Probability and statistics

V. Blobel
Universität Hamburg

Zeuthen, May 9 – 11, 2001

- Probability
- Frequentists and Bayesian view
- Bayes theorem
- Statistics
- Central limit theorem
- Transformations (Error propagation)
- Correlations beweeen data points
- Fitting correlated data

Probability

"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)

Historically the interest in statistical problems started with investigations of games of chance.

In games of chance there are often many different outcomes, which a-priori are equally likeli (tossing a coin, throwing a dice); this can be used to make a *combinatorical* definition of probability.

Similarly one can give an *empiricial* definition of probability/frequency.

Game of chance and statistics

Christiaan Huygens, "De Ratiociniis in Ludo Aleae", 1657

Jakob Bernoulli, "Ars Conjectandi, 1705

Pascal, in 'Oevres III', 307:

> " ... die Sache (d.h. das Glückspiel) war bisher ungewiß; nun aber, nachdem sie sich gegenüber dem Experiment unzugänglich gezeigt hatte, hat sie sich dem Verstand unterwerfen müssen. Wir haben es nämlich durch geometrische Argumentation auf eine berechenbare Kunst zurückgeführt, so daß es Teil an der Strenge der Geometrie hat und zuversichtlich vorangeht; es vereinigt die mathematische Gewißheit mit der Ungewißheit des Würfels, was vorher unvereinbar schien. Es nimmt nun den Namen beider an, und beansprucht den staunenerregenden Titel: Geometrie des Würfels."

" ... res hactenus erravit incerta; nunc autem quae experimento rebellis fuit rationis dominium effugere non potuit. Eam quippe tanta securitate in artem per geometricam reduximus, ut certitudinis eius particeps facta, iam audacter prodeat; et si metheseos demonstrationes cum aleae incertitudine iungendo, et quae contraria videntur conciliando, ab ultraque nominationem suam accipiens, stupendum hunc titulum iure sibi arrogat: aleae Geometrica."

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 often most natural for data analysis in physics.

Examples for frequencies:

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

The probability of hypotheses A, B, … is either 1 (true) of 0 (false).

Subjective probability – the Bayesian view

A, B, … are hypotheses, i.e. statements, that are either true of 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? – take an umbrella or not).

Bayes' theorem

It 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 $\rightarrow$ 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:

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

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

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

Expectation and variance

The expectation of a random variable x or mean is often called μ, it is the standard parameter describing the ”center of mass” of a distribution of a random variable.
The expectation is a linear operator:

$$E[a \cdot g(x) + b \cdot h(x)] = a \cdot E[g(x)] + b \cdot E[h(x)]$$

The variance of a random variable is non-negative; the square root is called the standard deviation σ and is the standard measure of the ”width” of a distribution. For the variance operator

$$V[a \cdot g(x)] = a^2 \cdot V[g(x)]$$

is valid.

Relation between standard distributions

The Binomial distribution follows from combinatorics.

$$\begin{tikzcd}[column sep=huge, row sep=huge]
\boxed{\begin{array}{c}\text{Binomial}\\ n,\, p\end{array}} \arrow[rr, "p \to 0"', "np = \mu"] \arrow[dr, "n \to \infty"'] & & \boxed{\begin{array}{c}\text{Poisson}\\ \mu\end{array}} \arrow[dl, "\mu \to \infty"] \\
& \boxed{\text{Normal}} &
\end{tikzcd}$$

Poisson and normal distribution follow from the Binomial distribution.

Binomial distribution

If the probability for a certain event is equal to p, the probability to observe in n experiments exactly k of the events, 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.

In a single event there are two possible outcomes, one (often called "success") with probability p and another (often called "failure") with probability $q = 1 - p$.

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

Law of large numbers

Estimate: If for an unknown value of p the values n and k are known from an experiment, the unknown value of the probability p can be estimated by $h = k/n$. An unbiased estimate of the variance of h is given by

$$V(h) = \frac{1}{n(n-1)} n \left(\frac{k}{n}\right) \left(1 - \frac{k}{n}\right)$$

The law of large numbers. The accuracy of an estimate h of the probability in an 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

Consider 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 $V[k] = \mu$

Rate of events, which is constant in time t:
number of events k follows a Poisson distribution, and time difference t follows exponential distribution:

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

Normal distribution

The normal distribution ist the most important continous distribution; almost anything which is a sum of many random contributions, follows (approximately) the normal distribution ($\rightarrow$ central limit theorem).

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

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 normal distribution!

χ^2 *distribution*

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

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

follows a χ^2 distribution $f_n(u) = f_n(\chi^2)$ with n degrees of freedom.

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

Important for statistical tests ("goodness of fit").

Characteristic functions

Given a random variable x with probability density $f(x)$, the characteristic function $\phi(t)$ (or Fourier transform) is defined as the expectation of e^{itx}

$$\phi(t) = E\left[e^{itx}\right] = \int_{-\infty}^{+\infty} e^{itx} \cdot f(x)\,\mathrm{d}x$$

with the inverse operation

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \cdot \phi(t)\,\mathrm{d}t\,.$$

The characteristic function is useful to calculate the distribution of a sum of random variables.

The sum of two random normal variables follows again a normal distribution.

The sum of two random Poisson variables follows again a Poisson distribution.

Sum of independent random variables

The characteristic function $\phi_z(t)$ for a *sum*

$$z = \sum_i x_i$$

of n independent random variables x_1, $x_2 \ldots x_n$ with probability densities $f_1(x_1)$, $f_2(x_2) \ldots f_n(x_n)$ with characteristic functions $\phi_1(t)$, $\phi_2(t) \ldots \phi_n(t)$ is given by the *product* of the individual characteristic functions:

$$\boxed{\phi_z(t) = \phi_1(t) \cdot \phi_2(t) \cdots \phi_n(t)}$$

This relation follows simply from the definition of the characteristic function:

$$\begin{aligned}\phi_z(t) = \int \ldots \int \exp\left(it\sum_{i=1}^{n} x_i\right) f_1(x_1)\, f_2(x_2) \cdots f_n(x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \cdots \mathrm{d}x_n \\ = \int e^{itx_1} f_1(x_1)\,\mathrm{d}x_1 \cdots \int e^{itx_n} f_n(x_n)\,\mathrm{d}x_n\end{aligned}$$

Central limit theorem

In the limit of large n the sum $\sum_i x_i$ becomes a random variable with mean $\sum_i \mu_i$ and with variance $\sum_i \sigma_i^2$, which follows the normal distribution, regardless of the distribution *of the individual variables x_i.*

$\rightarrow$ numerical examples

For finite n: theorem is valid, if sum is not dominated by one (few) terms

- velocity component v_x of gas molecules (good example)
- energy loss of charged particles traversing thin gas layer (bad example, Landau 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: $y = x^2$ $\quad x = \pm\sqrt{y} \quad \mathrm{d}x = \pm\frac{\mathrm{d}y}{2\sqrt{y}}$

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

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

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

Error propagation

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

Simpler: transform mean and variance based on expansion

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

Taking terms up to second order with correction term

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

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

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:

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

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} $\qquad \sigma_{ij} = \rho_{ij}\, \sigma_i \sigma_j \qquad -1 \leq \rho_{ij} \leq +1$

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

Generalization of the one-dimensional normal distribution to the case of n dimensions:

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

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

The n-dimensional normal distribution is often used as approximation to other n-dimensional distributions; the 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 $\boldsymbol{V}$).

Covariance matrix

Symmetric $n \times n$ covariance matrix $\boldsymbol{V} =$

$$\begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1n} \\ V_{21} & V_{22} & \cdots & V_{2n} \\ \cdots & \cdots & & \\ V_{n1} & V_{n2} & \cdots & V_{nn} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \cdots & \cdots & & \\ \rho_{1n}\sigma_n\sigma_1 & \rho_{2n}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{pmatrix}$$

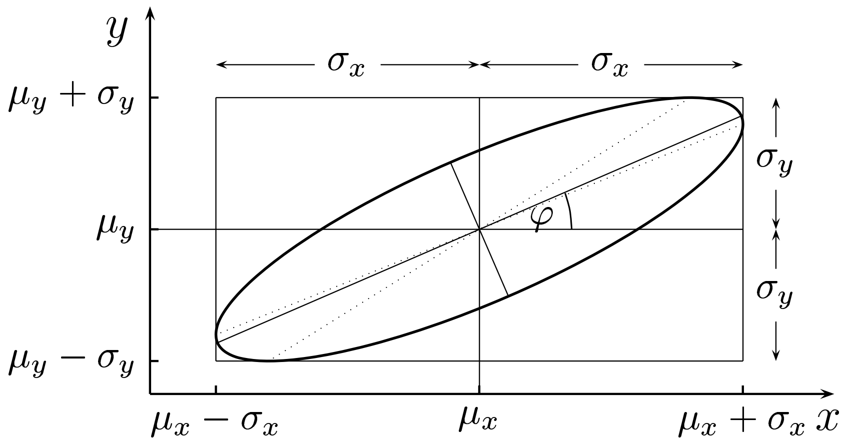
"Distance" $\boldsymbol{\Delta}$ from mean measured by

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

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

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 $\boldsymbol{y}$

$$\langle \boldsymbol{y} \rangle = \boldsymbol{\mu_y} E[\boldsymbol{y}] = E[\boldsymbol{Bx}] = \boldsymbol{B}E[\boldsymbol{x}] = \boldsymbol{B\mu_x}$$

(expectation is a linear operator)

Covariance matrix: "error" propagation

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

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:

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

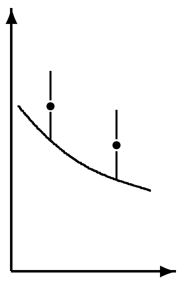
with the inverse of the covariance matrix

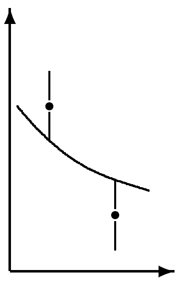
$$\boldsymbol{V}^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} 1/\sigma_1^2 & -\rho/\sigma_1\sigma_2 \\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2 \end{pmatrix}$$

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

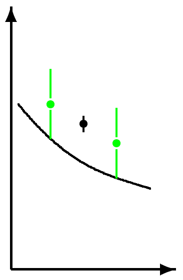
	same sign of Δ	opposite sign of Δ
$\rho = 0$	37 %	37 %
$\rho = +0.9$	59 %	0.0005 %
$\rho = -0.9$	0.0005 %	59 %

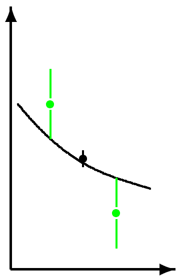
Same numerical Example - averaging of the two data points

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

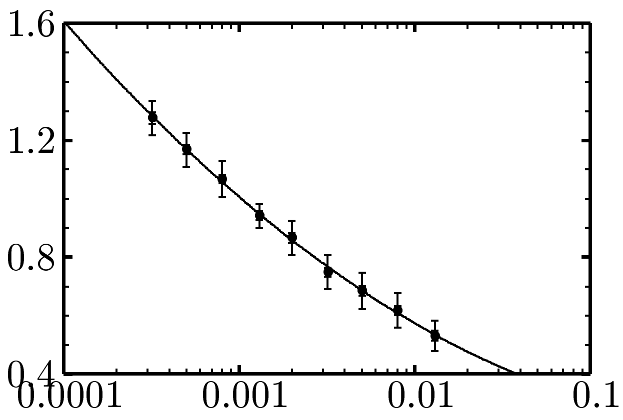




- Averaging of data points with large negative correlations results in very accurate values.
- Averaging of data points with large positive correlation reduces the number of data points without increasing their accuracy.

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?

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."*

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 Least Squares:

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

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:

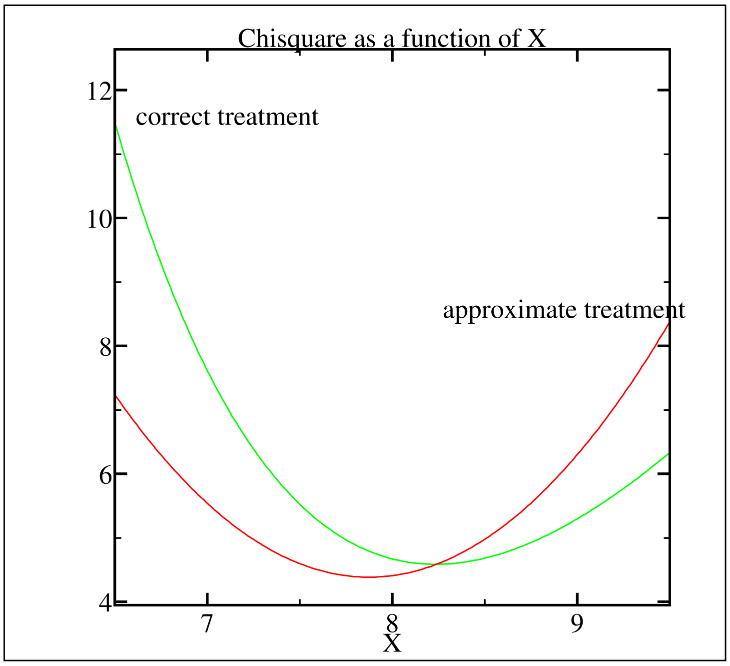
$$7.87 \pm 0.81 \qquad \rightarrow \text{ discrepancy}$$

Ansatz must be wrong, because normalization factor changed.

Geometrical explanation: angle φ of principal axis of error ellipse given by $\tan 2\varphi = 2V_{12}/(V_{22} - V_{11})$:

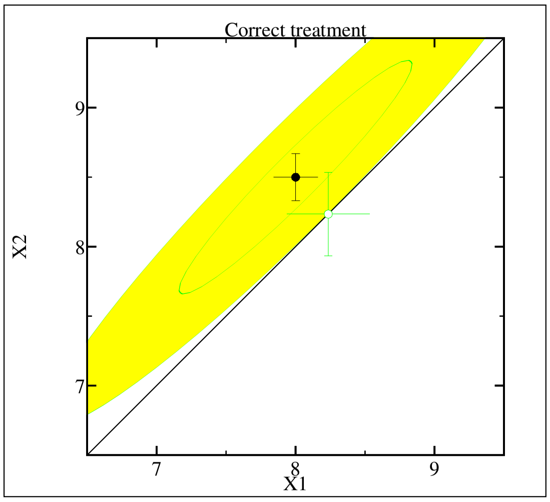
method	$\tan 2\varphi$	φ
correct treatment	250	44.8°
approximation	7.9	41.4°

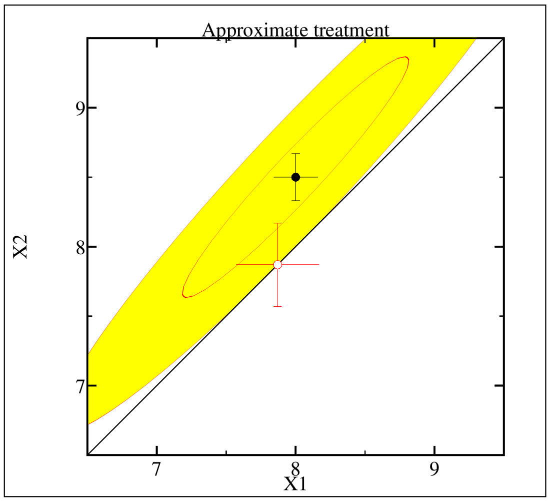
Parabolas



Ellipses

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.