon = 0.4$:

$$A^{-1} = \frac{1}{1-2\varepsilon} \begin{pmatrix} 1-\varepsilon & -\varepsilon \\ -\varepsilon & 1-\varepsilon \end{pmatrix} = \begin{pmatrix} 3 & -2 \\ -2 & 3 \end{pmatrix}$$

for example $\hat{f}_1 = 3 \cdot g_1 - 2 \cdot g_2 = g_1 + 2 \cdot (g_1 - g_2) = 300 - 162 = 138$

with $\sigma = \sqrt{3^2 \cdot 100 + 2^2 \cdot 81} = 35 \qquad (\gg 10)$

Bin-by-bin acceptance correction method

The bin-by-bin acceptance correction method work iteratively, starting from an initial estimate for the result $\boldsymbol{f}$:

1. Calculate (e.g. by Monte Carlo methods) for previous result $\boldsymbol{f}$

$$\boldsymbol{g}^{\text{rec}} = \boldsymbol{A}\,\boldsymbol{f}^{\text{gen}}$$

2. Correct observed data by ratio **gen:rec**:

for each bin i: $$f_i = \left(\frac{f_i^{\text{gen}}}{g_i^{\text{rec}}} \right) g_{i,\text{obs}}$$

Correlations between bins are ignored (bin-by-bin)

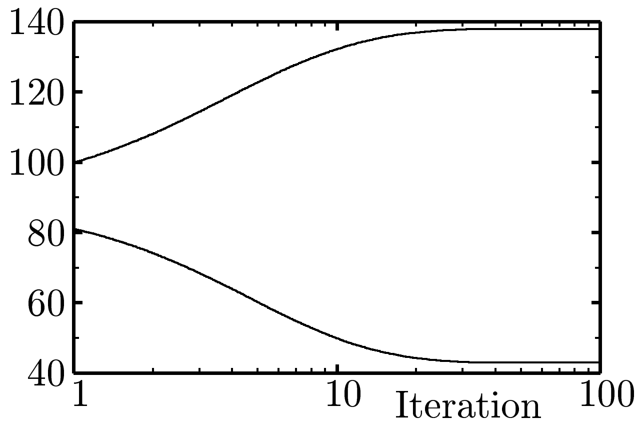
Often final values of factor $f_i^{\text{gen}}/g_i^{\text{rec}}$ are assumed to be error-free.

Properties? Convergence?

From a thesis: "The bin-by-bin unfolding method, which is used in this analysis, considers only net migration into or out of the bin under consideration. The bins are no longer correlated . . ." – this is not true!

. . . the bin-by-bin acceptance correction method

Parameter f_1 and f_2 versus iteration number (for $\varepsilon = 0.4$):



The method converges (of course) to the *same* values as the inversion method, but extremely slow (linear convergence).

$f_1 = 1.38\times(100\pm10) = 138\pm13.8$ $\qquad$ $f_2 = 0.53\times(81\pm9) = 43\pm4.8$

The errors are too small (correlation by finite resolution is ignored).

Correlations between data points

Case of two *correlated* data points:

$$\text{Covariance matrix} \quad \boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$\begin{aligned} \sigma_{1,2} &= \text{standard deviations of the variables} \\ \rho &= \text{correlation coefficient} \qquad -1 < \rho < +1 \end{aligned}$$

The χ^2 value for a deviation $\boldsymbol{\Delta}$ from a theoretical prediction is given by:

$$\boxed{\chi^2 = \boldsymbol{\Delta}^T \left(\boldsymbol{V}^{-1}\right) \boldsymbol{\Delta}}$$

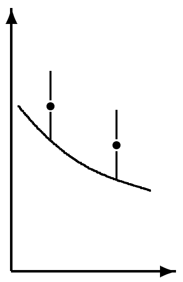
with the inverse of the covariance matrix

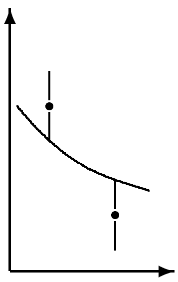
$$\boldsymbol{V}^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} 1/\sigma_1^2 & -\rho/\sigma_1\sigma_2 \\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2 \end{pmatrix}$$

Numerical Example for deviations

$\sigma_1 = \sigma_2 = \sigma$ and two deviations $\Delta = \pm 1\,\sigma$

$$\chi^2 = \begin{cases} 2/(1+\rho) & \text{same sign of } \Delta \\ 2/(1-\rho) & \text{opposite sign of } \Delta \end{cases}$$





Probability from χ^2

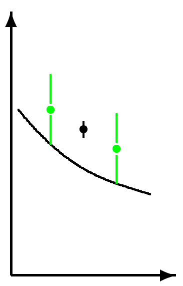
	same sign of Δ	opposite sign of Δ
$\rho = 0$	37 %	37 %
$\rho = +0.9$	59 %	0.0005 %
$\rho = -0.9$	0.0005 %	59 %

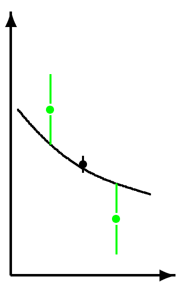
Same numerical Example - averaging of the two data points

Standard deviation of averaged data point:

	$\sqrt{\frac{1+\rho}{2}}\,\sigma$	comment
$\rho = 0$	0.707 σ	$1/\sqrt{2}$
$\rho = +0.9$	0.975 σ	same precison as single point
$\rho = -0.9$	0.224 σ	much more accurate

Averaging in case of negatively correlated data points

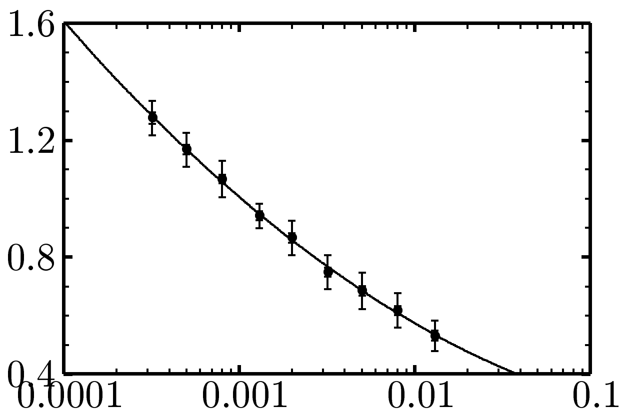




- Averaging of data points with large negative correlations results in very accurate values.
- Averaging of data points with large positive correlation reduces the number of data points without increasing their accuracy.

Example from a publication

Nine data points with relatively small fluctuation:



(My) fit of parabola (3 parameters) to the data shows: **either**

- errors to big (by factor > 5), **or**
- fluctuations too small.

Perhaps there is a large positive correlation between the data points?