1) EXPERIMENTAL ERROR

Aim of experiments:

1) to measure the numerical value of some physical quantity, e.g. c, α

"parameter determination"

2) to test a particular theory is consistent with data

"hypothesis testing"

reality is mixed, of course

a measurement of c; parameter measurement

and

test of c being constant with time

Why errors?

currently known value of $c = 2.99792458\times10^8$ m

a new experiment gives $c = 2.9900 \pm \sigma$

1) if $\sigma = 0.01\times10^8$ m

new result is consistent with the previous results

2) if $\sigma = 0.001\times10^8$ m

new result is inconsistent with the previous results;

-new discovery (c changes with time)

-either the new value or error is wrong

3) if $\sigma = 1.0\times10^8$ m

new result is irrelevant

So, depending on the experimental error, our reaction could be,
"the conventional theory is in good shape" or
"we have made a great discovery" or "we should find a better way to do an experiment".

Random and systematic errors

Random error: inability of any measuring device to give infinitely accurate answer

Systematic error: in the nature of mistake.

Example: determination of the decay constant λ of a radioactive source

counting number of decays within a given time interval → $-dn/dt$

weighting the sample → number of nuclei present n

$$\lambda = -\frac{\frac{dn}{dt}}{n}$$

random errors are due to
counting of decays (random process)
timing of the interval
weight measurement

systematic errors are due to
the counter used is not fully efficient and/or not surrounding the sample completely → lower counting than the true value
existence of other radioactive source, e.g. cosmic ray → higher counting than the true value
radioactive source is not pure → number of nuclei is less than the true number

random error can be estimated by repeating the measurement several times and comparing the results

2) DISTRIBUTION AND PROBABILITY

Distributions $n(x)$: describing how often a value of the variable x occurs in a definite sample

Discrete Distribution

range	x variable	$n(x)$
0 to 7	number of days	number of sunny days in a week
1 to ∞	integer x	number of working programme you produced after x compilations
–13.6 to 0eV	energy of ground and excited states of hydrogen atom	number of atoms with electrons in state of energy x in atomic hydrogen at 10°K

Continuous Distribution

range	x variable	n(x)
0 to 24 hours	hours of sleep each night	number of person sleeping for time x
0 to ∞	hours to understand statistics	number of person who understood statistics after x hours

Distributions are characterised by

1) the mean or expectation value = $E(x)$ $= \int_{-\infty}^{+\infty} x\, n(x)\, dx$

2) the mean of the square deviation from the mean or variance = $\sigma_x{}^2 = \int_{-\infty}^{+\infty} x^2 n(x)\, dx - \left[\int_{-\infty}^{+\infty} x\, n(x)\, dx\right]^2$

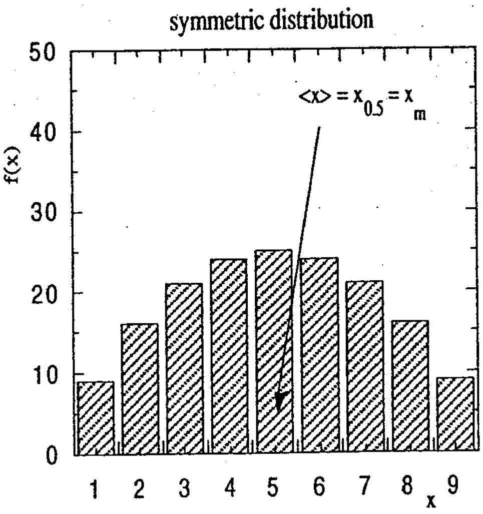
3) the median $= x_{0.5}$

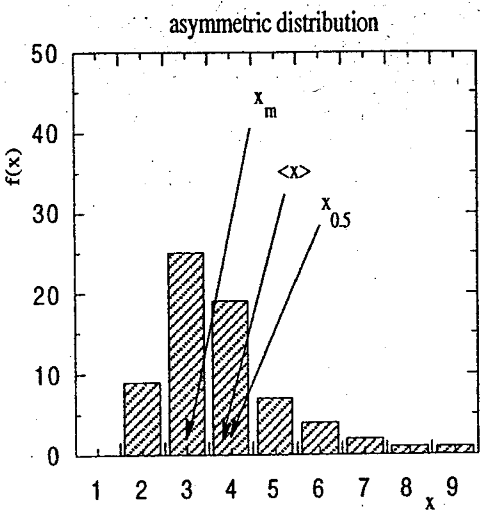
the value where the population for above that value and the population for below that value are identical

i.e. $\sum_{x_{\min}}^{x_{0.5}} f(x) = \sum_{x_{0.5}}^{x_{\max}} f(x)$ or $\int_{x_{\min}}^{x_{0.5}} \mathrm{d}x\, f(x) = \int_{x_{0.5}}^{x_{\max}} \mathrm{d}x\, f(x)$

4) the mode = x_{m}

the value which happens the most





symmetric distribution: $E(x) = x_{0.5} = x_{\mathrm{m}}$

Probabilities $P(x)$: with a sample of N measurements, the value x is obtained n_x times. Then the probability is defined to be

$$p(x) = \lim_{N \to \infty} \frac{n_x}{N}$$

This is equivalent to the normalised distribution, i.e.

$$p(x) = \frac{f(x)}{\sum_{x_{\min}}^{x_{\max}} f(x) \text{ or } \int_{x_{\min}}^{x_{\max}} \mathrm{d}x\, f(x)}$$

Note: $1 \geq p(x) \geq 0$

Expectation value and variance and can be written as

$$E(x) = \sum_{x_{\min}}^{x_{\max}} x\, p(x) \text{ or } \int_{x_{\min}}^{x_{\max}} \mathrm{d}x\, x\, p(x)$$

and

$$\begin{aligned} \sigma^2(x) &= \sum_{x_{\min}}^{x_{\max}} \left(x - E(x)\right)^2 p(x) \text{ or } \int_{x_{\min}}^{x_{\max}} \mathrm{d}x \left(x - E(x)\right)^2 p(x) \\ &= E\left(\{\, x - E(x)\,\}^2\right) = E(x^2) + E^2(x) - 2\,E\,(x\,E(x)) \\ &= E(x^2) + E^2(x) - 2\,E^2(x) \\ &= E(x^2) - E^2(x) \end{aligned}$$

3) SAMPLE OF EVENTS

For a set of N separate measurements of x, $\{x_1, x_2, ..., x_N\}$, how can we estimate the expectation value and variance ?

Estimation of expectation value:

$$\boxed{\hat{x} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

Note; since the expectation value of $\hat{x}$ is $E(x)$, $\hat{x}$ defined as above is unbiased. i.e, the expectation value doesn't depend on N, # of measurement

$$E(\hat{x}) = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = E(x)$$

How good is this estimate? → the variance of $\hat{x}$

$$\sigma^2(\hat{x}) = E\left(\{\hat{x} - E(x)\}^2\right) = \frac{1}{N^2} E\left(\left\{\sum_{i=1}^{N} x_i - N E(x)\right\}^2\right) = \frac{1}{N^2} E\left(\left\{\sum_{i=1}^{N}(x_i - E(x))\right\}^2\right) = \frac{1}{N^2}\left[E\left\{\left(\sum_{i=1}^{N} x_i\right)^2\right\} - N^2 E^2(x)\right]$$

$$= \frac{1}{N^2} E\left(\sum_{i=1}^{N}(x_i - E(x))^2\right) + \frac{1}{N^2} E\left(\sum_{i,j=1,\, i\neq j}^{N} (x_i - E(x))(x_j - E(x))\right)$$

if all x_i are independent (uncorrelated)

$$E\left(\left[x_i - E(x)\right]\left[x_j - E(x)\right]\right) = 0 \text{ for } i \neq j$$

and we obtain

$$\sigma^2(\hat{x}) = \frac{1}{N^2} E\left(\sum_{i=1}^{N}\{x_i - E(x)\}^2\right) = \frac{1}{N^2}\sum_{i=1}^{N} E\left(\{x_i - E(x)\}^2\right) = \frac{1}{N}\sigma^2(x)$$

i.e.

$$\boxed{\sigma^2(\hat{x}) = \frac{1}{N}\sigma^2(x)}$$

→ **the accuracy of the estimated mean value increases with increasing N**

Estimation of variance:

One may think a reasonable choice could be

$$\hat{\sigma}_x'^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x})^2$$

However, $\hat{\sigma}_x'^2$ is biased. This can be seen by evaluating the expectation value;

$$\begin{aligned} E\left(\sigma'^{2}_{x}\right) &= \frac{1}{N} E\left(\sum_{i=1}^{N} (x_i - \hat{x})^2\right) = \frac{1}{N} E\left(\sum_{i=1}^{N} \{[x_i - E(x)] + [E(x) - \hat{x}]\}^2\right) \\ &= \frac{1}{N} E\left(\sum_{i=1}^{N} [x_i - E(x)]^2 + \sum_{i=1}^{N} [E(x) - \hat{x}]^2 + 2 \sum_{i=1}^{N} [x_i - E(x)]\,[E(x) - \hat{x}]\right) \\ &= \frac{1}{N} E\left(\sum_{i=1}^{N} [x_i - E(x)]^2 + N\,[E(x) - \hat{x}]^2 - 2\,N\,[E(x) - \hat{x}]^2\right) \\ &= \frac{1}{N} \left\{E\left(\sum_{i=1}^{N} [x_i - E(x)]^2\right) - N\,E\left([E(x) - \hat{x}]^2\right)\right\} = \frac{1}{N} \left[N\,\sigma(x)^2 - N\,\sigma(\hat{x})^2\right] \\ &= \frac{N-1}{N}\,\sigma(x)^2 \end{aligned}$$

i.e.

$$E\left(\hat{\sigma}'^{2}_{x}\right) = \frac{N-1}{N}\,\sigma(x)^2$$

the estimation value depends on the number of events N → biased!

The unbiased estimation for the variance must be defined as

$$\boxed{\hat{\sigma}_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{x})^2}$$

$$\begin{aligned} E\left(\hat{\sigma}_x^2\right) &= \frac{1}{N-1} E\left[\sum_{i=1}^{N} (x_i - \hat{x})^2\right] \\ &= \frac{1}{N-1}\,(N-1)\,\sigma^2(x) \\ &= \sigma^2(x) \end{aligned}$$

4) IMPORTANT DISTRIBUTIONS

i) Binomial distribution

example: Tossing 4 coins. Probabilities for having

0 head	1 head	2 heads	3 heads	4 heads
0.5^4	4×0.5^4	6×0.5^4	4×0.5^4	0.5^4

Note the sum = $16\times0.5^4=1$

generalisation:

p= probability for success

q= probability for failure $p+q=1$

n= number of trial

probability to have m successes

$$\boxed{p_n(m)={}_nC_m\, p^m\, q^{m-n}}$$

where the combination ${}_nC_m=\dfrac{n!}{m!\,(n-m)!}$

Expectation value and variance:

$$E(m)=\sum_{m=0}^{n} m\,P_n(m)=\sum_{m=1}^{n}\frac{m\,n!}{m!\,(n-m)!}p^{m}q^{n-m}$$

$$=n\,p\sum_{m=1}^{n}\frac{(n-1)!}{(m-1)!\,(n-m)!}p^{m-1}q^{n-m}=n\,p\sum_{m'=0}^{n-1}\frac{(n-1)!}{m'!\,(n-m'-1)!}p^{m'}q^{n-m'-1}$$

$$=n\,p\sum_{m'=0}^{n'}\frac{n'!}{m'!\,(n'-m')!}p^{m'}q^{n'-m'}=n\,p\sum_{m=0}^{n'}P_{n'}(m)$$

$$=n\,p$$

and

$$E(m^2)=\sum_{m=0}^{n} m^2\,P_n(m)=\sum_{m=1}^{n}\frac{m^2\,n!}{m!\,(n-m)!}p^{m}q^{n-m}$$

$$=n\,p\sum_{m=1}^{n}\frac{m\,(n-1)!}{(m-1)!\,(n-m)!}p^{m-1}q^{n-m}=n\,p\sum_{m'=0}^{n-1}\frac{(m'-1)\,(n-1)!}{(m')!\,(n-m'-1)!}p^{m'}q^{n-m'-1}$$

$$=n\,p\sum_{m'=0}^{n'}\frac{(m'+1)\,n'!}{(m')!\,(n'-m')!}p^{m'}q^{n'-m'}=n\,p\left[\sum_{m=0}^{n'} m\,P_{n'}(m)+\sum_{m=0}^{n'}P_{n'}(m)\right]=n\,p\,(n'\,p+1)$$

$$=n\,p\,(n\,p+q)$$

expectation value: $E(m)=n\,p$

variance: $\sigma^2(m)=E\left(\{m-\mathrm{E}(m)\}^2\right)=E(m^2)-E^2(m)=n\,p\,q$

a) for p fixed, various n b) n fixed, various p c) pn fixed, various n

a) transition to the Gauss distribution

c) transition to the Poisson distribution

ii) Poisson distribution

The transition from the binomial to the Poisson distribution is done by $n \to \infty$, $p \to 0$ keeping $n\,p = \mu$(constant), **i.e. the distribution for events with small probability.**

$$P_n(m) = \frac{n!}{m!\left(n-m\right)!} p^m q^{n-m} = \frac{n!}{m!\left(n-m\right)!}\left(\frac{\mu}{n}\right)^m \frac{\left(1-\frac{\mu}{n}\right)^n}{\left(1-\frac{\mu}{n}\right)^m}$$

$$= \frac{\mu^m}{m!}\,\frac{n\left(n-1\right)\left(n-2\right)\cdots\left(n-m+1\right)}{n^m}\,\frac{\left(1-\frac{\mu}{n}\right)^n}{\left(1-\frac{\mu}{n}\right)^m}$$

$$= \frac{\mu^m}{m!}\left(1-\frac{\mu}{n}\right)^n \frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right)}{\left(1-\frac{\mu}{n}\right)^m}$$

by taking the limit $n \to \infty$

$$\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n = e^{-\mu} \quad \text{and} \quad \lim_{n\to\infty}\frac{1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{m-1}{n}\right)}{\left(1-\frac{\mu}{n}\right)^m} = 1$$

thus, **the Poisson distribution is defined as**

$$\boxed{\lim_{n\to\infty} P_n(m) = \frac{\mu^m}{m!}\,e^{-\mu} = f_\mu(m)}$$

Note that $f_\mu(m)$ is properly normalised, i.e.

$$\sum_{m=0}^{\infty} f_\mu(m) = e^{-\mu}\sum_{m=0}^{\infty}\frac{\mu^m}{m!} = e^{-\mu}\left(1+\frac{\mu}{1!}+\frac{\mu^2}{2!}+\cdots\right) = e^{-\mu}e^{\mu}$$
$$= 1$$

Expectation value and variance:

$$E(m) = \sum_{m=0}^{\infty} m\,f_\mu(m) = \sum_{m=1}^{\infty}\frac{\mu^m}{(m-1)!}e^{-\mu} = \mu\sum_{m=1}^{\infty}\frac{\mu^{m-1}}{(m-1)!}e^{-\mu} = \mu\sum_{m'=0}^{\infty} f_\mu(m')$$
$$= \mu$$

and

$$E\left(m^2\right) = \sum_{m=1}^{\infty}\frac{m\,\mu^m}{\left(m-1\right)!}e^{-\mu} = \mu\sum_{m'=0}^{\infty}\frac{\left(m'+1\right)\mu^{m-1}}{m'!}e^{-\mu}$$
$$= \mu\left\{\sum_{m'=0}^{\infty} m'\,f_\mu\!\left(m'\right) + \sum_{m'=0}^{\infty} f_\mu\!\left(m'\right)\right\} = \mu\left(\mu+1\right)$$

Thus,

expectation value: $E(m) = \mu$

variance: $\sigma^2(m) = E\left(\{m - \mathrm{E}(m)\}^2\right) = E(m^2) - E^2(m) = \mu$

Examples of the Poisson distribution:

-Consider the very large number of radioactive atomic nuclei n. The probability that one of the nuclei decays within the time interval Δt, $P(\Delta t)$ follows the Poisson distribution if $\Delta t \ll \mu$.

-Number of interactions in a given time interval Δt produced by a high intensity beam colliding on a thin target follows the Poisson distribution.

If the rate of the basic process changes during the measurement or events are correlated, the Poisson distribution cannot be applied.

-Number of interactions produced by a low intensity beam colliding on a thick target. The rate of the basic process decreases since the initial number of particles is limited.

-Number of particles detected in one second by a detector which has a dead time of 1msec and with a particle flux of more than 10^3/sec. In this case, no particle can be detected for 1msec once a particle is detected by the detector. If the flux is high, the detection of a particle is correlated to the previously detected particle.

Interesting example.

N particles are detected and n_+ of them are positive and n_- of them are negative, i.e. $n_+ + n_- = N$. The probability to obtain such an event is the product of the probability to observe N events (Poisson distribution) and the probability that n_+ in N events are positive (binomial distribution)

$$\frac{e^{-\mu}\mu^N}{N!} \times \frac{N!}{n_+!(N-n_+)!} p_+^{n_+}(1-p_+)^{N-n_+}$$

where μ is the expectation value for the total number of events and p_+ is the probability to obtain a positive particle. The above expression can be rewritten as

$$\frac{e^{-\mu}\mu^N}{N!} \times \frac{N!}{n_+!(N-n_+)!} p_+^{n_+}(1-p_+)^{N-n_+} = \frac{e^{-\mu}\mu^N}{n_+!n_-!} p_+^{n_+} p_-^{n_-}$$

$$= \frac{e^{-\mu p_+}(\mu p_+)^{n_+}}{n_+!} \times \frac{e^{-\mu p_-}(\mu p_-)^{n_-}}{n_-!}$$

where $p_- = 1 - p_+$. The last expression can be interpreted as the product of two independent Poisson distributions. The two Poisson distributions describe the probabilities for positive and negative particles respectively.

iii) Gauss distribution

Transition from the binomial to Gauss distributions can be obtained by taking the limit $n \to \infty$ for a finite p so that $np \to \infty$.-In this case, n, m and $(n-m)$ are all considered to be large, i.e. the Stirling's approximation can be used for their factorials:

$$n! = \sqrt{2\,\pi\,n}\,\left(\frac{n}{e}\right)^{n}\left(1+\frac{1}{12\,n}+\frac{1}{288\,n^{2}}+\cdots\right)$$
$$m! = \sqrt{2\,\pi\,m}\,\left(\frac{m}{e}\right)^{m}\left(1+\frac{1}{12\,m}+\frac{1}{288\,m^{2}}+\cdots\right)$$
$$(n-m)! = \sqrt{2\,\pi\,(n-m)}\,\left(\frac{n-m}{e}\right)^{n}\left[1+\frac{1}{12\,(n-m)}+\frac{1}{288\,(n-m)^{2}}+\cdots\right]$$

then,

$$\frac{n!}{m!\,(n-m)!} \approx \sqrt{\frac{n}{2\,\pi\,m\,(n-m)}}\;n^{n}\,m^{-m}\,(n-m)^{-n+m}$$
$$= \sqrt{\frac{1}{2\,\pi\,n}}\,\left(\frac{1}{n}\right)^{-(n+1)} m^{-\left(m+\frac{1}{2}\right)}\,(n-m)^{-\left(n-m+\frac{1}{2}\right)}$$
$$= \sqrt{\frac{1}{2\,\pi\,n}}\,\left(\frac{m}{n}\right)^{-\left(m+\frac{1}{2}\right)}\,\left(\frac{n-m}{n}\right)^{-\left(n-m+\frac{1}{2}\right)}$$

The binomial distribution is now approximated as

$$P_n(m) \approx \sqrt{\frac{1}{2\,\pi\,n}}\,\left(\frac{m}{n}\right)^{-\left(m+\frac{1}{2}\right)}\left(\frac{n-m}{n}\right)^{-\left(n-m+\frac{1}{2}\right)} p^{m}\,q^{n-m}$$
$$= \sqrt{\frac{1}{2\,\pi\,n}}\;\exp\left[-\left(m+\frac{1}{2}\right)\ln\frac{m}{n}-\left(n-m+\frac{1}{2}\right)\ln\frac{n-m}{n}+m\ln p+(n-m)\ln q\right]$$

By introducing m=np+ξ where $|\xi|<<|np|$, we obtain

$$\ln\frac{m}{n} = \ln\left(p+\frac{\xi}{n}\right) \approx \ln p+\frac{\xi}{p\,n}-\frac{1}{2}\left(\frac{\xi}{p\,n}\right)^{2}$$
$$\ln\frac{n-m}{n} = \ln\left(q-\frac{\xi}{n}\right) \approx \ln q-\frac{\xi}{q\,n}-\frac{1}{2}\left(\frac{\xi}{q\,n}\right)^{2}$$

Keeping the dominant term in ξ, $P_n(m)$ becomes

$$P_n(m) \approx \sqrt{\frac{1}{2\,\pi\,n\,p\,q}}\;\exp\left[-\frac{\xi}{2\,n\,p\,q}\right] = \frac{1}{\sqrt{2\,\pi}\,\sigma}\;\exp\left[-\frac{\xi^{2}}{2\,\sigma^{2}}\right]$$

where σ^2=npq. **In general, one formulates**

$$\boxed{G(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\;\exp\left[-\frac{(x-a)^{2}}{2\,\sigma^{2}}\right]}$$

which is often called a Gauss distribution. (As seen from the formula, x can be a continuous variable.

Note: By increasing μ, the Poisson distribution approaches to the Gauss distribution.

1) Some integrals...

$$I_0 = \int_0^{+\infty}\exp(-x^{2})\,\mathrm{d}x = \frac{\sqrt{\pi}}{2}\;,\; I_1 = \int_0^{+\infty}x\exp(-x^{2})\,\mathrm{d}x = -\frac{1}{2}\left[e^{-x^{2}}\right]_0^{+\infty} = \frac{1}{2}$$

then

$$I_2 = \int_0^{+\infty} x^2 \exp(-x^2)\,\mathrm{d}x = -\lim_{\alpha\to 1} \frac{\partial}{\partial\alpha} \int_0^{+\infty} \exp(-\alpha x^2)\,\mathrm{d}x$$

$$= -\lim_{\alpha\to 1} \frac{\partial}{\partial\alpha} \frac{1}{\sqrt{\alpha}} \int_0^{+\infty} \exp(-x'^2)\,\mathrm{d}x' = \lim_{\alpha\to 1} \frac{\sqrt{\pi}}{4\,\alpha\sqrt{\alpha}} = \frac{\sqrt{\pi}}{4}$$

$$\vdots$$

2) Expectation value

$$E(x) = \int_{-\infty}^{+\infty} x\,G(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{+\infty} x \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)\mathrm{d}x$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{+\infty} (x'+a) \exp\left(-\frac{x'^2}{2\sigma^2}\right)\mathrm{d}x' = a$$

3) Variance

$$E(x^2) = \int_{-\infty}^{+\infty} x^2\,G(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{+\infty} (x'+a)^2 \exp\left(-\frac{x'^2}{2\sigma^2}\right)\mathrm{d}x'$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma}\left(\int_{-\infty}^{+\infty} x'^2 \exp\left(-\frac{x'^2}{2\sigma^2}\right)\mathrm{d}x' + a^2 \int_{-\infty}^{+\infty} \exp\left(-\frac{x'^2}{2\sigma^2}\right)\mathrm{d}x' \right) = a^2 + \sigma^2$$

then

$$\mathrm{Variance}(\sigma) = E(x^2) - E^2(x) = \sigma^2$$

expectation value: a

variance: σ^2

4) The Gauss distribution has the following properties

$$\int_{a-\sigma}^{a+\sigma} G(x)\,\mathrm{d}x = 0.682, \quad \int_{a-2\sigma}^{a+2\sigma} G(x)\,\mathrm{d}x = 0.954, \quad \int_{a-3\sigma}^{a+3\sigma} G(x)\,\mathrm{d}x = 0.998$$

i.e. if 1000 **trials are made,** 998 **of the** x **values should in average be between** $a-3\sigma$ **and** $a+3\sigma$**.**

In a real experiment, neither the expectation value nor the variance are known.

→ **estimation of** $E(x)$**, and** $\sigma(x)$ **are required!!**

Example

m counts of radioactive decays were observed in a time interval of Δt

1) estimation of the expectation value by the measured count

$E(m)=m$

2) decay of radioactive nuclei → the Poisson distribution

$\sigma^2=E(m)=m$

3) if m is "sufficiently" (see Übungen II 3) large, Poisson→Gaussian

4) experimental value is given by

$m \pm \surd m$

0.682 (0.954, 0.998) probability that the true expectation value $E(m)$ is between $m \pm \surd m$, ($m \pm 2\surd m$, $m \pm 3\surd m$)

Note: Experimental errors have always a tail → the probability to be (for example) $\geq 3\sigma$ is larger than 0.002.

4) MULTI-VARIABLES DISTRIBUTION

Let us consider a case one measurement x consists of n variables, $\boldsymbol{x}=(x_1, x_2, \cdots, x_n)$:
Probability distribution for $\boldsymbol{x}$: $P(\boldsymbol{x})=P(x_1, x_2, \cdots, x_n)$
marginal distribution $g(x_r)$ is given by

$$\boxed{g(x_r)=\int_{-\infty}^{+} \mathrm{d}x_1 \int_{-\infty}^{+\infty} \mathrm{d}x_2 \cdots \int_{-\infty}^{+\infty} \mathrm{d}x_{r-1} \int_{-\infty}^{+\infty} \mathrm{d}x_{r+1} \cdots \int_{-\infty}^{+\infty} \mathrm{d}x_n P(x_1, x_1, \cdots, x_n)}$$

= probability distribution for a single variable x_r.

Expectation value of x_r is given by

$$E(x_r)=\int_{-\infty}^{+\infty} \mathrm{d}x_1 \int_{-\infty}^{+\infty} \mathrm{d}x_2 \cdots \int_{-\infty}^{+\infty} \mathrm{d}x_n\, x_r P(x_1, x_2, \cdots, x_n)$$
$$=\int_{-\infty}^{+\infty} \mathrm{d}x_r\, x_r\, g(x_r)$$

A distribution is described by

Variance: $\sigma_i{}^2 = E(\,[x_i - E(x_i)]^2)$

Covariance: $\mathrm{cov}(i, j)=E(\,[x_i - E(x_i)][x_j - E(x_j)])$

i.e. $\mathrm{cov}(i, i)= \sigma_i{}^2$

Note: $\mathrm{cov}(i,j)= \mathrm{cov}(j,i)$

If there is no correlation between the variables x_i and x_j,

$\mathrm{cov}(i,j)= 0$

i) Gaussian distribution with multi-variables

General formula is $\mathrm{G}(\boldsymbol{x}-\boldsymbol{a}) = k \exp\,[\,-(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{t}}\, W\,(\boldsymbol{x}-\boldsymbol{a})\,/\,2\,]$

$\boldsymbol{a}$: n component constant column vector

$\boldsymbol{x}$: n component column vector

$\boldsymbol{x}^{\mathrm{t}}$: transport

W: n×n matrix

k: normalisation factor $\left(\int_{-\infty}^{+\infty} dx\, G(\boldsymbol{x}-\boldsymbol{a})=1\right)$

Expectation value

$$E(\boldsymbol{x}-\boldsymbol{a})=\int_{-\infty}^{+\infty} \mathbf{d}\boldsymbol{x}\,(\boldsymbol{x}-\boldsymbol{a})\,G(\boldsymbol{x}-\boldsymbol{a})=k\int_{-\infty}^{+\infty} \mathbf{d}\boldsymbol{x}\,(\boldsymbol{x}-\boldsymbol{a})\exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{t}}\,W\,(\boldsymbol{x}-\boldsymbol{a})}{2}\right]=\boldsymbol{a}-\boldsymbol{a}=0$$

$E(\boldsymbol{x})=\boldsymbol{a}$

We also find

$$\frac{\partial\, E(\boldsymbol{x}-\boldsymbol{a})}{\partial \boldsymbol{a}}=\int_{-\infty}^{+\infty} \mathbf{d}\boldsymbol{x}\,\frac{\partial\,[(\boldsymbol{x}-\boldsymbol{a})\,G(\boldsymbol{x}-\boldsymbol{a})]}{\partial\,\boldsymbol{a}}=\int_{-\infty}^{+\infty} \mathbf{d}\boldsymbol{x}\,[-\boldsymbol{I}+(\boldsymbol{x}-\boldsymbol{a})\,(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{t}}\,W]\,G(\boldsymbol{x}-\boldsymbol{a})$$
$$=-\boldsymbol{I}+E\big((\boldsymbol{x}-\boldsymbol{a})\,(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{t}}\big)\,W=0$$

The second term is nothing but the covariance, i.e.

$$V \equiv W^{-1} = E\left((x-a)(x-a)^{\mathrm{t}}\right) = \begin{pmatrix} \sigma_1^2 & \mathrm{cov}(1,2) & \cdots \\ \mathrm{cov}(1\ 2) & \sigma_2^2 & \\ \vdots & & \ddots \end{pmatrix}$$

The matrix *V* is called the covariance matrix. (The matrix *W* is often called the weight matrix.)

Example: two variables

$$V \equiv W^{-1} = \begin{pmatrix} \sigma_1^2 & \mathrm{cov}(1,2) \\ \mathrm{cov}(1,2) & \sigma_2^2 \end{pmatrix}$$

Inverting *V*, we obtain

$$W = \frac{1}{\sigma_1^2\,\sigma_2^2 - \mathrm{cov}(2,\,1)^2}\begin{pmatrix} \sigma_2^2 & -\mathrm{cov}(1,2) \\ -\mathrm{cov}(1,2) & \sigma_1^2 \end{pmatrix}$$

If the two variables are uncorrelated, i.e. cov(1,2) = 0

$$W = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}$$

and

$$G = \frac{1}{2\,\pi\,\sigma_1\,\sigma_2}\exp\left[-\frac{(x_1-a_1)^2}{2\,\sigma_1^2}\right]\exp\left[-\frac{(x_2-a_2)^2}{2\,\sigma_2^2}\right]$$

i.e. **a product of two Gaussians** as expected.

$$\left(\text{the general normalization factor: } k = \frac{\sqrt{\det(W)}}{(2\,\pi)^{n/2}}\right)$$

When a correlation is present, by introducing the correlation coefficient ρ defined as

$$\rho = \frac{\mathrm{cov}(1,2)}{\sigma_1\,\sigma_2}$$

the matrix B can be written as

$$W = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\,\sigma_2} \\ \frac{-\rho}{\sigma_1\,\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}$$

Probability contour:

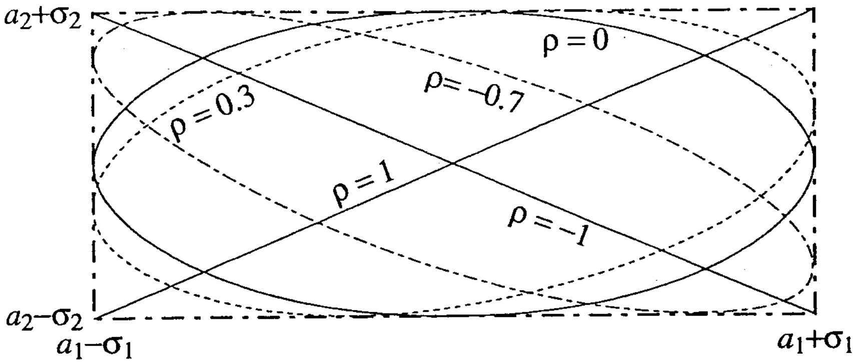
A probability contour on which the probability is down by $1/\sqrt{e}$ compared with the point ***x***=***a*** is given by

$$(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{t}}\,W\,(\boldsymbol{x}-\boldsymbol{a}) = 1$$

i.e.

$$\frac{(x_1-a_1)^2}{\sigma_1^2} + \frac{(x_2-a_2)^2}{\sigma_2^2} - \frac{2\,(x_1-a_1)(x_2-a_2)\,\rho}{\sigma_1\,\sigma_2} = 1-\rho^2$$

(note $|\rho| \leq 1$) which can be illustrated as



For a given value of $x=\tilde{x}_1$, the value of y giving the higest probability is given by

$$\frac{\partial G(\tilde{x}_1,\, x_2)}{\partial x_2}=0$$

$$\frac{(x_2-a_2)}{\sigma_2^2}-\frac{(\tilde{x}_1-a_1)}{\sigma_1\,\sigma_2}\rho=0$$

i.e.

$$x_2=\rho\frac{\sigma_2}{\sigma_1}\tilde{x}_1+\left(a_2-\rho\frac{\sigma_2}{\sigma_1}a_1\right)$$

If no correlation:

$x_1(x_2)$, which maximises the probability, is independent of $x_2(x_1)$, i.e. $a_2(a_1)$.

Positive correlation:

$x_1(x_2)$, which maximises the probability, increases with $x_2(x_1)$.

Negative correlation:

$x_1(x_2)$, which maximises the probability, decreases with $x_2(x_1)$.

5) TRANSFORMING VARIABLES AND ERROR PROPAGATION

Transformation of variables:

$$\boxed{x \rightarrow y \text{ with } y=y(x) \quad \Rightarrow \quad \mathrm{d}y = \left| \mathrm{d}y / \mathrm{d}x \right| \mathrm{d}x}$$

probability distributions transform as

$$\boxed{\Rightarrow \quad g(y) = \left| \mathrm{d}x / \mathrm{d}y \right| f(x)}$$

Note:

$$\int g(y)\,\mathrm{d}y = \int f(x) \left| \frac{dx}{dy} \right| \left| \frac{dy}{dx} \right| \mathrm{d}x = \int f(x)\,\mathrm{d}x$$

i.e. they are properly normalised.

Many variables:

$$\boldsymbol{x}=(x_1,x_2, \dots x_n) \rightarrow \boldsymbol{y}=(y_1,y_2, \dots y_n) \qquad \Rightarrow \qquad g(\boldsymbol{y}) = J(\boldsymbol{x}/\boldsymbol{y})\, f(\boldsymbol{x})$$

where

$$J\left(\frac{\boldsymbol{x}}{\boldsymbol{y}}\right) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \cdots \\ \frac{\partial x_2}{\partial y_1} & \ddots & \\ \vdots & & \end{vmatrix}$$

is called Jacobian.

Example

1) Linear transformation

$$\boldsymbol{y} = T\,\boldsymbol{x} + \boldsymbol{a} \qquad (\, y_i = T_{ij}\, x_j + a_i \,)\ \boldsymbol{a}\text{: constants}$$

Expectation value is given by

$$E(\boldsymbol{y}) = T\,E(\boldsymbol{x}) + \boldsymbol{a}$$

Covariance matrix is given by

$$\begin{aligned} V_y &= E\{\ [y-E(y)]\,[y-E(y)]^{\mathrm{t}}\ \} \\ &= E\{\ [T\boldsymbol{x} + \boldsymbol{a} - T\,E(\boldsymbol{x}) + \boldsymbol{a}]\,[T\boldsymbol{x} + \boldsymbol{a} - T\,E(\boldsymbol{x}) + \boldsymbol{a}]^{\mathrm{t}}\ \} \\ &= T\,E\{\ [\boldsymbol{x} - E(\boldsymbol{x})]\,[\boldsymbol{x} - E(\boldsymbol{x})]^{\mathrm{t}}\ \}\ T^{\mathrm{t}} \\ &= T\,V_x\,T^{\mathrm{t}} \end{aligned}$$

Transformation of the covariance matrix is given by

$$V_y = T\,V_x\,T^{\mathrm{t}}$$

2) Nonlinear case

Assume that ***x*** is not far from *E*(***x***), i.e.

$$y_i = y_i\big(E(x)\big) + \sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{x=E(x)} \delta x_j + \frac{1}{2}\sum_{j,k}\left(\frac{\partial^2 y_i}{\partial x_j \partial x_k}\right)_{x=E(x)} \delta x_j\, \delta x_k$$

where, $\delta_j = x_j - E(x_j)$, i.e. $E(\delta_j) = E\big(x_j - E(x_j)\big) = 0$.

$$E(y_i) = y_i\big(E(x)\big) + \frac{1}{2}\sum_{j,k}\left(\frac{\partial^2 y_i}{\partial x_j \partial x_k}\right)_{x=E(x)} Cov(j,k) \; + \cdots$$

and

$$y_i^2 = \left\{ y_i\big(E(x)\big) + \sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{x=E(x)} \delta x_j + \frac{1}{2}\sum_{j,k}\left(\frac{\partial^2 y_i}{\partial x_j \partial x_k}\right)_{x=E(x)} \delta x_j\, \delta x_k \right\}$$

$$\times \left\{ y_i\big(E(x)\big) + \sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{x=E(x)} \delta x_j + \frac{1}{2}\sum_{j,k}\left(\frac{\partial^2 y_i}{\partial x_j \partial x_k}\right)_{x=E(x)} \delta x_j\, \delta x_k \right\}$$

$$= y_i^2\big(E(x)\big) + 2y_i\big(E(x)\big)\sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{x=E(x)} \delta x_j + \sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{x=E(x)} \delta x_j \times \sum_j \left(\frac{\partial y_i}{\partial x_j}\right)_{x=E(x)} \delta x_j$$

$$+\, y_i\big(E(x)\big)\sum_{j,k}\left(\frac{\partial^2 y_i}{\partial x_j \partial x_k}\right)_{x=E(x)} \delta x_j\, \delta x_k$$

thus

$$E(y_i^2) = y_i^2\big(E(x)\big) + y_i\big(E(x)\big)\sum_{j,k}\left(\frac{\partial^2 y_i}{\partial x_j \partial x_k}\right)_{x=E(x)} Cov(j,k) + \sum_{j,k}\left(\frac{\partial y_i}{\partial x_j}\right)_{x=E(x)}\left(\frac{\partial y_i}{\partial x_k}\right)_{x=E(x)} Cov(j,k)$$

The variance can be given by

$$\sigma^2(y_i) = E(y_i^2) - E^2(y_i) = \sum_{k,l}\left(\frac{\partial y_i}{\partial x_k}\right)_{x=E(x)}\left(\frac{\partial y_i}{\partial x_l}\right)_{x=E(x)} \mathrm{Cov}(k,l)$$

or more generally

$$\boxed{V_y = T\, V_x\, T^{\mathrm{t}}}$$

with

$$T = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots \\ \frac{\partial y_2}{\partial x_1} & \ddots & \\ \vdots & & \end{pmatrix}$$

If there is **no correlation** between x_i and x_j, we obtain

$$\boxed{\sigma^2(y_i) = E(y_i^2) - E^2(y_i) = \sum_k \left(\frac{\partial y_i}{\partial x_k}\right)^2_{x=E(x)} \sigma^2(x_k)}$$

which is usually referred as **"error propagation"**.

An example of error propagation

A detector measures the x and y coordinates of a point with errors σ_x and σ_y independently

$\rightarrow$ polar coordinate

$$r^2 = x^2+y^2, \quad \tan\theta = y/x$$

Using

$$\partial r/\partial x = x/r, \quad \partial r/\partial y = y/r$$

$$\partial\theta/\partial x = -y/r^2, \quad \partial\theta/\partial y = x/r^2$$

variances become

$$\sigma_r{}^2 = (\partial r/\partial x)^2\, \sigma_x{}^2 + (\partial r/\partial y)^2\, \sigma_y{}^2 = (x^2\sigma_x{}^2+y^2\sigma_y{}^2)/r^2$$

$$\sigma_\theta{}^2 = (y^2\sigma_x{}^2+x^2\sigma_y{}^2)/r^4$$

On the other hand

$$x = r\cos\theta, \; y = r\sin\theta$$

i.e.

$$\partial x/\partial r = \cos\theta, \; \partial x/\partial\theta = -r\sin\theta$$

and

$$\sigma_x^2 = \left(\frac{\partial x}{\partial r}\right)^2 \sigma_r^2 + \left(\frac{\partial x}{\partial \theta}\right)^2 \sigma_\theta^2 + 2\left(\frac{\partial x}{\partial r}\right)\left(\frac{\partial x}{\partial \theta}\right) \mathrm{Cov}(r,\theta)$$

$$= \frac{1}{r^2}\left(x^2\cos^2\theta + y^2\sin^2\theta\right)\sigma_x^2 + \frac{1}{r^2}\left(y^2\cos^2\theta + x^2\sin^2\theta\right)\sigma_y^2 - 2\,r\sin\theta\cos\theta\,\mathrm{Cov}(r,\theta)$$

which gives

$$\mathrm{Cov}(r,\theta) = \frac{x^2\cos^2\theta + y^2\sin^2\theta - r^2}{2\,r^3\sin\theta\cos\theta}\,\sigma_x^2 + \frac{y^2\cos^2\theta + x^2\sin^2\theta}{2\,r^3\sin\theta\cos\theta}\,\sigma_y^2$$

$$= \frac{y^2\cos^2\theta + x^2\sin^2\theta}{2\,r^3\sin\theta\cos\theta}\left(\sigma_y^2 - \sigma_x^2\right) = \frac{y^2\,x^2}{r^5\sin\theta\cos\theta}\left(\sigma_y^2 - \sigma_x^2\right)$$

$$= \frac{x\,y}{r^3}\left(\sigma_y^2 - \sigma_x^2\right)$$

This should be compared with

$$\begin{pmatrix} \sigma_r^2 & \mathrm{Cov}(r,\theta) \\ \mathrm{Cov}(r,\theta) & \sigma_\theta^2 \end{pmatrix} = \begin{pmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{pmatrix} \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y \end{pmatrix} \begin{pmatrix} \frac{\partial r}{\partial x} & \frac{\partial \theta}{\partial x} \\ \frac{\partial r}{\partial y} & \frac{\partial \theta}{\partial y} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{x}{r} & \frac{y}{r} \\ -\frac{y}{r^2} & \frac{x}{r^2} \end{pmatrix} \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \begin{pmatrix} \frac{x}{r} & -\frac{y}{r^2} \\ \frac{y}{r} & \frac{x}{r^2} \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{x^2\sigma_x^2 + y^2\sigma_y^2}{r^2} & \dfrac{x\,y}{r^3}\left(\sigma_y^2 - \sigma_x^2\right) \\ \dfrac{x\,y}{r^3}\left(\sigma_y^2 - \sigma_x^2\right) & \dfrac{y^2\sigma_x^2 + x^2\sigma_y^2}{r^4} \end{pmatrix}$$

Note:

if $\sigma_x=\sigma_y$, no correlation between σ_r and σ_θ

if $\sigma_x<\sigma_y$, the correlation is positive for $x>0$, $y>0$

6) CENTRAL LIMIT THEOREM

The central limit theorem is:

If the $\boldsymbol{x}=(x_1, x_2, ...x_n)$ are a set of n independent variables all following an arbitray distribution with mean a and variance σ^2, then in the limit $n \to \infty$ their arithmetic mean

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

follows a **Gauss distribution with mean a and variance σ^2/n.**

If $\boldsymbol{x}=(x_1, x_2, ...x_n)$ are from n distributions with different mean a_i and variance $\sigma^2{}_i$, $\hat{x}$ still follows a Gauss distribution of mean $(1/n)\Sigma a_i$ and variance $(1/n)\Sigma\sigma^2{}_i$. The distribution of each x_i is irrelevant.

If x_i are already Gaussian, $\hat{x}$ is a Gaussian for $n \geq 1$, else for some "large" n.

(how large? see Übungen V)

7) CHI SQUARE DISTRIBUTION

Assume x follows a Gauss distribution with a mean 0 and a variance 1 (the standard Gaussian); Draw a sample $x_1, x_2, x_3, \cdots x_n$. A sum of squares

$$\chi^2 = x_1^2 + x_2^2 + x_3^2 \cdots + x_n^2$$

follows the probability function

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

which is referred as chi square distribution and n is the number of degrees of freedom.

Note:

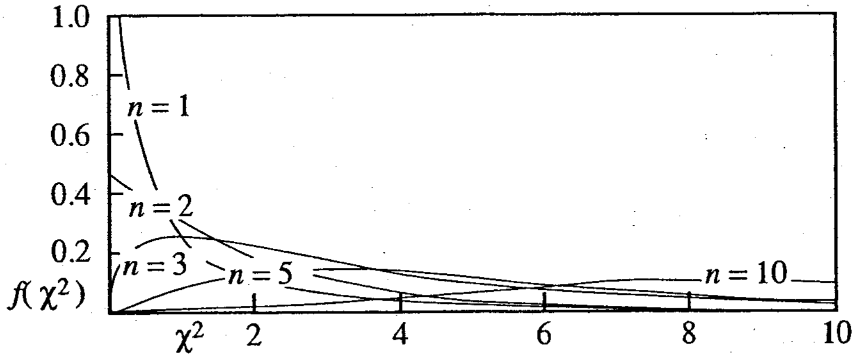
$$\boxed{\begin{array}{l} \boldsymbol{E(\chi^2)=n},\ \boldsymbol{E\{(\chi^2)^2\}=n^2+2n}, \\ \boldsymbol{\sigma^2(\chi^2)=2n} \end{array}}$$

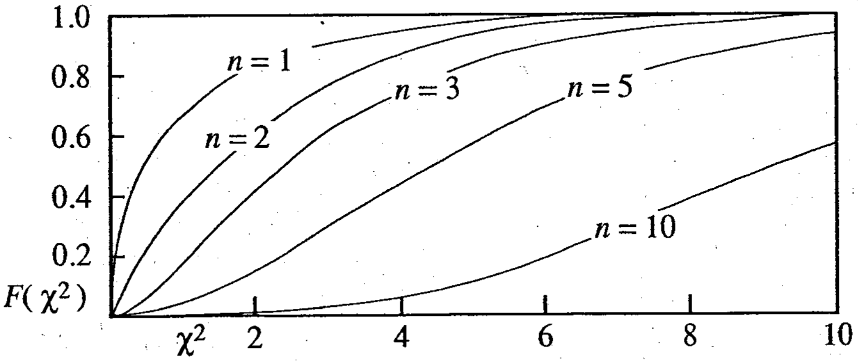
and

$$f(\chi^2=0) = \begin{array}{ll} \infty & n=1 \\ 0.5 & n=2 \\ 0.0 & n\geq 3 \end{array}$$

The probability that the random variable χ^2 does not exceed χ_0^2 is given by

$$F(\chi_0^2) = \frac{1}{\Gamma(n/2)\, 2^{n/2}} \int_0^{\chi_0^2} u^{n/2-1} e^{-u/2}\, \mathrm{d}u\ .$$





General definition of χ^2

$$\chi^2 = \frac{(x_1-a)^2+(x_2-a)^2+\cdots+(x_n-a)^2}{\sigma^2}$$

where x_i follows a Gauss distribution with an expectation value a and variance σ^2.

Recall previous estimation for the variance (Chapter 3)

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{x})^2$$

where x_i is drawn from a normal distribution with variance σ^2. Then, $(N{-}1)\hat{\sigma}_x^2/\sigma^2$ follows the χ^2 distribution with $(N{-}1)$ degrees of freedom.

Proof: if we define

$$y_1 = \frac{1}{\sqrt{1\times 2}}(x_1-x_2),\ y_2 = \frac{1}{\sqrt{2\times 3}}(x_1+x_2-2x_3),\ y_3 = \frac{1}{\sqrt{3\times 4}}(x_1+x_2+x_3-3x_4)$$

$$\vdots$$

$$y_{n-1} = \frac{1}{\sqrt{(n-1)\times n}}(x_1+x_2+x_3\cdots+x_{n-1}-(n-1)x_n)$$

$$y_n = \frac{1}{\sqrt{n}}(x_1+x_2+x_3\cdots+x_n) = \sqrt{n}\,\hat{x}$$

y_i follows a normal distribution with variance σ^2 and expectation value $E(y)$=0.

Then $\Sigma y_i{}^2 = \Sigma x_i{}^2$.

It follows that

$$(n-1)\hat{\sigma}_x^2 = \sum_{i=1}^{n}(x_i-\hat{x})^2 = \sum_{i=1}^{n}x_i^2 - n\,\hat{x}^2 = \sum_{i=}^{n}y_i^2 - y_n^2 = \sum_{i=1}^{n-1}$$

i.e.

$$\frac{(n-1)\hat{\sigma}_x^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n-1}y_i^2$$

follows χ^2 distribution with $(n{-}1)$ degrees of freedom.

Why $n{-}1$ degrees of freedom ?

due to the constraint

$$\sum_{i=1}^{n}x_i = n\,\hat{x}$$

the degrees of freedom is reduced from n to $n{-}1$.

8) THE METHOD OF MAXIMUM LIKELIHOOD

A) Likelihood function

Function $f(\boldsymbol{x};\boldsymbol{\lambda})$ describes the probability of the random variable $\boldsymbol{x}=(x_1, x_2, \cdots x_\nu)$ with a specific set of parameters $\boldsymbol{\lambda}=(\lambda_1, \lambda_2, \cdots \lambda_p)$. When N measurements are made with $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots \boldsymbol{x}^{(N)}$, the probability to have such N events is given by

$$\boxed{L=\prod_{j=1}^{N} f\left(\boldsymbol{x}^{(j)};\boldsymbol{\lambda}\right)}$$

which is a function of $\boldsymbol{\lambda}$. which is called likelihood function.

Note: $f(\boldsymbol{x};\boldsymbol{\lambda})$ must be properly normalised, i.e. $\int_{x_{\min}}^{x_{\max}} dx\, f(x;\boldsymbol{\lambda})=1$

Example:

We have an asymmetric coin and want to decide whether this belong to class A or B.

	A	B
heads	1/3	2/3
tails	2/3	1/3

After 5 tosses, one heads and four tails were obtained. Likelihood functions are

A: L= $(1/3)\times(2/3)^4 \approx 0.0658$

B: L= $(2/3)\times(1/3)^4 \approx 0.00823$

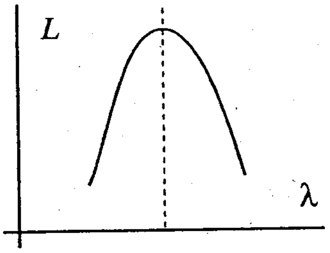
i.e. class A is the likely solution.

b)Maximum likelihood

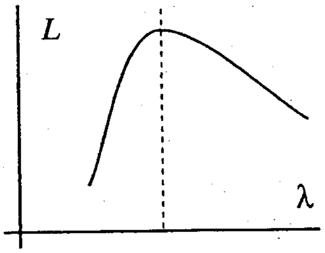
general extension of the example

→ a set of parameters $\boldsymbol{\lambda}$ which gives the highest value of the likelihood function have the highest confidence as the estimation.

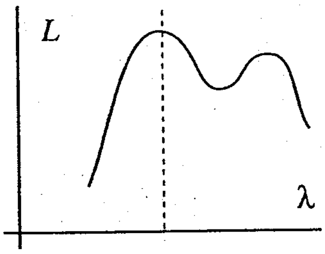
$$\boxed{\text{Find } \boldsymbol{\lambda} \text{ which makes } L \text{ maximum!}}$$



uniquely defined, symmetric



uniquely defined, asymmetric



???????????

For computational reasons, it is easy to use a "log likelihood function" defined as

$$l=\ln\left(\prod_{j=1}^{N} f\left(\boldsymbol{x}^{(j)};\boldsymbol{\lambda}\right)\right)$$

maximising L = maximising l

Single parameter case

At l maximum, $dl/d\lambda=0$;

$$\frac{\mathrm{d}l}{\mathrm{d}\lambda} = \sum_{j=1}^{N} \frac{\mathrm{d}}{\mathrm{d}\lambda}\, ln\left[f\left(x^{(j)};\lambda\right)\right] = \sum_{j=1}^{N} \frac{\frac{d}{d\lambda} f\left(x^{(j)};\lambda\right)}{f\left(x^{(j)};\lambda\right)} = \sum_{j=1}^{N} \frac{f'\left(x^{(j)};\lambda\right)}{f\left(x^{(j)};\lambda\right)}$$

and obtain the likelihood equation

$$\boxed{\sum_{j=1}^{N} \frac{f'\left(x^{(j)};\lambda\right)}{f\left(x^{(j)};\lambda\right)} = 0}$$

Example: Repeated measurements with different accuracy's.

Assume that the accuracy of each experiment can be expressed by a normal distribution with variance σ_j^2. The likelihood function is given as

$$L = \prod_{j=1}^{N} f\left(x^{(j)};\lambda\right) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_j} \exp\left[-\frac{\left(x^{(j)}-\lambda\right)^2}{2\sigma_j^2}\right]$$

and the log likelihood function is

$$l = -\frac{1}{2}\sum_{j=1}^{N} \frac{\left(x^{(j)}-\lambda\right)^2}{\sigma_j^2} + \text{constant}$$

From the likelihood equation

$$\frac{\mathrm{d}l}{\mathrm{d}\lambda} = \sum_{j=1}^{N} \frac{\left(x^{(j)}-\lambda\right)}{\sigma_j^2} = \sum_{j=1}^{N} \frac{x^{(j)}}{\sigma_j^2} - \lambda \sum_{j=1}^{N} \frac{1}{\sigma_j^2} = 0$$

we obtain

$$\lambda = \frac{\sum_{j=1}^{N} \frac{x^{(j)}}{\sigma_j^2}}{\sum_{j=1}^{N} \frac{1}{\sigma_j^2}} \qquad \textbf{weighted average}$$

B) Information inequality

A good estimation of a parameter λ, S must be "unbiased" and its variance $\sigma^2(S)$ must be as small as possible.

--there exist an optimised relation between these two requirements--

information inequality

$$E(S) = \int S\, f\left(x^{(1)};\lambda\right) f\left(x^{(2)};\lambda\right) \cdots f\left(x^{(N)};\lambda\right) \mathrm{d}x^{(1)}\,\mathrm{d}x^{(2)} \cdots dx^{(N)}$$

and

$$E(\mathrm{S}) = B(\lambda) + \lambda$$

$B(\lambda)$: possible bias

$$B(\lambda) + \lambda = \int S f\left(x^{(1)};\lambda\right) f\left(x^{(2)};\lambda\right) \cdots f\left(x^{(N)};\lambda\right) \mathrm{d}x^{(1)}\,\mathrm{d}x^{(2)} \cdots dx^{(N)}$$

By differentiating respect to λ, we obtain

$$B'(\lambda)+1=\int S\left[\sum_{i=1}^{N}\frac{f'(x^{(i)};\lambda)}{f(x^{(i)};\lambda)}\right]f(x^{(1)};\lambda)\,f(x^{(2)};\lambda)\cdots f(x^{(N)};\lambda)\,\mathrm{d}x^{(1)}\,\mathrm{d}x^{(2)}\cdots\mathrm{d}x^{(N)}$$

$$=E\left(S\left[\sum_{i=1}^{N}\frac{f'(x^{(i)};\lambda)}{f(x^{(i)};\lambda)}\right]\right)=E(S\,l')$$

Since clearly $E(l')=0$, we can derive

$$B'(\lambda)+1=E(S\,l')-E(S)\,E(l')=E\{[S-E(S)]\,l'\}$$

and using $\{E(x\,y)\}^2\le \mathrm{E}(x^2)\,E(y^2)$, it follows that

$$[B'(\lambda)+1]^2\le E\{[S-E(S)]^2\}\,E(l'^2)$$

We can also derive

$$E(l'^2)=E\left\{\left[\sum_{i=1}^{N}\frac{f'(x^{(i)};\lambda)}{f(x^{(i)};\lambda)}\right]^2\right\}=E\left\{\sum_{i}\left[\frac{f'(x^{(i)};\lambda)}{f(x^{(i)};\lambda)}\right]^2\right\}+E\left\{\sum_{i\neq j}\left[\frac{f'(x^{(i)};\lambda)}{f(x^{(i)};\lambda)}\right]\left[\frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)}\right]\right\}$$

$$=\sum E\left[\left(\frac{f'(x^{(i)};\lambda)}{f(x^{(i)};\lambda)}\right)^2\right]+\sum_{i\neq j}E\left[\frac{f'(x^{(i)};\lambda)}{f(x^{(i)};\lambda)}\right]E\left[\frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)}\right]=N\,E\left[\left(\frac{f'(x;\lambda)}{f(x;\lambda)}\right)^2\right]$$

where the second term vanishes using $E\,[\,f'(x^{(i)};\lambda)/f(x^{(i)};\lambda)\,]=\int f'(x^{(i)};\lambda)\,\mathrm{d}x^{(i)}=0$.

By introducing

$$\boxed{I(\lambda)=E(l'^2)=N\,E\left[\left(\frac{f'(x;\lambda)}{f(x;\lambda)}\right)^2\right]=N\int dx\,\frac{\left(\frac{\partial f}{\partial\lambda}\right)^2}{f}}$$

where $I(\lambda)$ is called "information", we have

$$\sigma^2(S)\ge\frac{[1+B'(\lambda)]^2}{I(\lambda)}$$

which is referred as the "information inequality". In the case of vanishing bias,

$$\boxed{\sigma^2(S)\ge\frac{1}{I(\lambda)}}$$

C) Error on the estimated parameter value

1) Single parameter

a) Asymptotic case:

For the limit of $N\to\infty$, i.e. a large number of measurements;

Likelihood function L becomes a Gauss distribution respect to the parameter λ

$$L=\prod_{j=1}^{N}f(x^{(j)};\lambda)\to L(\lambda)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(\lambda-\bar{\lambda})^2}{2\,\sigma^2}\right]$$

where

$$\bar{\lambda}=E(\lambda)$$
$$\sigma=\mathrm{Var}(\lambda)$$

The log likelihood function becomes

$$l(\lambda) = \ln L(\lambda) = -\frac{(\lambda-\bar{\lambda})^2}{2\sigma^2} + \text{constant}$$

The σ can be obtained as

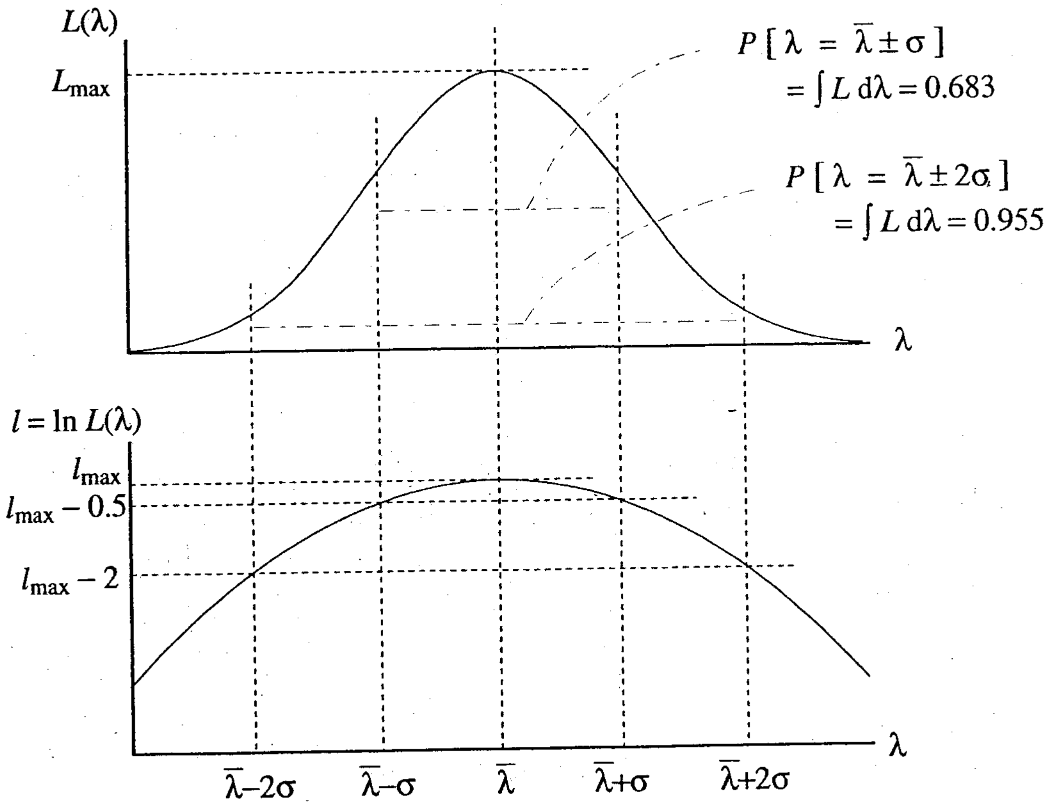
$$\sigma = \frac{1}{\sqrt{-\dfrac{\partial^2 \ln L(\lambda)}{\partial\lambda^2}}}$$

The 68 (96)% confidence interval for the estimated parameter λ is given by between $\bar{\lambda}-\sigma(2\sigma)$ and $\bar{\lambda}+\sigma(2\sigma)$, or $\lambda = \bar{\lambda}\pm\sigma(2\sigma)$; i.e.

$$\mathrm{P}\,[\, \bar{\lambda}-\sigma(2\sigma) < \lambda < \bar{\lambda}+\sigma(2\sigma)\,] = 0.683 \;\; (0.955).$$

Note: at $\lambda=\bar{\lambda}\pm\sigma(2\sigma)$, log likelihood function is reduced by 0.5 (2) from its maximum

$l(\lambda=\bar{\lambda}\pm\sigma) = l_{\max}-0.5$, $l(\lambda=\bar{\lambda}\pm 2\sigma) = l_{\max}-2$ with $l_{\max} = l(\bar{\lambda})$



Thus, procedure for the maximum likelihood method is;

i) build the log likelihood function $l(\lambda) = \ln L(\lambda)$

ii) find the parameter value $\bar{\lambda}$ which maximises l, i.e. $l_{\max} = l(\bar{\lambda})$

iii) determine the error on $\bar{\lambda}$ by

a) calculating

$$\sigma = \frac{1}{\sqrt{-\dfrac{\partial^2 \ln L(\lambda)}{\partial \lambda^2}}}$$

at $\lambda = \bar{\lambda}$.

or

b) calculating two points where

$l(\lambda_-) = l(\lambda_+) = l_{\max} - 0.5$: $\lambda_- < \lambda_+$

Then $\sigma = \bar{\lambda} - \lambda_- = \lambda_+ + \bar{\lambda}$.

The probability that the true $E(\lambda)$ is between $\bar{\lambda}+\sigma$ and $\bar{\lambda}-\sigma$ is 68% and between $\bar{\lambda}+2\sigma$ and $\bar{\lambda}-2\sigma$ is 96%.

b) General case:

Assume that the log likelihood function $l = \ln L(\boldsymbol{x},\lambda)$ is a continuous function of λ and has its maximum $l_{\max}$ at $\lambda = \bar{\lambda}$. The expectation value of λ is the estimated to be $\bar{\lambda}$. At around $l = l_{\max}$, there exists a transformation $g = g(\boldsymbol{x},\lambda)$ which transform l into a parabolic form;

$$l_g = \ln L(x,g) = -\frac{[g - G(x)]^2}{2\,\sigma_g^2} + \text{constant}$$

where the estimated expectation value of g, $G(\boldsymbol{x})$ is given by $G(\boldsymbol{x}) = g(\boldsymbol{x},\bar{\lambda})$. It can be shown that G does not depend on $\boldsymbol{x}$ for a large number of measurements N. As shown previously, the 68% and 96% confidence intervals in the g parameter space are given by $G-\sigma_g<g<G+\sigma_g$ and $G-2\sigma_g<g<G+2\sigma_g$ respectively and we have

$$l_g(G \pm \sigma_g) = l_g(G) - \frac{1}{2}$$

and

$$l_g(G \pm 2\sigma_g) = l_g(G) - 2$$

Transformations of the confidence intervals into the original λ parameter space, $\lambda_- < \lambda < \lambda_+$ and $\lambda_{--} < \lambda < \lambda_{++}$ are given by

$$l(\lambda_+) = l(\lambda_-) = l_{\max} - \frac{1}{2}$$

and

$$l(\lambda_{++}) = l(\lambda_{--}) = l_{\max} - 2$$

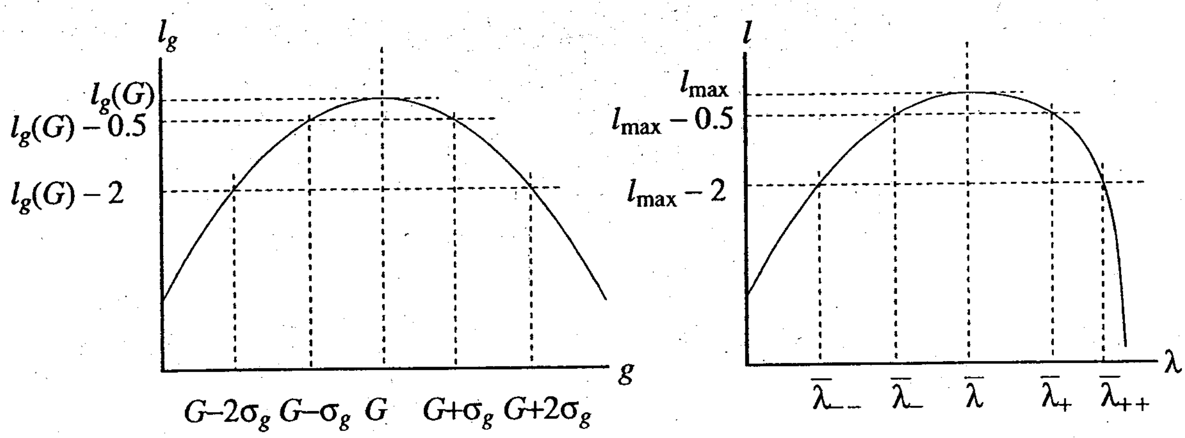
So, we do not need know the actual transformation $g = g(\boldsymbol{x},\lambda)$ in order to obtain the confidence interval on λ. The procedure for the maximum likelihood method is identical to the asymptotic case up to the second step. The error determination is done by the method b).

Note: in the most general case

$\bar{\lambda} - \lambda_- \neq \lambda_+ - \bar{\lambda}$, i.e. errors are **asymmetric** and

$\bar{\lambda}_{++}-\bar{\lambda} \neq 2(\bar{\lambda}_{+}-\bar{\lambda})$, i.e. 68% confidence interval is

not related to 96% confidence interval



2) Multi-parameters

The log likelihood function with p parameters $\boldsymbol{\lambda}=(\lambda_1, \lambda_2, \ldots \lambda_p)$ can be expanded around the expectation values as

$$l(\boldsymbol{\lambda})=l(\bar{\boldsymbol{\lambda}})+\sum_{k=1}^{p}\left(\frac{\partial l}{\partial \lambda_k}\right)_{\boldsymbol{\lambda}=\bar{\boldsymbol{\lambda}}}(\lambda_k-\bar{\lambda}_k)-\frac{1}{2}\sum_{l=1}^{p}\sum_{m=1}^{p}\left(\frac{\partial^2 l}{\partial \lambda_l \partial \lambda_m}\right)_{\boldsymbol{\lambda}=\bar{\boldsymbol{\lambda}}}(\lambda_l-\bar{\lambda}_l)(\lambda_m-\bar{\lambda}_m)+\cdots$$

Since $\partial l(\boldsymbol{\lambda}) / \partial\lambda_i=0$ at $\lambda_i=\bar{\lambda}_i$, we obtain

$$l(\boldsymbol{\lambda})=l(\bar{\boldsymbol{\lambda}})-\frac{1}{2}(\boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}})^{\mathrm{t}}H(\boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}})+\cdots$$

where

$$H=\begin{pmatrix}\dfrac{\partial^2 l}{\partial\lambda_1^2} & \dfrac{\partial^2 l}{\partial\lambda_1\partial\lambda_2} & \cdots \\ \dfrac{\partial^2 l}{\partial\lambda_1\partial\lambda_2} & \dfrac{\partial^2 l}{\partial\lambda_2^2} & \\ \vdots & & \ddots\end{pmatrix}_{\boldsymbol{\lambda}=\bar{\boldsymbol{\lambda}}}$$

The higher order terms can be neglected in the region where $\boldsymbol{\lambda}$ is very close to their expectation values. The likelihood function is then given by

$$L=k\exp\left[-\frac{1}{2}(\boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}})^{\mathrm{t}}H(\boldsymbol{\lambda}-\bar{\boldsymbol{\lambda}})\right]$$

where k is the normalisation constant. As discussed in section 4-i, H^{-1} can be identified as the covariant matrix given by

$$C=H^{-1}=\begin{pmatrix}\sigma_1^2 & \mathrm{cov}(1,2) & \cdots \\ \mathrm{cov}(1,2) & \sigma_2^2 & \\ \vdots & & \ddots\end{pmatrix}$$

a)Confidence region for two parameter case

A simple demonstration for the confidence region can be given for a case with uncorrelated two parameters having the same standard deviation σ. The likelihood function is given by

$$L(\lambda_1, \lambda_2) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{(\lambda_1-\bar{\lambda}_1)^2}{2\sigma^2}\right]\exp\left[-\frac{(\lambda_2-\bar{\lambda}_2)^2}{2\sigma^2}\right]$$

The probability to have λ_1 between $\bar{\lambda}_1-\sigma$ and $\bar{\lambda}_1+\sigma$ and λ_2 between $-\infty$ and $+\infty$ is given by

$$P(\bar{\lambda}_1-\sigma<\lambda_1<\bar{\lambda}_1+\sigma, -\infty<\lambda_2<+\infty) = \int_{\bar{\lambda}_1-\sigma}^{\bar{\lambda}_1+\sigma} d\lambda_1 \int_{-\infty}^{+\infty} d\lambda_2\, L(\lambda_1, \lambda_2)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma}\int_{\bar{\lambda}_1-\sigma}^{\bar{\lambda}_1+\sigma} d\lambda_1 \exp\left[-\frac{(\lambda_1-\bar{\lambda}_1)^2}{2\sigma^2}\right] = 0.683$$

The probability to have λ_1 between $\bar{\lambda}_1-\sigma$ and $\bar{\lambda}_1+\sigma$ and λ_2 between $\bar{\lambda}_2-\sigma$ and $\bar{\lambda}_2+\sigma$ is given by

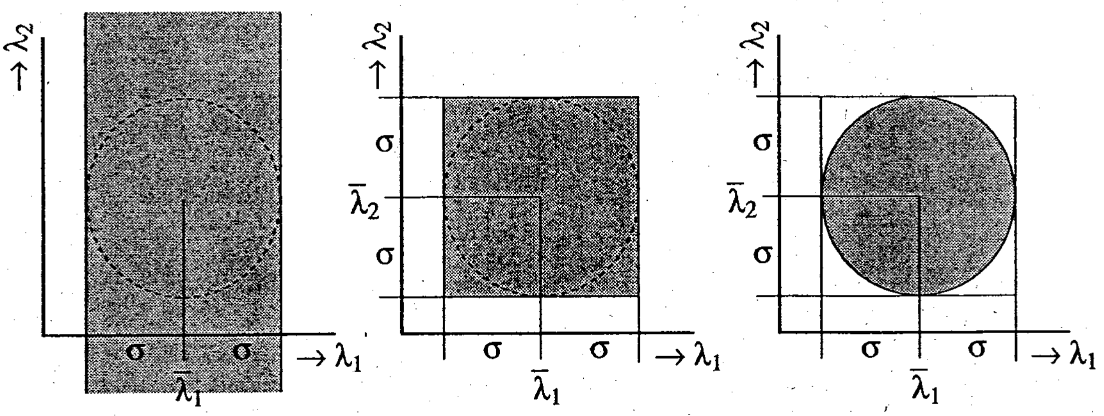
$$P(\bar{\lambda}_1-\sigma<\lambda_1<\bar{\lambda}_1+\sigma, \bar{\lambda}_2-\sigma<\lambda_2<\bar{\lambda}_2+\sigma) = \int_{\bar{\lambda}_1-\sigma}^{\bar{\lambda}_1+\sigma} d\lambda_1 \int_{\bar{\lambda}_2-\sigma}^{\bar{\lambda}_2+\sigma} d\lambda_2\, L(\lambda_1, \lambda_2) = 0.466$$

Finally, the probability to have λ_1 and λ_2 within the one standard deviation contour is obtained by

$$P[(\lambda_1, \lambda_2) < (\bar{\lambda}_1 \pm \sigma, \bar{\lambda}_2 \pm \sigma)] = \int_{(\lambda-\bar{\lambda}_1)^2+(\lambda-\bar{\lambda}_2)^2<\sigma^2} d\lambda_1\, d\lambda_2\, L(\lambda_1, \lambda_2)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma^2}\int_0^{2\pi} d\phi \int_0^{\sigma} r\, dr \exp\left(-\frac{r^2}{2\sigma^2}\right) = 0.393$$

This means that when the maximum likelihood method yields estimations for the expectation values for λ_1 and λ_2 and their variance σ, the probability that the true expectation value for λ_1 lies within ±σ while λ_2 can be anywhere is 68%. The probability that the true expectation values for both parameters lie within ±σ is only 47% and that the true expectation values for both parameters lie within the circle with a radius σ is 39%.

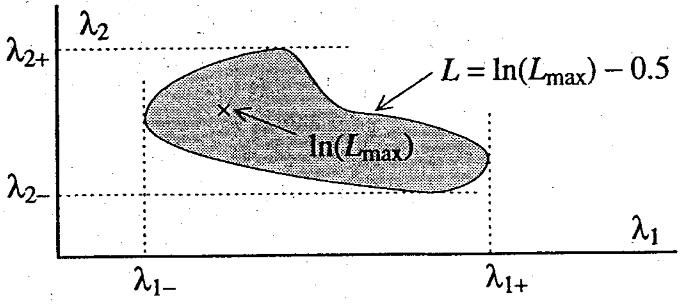


Note that the one standard deviation (λ_1, λ_2) contour is identical to the contour for

$\ln\{L(\lambda_1, \lambda_2)\} = \ln(L)_{\max} - 0.5.$

General Case:

⇒ As in the case of the single parameter, when the likelihood function is no longer Gaussian, the confidence regions is given by the equal likelihood (λ_1, λ_2) contour with $\ln\{L(\lambda_1, \lambda_2)\} = \ln(L)_{\max} - 0.5$,



where the probability that the true expectation value of λ_1 lies between λ_{1-} and λ_{1+} is 68% and that the true expectation values of λ_1 and λ_2 lie inside of the contour is 39%. The region given by $\lambda_{1-}<\lambda_1<\lambda_{1+}$ and $\lambda_{2-}<\lambda_2<\lambda_{2+}$ (rectangle shape) is usually not used since its statistical meaning is unclear.

b)Confidence region for many parameters

-Single parameter confidence intervals are identical to the two parameter case and λ_{1-} and λ_{1+} are given by

$$\ln\{L(\lambda_{1\pm}, \lambda_2 \cdots \lambda_n)\}=\ln(L_{\max})-0.5$$

where $L(\lambda_{1\pm}, \lambda_2 \cdots \lambda_n)$ is maximised respect all λ's except λ_1. Then the probability that the true expectation value of λ_1 lies between λ_{1-} and λ_{1+} irrespective of all the other parameters is 68%.

-The probability that all n parameters are within the equal likelihood contour drawn by $\ln\{L(\lambda_1, \lambda_2 \cdots \lambda_p)\}=\ln(L_{\max})-down$ is given by $F(2down)$ where $F(2down)$ is the integral of chi square distribution with n degrees of freedom from 0 to 2×*down*;

$$F(2\times down) = \frac{1}{\Gamma(n/2)\,2^{n/2}} \int_0^{2\times down} u^{n/2-1}\, e^{-u/2}\, du$$

For example,

No. of parameters	down	probability	down	probability
2	0.5	0.393	1.0	0.632
3	0.5	0.199	1.0	0.428
4	0.5	0.090	1.0	0.264

9) THE METHOD OF LEAST SQUARES

A) Simple example: direct measurement

A set of n measurements

$a_i = E(a) + \varepsilon_i$; E expectation value, i=1 to N, ε_i error distributed normally around 0

$$E(\varepsilon_i) = 0,\ E(\varepsilon_i^2) = \sigma_i^2$$

The probability distribution for a_i is given by

$$\frac{1}{\sqrt{2\pi}\,\sigma_i}\exp\left[-\frac{(a_i - E(a))^2}{2\,\sigma_i^2}\right]$$

The log likelihood function is then given by

$$\ln L = \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma_i}\exp\left[-\frac{(a_i - E(a))^2}{2\,\sigma_i^2}\right] = -\frac{1}{2}\sum_{i=1}^{N}\frac{(a_i - E(a))^2}{\sigma_i^2} + \text{constant}$$

maximising $\ln L \longrightarrow$

minimising

$$M = \sum_{i=1}^{N}\frac{(a_i - E(a))^2}{\sigma_i^2}$$

The method of least squares:

$\hat{a}$: the best estimate for the expectation value of a gives minimum M

$$\left[\frac{\partial M}{\partial a}\right]_{a=\hat{a}} = -2\sum_{i=1}^{n}\frac{(a_i - \hat{a})}{\sigma_i^2} = 0$$

$$\sum_{i=1}^{N}\frac{a_i}{\sigma_i^2} = \hat{a}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}$$

i.e.

$$\boxed{\hat{a} = \frac{\sum_{i=1}^{N}\frac{a_i}{\sigma_i^2}}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}}$$

and the variance of $\hat{a}$ is given by

$$\boxed{\sigma^2(\hat{a}) = \frac{1}{\sum_{i=1}^{N}\frac{1}{\sigma_i^2}}}$$

Note that the newly estimated error $\varepsilon_i' = a_i - \hat{a}$ follows a normal distribution with the mean 0 and the variance σ_i^2. Then the minimised M

$$M_{\text{minimum}} = \sum_{i=1}^{N}\frac{(a_i - \hat{a})^2}{\sigma_i^2}$$

follows the chi-square distribution with N–1 degrees of freedom, i.e.

$$E(M_{\text{minimum}}) = N-1$$

$\longrightarrow \chi^2$ test (see the next chapter) example: two experiments E731 at FNAL and NA31 at CERN measured a CP violation parameter Re(ε'/ε) to be

$$\mathrm{Re}\left(\frac{\varepsilon'}{\varepsilon}\right)=\left(7.4\pm 5.9\right)\times 10^{-4}\ \text{(E731)},\ \left(23.0\pm 6.5\right)\times 10^{-4}\ \text{(NA31)}$$

The method of least squares gives

$$\mathrm{Re}\left(\frac{\varepsilon'}{\varepsilon}\right)=\left(14.4\pm 4.4\right)\times 10^{-4}$$

and

$$M_{\text{minimum}}=3.16 \quad \text{(expectation value = 1)}$$

B) Indirect Measurement

In general, a measured quantity is a function of unknown parameters. A set of n observations

$$\boldsymbol{a}=\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

is obtained from a distribution with expectations

$$E(\boldsymbol{a})=\begin{pmatrix} f_1\,[E(\boldsymbol{\lambda})] \\ \vdots \\ f_n\,[E(\boldsymbol{\lambda})] \end{pmatrix}$$

where f_i is a function of p unknown parameters

$$\boldsymbol{\lambda}=\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_p \end{pmatrix}$$

The method of least squares is to estimate the expectation values of the unknown parameters $\boldsymbol{\lambda}$ by minimising

$$M=[\boldsymbol{a}-\boldsymbol{f}\,(\boldsymbol{\lambda})]^{\mathrm{t}}\,W[\boldsymbol{a}-\boldsymbol{f}\,(\boldsymbol{\lambda})]$$

with $W=V^{-1}$ where V is the covariant matrix of $\boldsymbol{a}$, i.e.

$$V=\begin{pmatrix} \sigma_1^2 & \mathrm{cov}(1,2) & \cdots \\ \mathrm{cov}(1,2) & \sigma_2^2 & \\ \vdots & & \ddots \end{pmatrix}$$

i) Linear case

The function f_i depends linearly on $\boldsymbol{\lambda}$, i.e.

$$E(a_i)=\sum_{j=1}^{p} c_{ij}\,\hat{\lambda}_j \quad \text{or} \quad \hat{\boldsymbol{a}}=C\,\hat{\boldsymbol{\lambda}}$$

The best estimate for the expectation values of $\boldsymbol{\lambda}$ can be obtained by minimising

$$M=(\boldsymbol{a}-C\,\boldsymbol{\lambda})^{\mathrm{t}}\,W\,(\boldsymbol{a}-C\,\boldsymbol{\lambda})$$

At its minimum, we have

$$\delta M=-\left(\boldsymbol{a}-C\,\hat{\boldsymbol{\lambda}}\right)^{\mathrm{t}}\,W\,C\,\delta\hat{\boldsymbol{\lambda}}=-\left(\boldsymbol{a}^{\mathrm{t}}\,W\,C-\hat{\boldsymbol{\lambda}}^{\mathrm{t}}\,C^{\mathrm{t}}\,W\,C\right)\delta\hat{\boldsymbol{\lambda}}=0$$

By noting $W^{\mathrm{t}}=W$, it follows that

$$\hat{\boldsymbol{\lambda}}=(C^{\mathrm{t}}\,W\,C)^{-1}\,C^{\mathrm{t}}\,W\,\boldsymbol{a}$$

and using the error propagation shown in Chapter 5, it follows that

$$V(\hat{\boldsymbol{\lambda}}) = \left[(C^{\mathrm{t}}\, W\, C)^{-1}\, C^{\mathrm{t}}\, W\right] V(\boldsymbol{a}) \left[(C^{\mathrm{t}}\, W\, C)^{-1}\, C^{\mathrm{t}}\, W\right]^{\mathrm{t}} = \left(C^{\mathrm{t}}\, V^{-1}(\boldsymbol{a})\, C\right)^{-1}$$

where we used $W=V^{-1}(\boldsymbol{a})$, $V^t=V$ and $W^t=W$. The improved measurements and their errors are given by

$$\hat{\boldsymbol{a}} = C\,\hat{\boldsymbol{\lambda}} \text{ and } V(\hat{\boldsymbol{a}}) = C\, V(\hat{\boldsymbol{\lambda}})\, C^{\mathrm{t}}$$

Simple example: The case of direct measurement can be obtained by

$$\boldsymbol{\lambda} = \lambda,\ \boldsymbol{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix},\ W = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots \\ 0 & \frac{1}{\sigma_2^2} & \\ \vdots & & \ddots \end{pmatrix} \text{ and } C = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

In this case, we obtain

$$\hat{\lambda} = (C^t\, W\, C)^{-1}\, C^{\mathrm{t}}\, W\, \boldsymbol{a} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots \frac{1}{\sigma_n^2}\right)^{-1} \left(\frac{a_1}{\sigma_1^2} + \frac{a_2}{\sigma_2^2} + \cdots \frac{a_N}{\sigma_N^2}\right) = \frac{\sum_{i=1}^{N} \frac{a_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

and

$$V(\hat{\lambda}) = (C^{\mathrm{t}}\, W\, C)^{-1} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \cdots \frac{1}{\sigma_N^2}\right)^{-1} = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

which agrees with the case of a direct measurement.

ii) Non linear case

Even if the function f_i no longer depends linearly on $\boldsymbol{\lambda}$, it can be expanded as

$$f_i(\boldsymbol{\lambda}) = f_i(\boldsymbol{\lambda}_0) + \sum_{j=1}^{p} \left(\frac{\partial f_i(\boldsymbol{\lambda})}{\partial \lambda_j}\right)_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)_j + \cdots \equiv f_i^{\,0} + \sum_{j=1}^{p} c_{i\,j}\, \delta\lambda_j + \cdots$$

where $\boldsymbol{\lambda}_0$ is a starting value for estimating the expectation value of $\boldsymbol{\lambda}$. The function M is then given by

$$M = \left(\boldsymbol{a} - C\,\delta\boldsymbol{\lambda} - \boldsymbol{f}^{\,0} - \cdots\right)^{\mathrm{t}} W \left(\boldsymbol{a} - C\,\delta\boldsymbol{\lambda} - \boldsymbol{f}^{\,0} - \cdots\right)$$

where

$$C = \begin{pmatrix} \left(\frac{\partial f_1(\boldsymbol{\lambda})}{\partial \lambda_1}\right)_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} & \left(\frac{\partial f_1(\boldsymbol{\lambda})}{\partial \lambda_2}\right)_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} & \cdots \\ \left(\frac{\partial f_2(\boldsymbol{\lambda})}{\partial \lambda_1}\right)_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} & \left(\frac{\partial f_2(\boldsymbol{\lambda})}{\partial \lambda_2}\right)_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} & \vdots \\ \vdots & \cdots & \ddots \end{pmatrix} \text{ and } \boldsymbol{f}^{\,0} = \begin{pmatrix} f_1^{\,0} \\ f_2^{\,0} \\ \vdots \end{pmatrix}$$

If $\boldsymbol{\lambda}_0$ is close to its expectation value, the higher orders in $\delta\boldsymbol{\lambda}$ can be neglected (linear approximation). At $M=M_{\mathrm{minimum}}$, we have

$$\delta M = -\left(a - C\,\delta\boldsymbol{\lambda} - f^{0}\right)^{\mathrm{t}} W\, C\, \delta\boldsymbol{\lambda} =$$

$$-\left[\left(a - f^{0}\right)^{\mathrm{t}} W\, C - \delta\boldsymbol{\lambda}^{\mathrm{t}}\, C^{\mathrm{t}}\, W\, C\right] \delta\boldsymbol{\lambda} = 0$$

and it follows that

$$\boxed{\delta\hat{\boldsymbol{\lambda}} = \left(C^{\mathrm{t}}\, W\, C\right)^{-1} C^{\mathrm{t}}\, W\left(a - f^{0}\right)}$$

and the estimation of $E(\boldsymbol{\lambda}\,)$ is given by

$$\boldsymbol{\lambda}_{1} = \boldsymbol{\lambda}_{0} + \delta\hat{\boldsymbol{\lambda}}$$

Then we use $\boldsymbol{\lambda}_{1}$ as a starting value and repeat the process again with newly calculated C and f. When the change in $\boldsymbol{\lambda}$ becomes "sufficiently" small after m iteration, the process is considered to be converged and $\boldsymbol{\lambda}_{m}$ is taken as the best estimate. A rapid convergence is obtained if the starting value is close to the true value and the linear approximation is valid. The covariance matrix of $\boldsymbol{\lambda}_{m}$ is given by

$$\boxed{V(\boldsymbol{\lambda}_{m}) = \left(C^{\mathrm{t}}_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{m-1}}\, V^{-1}\, C_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{m-1}}\right)^{-1}}$$

and "improved" measurements and their errors are given by

$$\boxed{a = f^{0}_{m-1} + C_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{m-1}}\left(\boldsymbol{\lambda}_{m} - \boldsymbol{\lambda}_{m-1}\right)}$$

and

$$\boxed{V(\hat{a}) = C^{\mathrm{t}}_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{m-1}}\, V(\boldsymbol{\lambda}_{m})\, C_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_{m-1}}}$$

It must be noted that M_{minimum} follows the chi-square distribution with

$N - p$ degrees of freedom

$E(M_{\text{minimum}}) = N - p$

C) Measurements with constraints and no unknown parameter

Lagrangian multipliers are a commonly used mathematical tool for the minimisation of a function with constraints. Suppose we want ti minimise a function $f(x, y)$ under the constraint $g(x,y)$=constant. It follows that

$$\mathrm{d}f(x,y) = \frac{\partial f}{\partial x}\mathrm{d}x + \frac{\partial f}{\partial y}\mathrm{d}y = 0, \quad \text{and} \quad \mathrm{d}g(x,y) = \frac{\partial g}{\partial x}\mathrm{d}x + \frac{\partial g}{\partial y}\mathrm{d}y = 0$$

which can be satisfied by having

$$\frac{\dfrac{\partial f}{\partial x}}{\dfrac{\partial f}{\partial x}} = \frac{\dfrac{\partial f}{\partial y}}{\dfrac{\partial g}{\partial y}} \equiv \mu$$

This leads to

$$\frac{\partial f}{\partial x} - \mu\frac{\partial g}{\partial x} = 0, \quad \text{and} \quad \frac{\partial f}{\partial y} - \mu\frac{\partial g}{\partial y} = 0$$

which is equivalent to minimising

$$\boxed{L = f - \mu g}$$

where μ is the Lagrangian multiplier.

We consider a case where the estimated expectation values of N measurements $a^{\mathrm{t}} = (a_1, a_2, \ldots a_n)$ must fulfil q constraints,

$$\begin{pmatrix} g_1(\hat{a}) \\ g_2(\hat{a}) \\ \vdots \\ g_q(\hat{a}) \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_q \end{pmatrix} .$$

The estimates $\hat{a}$ are obtained by minimising

$$L = (a - \hat{a})^{\mathrm{t}}\, W\, (a - \hat{a}) + \mu^{\mathrm{t}}\, g\, (\hat{a})$$

i.e.

$$\delta L = -(a - \hat{a})^{\mathrm{t}}\, W\, \delta\hat{a} + \mu^{\mathrm{t}}\, \delta g\, (\hat{a}) = 0$$

where $\mu^{\mathrm{t}} = (\mu_1, \mu_2, \ldots \mu_q)$ are the q Lagrangian multipliers.

i) When the constraints are linear functions of $\boldsymbol{a}$, i.e.

$$g\,(\hat{a}) + d = B\, a + d,$$

it follows that

$$L = (a - \hat{a})^{\mathrm{t}}\, W\, (a - \hat{a}) + \mu^{\mathrm{t}}\, (B\, \hat{a} + d)$$

and where L is minimum at $a = \hat{a}$, i.e.

$$\delta L = -\ (a - \hat{a})^{\mathrm{t}}\, W\, \delta\hat{a} + \mu^{\mathrm{t}}\, B\, \delta\hat{a} = 0$$

This requires that

$$(a - \hat{a})^{\mathrm{t}}\, W = \mu^{\mathrm{t}}\, B$$

i.e.

$$\hat{a} = a - W^{-1} B^{\mathrm{t}}\, \mu$$

The constraints are now

$$B\, \hat{a} - d = B\, (a - W^{-1} B^{\mathrm{t}}\, \mu\,) - d = 0$$

which gives us the estimation of μ to be

$$\hat{\mu} = (B\, W^{-1} B^{\mathrm{t}})^{-1}\, (B\, a - d\,).$$

Thus the estimation $\hat{a}$ is given by

$$\boxed{\hat{a} = a - W^{-1} B^{\mathrm{t}}\, (B\, W^{-1} B^{\mathrm{t}})^{-1}\, (B\, a - d\,).}$$

The covariance matrix for $\hat{a}$ is obtained by propagating errors and

$$\boxed{C\,(\hat{a}) = [\, W^{-1} - W^{-1}\, B^{\mathrm{t}}\, (B\, W^{-1} B^{\mathrm{t}})^{-1}\, B\, W^{-1}\,]^{-1}}$$

Example: measurements of three angles of a triangle, $a_1 \pm \sigma_1$, $a_2 \pm \sigma_2$, $a_3 \pm \sigma_3$ which should satisfy $a_1 + a_2 + a_3 = 180°$. The improved measurements which fulfil this constraint can be obtained using the method of least squares shown above with

$$a=\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix},\ B=\begin{pmatrix} 1 & 1 & 1 \end{pmatrix},\ d=180^\circ,\ W=\begin{pmatrix} 1/\sigma_1^2 & 0 & 0 \\ 0 & 1/\sigma_2^2 & 0 \\ 0 & 0 & 1/\sigma_3^2 \end{pmatrix}$$

ii)When constrains become non-linear, we need to expand $g\,(\hat{a})$ by introducing

$$g_i(\hat{a}) = g_i(a_0) + \sum_{j=1}^{N} \left(\frac{\partial f_i(a)}{\partial a_j}\right)_{a=a_0} (\hat{a}-a_0)_j + \cdots \equiv g_i^0 + \sum_{j=1}^{N} c_{ij}\,\delta a_j + \cdots$$

where a_0 are initial values for $\hat{a}$ (measured values of a for example). It follows that

$$\boxed{\delta L = -\ (a - a_0 - \delta a\ )^{t}\, W\, \delta a + \ \mu^{t}\, B\delta a}$$

where

$$B=\begin{pmatrix} \left(\frac{\partial g_1(a)}{\partial a_1}\right)_{a=a_0} & \left(\frac{\partial g_1(a)}{\partial a_2}\right)_{a=a_0} & \cdots \\ \left(\frac{\partial g_2(a)}{\partial a_1}\right)_{a=a_0} & \left(\frac{\partial g_2(a)}{\partial a_2}\right)_{a=a_0} & \vdots \\ \vdots & \cdots & \ddots \end{pmatrix} \text{ and } g_0=\begin{pmatrix} g_1^0 \\ g_2^0 \\ \vdots \\ g_q^0 \end{pmatrix}$$

Using $\delta L = 0$

$$(\hat{a} - a)^{t}\, W = -\, \mu^{t}\, B$$

with

$$\hat{a} = a_0 + \delta a$$

we obtain

$$\delta a = (a - a_0) - W^{-1} B^{t}\, \mu$$

and from the constraint equation

$$g^0 + B\ \delta\hat{a} - d = g^0 + B\ [(a - a_0) - W^{-1} B^{t}\, \mu\ \ ] - d = 0$$

a estimate for the multiplier is given as

$$\hat{\mu} = (B\ W^{-1} B^{t})^{-1}\ [g^0 + B\ (a - a_0) - d\ ]$$

thus improved measurements are give by

$$\boxed{\hat{a} = a_0 + \delta a = a\ -\ W^{-1} B^{t}\ (B\ W^{-1} B^{t})^{-1}\ \ [g^0 + B\ (a - a_0) - d\ ].}$$

By replacing a_0 with $\hat{a}$ and recalculating g^0 and B, we repeat m times this procedure till δa becomes sufficiently small. Once this procedure converges, error propagation gives

$$\boxed{C\,(\hat{a}) = [\ W^{-1}\ -\ W^{-1}\ B^{t}\ (B\ W^{-1} B^{t})^{-1}\ B\ W^{-1}\ ]^{\ -1}}$$

where B is calculated by $\hat{a}$ obtained from the m–1-th iteration.

It must be noted that M_{minimum} follows the chi-square distribution with

q degrees of freedom.

$$E(M_{\text{minimum}}) = q$$

D) Measurements with constraints and unknown parameters

Finally, we consider a system with N measurements, $\boldsymbol{a}$, and r unknowns, $\boldsymbol{y}$, and q constraints, $\boldsymbol{g}$, which are the functions of $\boldsymbol{a}$ and $\boldsymbol{y}$;

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{pmatrix} \boldsymbol{g} = \begin{pmatrix} g_1(\boldsymbol{a}, \boldsymbol{y}) \\ g_2(\boldsymbol{a}, \boldsymbol{y}) \\ \vdots \\ g_q(\boldsymbol{a}, \boldsymbol{y}) \end{pmatrix} .$$

An example is given in Übungen XI. The best estimate for $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and the improved measurements $\hat{\boldsymbol{a}}$ are obtained by minimising

$$L = (\boldsymbol{a} - \hat{\boldsymbol{a}})^t W (\boldsymbol{a} - \hat{\boldsymbol{a}}) + \boldsymbol{\mu}^t \boldsymbol{g}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{y}})$$

i.e.

$$\delta L = -(\boldsymbol{a} - \hat{\boldsymbol{a}})^t W \delta\hat{\boldsymbol{a}} + \boldsymbol{\mu}^t \delta\boldsymbol{g}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{y}}) = 0$$

under the constraints

$$\boldsymbol{g}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{y}}) = \boldsymbol{d}$$

Linearization of $\boldsymbol{g}$ gives

$$\boldsymbol{g}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{y}}) = \boldsymbol{g}^0 + A\,\delta\boldsymbol{y} + B\,\delta\boldsymbol{a}$$

with

$$B = \begin{pmatrix} \left(\dfrac{\partial g_1(\boldsymbol{a}, \boldsymbol{y})}{\partial a_1}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} & \left(\dfrac{\partial g_1(\boldsymbol{a}, \boldsymbol{y})}{\partial a_2}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} & \cdots \\ \left(\dfrac{\partial g_2(\boldsymbol{a}, \boldsymbol{y})}{\partial a_1}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} & \left(\dfrac{\partial g_2(\boldsymbol{a}, \boldsymbol{y})}{\partial a_2}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} \ddots & \vdots \\ \vdots & \cdots & \left(\dfrac{\partial g_q(\boldsymbol{a}, \boldsymbol{y})}{\partial a_N}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} \end{pmatrix}$$

and

$$A = \begin{pmatrix} \left(\dfrac{\partial g_1(\boldsymbol{a}, \boldsymbol{y})}{\partial y_1}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} & \left(\dfrac{\partial g_1(\boldsymbol{a}, \boldsymbol{y})}{\partial y_2}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} & \cdots \\ \left(\dfrac{\partial g_2(\boldsymbol{a}, \boldsymbol{y})}{\partial y_1}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} & \left(\dfrac{\partial g_2(\boldsymbol{a}, \boldsymbol{y})}{\partial y_2}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} \ddots & \vdots \\ \vdots & \cdots & \left(\dfrac{\partial g_r(\boldsymbol{a}, \boldsymbol{y})}{\partial y_p}\right)_{\substack{\boldsymbol{a}=\boldsymbol{a}_0 \\ \boldsymbol{y}=\boldsymbol{y}_0}} \end{pmatrix}$$

and

$$g_0 = \begin{pmatrix} g_1(\boldsymbol{a}_0, y_0) \\ g_2(\boldsymbol{a}_0, y_0) \\ \vdots \\ g_q(\boldsymbol{a}_0, y_0) \end{pmatrix} .$$

As before, $\boldsymbol{a}_0$ and y_0 are the initial values for $\hat{\boldsymbol{a}}$ and $\hat{y}$, i.e.

$$\hat{\boldsymbol{a}} = \boldsymbol{a}_0 + \delta\boldsymbol{a} \quad \text{and} \quad \hat{y} = y_0 + \delta y \quad .$$

Thus, the problem is reduced to solve

$$\boxed{\begin{array}{l}(\boldsymbol{a}-\boldsymbol{a}_0-\delta\boldsymbol{a})W^{\mathrm{t}}-\mu^{\mathrm{t}}B=0\\ \mu^{\mathrm{t}}A=0\\ \boldsymbol{g}^0+A\delta\boldsymbol{y}+B\delta\boldsymbol{a}=\boldsymbol{d}\end{array}}$$

for $\delta\boldsymbol{a}$ and δy.

It must be noted that M_{minimum} follows the chi-square distribution with $q-r$ degrees of freedom.

$$E(M_{\mathrm{minimum}})=q-r$$

10) HYPOTHESIS TESTING

Let us define:

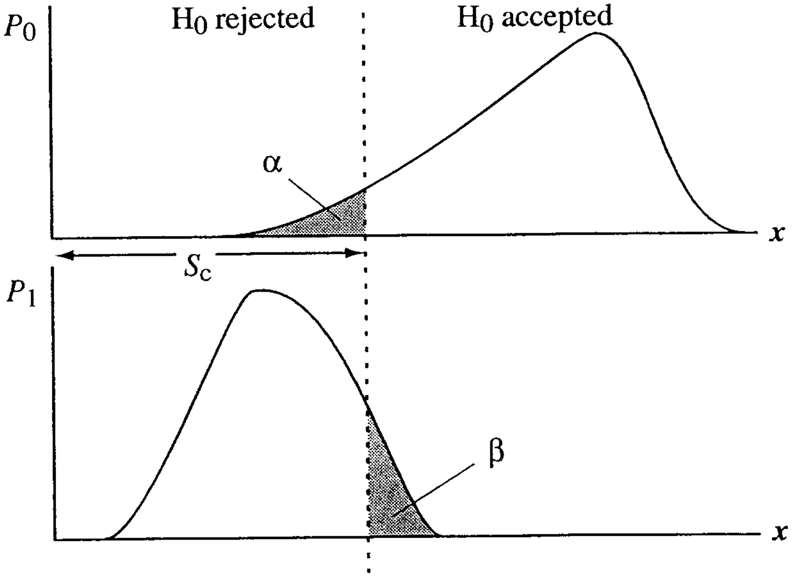
x: sample of variables, $\boldsymbol{\lambda}$: parameters

and

$\mathrm{H}_0(x,\lambda_0)$: null hypothesis, hypothesis to be tested

$\mathrm{H}_1(x,\lambda_1)$: alternative hypothesis

When λ_1 are completely fixed parameters, H_1 is called "simple" and else "composite". The critical region S_c is the region of x for the null hypothesis where we reject the H_0. The probability distribution of x for H_0 and H_1 are given by $P_0(x,\lambda_0)$ and $P_1(x,\lambda_1)$ respectively. The following figure illustrate the situation for the single variable case:



The area α is the probability for x to fall inside S_c under the null hypothesis: $P_0(x\in S_c, \lambda_0)$. The area β is the probability for x to fall outside of S_c under the alternative hypothesis: $P_1(x\notin S_c, \lambda_1)$.

Type-I error: Rejecting the null hypothesis although it is the correct one, since observed x is in S_c. The probability for Type-I error to occur is α, which is called "significance"

Type-II error: Accepting the null hypothesis although the correct one is the alternative hypothesis since observed x is outside of S_c. The probability for Type-I error to occur is β, where 1−β is called "power".

The critical region S_c has to be choose so that

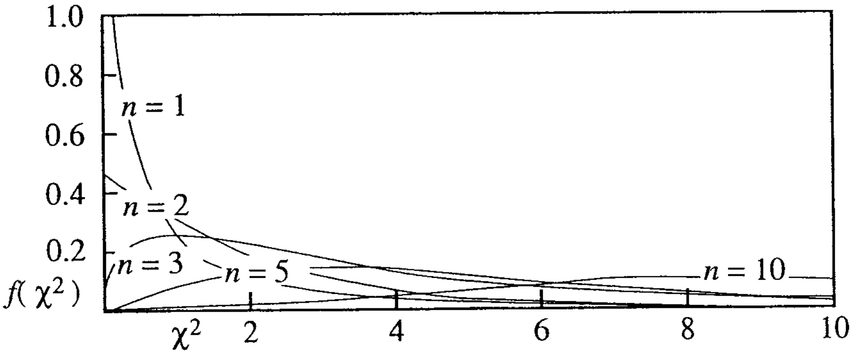
$$\boxed{\begin{array}{c}\alpha\text{: as small as possible}\\ 1-\beta\text{: as large as possible}\end{array}}$$

for an optimal hypothesis testing.

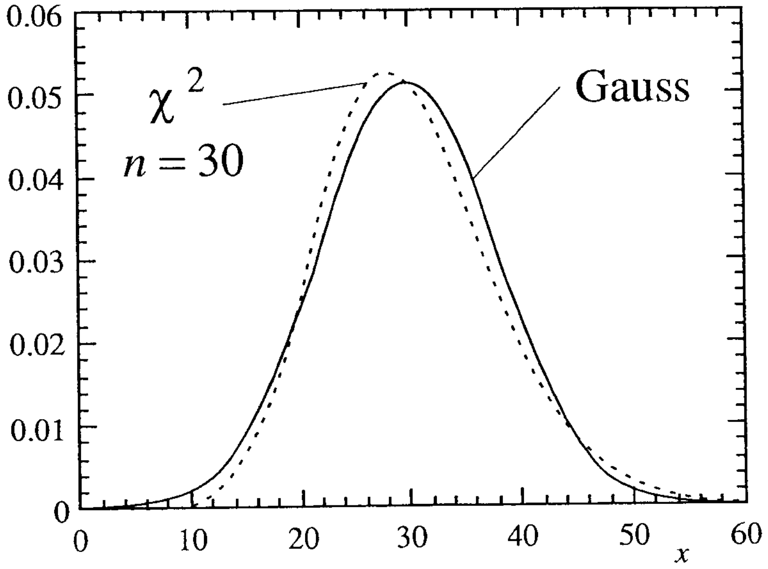
11) CHI-SQUARE TEST

The minimum value of the least square, often called "chi-square", is expected to follow the χ^2 probability distribution with n degrees of freedom, $f_n(\chi^2)$, if the measurements are described by the Gauss distribution. The degrees of freedom n is given by

N(number of measurements)	direct measurement
$N - p$(number of parameters)	indirect measurement
q(number of constraints) – r(number of unknowns)	with constraint.



The expectation value and variance are given by n and $2n$. For very large values of n, the χ^2 distribution can be described by a Gauss distribution with an expectation value n and σ^2=$2n$.



Let us denote s to be a value of chi-square obtained from the method of least squares. The probability that a value of chi-square will exceed s is given by

$$F_n(s) = 1 - \int_0^s f_n(\chi^2)\,\mathrm{d}\chi^2$$

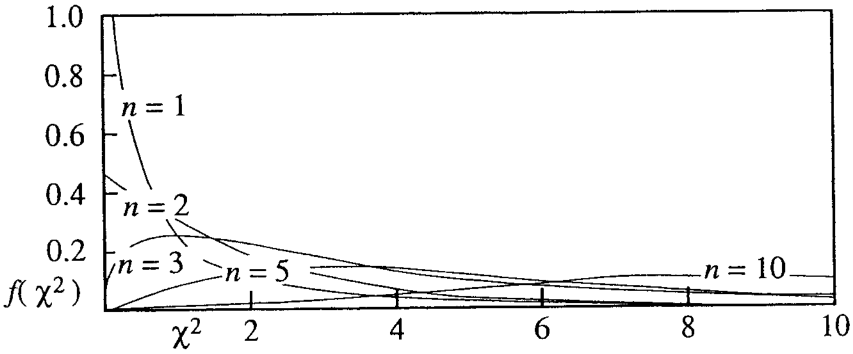
The value of $F_n(s)$ indicates the probability that the hypothesis applied is consistent with the data. When $F_n(s)$ is too small, we may conclude that the hypothesis applied to fit the data is not the right one. How small is too small? This is, unfortunately, up to you...

In the previous example of the two measurements of Re(ε' / ε), it gives

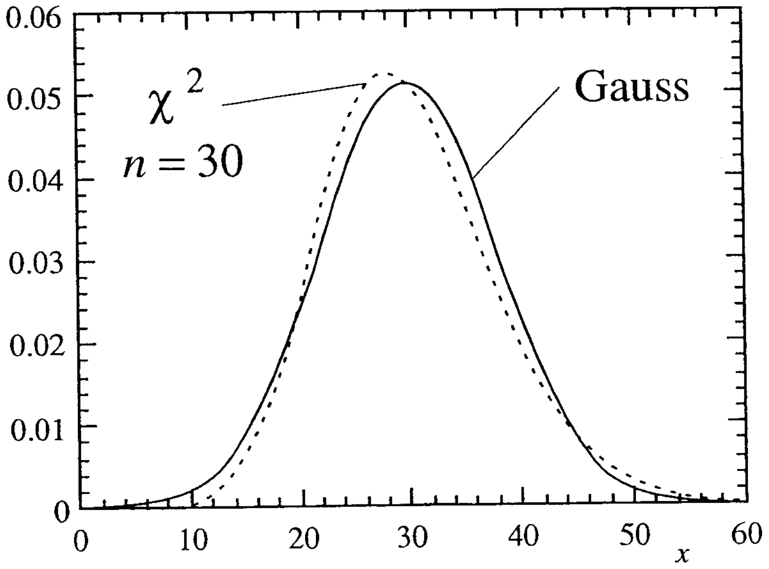
11) CHI-SQUARE TEST

The minimum value of the least square, often called "chi-square", is expected to follow the χ^2 probability distribution with n degrees of freedom, $f_n(\chi^2)$, if the measurements are described by the Gauss distribution. The degrees of freedom n is given by

N(number of measurements)	direct measurement
$N - p$(number of parameters)	indirect measurement
q(number of constraints) – r(number of unknowns)	with constraint.



The expectation value and variance are given by n and $2n$. For very large values of n, the χ^2 distribution can be described by a Gauss distribution with an expectation value n and σ^2=$2n$.



Let us denote s to be a value of chi-square obtained from the method of least squares. The probability that a value of chi-square will exceed s is given by

$$F_n(s) = 1 - \int_0^s f_n(\chi^2)\, d\chi^2$$

The value of $F_n(s)$ indicates ~~that~~ the probability that the hypothesis applied is consistent with the data. When $F_n(s)$ is too small, we may conclude that the hypothesis applied to fit the data is not the right one. How small is too small? This is, unfortunately, up to you...

In the previous example of the two measurements of Re(ε' / ε), it gives

$$1-\int_0^{3.16} f_{n=1}(\chi^2)\, d\chi^2 = 0.076$$

i.e. the probability that the two measurments are compatible is 7.6%.

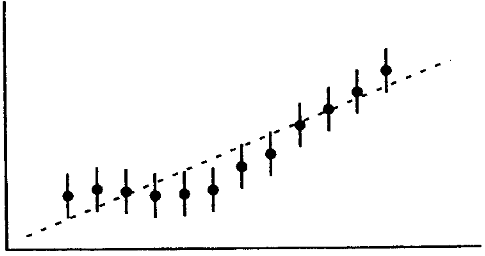
It must be noted that a too small value of chi-square is also suspicious. As seen from the χ^2 distribution for various n, not only too large values but also too small values of χ^2, i.e. much smaller that n, are improbable for n>3. A too mall value of χ^2 is often due to the overestimation of the experimental errors.

For the case where the fits are repeated many times, such as fitting track parameters seen by a spectrometer, the distribution of the obtained values of chi-square must be examined whether they indeed follow the χ^2 distribution with the corresponding number of degrees of freedom. An easy way is to plot the distribution of probabilities f_n(chi-square). If the values of chi-square indeed follow the χ^2 distribution, this probability distribution should be flat between 0 and 1.

Another useful application of the chi-square test is to choose the right hypothesis by selecting one which gives the largest $F_n(s)$. This does not avoid, of course, selecting a wrong hypothesis because the right one was not tested!!!!

12) RUN TEST

Consider the data and errors shown in the following figure:



The straight line obtained by the χ^2 test is also shown. The visual inspection indicates a systematic deviation of the data from the straight line hypothesis. However, the obtained vale of χ^2 is reasonably small, which could be due to overestimated errors. In this case, "the run test" provide an extra information.

There are six points above (A) and below (B) the straight line in a sequence, AAABBBBBBAAA., i.e. a run of A, a run of B then a run of B giving a total of three runs. What is the probability for having a particular number of runs, *r*, for a given number of A, N_A and B, N_B. The number of ways to have N_A and N_B for a given *N*, $N= N_A+N_B$ is given by,

$$C(N, N_A) = \frac{N!}{N_A! N_B!}$$

Next, suppose *r* is even and the sequence starts with A, for example

AA BBB AAAA BB AAA B

i.e. A's are divided in to

AA I AAAA I AAA

and there are $r/2-1$ division. There are N_A-1 places to put the first division, N_A-2 places for the second division etc. giving $C(N_A-1, r/2-1)$ possibilities for the sequences. A similar possibilities are valid for B's. The probability to have *r* (even) run is given by

$$P_r = 2\frac{C(N_A - 1, r/2 - 1) \times C(N_B - 1, r/2 - 1)}{C(N, N_A)} \qquad r\text{: even}$$

where 2 takes the cases starting with B into account. For *r* odd, we have

$$P_r = \frac{C(N_A - 1, r - 3/2) \times C(N_B - 1, r - 1/2) + C(N_A - 1, r - 1/2) \times C(N_B - 1, r - 3/2)}{C(N, N_A)} \qquad r\text{: odd}$$

From these, we obtain

$$E(r) = 1 + \frac{2N_A N_B}{N}$$

$$V(r) = \frac{2N_A N_B(2N_A N_B - N)}{N^2(N-1)}$$

When N_A and N_B are sufficiently large, >10 to 15, one can use a Gaussian approximation.

The example shown in the figure, we have $N = 12$ and $N_A = 6$. It follows $E(r) = 7$ and $V(r)=2.73$. To have $r = 3$ has a significance of ~1% which is rather low. Therefore, the straight line hypothesis can be rejected although χ^2 test accepts the hypothesis.

13) KOLMOGOROV TEST

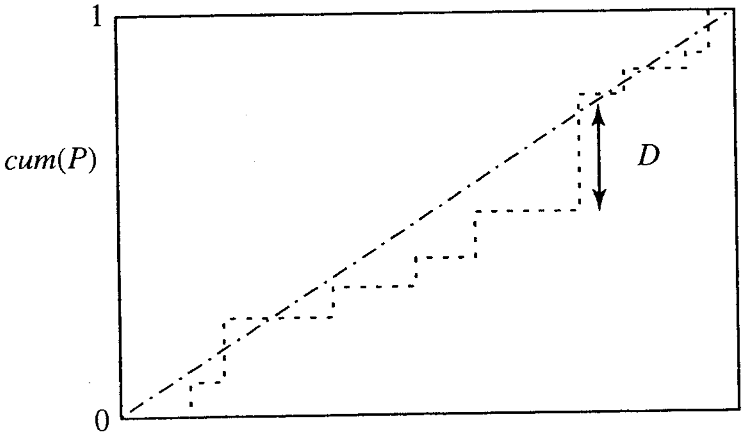
We take the values of the measured variable and arrange them in increasing order. Then we plot the cumulative distribution *cum*(x), divided by the number of measurements N. We also plot the cumulative distribution *cum*[$P(x)$] for a probability distribution of x with the considered hypothesis. By defining

D: absolute difference between the two plots,

$d = D \times N^{1/2}$

d provides the hypothesis test. For large N, we have

S_c:	d>1.63	d>1.36	d>1.22	d>1.07
Significance:	1%	5%	10%	20%



The broken line is for data and the other line corresponds to the cumulative distribution for the probability distribution with a hypothesis that variables are distributed evenly.

14) MONTE CARLO METHOD

1) Uniform distribution

The probability is constant in the interval $a \leq x \leq b$ and 0 outside of this region:

$$f(x) = \frac{1}{b-a} : a \leq x \leq b, \qquad f(x) = 0 : x < a,\ x > b$$

where f(x) is properly normalised, i.e.

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

The distribution function F(*x*) is given by

$$F(x) = \int_{-\infty}^{x} f(x)dx = \begin{array}{l} 1 :\ x > b \\ \frac{x-a}{b-a} :\ a \leq x \leq b \\ 0 :\ x < a \end{array}$$

and the expectation value and variance are given by

$$E(x) = \frac{a+b}{2}, \quad V(x) = \frac{(b+a)^2}{12}$$

Computers can generate "random numbers", a chain of values of randomly and evenly distributed between 0 and 1.

2) Generation of any distribution by transformation of the uniform distribution

Let us consider that *x* is described by a uniform distribution of the type

$$f(x) = 1 : 0 \leq x \leq 1, \qquad f(x) = 0 : x < 0,\ x > 1$$

and *y* is a random variable described by the probability distribution *g*(*y*). The variable transformation gives, *g*(*y*)d*y* = d*x*. By integration both side, we obtain

$$x = \int_{-\infty}^{y} g(z)dz \equiv G(y).$$

From this equation, we conclude the following:

i) draw *x*, ii)invert x=*G*(*y*), i.e. $y=G^{-1}(x)$.

The random variable *y* is distributed following the probability distribution *g*(*y*).

Even *G*(*y*) is not known, random numbers distributed as g(y) in the interval $a \leq y \leq b$ can be obtained in the following way:

i) Find then maximum, $g_{\max}$, of *g*(*y*) in the interval $a \leq y \leq b$.

ii) Draw a random number, y_r, from a uniform distribution of the rage $a \leq y \leq b$.

iii) Draw a random number, g_r, from a uniform distribution of the rage $0 \leq g_r \leq g_{\max}$.

iv) Keep g_r only if $g_r < g(y_r)$.

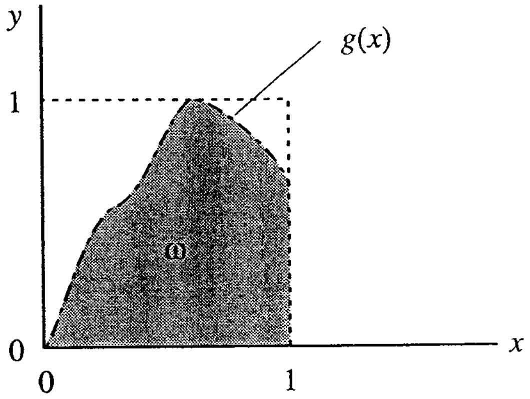
Repeating this, the distribution of y_r kept follow *g*(*y*).

3) Monte Carlo integration method

An integration of

$$I = \int_0^1 g(x)dx$$

where the integrand $g(x)$ varies in the region $0 \le g(x) \le 1$. The integral I is identical to the surface of the area ω shown in the figure.



Now we generate N pairs of random numbers (x, y) where x and y are evenly distributed between 0 and 1. If we find n pairs in the area ω, i.e. $[x, y < g(x)]$, the integral I is given by

$$I = \lim_{N\to\infty} \frac{n}{N}$$

The integral can be generalised easily to any ra‸nge with many variables.

Bibliography

Introducory

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

W. T. Eadie et al., *Statistical Methods in Experimental Physics*, North Holand Publishing Co.

S. L. Meyer, *Data Analysis for Scientists and Engineers*, Wiley

Advanced

S. Brandt, *Statistical and Computational Methods in Data Analysis*, North Holland Publishing Co.