6. Errors

- same precision of measured value and error
  - e.g. 1.5 ± 0.5 cm.
  - inconsistent: x=5±0.2 or x=1.07±0.1
- ideally: measurement should be free of bias (Verfälschung), i.e., the true value should be the mean of the measured values in the limit of infinite measurements (for a more precise definition, see Chap. 7)
- for measurements with results which follow a known distribution, the error is chosen as the corresponding standard deviation (assuming the measurement value to be the expectation value)
- Examples
  - Poisson-distributed events: We measure 150 photons. The result is $150\pm\sqrt{150}=150\pm12$
  - uniformly distributed processes: We measure the time with a digital clock which displays seconds. The error is $1/\sqrt{12}$ s = 0.3s.
  - binomial distribution: We detect N=45 from N_0=60 particles which pass a detector. The detection rate is N/N_0=0.75.
    The error of the rate is $\delta N/N_0=\sqrt{[N_0*p*(1-p)]}/N_0=\sqrt{[p*(1-p)/N_0]}=\sqrt{[0.75*0.25/60]}=0.06$

- Empirical derivation from a series of measurements
  - repeat measurement and derive error from fluctuations
  - for details, see Chap. 7 (parameter estimation)
  - justified if purely statistical errors, i.e., the results are independent of each other and there are no (correlated) systematic errors (otherwise, see below)
  - recipe:

$$x_m = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{remember CLT: } E(\bar{x}) = \mu_x, \;\; \mathrm{Var}(\bar{x}) = \frac{\sigma_x^2}{N}$$

with μ_x and σ_x the "true" parameters of the underlying distribution

$$\delta_x^2 := s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad \text{"empirical variance", denoted by } s_x^2 \text{ to discriminate from } \sigma_x^2$$

$$\delta_{\bar{x}}^2 := s_{\bar{x}}^2 = Var(\bar{x}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} Var(x_i) = \frac{1}{N} Var(x_i) \rightarrow$$

$$\rightarrow \frac{s_x^2}{N} = \frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad \text{"(empirical) variance of the mean"}$$

for the result, quote $x_m \pm \delta_{\bar{x}}$, with individual errors δ_x

The (N-1)-term in the denominator of s_x^2 will be derived in Chap. 7.
It is caused by the fact that we need to use the estimate for the mean, x_m,
instead of the true mean μ (which is unknown).
Note that for $N = 1$ measurements the variance remains unspecified, which makes sense.
If there would be an N in the denominator, we would find $s_x^2 = 0$ for $N = 1$ (since $x_m = x_i$)!

Averaging weighted measurements

- Important to combine the results of the individual measurements in such a way as to find the highest precision
- in the following, we assume independent measurements

Example: two measurements with results x_1, x_2 and errors δ_1, δ_2. The variance of the weighted sum (with $w_1 + w_2 = 1$)

$\bar{x} = w_1 x_1 + w_2 x_2$ is given by

$Var(\bar{x}) =: \delta^2 = w_1^2 \delta_1^2 + w_2^2 \delta_2^2$,

and we choose the weights in such a way as to minimize the variance. This is obtained with (prove by yourself)

$$w_i = \frac{1/\delta_i^2}{1/\delta_1^2 + 1/\delta_2^2} \quad \text{such that} \quad \frac{1}{\delta^2} = \frac{1}{\delta_1^2} + \frac{1}{\delta_2^2}.$$

The measurment with the smaller error obtains a higher weight!

For N measurements, we obtain the general expressions

$$\bar{x} = \frac{\sum_i x_i / \delta_i^2}{\sum_i 1/\delta_i^2} \quad \text{and} \quad Var(\bar{x}) = \delta^2 = \frac{1}{\sum_i 1/\delta_i^2}$$

If all errors are identical, we obtain the previous expression for $\bar{x}$ and δ, where the latter is a factor of $1/\sqrt{N}$ smaller than the individual errors.

Error propagation

In the following, we identify the measurement error with the positive root of the mean square deviation of the individual measurements, i.e., with the standard deviation. Suppose that we measure the quantity x (in the remainder of this chapter, we don't make a distinction between x and x) which follows a certain distribution with variance $Var(x)$ and "error" σ_x (because of the CLT, this distribution will be often a normal one). We like to find now the corresponding error of $f(x)$, where f is a function.

1. One variable, linear function

Let $f = ax + b$ with constants a and b. Then,

$Var(f) = a^2 Var(x)$, i.e.,

$\sigma_f = |a| \ \sigma_x$

2. One variable, arbitrary function. Taylor expansion around the measured (mean) value x_m with individual variance $Var(x) = E\left[(x - x_m)^2\right] = E(\Delta x^2) = \sigma_x^2$

$$f(x) \approx f(x_m) + (x - x_m)\left.\frac{df}{dx}\right|_{x_m}$$

$$Var(f) = \left(\frac{df}{dx}\right)^2 Var(x) \quad \text{and} \quad \sigma_f = \left|\frac{df}{dx}\right| \sigma_x$$

Note that this approximation is valid for "small" errors, where "small" means that the differential should not change much over a few σ.

3. One function of two variables

a) linear function

$$f = ax + by + c \quad \Rightarrow \quad Var(f) = a^2 Var(x) + b^2 Var(y) + 2ab\,\mathrm{cov}(x,y)$$

b) general function. Taylor expansion

$$f(x,y) \approx f(x_m, y_m) + (x - x_m)\frac{\partial f}{\partial x} + (y - y_m)\frac{\partial f}{\partial y}$$

$f_m = E\left[f(x,y)\right] = f(x_m, y_m)$ to first order, since $E(\Delta x) = E(\Delta y) = 0$

But note: to 2nd order, $E\left[f(x,y)\right] = f(x_m, y_m) + \frac{\partial^2 f}{\partial x \partial y}\mathrm{cov}(x,y) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}Var(x) + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}Var(y)$,

i.e., also the expectation value might be influenced. This is the generalization of $E(xy) = E(x)E(y) + \mathrm{cov}(x,y)$

$$Var(f) = E\left[\left(f(x,y) - f_m\right)^2\right] = E\left[\Delta x^2\left(\frac{\partial f}{\partial x}\right)^2 + \Delta y^2\left(\frac{\partial f}{\partial y}\right)^2 + 2\Delta x \Delta y\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\right] =$$

$$= \left(\frac{\partial f}{\partial x}\right)^2 E(\Delta x^2) + \left(\frac{\partial f}{\partial y}\right)^2 E(\Delta y^2) + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)E(\Delta x \Delta y) =$$

$$= \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\mathrm{cov}(x,y) =$$

$$= \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\rho(x,y)\sigma_x\sigma_y$$

This reduces to the "standard" law of error-propagation (without the covariance term)

if and only if the covariance term vanishes, i.e., if the x and y are uncorrelated

Examples for error propagation

$$f = ax^n \quad Var(f) = a^2(nx^{n-1})^2 Var(x) \quad \sigma_f = \left|anx^{n-1}\right| \sigma_x \quad \left|\frac{\sigma_f}{f}\right| = \left|n\frac{\sigma_x}{x}\right|$$

$$f = ae^{bx} \quad Var(f) = a^2(be^{bx})^2 Var(x) \quad \sigma_f = \left|abe^{bx}\right| \sigma_x \quad \left|\frac{\sigma_f}{f}\right| = |b|\sigma_x$$

$$f = ax \pm by \quad Var(f) = a^2 Var(x) + b^2 Var(y)\left[\pm 2ab\,\mathrm{cov}(x,y)\right]$$

$$\sigma_f = \sqrt{a^2\sigma_x^{\,2} + b^2\sigma_y^2\left[\pm 2ab\,\mathrm{cov}(x,y)\right]}$$

$$f = axy \quad Var(f) = a^2\left(y^2 Var(x) + x^2 Var(y)\left[+2xy\,\mathrm{cov}(x,y)\right]\right)$$

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2\left[+2\rho\left(\frac{\sigma_x}{x}\right)\left(\frac{\sigma_y}{y}\right)\right]$$

$\left[\text{note covariance: } E(f) = a\left[E(x)\cdot E(y) + \mathrm{cov}(x,y)\right]\right]$

Similar result for $f = ax/y$ [but with $-\mathrm{cov}$ term]:

$$f = a\frac{x}{y} \qquad \left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2\left[-2\rho\left(\frac{\sigma_x}{x}\right)\left(\frac{\sigma_y}{y}\right)\right]$$

4. One function of several variables

With covariance matrix ("error matrix") **C**,

$$\mathbf{C} = E\left[(\mathbf{x}-\mathbf{x_m})(\mathbf{x}-\mathbf{x_m})^{\mathbf{T}}\right] = \begin{pmatrix} E(\Delta x_1^2) & E(\Delta x_1 \Delta x_2) & \dots & E(\Delta x_1 \Delta x_n) \\ E(\Delta x_2 \Delta x_1) & E(\Delta x_2^2) & \dots & E(\Delta x_2 \Delta x_n) \\ \dots & \dots & \dots & \\ E(\Delta x_n \Delta x_1) & E(\Delta x_n \Delta x_2) & \dots & E(\Delta x_n^2) \end{pmatrix}$$

$$C_{ij} = \rho_{ij}\sigma_i\sigma_j$$

we find from generalizing the previous results for two variables

$$Var(f) = \sigma_f^2 = \sum_{i,j=1}^{n} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j} C_{ij}$$ or in vector notation

$$Var(f) = \sigma_f^2 = (\nabla f)^{\mathbf{T}} \cdot \mathbf{C} \cdot \nabla f$$

If the variables are uncorrelated, only the diagonal terms "survive", and we obtain (again) the "standard" law of error propagation,

$$Var(f) = \sigma_f^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2$$

5. Several functions of several variables

Last generalization. Suppose there are now m functions $f_1, f_2, ..., f_m$ of n variables $x_1, x_2, ..., x_n$.

In this case, there will be always a correlation of the f_k, since they share the same variables x_i.

Taylor expansion of the individual functions gives

$$Var(f_k) = \sigma_{f_k}^2 = \sum_{i,j=1}^{n} \frac{\partial f_k}{\partial x_i} \frac{\partial f_k}{\partial x_j} C_{ij} \qquad \text{or in vector notation}$$

$$Var(f_k) = \sigma_{f_k}^2 = (\nabla f_k)^{\mathbf{T}} \cdot \mathbf{C} \cdot \nabla f_k \qquad \text{(as before),}$$

whereas for the covariances between the functions we obtain

$$\operatorname{cov}(f_k, f_l) = E(f_k f_l) - E(f_k)E(f_l) \approx$$

$$\approx E\left[(x_1 - x_{1m})(x_1 - x_{1m})\right]\left(\frac{\partial f_k}{\partial x_1}\right)\left(\frac{\partial f_l}{\partial x_1}\right) + ... + E\left[(x_1 - x_{1m})(x_2 - x_{2m})\right]\left(\frac{\partial f_k}{\partial x_1}\right)\left(\frac{\partial f_l}{\partial x_2}\right) + ...$$

$$\operatorname{cov}(f_k, f_l) = \sum_{i,j=1}^{n} \left(\frac{\partial f_k}{\partial x_i}\right)\left(\frac{\partial f_l}{\partial x_j}\right) \operatorname{cov}(x_i, x_j),$$

which includes the above expression for the variances in case of $k = l$.

The most general law of error propagations

Note that even for uncorrelated variables, at least the term involving $i = j$,

$$\mathrm{cov}(f_k, f_l) = \sum_{i=1}^{n} \left(\frac{\partial f_k}{\partial x_i} \right) \left(\frac{\partial f_l}{\partial x_i} \right) \mathrm{cov}(x_i, x_i) = \sum_{i=1}^{n} \left(\frac{\partial f_k}{\partial x_i} \right) \left(\frac{\partial f_l}{\partial x_i} \right) Var(x_i)$$

will always be present, i.e., the functions will be always correlated!

Denoting the matrix of partial derivatives (which is nothing else than the transformation matrix, see Chap. 4) by

$$\mathbf{T} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}_{\mathbf{x}=\mathbf{x_m}} \in \Re^{mxn}, \quad \mathrm{T_{ki}} = \frac{\partial f_k}{\partial x_i},$$

we can write the most general law of error propagation as

$$\mathbf{C_f} = \mathbf{T} \mathbf{C_x} \mathbf{T}^{\mathbf{T}}$$

with $\mathbf{C_x} \in \Re^{nxn}$ and $\mathbf{C_f} \in \Re^{mxm}$. Both matrices are symmetric.

This expression contains everything one has to know about error propagation.

Example

In a carthesian coordinate system a point (x, y) is measured, where the measurements should be independent, and the error in y is three times larger than the one in x. E.g.,

$\mathbf{C}_x = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix}$.

Let's calculate the errors in polar coordinates, $r = \sqrt{x^2 + y^2}$ and $\varphi = \text{atan}\frac{y}{x}$.

The transformation matrix (matrix of partial derivatives) is

$$\mathbf{T} = \begin{pmatrix} \frac{x}{r} & \frac{y}{r} \\ \frac{-y}{r^2} & \frac{x}{r^2} \end{pmatrix}_{\mathbf{x}=\mathbf{x}_m}, \quad \text{and we consider the errors at } \mathbf{x}_m = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \text{ Then, } \mathbf{T} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix},$$

and

$$C_{r,\varphi} = \mathbf{T}\mathbf{C}_x\mathbf{T}^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 5 & \frac{4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} & \frac{5}{2} \end{pmatrix}$$

The errors in r and φ are the square root of the diagonal, $\sqrt{5}$ and $\sqrt{5/2}$, respectively, and the covariance between (r, φ) is $\text{cov}(r, \varphi) = 4/\sqrt{2}$.

Only if these terms are considered, the back-transformation gives the original results!

... and any function involving both (r, φ) needs this covariance term for the errors.

Systematic errors

- the effect of systematic errors does not decrease with the number of measurements, since all measurements share the same effect, and thus are not independent
- sometimes, statistical errors and systematic ones are stated separately, e.g.

  $x=5.0\pm1.4\pm2.3$

  where the 2nd term is the systematic error.
- since statistical errors and systematic ones are independent of each other, they add quadratically (see below)
- treatment: split errors in random and systematic ones, and calculate covariance matrix

Assume you measure two quantities x_1 and x_2, which have a common systematic error S and individual random errors σ_1, σ_2.

Example:

you measure the length and the width of an rectangle, where the measuring tape gives systematically too high values (the tape might be stretched).
In this case, both measurements are affected by an individual random error (reading) and a common systematic one.
When calculating the area of the rectangle, the covariance term (resulting from the systematic error) usually plays the dominating role.

Split up the individual measurements in two parts,

$x_1 = x_1^R + x_1^S, \qquad x_2 = x_2^R + x_2^S$

with corresponding random (R) and systematic (S) errors. Then, x_1^R and x_2^R are independent of each other and of x_1^S and x_2^S, whereas x_1^S and x_2^S are identical.

$$Var(x_1) = E(x_1^2) - E(x_1) \cdot E(x_1) = E\left[(x_1^R + x_1^S)^2\right] - E^2(x_1^R + x_1^S) =$$
$$= E\left[(x_1^R)^2\right] - E^2(x_1^R) + E\left[(x_1^S)^2\right] - E^2(x_1^S) = \sigma_1^2 + S^2,$$

since $E(x_1^R \cdot x_1^S) = E(x_1^R) \cdot E(x_1^S)$ because they are independent (Chap. 4).
The above equation proves that systematic and random errors add quadratically.

Similarily,

$Var(x_2) = \sigma_2^2 + S^2$ and

$$\begin{aligned}\mathrm{cov}(x_1, x_2) &= E(x_1 x_2) - E(x_1)E(x_2) = \\ &= E\left[(x_1^R + x_1^S)(x_2^R + x_2^S)\right] - E(x_1^R + x_1^S)E(x_2^R + x_2^S) = \\ &= E\left[x_1^S x_2^S\right] - E(x_1^S)E(x_2^S) = \\ &= \mathrm{cov}(x_1^S, x_2^S) = Var(x_1^S) = Var(x_2^S) = S^2\end{aligned}$$

since all products involving x^R cancel, because they are independent of anything else.

Thus, the complete covariance matrix reads

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 + S^2 & S^2 \\ S^2 & \sigma_2^2 + S^2 \end{pmatrix}$$

Often, the systematic error is proportional to the measurement,

$S = \varepsilon x$

(e.g., in case of the measurement tape, if the tape is stretched). In this case, the error matrix is

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 + \varepsilon^2 x_1^2 & \varepsilon^2 x_1 x_2 \\ \varepsilon^2 x_1 x_2 & \sigma_2^2 + \varepsilon^2 x_2^2 \end{pmatrix}$$

Examples

Example 1

By means of the well-known linear regression (Chap. 8), one can, e.g., fit a straight line $y = mx + b$ to a certain data-set. Fit-parameters are the offset b and the slope m, with errors σ_b and σ_m, respectively. As we will show, there is (almost) always a correlation between both quantities, which

a) is negative when the "sample" mean of the abscissae, $\overline{x} = \left(\sum_{i=1}^{N} x_i \right) / N$, is positive,

b) or positive, when the sample mean is negative.

c) for $\overline{x} = 0$, the correlation vanishes.

If one wants, e.g., to use the derived relation for extra- or interpolation, the accuracy of the extra-/interpolated value depends on this correlation.

Assume that we want to derive the y-value for the abscissa x_1, and that $\overline{x} > 0$.

$y_1 = mx_1 + b;$ then

$\sigma_{y_1}^2 = x_1^2 \sigma_m^2 + \sigma_b^2 + 2x_1 \operatorname{cov}(m, b)$

which is smaller (larger) for positive (negative) x_1 than if neglecting (forgetting!) the correlation.

Example 2

A current I is determined by measuring the voltage U, using a meter of resolution σ_U, across a resistance $R \pm \sigma_R$. The error on $I = U/R$ results as

$$\left(\frac{\sigma_I}{I}\right)^2 = \left(\frac{\sigma_U}{U}\right)^2 + \left(\frac{\sigma_R}{R}\right)^2, \text{ i.e., } \sigma_I^2 = \frac{\sigma_U^2 + I^2\sigma_R^2}{R^2}.$$

If two currents, I_1 and I_2, are measured using the same resistance, there will be a *systematic* error, and both currents become correlated. The covariance can be calculated from our general formula of error propagation,

$$\text{cov}(I_1, I_2) = \frac{\partial I_1}{\partial R}\frac{\partial I_2}{\partial R}\sigma_R^2 = \left(-\frac{U_1}{R^2}\right)\left(-\frac{U_2}{R^2}\right)\sigma_R^2 = \frac{I_1 I_2}{R^2}\sigma_R^2.$$

The errors on I_1 and I_2 are not influenced by this covariance, of course. However, if one calculates functions of I_1 *and* I_2, the corresponding error *is* affected. E.g., the variance of $(I_1 - I_2)$ is given by (again using our general formula)

$$Var(I_1 - I_2) = 1^2\sigma_{I_1}^2 + (-1)^2\sigma_{I_2}^2 - 2\,\text{cov}(I_1, I_2) = \frac{\sigma_U^2 + I_1^2\sigma_R^2 + \sigma_U^2 + I_2^2\sigma_R^2 - 2I_1 I_2\sigma_R^2}{R^2} =$$

$$= \frac{2\sigma_U^2 + (I_1 - I_2)^2\sigma_R^2}{R^2}$$

which can be significantly smaller than if forgetting the correlation

$$\left[(I_1 - I_2)^2\sigma_R^2 \text{ vs. } (I_1^2 + I_2^2)\sigma_R^2\right]$$

Example 3

Given are the sides of a rectangle, $a, b,$ with reading error δ_1 and scaling error δ_2 (stretched tape). Calculate the error on the area F. The error matrix is

$$\mathbf{C} = \begin{pmatrix} \delta_1^2 + \delta_2^2 a^2 & \delta_2^2 ab \\ \delta_2^2 ab & \delta_1^2 + \delta_2^2 b^2 \end{pmatrix}$$

$F = ab$. From our generalized law of error propagation, we find

$$(\delta F)^2 = b^2 (\delta a)^2 + a^2 (\delta b)^2 + 2ab \operatorname{cov}(a,b)$$

$$\left(\frac{\delta F}{F}\right)^2 = \left(\frac{\delta a}{a}\right)^2 + \left(\frac{\delta b}{b}\right)^2 + 2\frac{\operatorname{cov}(a,b)}{ab} = \left(\frac{\delta_1^2}{a^2} + \delta_2^2\right) + \left(\frac{\delta_1^2}{b^2} + \delta_2^2\right) + 2\frac{\delta_2^2 ab}{ab} = \delta_1^2\left(\frac{1}{a^2} + \frac{1}{b^2}\right) + 2\delta_2^2 + 2\delta_2^2$$

Due to the covariance, the relative error on the area becomes larger by $2\delta_2^2$ compared to the case of neglecting (forgetting) the correlation. Note that also the area itself is affected by the correlation,

$$E(F) = ab + \operatorname{cov}(ab) = F(1 + \delta_2^2)$$

7. Estimation

Random sampling -- a few definitions

Distribution functions give probabilities for $x < \mathrm{x} \leq x + dx$, which moreover depend on certain parameters λ which are usually unknown. To "measure" the pdf and λ, one has to approximate it by a *frequency distribution* obtained experimentally.

The number of experiments performed, called a sample, is necessarily finite. Each sample is obtained from a set of elements which is usually of infinite size (composed of all conceivable events), which is called the (parent) population. If a sample of N elements is drawn, the sample has size N. The sample can be described by a N – dimensional random variable

$\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_N)$

The sample random variable **x** follows a pdf

$g(\mathbf{x}) = g(x_1, x_2, ..., x_N)$,

and has to fulfill two conditions in order to describe the process of *random sampling*.

a) the x_i have to be independent, i.e., $\mathrm{g}(\mathbf{x}) = g_1(x_1) g_2(x_2) ... g_N(x_N)$,

b) the individual distributions have to be equal and identical to the pdf of the parent population $f(x)$, i.e., $g_i(x_i) = f(x) \quad \forall i$

A function of a sample **x**, which itself is a r.v., is called a statistic. A well known example is the *sample mean* defined as the arithmetic mean of the x_i, $\bar{\mathrm{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{x}_i$.

Estimators

- A typical problem of data analysis is the following:
  The general pdf of the parent population is known. The numerical value of one or several parameters shall be obtained from a sample. Thus, we are dealing with the estimation of parameters. Since the estimated value is obtained by means of sampling, it is a statistic, called an *estimator.*

- an *estimator* is a statistic,

  $S=S(x_1, x_2,\ldots,x_N)$

  i.e., a procedure (function) applied to the data sample which gives a (numerical) property of the parent population or a property or parameter of the parent distribution function.

- for a given sample, there might be different estimators.

- the quality of an estimator can be described in three terms,
  consistency, bias and efficiency
  - a 'good' estimator has to be consistent, unbiased and efficient.
  - to compare two estimators which are consistent and unbiased, the better one has to be more efficient .
  - a 'bad' estimator is inconsistent, biased and inefficient.
  - estimators should be at least consistent. Some estimators are inevitably biased.

Bias, consistency and efficiency

An estimator is unbiased if for any size of the sample its expectation value is equal to the parameter to be estimated,

$E(\mathrm{S}(\mathrm{x}_1,\mathrm{x}_2,...,\mathrm{x}_N)) = \lambda$ for any N

If a bias is found, it is often easy to correct for. If , e.g., $E(\mathrm{S}(\mathrm{x}_1,\mathrm{x}_2,...,\mathrm{x}_N)) = \lambda + b$,

then $\mathrm{S}(\mathrm{x}_1,\mathrm{x}_2,...,\mathrm{x}_N) - b$ is an unbiased estimator.

An estimator is consistent if the result becomes increasingly accurate for increasing sample size,

$\lim_{N\to\infty} \sigma(\mathrm{S}) = 0$ and $\lim_{N\to\infty} \mathrm{S} = \lambda.$

If an estimator is consistent, its bias (if any) vanishes for $N \to \infty$: consistent = asympt. unbiased

To compare the relative efficiency of two estimators, one can use the quotient

$$\eta = \frac{\sigma^2(\mathrm{S}_1)}{\sigma^2(\mathrm{S}_2)}.$$

Often, the efficiency can be quantified in terms of a lower limit, the so-called minimum variance bound (MVB, see "maximum likelihood"). If $Var(\mathrm{S}) = MVB,$ the estimator is called efficient.

Example for bias

The Malmquist bias is a selection effect in observational astronomy. Specifically, if a sample of objects (galaxies, quasars, stars, etc.) is flux (“magnitude”)-limited, then the observer will see an increase in average luminosity with distance. This is, of course, because the less luminous sources at large distances will not be detected. The solution is then to use a sample that is not magnitude limited (for example, one that is volume limited.)

Basic estimators: sample mean and variance

As already shown, the expectation value of the sample (arithmetic) mean of a r.v. x is its expectation value,

$$E(\bar{x}) = E(\frac{1}{N}\sum_{i=1}^{N} x_i) = \frac{1}{N}(E(x_1) + E(x_2) + ... + E(x_N)) = \mu_x.$$

Since this is true for any sample size, the sample mean is an unbiased estimator for the expectation value of x in the parent population, the *population mean.*

The variance of the sample mean is

$$\sigma^2(\bar{x}) = E\left\{(\bar{x} - E(\bar{x}))^2\right\} = E\left\{\left(\left(\frac{\sum x_i}{N}\right) - \mu_x\right)^2\right\} = E\left\{\frac{1}{N^2}\left(\left(\sum x_i\right) - N\mu_x\right)^2\right\} = \frac{1}{N^2}E\left\{\left(\sum(x_i - \mu_x)\right)^2\right\} =$$

$$= \frac{1}{N^2}E\left\{(x_1 - \mu_x)^2 + ... + (x_N - \mu_x)^2\right\} = \frac{\sigma^2(x)}{N},$$ since all mixed terms of the type

$E\left\{(x_i - \mu_x)(x_j - \mu_x)\right\}$, i.e., the covariances, vanish (independent r.v.).

Thus, the sample mean is a consistent estimator for μ_x, $\lim_{N\to\infty}\sigma(\bar{x}) = \lim_{N\to\infty}\frac{\sigma(x)}{\sqrt{N}} = 0.$

Now, let's estimate the variance of the population (see also Chap. 6). At first we assume that μ_x is known. An obvious estimator for the variance is

$S(Var) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2 =: s_1^2$ which is consistent (prove yourself) and unbiased, since

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

Usually, however, the "true" population mean is not kown. An obvious remedy is to replace it by the sample mean. Let's call the corresponding estimator s',

$$s'^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - 2\bar{x}\frac{\sum_{i=1}^{N}x_i}{N} + \frac{N\bar{x}^2}{N} = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

with expectation value

$$E(s'^2) = E\left\{\frac{1}{N}\sum_{i=1}^{N}(x_i^2 - \bar{x}^2)\right\} = \frac{1}{N}NE(x^2) - E(\bar{x}^2) = E(x^2) - E(\bar{x}^2) =$$

$$= E(x^2) - \left(E(x)\right)^2 + \left(E(x)\right)^2 - E(\bar{x}^2).$$

As we have shown above, $E(x) = E(\bar{x})$, and thus

$$E(s'^2) = E(x^2) - \left(E(x)\right)^2 - \left[E(\bar{x}^2) - \left(E(\bar{x})\right)^2\right] = Var(x) - Var(\bar{x}) = \left(1 - \frac{1}{N}\right)Var(x).$$

Thus, s' is biased, where the bias vanishes for $N \to \infty$! The reason for this bias is that we have not used the true but the sample mean. Since the sample mean, by construction, lies somewhat closer to the data than the true mean, the corresponding variance is smaller.

The bias can be corrected for by applying "Bessel's correction", namely

$$E(s'^2) = \left(1 - \frac{1}{N}\right) Var(x) = \frac{N-1}{N} Var(x) \quad \Rightarrow \quad Var(x) = \frac{N}{N-1} E(s'^2) = E\left(\frac{N}{N-1} s'^2\right)$$

by multiplying with $N/(N-1)$.

Thus, a bias free estimator for the variance of a sample, s, is given by

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

The denominator can be also understood as follows: Some information about the sample has been used for calculating the sample mean, which is lost when calculating the sample variance. The effective number of the sample elements is thus reduced.

For large N, the variance of these estimators can be calculated (with some effort) via

$$Var(s'^2) \approx \frac{1}{N}\left[E\left\{(x - \mu_x)^4\right\} - \left(E\left\{(x - \mu_x)^2\right\}\right)^2 \right], \quad Var(s^2) = \left[\frac{N}{N-1}\right]^2 Var(s'^2)$$

which vanishes for $N \to \infty$. Thus, both estimators are consistent. For a Gaussian distribution (conventional measurement errors), this reduces to (cf. Chap. 5)

$$Var(s'^2, \text{ large } N) \approx \frac{3\sigma^4 - \sigma^4}{N} = \frac{2\sigma^4}{N} \quad [\text{for arbitrary } N,\ Var(s'^2) = \frac{2\sigma^4 (N-1)}{N^2}]$$

$$Var(s^2) = \frac{2\sigma^4}{N-1} \text{ (arbitrary } N) \qquad \text{(Eq. 7.1)}$$

Estimating σ

So far, we have calculated various estimators for the variance, s_1^2, s'^2 and s^2.
The obvious way to estimate the standard deviation is to take the square root of these estimates, i.e.

$$\mathrm{S}(\sigma) = \sqrt{\mathrm{S}(Var)} = \begin{cases} \mathrm{s}_1 \\ \mathrm{s}' \\ \mathrm{s} \end{cases}$$

Though the law of large numbers guarantees consistency, the square root of an unbiased quantity (s_1^2, s^2) not necessarily needs to be unbiased itself. Fortunately, in all calculations of error propagation or significance the standard deviation appears as σ^2.

The variance of $\mathrm{S}(\sigma)$ can be calculated from the law of error propagation,

$$Var(\mathrm{S}(\sigma^2)) = \left(\frac{d\sigma^2}{d\sigma}\right)^2 Var(\mathrm{S}(\sigma)) = 4\sigma^2 Var(\mathrm{S}(\sigma)).$$

Thus, for large N and a Gaussian distribution, we obtain

$$Var(\mathrm{s}') \approx \frac{\sigma^2}{2N}, \qquad Var(\mathrm{s}) = \frac{\sigma^2}{2(N-1)} \qquad \text{or} \qquad \sigma_{\mathrm{s}'} \approx \frac{\sigma}{\sqrt{2N}}, \qquad \sigma_{\mathrm{s}} = \frac{\sigma}{\sqrt{2(N-1)}} \qquad \text{(Eq. 7.2)}$$

If the population variance is not known or cannot be guessed, then σ needs to be replaced by the corresponding estimators from above.

Estimating the correlation coefficient

A simple estimator for the correlation coefficient within a sample can be constructed via

$$S(\rho) = r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad \text{(Pearson's r)}$$

(independent of N vs. $N-1$), and with corresponding error *for large* N (>500)

$$\sigma_\rho \approx \frac{1-\rho^2}{\sqrt{N-1}}.$$

For *moderate* N, it is better to transform to a variable z (Fisher's z-transformation),

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

not to be confused with Fischer-Z

which has a standard deviation $\sigma_z = 1/\sqrt{(N-3)}$.

Example: 13 physics students were given an essay to write. The correlation between the essay mark and their end-of-the-semester average physics mark was found to be S(ρ) = -0.16. If this correlation were really negative, this would imply that literate students are bad at physics and vice versa. Is there any support for this?

Transforming from $r = S(\rho)$ to z, we obtain

$$z = 0.5\ln\frac{1-0.16}{1+0.16} = -0.1613.$$

The error $\sigma_z = 1/\sqrt{10} = 0.316$. The deviation from zero correlation is only half a standard deviation, so not significant.

Summary on basic estimators

- all estimators discussed so far
  - are consistent and
  - do not depend on the distribution of the parent population

    (except for expressions Eq. 7.1/7.2 that assume a normal distribution)
- furthermore, the basic estimators for the mean and the variance (“sample mean, sample variance”) are unbiased estimators of the mean and the variance of the parent population

Parameter estimation: an example

Estimate the parameters of a correlated binormal distribution from random sampling (and compare with the parameters from the parent population)

Create correlated sample from uncorrelated random numbers:

calculate 2N independent, normally distributed random numbers (either from intrinsic generator if present or from uniformly distributed numbers and Box-Muller algorithm);

the generated numbers are reduced r.v.: $u_i = \frac{x_i - \mu}{\sigma} \quad \Rightarrow \quad x_i = \sigma u_i + \mu$:

scale with individual σ_1, σ_2 (for each of the N r.v. pairs in direction 1,2);

rotate coordinate system to obtain correlated x,y pairs; add means μ_x, μ_y

Here, we use the example from Chap. 5 with $\boldsymbol{\mu} = (2,2)^T, \sigma_1 = 0.6252$ and $\sigma_2 = 1.6152$:

For a rotation angle of $\theta = -31.6^\circ$, we should obtain

ρ=0.7, $\sigma_x = 1.0$, $\sigma_y = 1.4142$ and $z = 0.8673$

u=randomn(seed, N)*σ_1

v=randomn(seed, N)*σ_2 2 times N independent, normally distr. r.v. with $\sigma_{1,2}$

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u_i \\ v_i \end{pmatrix}$$

note that the rotation matrix has been inverted (= transposed because of its orthogonality)

$x_i = x_i + \mu_x \qquad y_i = y_i + \mu_y$ add mean (shift center)

Calculate estimates and corresponding errors

(by using the estimates instead of the actual, but unknown quantities)

$$S(\mu_x) = \bar{x} = \frac{1}{N}\sum x_i \qquad S(\mu_y) = \bar{y} = \frac{1}{N}\sum y_i$$

$$S(\sigma_x) = s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}} \qquad S(\sigma_y) = s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{N-1}}$$

$$\sigma_{\mu_x} = \frac{s_x}{\sqrt{N}} \qquad \sigma_{\mu_y} = \frac{s_y}{\sqrt{N}}$$

$$\sigma_{s_x} = \frac{s_x}{\sqrt{2(N-1)}} \qquad \sigma_{s_y} = \frac{s_y}{\sqrt{2(N-1)}}$$

$$S(\rho) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y} \qquad \sigma_\rho = \frac{1-(S(\rho))^2}{\sqrt{N-1}}$$

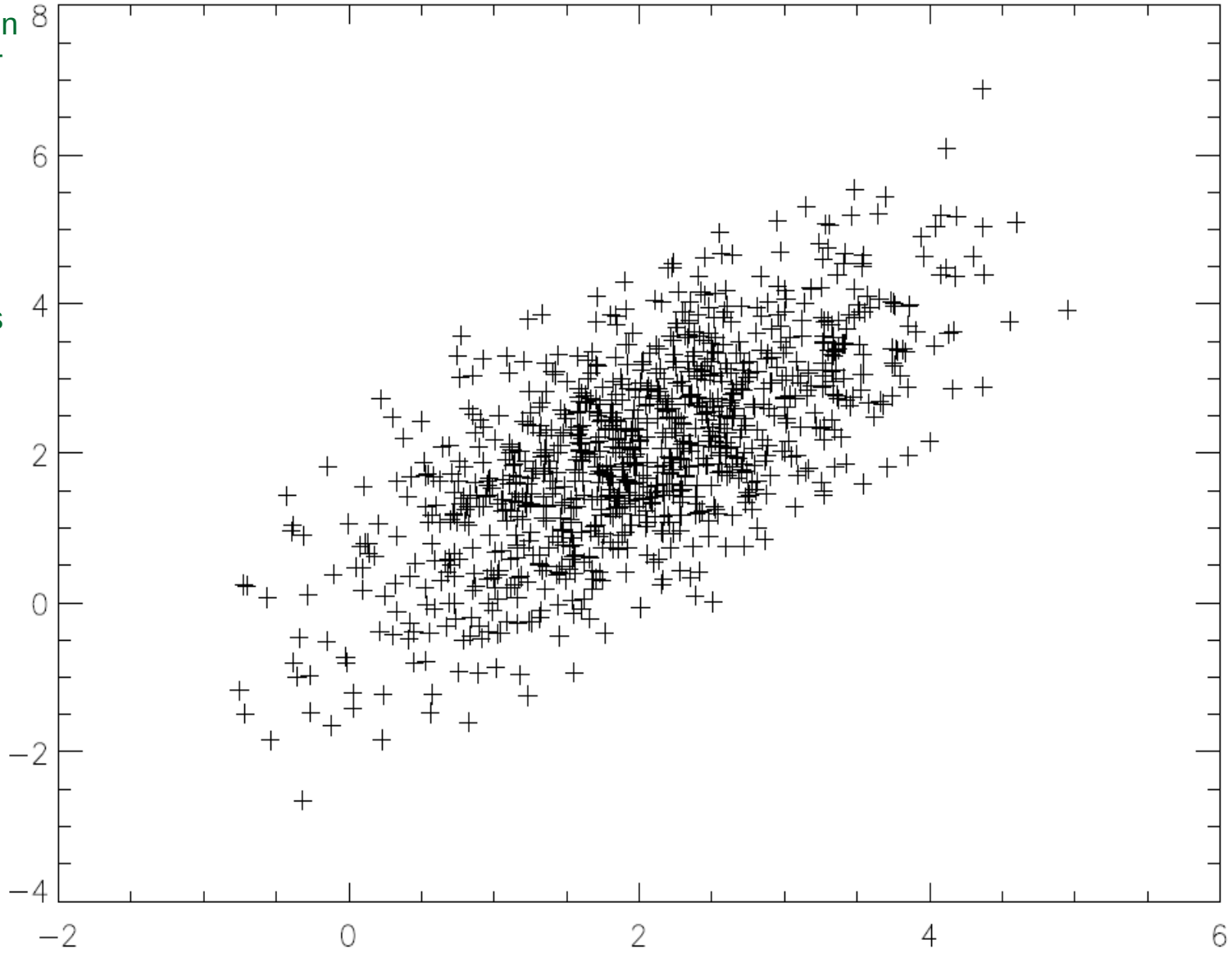
$$z = \frac{1}{2}\ln\frac{1+S(\rho)}{1-S(\rho)} \qquad \sigma_z = 1/\sqrt{(N-3)}$$

all sums extend from $i = 1, N$

NOTE again: All errors scale with $\frac{1}{\sqrt{N-k}}$, k $\in$ [0,3], i.e., **all estimators are consistent**

binormal distribution with correlation, for parameters as described before (N=1000 x,y-pairs)

Compare with the covariance ellipses from Chap. 5



Results for the sample estimates for different N

```
N=10
mu_x:       1.74463   +/- 0.301419
mu_y:       1.69644   +/- 0.340753
sig_x:      0.95317   +/- 0.224664
sig_y:      1.07756   +/- 0.253982
rho:        0.649858  +/- 0.192562
z:          0.775052  +/- 0.377964

N=100
mu_x:       1.81596   +/- 0.100339
mu_y:       1.80564   +/- 0.145755
sig_x:      1.00339   +/- 0.0713076
sig_y:      1.45755   +/- 0.103584
rho:        0.631415  +/- 0.0604344
z:          0.743767  +/- 0.101535

N=1000
mu_x:       2.00010   +/- 0.0306488
mu_y:       2.05395   +/- 0.0419521
sig_x:      0.969200  +/- 0.0216828
sig_y:      1.32664   +/- 0.0296795
rho:        0.682793  +/- 0.0168885
z:          0.834329  +/- 0.0316703
```

```
N=10000:
mu_x:       2.00184   +/- 0.00989791
mu_y:       2.00837   +/- 0.0140576
sig_x:      0.989791  +/- 0.00699923
sig_y:      1.40576   +/- 0.00994069
rho:        0.700583  +/- 0.00509210
z:          0.868444  +/- 0.0100015
```

to be compared with the population parameters

```
mu_x:       2.0
mu_y:       2.0
sig_x:      1.0
sig_y:      1.4142
rho:        0.7
z:          0.8673
```

note that all errors decrease with ≈ $N^{-0.5}$. Even for N=10, there is a significant indication that the (x,y) data are correlated!

Stratified sampling ('beating' 1/√N)

- suppose you want to estimate a certain quantity of a parent population based on a smaller sample, e.g., the average weight of students at your university.
- the most simple method is to make *N* measurements from a random sample, and to quote your results as

$$\bar{x} \pm s_x / \sqrt{N}\text{, with sample mean } \bar{x} \text{ and sample standard dev. } s_x$$

- but, you can do better, exploiting the fact that male and female students have different average weights, and *if* you know the relative proportions of male and female students at your university (consult the corresponding records!)
- If you perform the simple estimate from above, the ratio of male to female students in your sample will scatter about the actual ratio (e.g., if you have bad luck, your sample contains much more male than female students), and this scatter adds to the scatter in the average weight.
- This can be avoided by measuring the average weights in a male and female subsample, and adding up the results accounting for the specified ratio. In this way, the error on the total average can be significantly reduced!
- This method is called *stratified* (or *partitioned*) sampling *('geschichtetes Stichprobenverfahren')*

Let's consider the more general case that the parent population T can be divided into k (exhaustive!) subpopulations $T_1 ... T_k$.
The quantity x shall be described by correponding pdf's $f_1(x) ... f_k(x)$, with distribution function

$$F_i(x) = \int_{-\infty}^{x} f_i(x)dx = P(\mathrm{x} < x \,|\, \mathrm{x} \in T_i)$$

that is a conditional probability $\left(\mathrm{x}\right.$ has to be part of a certain subpopulation$\left.\right)$.

We now use the rule of total probability (Chap. 2) to obtain the distribution function for the total population T,

$$F(x) = P(\mathrm{x} < x \,|\, \mathrm{x} \in T) = \sum_{i=1}^{k} P(\mathrm{x} < x \,|\, \mathrm{x} \in T_i) P(\mathrm{x} \in T_i)$$

Denoting $P(\mathrm{x} \in T_i) =: p_i$, we find

$$F(x) = \sum_{i=1}^{k} p_i F_i(x) \text{ , and likewise } f(x) = \sum_{i=1}^{k} p_i f_i(x)$$

Thus, the population mean can be expressed by

$$\hat{x} = E(\mathrm{x}) = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} \sum_{i=1}^{k} xp_i f_i(x)\, dx = \sum_{i=1}^{k} p_i \int_{-\infty}^{\infty} xf_i(x)dx = \sum_{i=1}^{k} p_i \hat{x}_i \qquad \text{(Eq.7.3)}$$

The mean of the total population is the mean of the subpopulations, weighted by their probabilities of occurring.

The variance of the population can be calculated from its definition

$$Var(\mathrm{x}) = \sigma^2(\mathrm{x}) = \int_{-\infty}^{\infty} (x - \hat{x})^2 \, f(x) dx = \int_{-\infty}^{\infty} (x - \hat{x})^2 \sum_{i=1}^{k} p_i \, f_i(x) \, dx =$$

$$= \sum_{i=1}^{k} p_i \int_{-\infty}^{\infty} \left[(x - \hat{x}_i) + (\hat{x}_i - \hat{x}) \right]^2 \, f_i(x) dx$$

Since the x_i are independent, all mixed terms (covariances) vanish, and we obtain

$$\sigma^2(\mathrm{x}) = \sum_{i=1}^{k} p_i \left[\int_{-\infty}^{\infty} (x - \hat{x}_i)^2 \, f_i(x) dx + (\hat{x}_i - \hat{x})^2 \int_{-\infty}^{\infty} f_i(x) dx \right],$$

$$\sigma^2(\mathrm{x}) = \sum_{i=1}^{k} p_i \left[\sigma_i^2 + (\hat{x}_i - \hat{x})^2 \right] \qquad \text{(Eq. 7.4)}$$

The variance of the total population is the weighted variance of the subpopulations, plus the weighted variance of the subpopulation mean about the population mean

(The latter term corresponds to the 'additional' scatter mentioned earlier)

For $k = 2$ (corresponding to the previous example), we can express this alternatively as

$$\sigma^2(\mathrm{x}) = \sum_{i=1}^{2} p_i \left[\sigma_i^2 + (\hat{x}_i - \hat{x})^2 \right] = p_1\sigma_1^2 + p_2\sigma_2^2 + p_1 p_2 (\hat{x}_1 - \hat{x}_2)^2$$

accounting for $p_2 = (1 - p_1)$ in this case (prove yourself).

Thus, if we draw a random sample from the total population of size N, the variance of the corresponding sample mean, $\bar{x}$, will be

$$Var(\bar{x}) = \sigma^2(\bar{x}) = \frac{\sigma^2(x)}{N} = \frac{1}{N} \sum_{i=1}^{k} p_i \left[\sigma_i^2 + (\hat{x}_i - \hat{x})^2 \right] \qquad \text{(Eq. 7.5)}$$

Alternatively, we can perform a *stratified sampling*. In this case, we estimate the total population mean from the estimates of the subpopulation means,

$$S(\hat{x}) = \overline{x} = \sum_{i=1}^{k} p_i \overline{x}_i \qquad \text{(Eq. 7.6)}$$

(Remember that the p_i's need to be known from external resources or from preliminary sampling). This is an unbiased estimator, since (cf. Eq. 7.3)

$$E(\overline{x}) = \sum_{i=1}^{k} p_i E(\overline{x}_i) = \sum_{i=1}^{k} p_i \hat{x}_i = \hat{x}$$

So far, the subsample sizes n_i, with $\sum_{i=1}^{k} n_i = N$, did not play any role (though they should be large enough to ensure low errors on $\overline{x}_i$). Let's assume now that the n_i's are arbitrary, and that we want to calculate $\hat{x}$ from the arithmetic mean of the total *partitioned* sample, without using the p_i's. In this case,

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{N} \sum_{i=1}^{k} n_i \overline{x}_i ,$$

with expectation value $E(\overline{x}) = \sum_{i=1}^{k} \frac{n_i}{N} \hat{x}_i$. Comparison with Eq. (7.6) shows that only for $\frac{n_i}{N} = p_i$ the correct result is obtained, whereas otherwise the arithmetic mean cannot be used as an estimator for $\hat{x}$.

The variance of the estimator for the population mean (Eq. 7.6) can be calculated from the usual calculation rules,

$$S(\hat{x}) = \bar{x} = \sum_{i=1}^{k} p_i \bar{x}_i \quad \Rightarrow \quad Var(S(\hat{x})) = Var(\bar{x}) = \sigma^2(\bar{x}) = \sum_{i=1}^{k} p_i^2 Var(\bar{x}_i) = \sum_{i=1}^{k} p_i^2 \frac{\sigma_i^2}{n_i}.$$

i) If we would sample according to the ratios, $n_i = Np_i$, this would result in

$$\sigma^2(\bar{x}) = \frac{1}{N} \sum_{i=1}^{k} p_i \sigma_i^2, \qquad \text{(Eq. 7.7)}$$

which indeed is lower than the variance resulting from a 'simple' sampling from the total population, Eq. (7.5).

Only if the indiviual means $\hat{x}_i$ differ largely from the total mean $\hat{x}$, however, the difference is significant, and stratified sampling is worth doing. Otherwise, much time (for defining the p_i and estimating the individual $\bar{x}_i$) is spent for almost nothing.

ii) Interestingly, we can reduce $Var(\bar{x})$ even further, by choosing optimum n_i's. To this end, we minimize

$$Var(\bar{x}) = \sum_{i=1}^{k} p_i^2 Var(\bar{x}_i) = \sum_{i=1}^{k} p_i^2 \frac{\sigma_i^2}{n_i}$$

with respect to n_i and the condition $\sum_{i=1}^{k} n_i = N$ by means of the method of Langrangian multipliers.

We define the function to be minimized as

$$L = Var(\bar{x}) + \mu\left(\left(\sum_{i=1}^{k} n_i\right) - N\right),$$

and solve for $\frac{\partial L}{\partial n_i} = 0$ and $\frac{\partial L}{\partial \mu} = 0$ in parallel.

$$\frac{\partial L}{\partial n_i} = -\frac{p_i^2 \sigma_i^2}{n_i^2} + \mu = 0 \quad \Rightarrow \quad n_i = \frac{p_i \sigma_i}{\sqrt{\mu}}$$

$$\frac{\partial L}{\partial \mu} = \left(\sum_{i=1}^{k} n_i\right) - N = 0 \quad \Rightarrow \quad \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \frac{p_i \sigma_i}{\sqrt{\mu}} = N \quad \Rightarrow \quad \sqrt{\mu} = \frac{1}{N} \sum_{i=1}^{k} p_i \sigma_i \quad \Rightarrow$$

$$n_i = \frac{N p_i \sigma_i}{\sum_{i=1}^{k} p_i \sigma_i} \quad \text{and} \quad Var_{\text{opt}}(\bar{x}) = \frac{1}{N}\left(\sum_{i=1}^{k} p_i \sigma_i\right)^2, \text{ i.e., } \sigma_{\text{opt}}(\bar{x}) = \frac{\sum_{i=1}^{k} p_i \sigma_i}{\sqrt{N}} \quad \text{(Eq. 7.8)}$$

Thus, the *optimum* n_i and $\sigma(\bar{x})$ depend on the p_i AND on the variance of the individual subsamples. If all σ_i are identical, then $n_i \to N p_i$, and we recover Eq. (7.7).

For significantly different σ_i, on the other hand, *stratified sampling* with n_i according to Eq. (7.8) reduces the fluctuations of $\bar{x}$ considerably.

Note: The individual σ_i within the subpopulations need to be known or have to be estimated in parallel with the estimates $\bar{x}_i$.

Example for stratified sampling

The parent population shall consist of two Gaussian subpopulations, with

$p_1 = 0.8, \;\; \hat{x}_1 = 1, \;\;\; \sigma_1 = 0.5$

$p_2 = 0.2, \;\; \hat{x}_2 = 10, \;\; \sigma_2 = 4.0$

According to Eq. (7.3), $\hat{x} = 0.8 \cdot 1 + 0.2 \cdot 10 = 2.8$,

and the variance of the total sample is (Eq. 7.4 and below)

$Var(x) = 0.8 \cdot \left(0.5^2 + (1 - 2.8)^2\right) + 0.2 \cdot \left(4.0^2 + (10 - 2.8)^2\right) = 16.36$ or

$Var(x) = 0.8 \cdot 0.5^2 + 0.2 \cdot 4.0^2 + 0.8 \cdot 0.2 \cdot (10 - 1)^2 = 16.36$

NOTE: Variance dominated by the fluctuations of the individual means about the total mean

For stratified sampling, the optimum n_i are

$$\sum p_i \sigma_i = 0.8 \cdot 0.5 + 0.2 \cdot 4 = 1.2 \quad \Rightarrow \quad n_1 = N \frac{0.8 \cdot 0.5}{1.2} = \frac{N}{3}, \;\; n_2 = N \frac{0.2 \cdot 4}{1.2} = \frac{2N}{3},$$

i.e., $\frac{n_1}{n_2} = \frac{1}{2}$

Let's simulate this now, by sampling from a population of Gaussian random numbers. For a parent sample size of 800000 + 200000 numbers, we obtain $\bar{x}_p = 2.7998$ and $Var(x_p) = 16.3245$, very close to the theoretical values.

From *simple* sampling of the parent population with $N = 300$, we estimate

$\bar{x}_p = 3.20 \pm 0.25$, i.e., the estimated variance of the parent population

$Var(x_p) \approx N\sigma^2(\bar{x}_p) = 300 \cdot 0.25^2 = 18.75$, to be compared with 16.36 (not too bad)

From *stratified* sampling with $n_1 = 100$ and $n_2 = 200$, on the other hand, we find

$$\left.\begin{aligned} \bar{x}_1 &= 0.976,\ s(x_1) = 0.456 \\ \bar{x}_2 &= 9.960,\ s(x_2) = 4.305 \end{aligned}\right\} \text{ in agreement with the parameters for the subpopulations}$$

and thus

$\bar{x}_p \approx 0.8 \cdot 0.976 + 0.2 \cdot 9.96 = 2.77$

$$\sigma^2_{\text{opt}}(\bar{x}_p) \approx \frac{(0.8 \cdot 0.456)^2}{100} + \frac{(0.2 \cdot 4.305)^2}{200} \overset{\text{Eq. 7.8}}{\approx} \frac{(0.8 \cdot 0.456 + 0.2 \cdot 4.305)^2}{300} = 0.005, \text{ resulting in}$$

$\bar{x}_p = 2.77 \pm 0.07$ (Note: Equality in the two expressions above is only achieved if $s = \sigma$)

Other subsample sizes gave the following results

$$\left.\begin{aligned} n_1 : n_2 &= 1, \text{ i.e., } n_1 = 150 \text{ and } n_2 = 150 : \ \bar{x}_p = 2.85 \pm 0.08 \\ n_1 : n_2 &= 4, \text{ i.e., } n_1 = 240 \text{ and } n_2 = 60 : \ \bar{x}_p = 2.78 \pm 0.12 \end{aligned}\right\} \text{larger error than for } n_1 : n_2 = 0.5,$$

but still better compared to simple sampling

Obviously, the result from stratified sampling has a much better quality than from simple sampling, though the total sample size is identical.

For comparison, the results for $N = 3000$ ($n_1 = 1000$ and $n_2 = 2000$) are

$$\left.\begin{aligned} &\text{simple sampling:} \quad \bar{x}_p = 2.876 \pm 0.076 \\ &\text{stratified sampling:} \quad \bar{x}_p = 2.807 \pm 0.023 \end{aligned}\right\} \left(1/\sqrt{\text{N}} \text{ scaling of error in both cases}\right)$$

Sampling from finite populations without replacement

- finite (i.e., also discrete) population: independence of consecutive drawings is lost when individual elements are not replaced
- thus, no genuine random sampling
  - should be no problem when number of elements, N, is very large compared to sample size n

Let the population be composed of N elements, $y_1 ... y_N$. At first, we need to define the population mean $\hat{y}$ and the variance $\sigma^2(y)$. Since each element has the same probability to be drawn,

$$\hat{y} = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \, .$$

Thus, $\hat{y}$ is just the arithmetic mean of the population elements.

Here (but see also Chap. 1), we define the population variance as

$$\sigma^2(y) = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 ,$$

since the number of degrees of freedom ($\rightarrow$ Chap. 8) of the sum of squares is $(N-1)$.

In the above sum, the first term can take any value, the 2nd one as well and so on until the $(N-1)$th term. The Nth term, however, is completely determined, due to the restriction $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

Now, we draw a sample $x_1 ... x_n$ from the population $y_1 ... y_N$, with $n \le N$,

and determine sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and sample variance,

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 .$$

Without proof $\left(\text{see Brandt, Chap. 6.4}\right)$, we quote the following properties

(remember: finite parent population, no replacement)

i) $E(\bar{x}) = \bar{y}$

The sample mean is an unbiased estimator of the population mean.

ii) $E(s_x^2) = \sigma^2(y)$

The sample variance is an unbiased estimator of the population variance.

iii) $Var(\bar{x}) = \frac{\sigma^2(y)}{n}\left(1 - \frac{n}{N}\right)$

For $n \ll N$, similar to the case of an infinite population.

For $n < N$, smaller than for the case of an infinite population.

For n=N, the variance of the mean becomes zero, since sample and population mean are identical.

SUMMARY: mean and variance for a sample drawn from a finite population without replacement have similar properties as if drawn from an infinite population.

→ mean of 'Lotto' numbers, 'Feynman's restaurant problem'

Likelihood

- What has been discussed so far, can be generalized as follows
  - Given a sample of observations/measurements, we like to find the appropriate theoretical description of the properties of the underlying population.
  - Examples

case 1	given: wanted:	N alternative hypotheses H_i relative probabilities for the validity of the H_i
case 2	given: wanted:	one hypothesis H_o a statement about the validity of $H_o \rightarrow$ (Chap. 10)
case 3	given: wanted:	a valid hypothesis H(λ), where λ is a set of unknown, continuous parameters “best” value for λ and error

The Bayes theorem revisited

- discrete hypotheses
  - Bayes theorem (see Chap. 2)

    $$P(A\cdot B) = P(A\,|\,B)P(B) = P(B\,|\,A)P(A)$$

  - apply to an observed value $k^{(1)}$ and hypotheses H_i . Here and in the following, variables with superscripts denote observed/measured quantities, and the "1" denotes that it is the first (and only) measurement of the r.v. k. (from here on, we don't distinguish $k^{(1)}$ from $k^{(1)}$).
  - Let's assume that there is a restricted number of hypotheses which can explain the observation, and that we know the probability distribution $P(k|\,H_i)$ for the r.v. k. The probability for the validity of H_i given k is

    $$P(H_i\,|\,k) = \frac{P(k\,|\,H_i)P(H_i)}{P(k)}$$

    and for the specific observation $k^{(1)}$

    $$P(H_i\,|\,\mathrm{k} = k^{(1)}) = \frac{P(k^{(1)}\,|\,H_i)P(H_i)}{P(k^{(1)})} = \frac{P(k^{(1)}\,|\,H_i)P(H_i)}{\sum_j P(k^{(1)}\,|\,H_j)P(H_j)}$$

    The 2nd equality follows from the rule of total probability, or (if the H_i are not exhaustive and not mutually exclusive), from normalizing $P(H_i\,|\,k^{(1)})$ in such a way that the probability for the validity of any of the hypothesis is equal to unity, i.e., $\sum_i P(H_i\,|\,k^{(1)}) = 1$

  - $P(H_i|\,k^{(1)})$ is the *a posteriori* probability (briefly called *posterior*) for the validity of the hypothesis i after the event $k^{(1)}$ has occurred, and is the quantity we are interested in.
  - What can be easily calculated is the probability $P(k^{(1)}|H_i)$, since the hypothesis (theory) is known.
  - $P(H_i)$ is the probability for the validity of the hypothesis i *before* the observation, the *prior.* Generally, this is the cumbersome quantity.

- The relative probabilities for two hypotheses i and j for a specific measurement $k^{(1)}$ is given by

$$\frac{P(H_i \mid k^{(1)})}{P(H_j \mid k^{(1)})} = \frac{P(k^{(1)} \mid H_i)P(H_i)}{P(k^{(1)} \mid H_j)P(H_j)}$$

- Note: this ratio does not depend on $P(k^{(1)})$ [which can be also very cumbersome, even if all priors were known, due to a large number of potential hypotheses or parameters]
- In MCMC (Markov Chains Monte Carlo, see Sect. 9), this ratio plays a crucial role.

example:

The decay probabilities, μ, of pions and kaons into myons in a detector are $P(\mu \mid \pi) = 0.02$ and $P(\mu \mid K) = 0.10$, respectively. The relative abundances of pions and kaons are 3:1.

A myon has been detected. Does it originate from a pion or a kaon?

$$\frac{P(\pi \mid \mu)}{P(K \mid \mu)} = \frac{P(\mu \mid \pi)P(\pi)}{P(\mu \mid K)P(K)} = \frac{0.02 \cdot 3}{0.10 \cdot 1} = 0.6$$

A decay from a pion is 60% as probable as a decay from a kaon.

Continuous parameters

- Now, we investigate the case that a parameter λ of a hypothesis is looked for.
  In this case, we deal with probability densities, and for two r.v., the Bayes theorem reads (see Chap. 4)

$$f(x,\lambda) = f(x\,|\,\lambda)h(\lambda) = f(\lambda\,|\,x)g(x)$$

  where h and g are the marginal distributions of f $\left[h(\lambda) = \int f(x,\lambda)\mathrm{d}x; \quad g(x) = \int f(x,\lambda)\mathrm{d}\lambda\right]$.

- If the observation gives the result $x^{(1)}$ and the parameter of the hypothesis (theory!) is λ, the corresponding probability density reads

$$f(\lambda\,|\,x^{(1)}) = \frac{f(x^{(1)}\,|\,\lambda)h(\lambda)}{g(x^{(1)})}$$

  Thus, the probability density for the parameter λ given a measurement $x^{(1)}$ (wanted) depends on the probability density for the measurement of $x^{(1)}$ given the parameter λ (calculatable) times the prior $h(\lambda)$ (cumbersome), divided by a normalization factor (the evidence)

$$f(\lambda\,|\,x^{(1)}) = \frac{f(x^{(1)}\,|\,\lambda)h(\lambda)}{\int_{-\infty}^{\infty} f(x^{(1)}\,|\,\lambda)h(\lambda)d\lambda}$$

- see also "conjugate priors"

Example:

We measure the time $t^{(1)}$ when a particle has decayed. The measurement error should be Gaussian, with resolution s. Thus, the probability to measure an arbitrary t for an actual decay time T is $f(t\,|\,T) \propto \exp\left(-\frac{(t-T)^2}{2s^2}\right)$. The pdf for the decay time before the measurement (the prior) follows from the decay law, $h(T) = \frac{\exp(-T/\tau)}{\tau}$.

Because of the normalization, constant factors can be neglected. The probability density that the actual decay time is T when we measure $t^{(1)}$ is thus given by

$$f(T\,|\,t^{(1)}) = \frac{\exp\left(-\frac{(t^{(1)}-T)^2}{2s^2}\right)\exp\left(-\frac{T}{\tau}\right)}{\int_0^\infty \exp\left(-\frac{(t^{(1)}-T)^2}{2s^2}\right)\exp\left(-\frac{T}{\tau}\right)dT} =$$

Example evaluated with Mathematica

```
In[1]:= f[T_, t1_] := Exp[-(t1 - T)^2/2/s^2] * Exp[-T/tau]
```
nominator

```
In[2]:= g[t1_] := Integrate[f[T, t1], {T, 0, Infinity}, Assumptions → Re[s^2] > 0]
```
denominator

```
In[3]:= g[t1]
```

$$\text{Out[3]}= e^{\frac{s^2-2\,t1\,tau}{2\,tau^2}}\sqrt{\frac{\pi}{2}}\;s\left(-1+\sqrt{\frac{1}{s^2}}\;s+\text{Erfc}\left[\frac{s^2-t1\;tau}{\sqrt{2}\;s\;tau}\right]\right) = \int_0^\infty f(t^{(1)}\,|\,T)h(T)dT$$

```
In[4]:= D[f[T, t1]/g[t1], T]
```
derivative of $f(T\,|\,t^{(1)})$ with respect to T

$$\text{Out[4]}= \frac{e^{-\frac{(-T+t1)^2}{2s^2}-\frac{T}{tau}-\frac{s^2-2\,t1\,tau}{2\,tau^2}}\sqrt{\frac{2}{\pi}}\left(\frac{-T+t1}{s^2}-\frac{1}{tau}\right)}{s\left(-1+\sqrt{\frac{1}{s^2}}\;s+\text{Erfc}\left[\frac{s^2-t1\;tau}{\sqrt{2}\;s\;tau}\right]\right)}$$

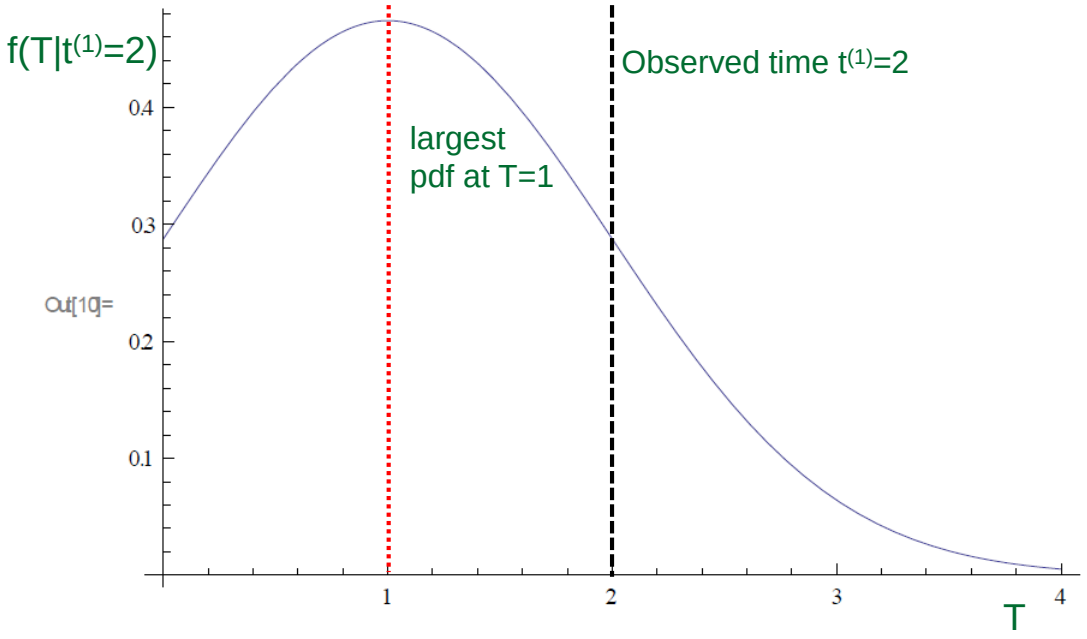
```
In[5]:= Solve[D[f[T, t1]/g (t1), T] == 0, T]
```
determine maximum of $f(T\,|\,t^{(1)})$ by solving for $\left.\frac{df(T\,|\,t^{(1)})}{dT}\right|_T = 0$

$$\text{Out[5]}= \left\{\left\{T \to \frac{-s^2+t1\;tau}{tau}\right\}\right\} \quad T_{\text{most prob.}} = t^{(1)} - \frac{s^2}{\tau},$$

```
In[6]:=
tau = 1.
s = 1.
t1 = 2.
x = Re[g[t1]]
Out[6]= 1.
Out[7]= 1.
Out[8]= 2.
Out[9]= 0.470568
In[10]:= Plot[f[T, t1] / x, {T, 0., 4.}]
```

Given an observed decay time $t^{(1)}$, the highest probability density for the actual decay time is located at earlier times, due to the exponential prior. E.g., if $s = 1$, $\tau = 1$, then $T_{\text{most prob.}} = t^{(1)} - 1$.

Out[10]=



Likelihood quotient and function

- usually, the probability (density) for the prior is unknown.
  - What are the prior probabilities for two alternative hypothesis?
  - What is the pdf for a certain parameter? (As outlined in Chap. 2, this is often assumed as being uniform, but then, e.g., f(m) and f(m^2) give different results for f(m|x).)
  - Thus, for alternative hypotheses i,j or distinct parameters λ_i, λ_j, one usually quotes the "likelihood ratio"

$$Q = \frac{P(k^{(1)} \mid H_i)}{P(k^{(1)} \mid H_j)} \quad \text{or} \quad Q = \frac{f(x^{(1)} \mid \lambda_i)}{f(x^{(1)} \mid \lambda_j)}$$

    which contains the full information of the observation. Neyman (1937, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", Phil. Trans. A236, 333) has shown that to discriminate between two alternative hypotheses there is no other parameter which is more effective.
  - The result of the measurement can be (very loosely) expressed by saying that the hypothesis/parameter i is Q times more probable than the hypothesis/parameter j (strictly speaking, such a statement of absolute probabilities assumes equal priors).
  - Better to quote "only" the likelihood ratio, i.e., by saying that the likelihood of hypothesis/parameter i is Q times larger than the likelihood of hypothesis/parameter j.

- definition: likelihood
  (in German: “Wahrscheinlichkeit”, same word as for “probability”. Thus, also in German called “likelihood”, to allow for a discrimination)

  The likelihood L_i for a hypothesis H_i with pdf $f_i(x)$ or a discrete probability distribution $P_i(k)$ and observations $x^{(1)}$ or $k^{(1)}$, respectively, is given by

  $$L_i(x^{(1)}) = f_i(x^{(1)}) \qquad \text{or} \qquad L_i(k^{(1)}) = P_i(k^{(1)}).$$

  For probability densities $f(x \mid \lambda)$ or probabilities $P(k \mid \lambda)$ and observations $x^{(1)}$ or $k^{(1)}$, the likelihood is given by

  $$L(\lambda \mid x^{(1)}) = f(x^{(1)} \mid \lambda) \quad \text{or } L(\lambda \mid k^{(1)}) = P(k^{(1)} \mid \lambda).$$

  The likelihood quantifies the validity of a hypothesis for a given observation, whereas the pdf relates the r.v. with the hypothesis. The consideration of a likelihood makes only sense if more than one hypothesis is given, or if the hypothesis depends on parameters. If the likelihood depends on parameters, it is called a likelihood function.

- The likelihood is an a posteriori probability, in contrast to “normal” a priori probability, and must not be confused with a conventional pdf.
- The likelihood is large when the occurrence of a specific observation for a given hypothesis is likely. The likelihood quantifies in how much a hypothesis is supported by the data.
- If a specific observation is extremely unlikely (i.e., L is very small), the validity of the hypothesis is more than doubtful, but only if other hypotheses with larger L are available:
  → for parameter estimation, maximize $L(\lambda)$

Likelihood for a sample

- Usually, an experiment results in a *sample* of N independent values $x^{(j)}$,j=1,N, which are independently, identically distributed (i.i.d.), following the pdf f(x).

The combined pdf is then the product of the individual pdfs,

$$g(x_1,...,x_N) = \prod_{j=1}^{N} f(x_j) \qquad \text{(here, the } x_j \text{ are still r.v.)}$$

whereas g evaluated for the observed sample is the sample likelihood,

$$L(x^{(1)},...,x^{(N)}) = \prod_{j=1}^{N} L(x^{(j)}) = \prod_{j=1}^{N} f(x^{(j)})$$

For discrete variables, we have

$$L(k^{(1)},...,k^{(N)}) = \prod_{j=1}^{N} L(k^{(j)}) = \prod_{j=1}^{N} P(k^{(j)} \mid H),$$

and for pdf's which depend on a parameter λ (or parameter-set $\boldsymbol{\lambda}$)

$$L(\boldsymbol{\lambda} \mid x^{(1)},...,x^{(N)}) = \prod_{j=1}^{N} L(\boldsymbol{\lambda} \mid x^{(j)}) = \prod_{j=1}^{N} f(x^{(j)} \mid \boldsymbol{\lambda}),$$

$$L(\boldsymbol{\lambda} \mid k^{(1)},...,k^{(N)}) = \prod_{j=1}^{N} L(\boldsymbol{\lambda} \mid k^{(j)}) = \prod_{j=1}^{N} P(k^{(j)} \mid \boldsymbol{\lambda})$$

log-likelihood

- in the following, we will concentrate on the last case, i.e., probabilities/probability densities which depend on a parameter

$$L(\boldsymbol{\lambda} \mid x^{(1)},\ldots,x^{(N)}) = \prod_{j=1}^{N} f(x^{(j)} \mid \boldsymbol{\lambda}).$$

Since for many reasons (e.g., to find the maximum of the likelihood) one has to calculate the derivative of L, it is convenient to consider the log-likelihood,

$$\ln\ L(\lambda) = \sum_{j=1}^{N} \ln\Big[\, f(x^{(j)} \mid \lambda)\Big]$$

log-likelihood – example

In two equidistant time-intervals, we detect 5 and 10 X-ray photons from an X-ray source. Two competing theories predict a mean number of either 2 or 12 photons per interval. Which theory is "more likely"?

$$P(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}; \quad \ln L(\lambda \mid k^{(1)}, k^{(2)}) = \ln L(\lambda) = -2\lambda + \sum_{j=1}^{2} (k_j \ln \lambda - \ln k_j!)$$

$$\ln L(2) = -4 + \left[(5 \ln 2 - \ln 5!) + (10 \ln 2 - \ln 10!) \right] = -13.49; \quad \ln L(12) = -6.82$$

The log-likelihood for λ=12 is much larger than the one for λ=2, with a likelihood ratio of

$$\frac{L(2)}{L(12)} = \exp(-13.49 + 6.82) = 0.001.$$

The observed sample indicates that hypothesis 1 might be excluded. (The significance of such results will be discussed in Chap. 10).

Remember that to obtain actual probability (density) ratios, one would have to multiply with the prior ratios. If both hypotheses were equally probable, then one could say that hypothesis/theory 1 is a factor of 0.001 less probable than hypothesis/theory 2.

Maximum likelihood

- Generalizing the foregoing concept, a variation of the (set of) parameter(s) in the likelihood function allows for infinite hypotheses.
- Highest confidence should be in that set $\tilde{\boldsymbol{\lambda}}$ which has the maximum likelihood, which is then the best estimate of $\boldsymbol{\lambda}$ (“most likely value of $\boldsymbol{\lambda}$”, but actually the value of $\boldsymbol{\lambda}$ which makes the data most likely).
- The error of $\tilde{\boldsymbol{\lambda}}$ can be derived from the distribution of L about $\tilde{\boldsymbol{\lambda}}$.
- for many parameters, maximum needs to be derived numerically
- if more than one maximum, prefer the one with the largest L
- caution if several maxima with almost equal L are present

Let's first consider the case with a single parameter, λ. Since the maxima of L and $\ln L$ occur at the same λ, the "most likely value of λ" can be calculated by solving the likelihood equation

$$\left.\frac{d\ln L}{d\lambda}\right|_{\tilde{\lambda}} = 0 = \sum_{j=1}^{N}\frac{d}{d\lambda}\ln\left[f(x^{(j)};\lambda)\right] = \sum_{j=1}^{N}\frac{d\left[f(x^{(j)};\lambda)\right]/d\lambda}{f(x^{(j)};\lambda)} = \sum_{j=1}^{N}\frac{f'}{f}$$

[Here and in the following, we write for brevity $f(x^{(j)};\lambda) =: f(x^{(j)}\,|\,\lambda)$, where this (conditional) pdf is normalized with respect to all $x^{(j)}$ (see Chap. 4), i.e.
$\int f(x^{(1)}, x^{(2)}, \ldots x^{(N)};\lambda)\mathrm{d}x^{(1)}\mathrm{d}x^{(2)}\ldots\mathrm{d}x^{(N)} = 1$]

In the general case of p parameters, the likelihood equation is replaced by a system of p equations which have to be solved simultaneously

$$\left.\frac{\partial\ln L}{\partial\lambda_i}\right|_{\tilde{\lambda}} = 0, \quad i = 1, p$$ [evaluated at the parameter set $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots\tilde{\lambda}_p)$, see below]

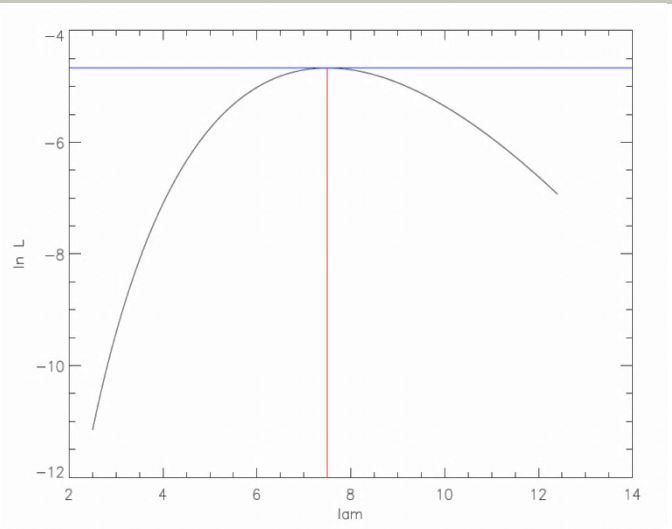
continue with last example:

Consider now the same measurements, but without a given hypothesis for the mean value. Instead, derive the "most likely value of λ" given the data.

$$\ln L(\lambda) = -N\lambda + \sum_{j=1}^{N}(k^{(j)}\ln\lambda - \ln k^{(j)}!) \;\Rightarrow\; \left.\frac{d\ln L}{d\lambda}\right|_{\tilde{\lambda}} = 0 = -N + \sum_{j=1}^{N}\frac{k^{(j)}}{\tilde{\lambda}} \;\Rightarrow\; \tilde{\lambda} = \frac{1}{N}\sum_{j=1}^{N}k^{(j)}$$

The maximum likelihood value for the mean is the sample mean (which was to be expected)

For the actual example, we find $\tilde{\lambda} = \frac{5+10}{2} = 7.5$. For this value, $\ln L(7.5) = -4.66$



log-likelihood function ln L(λ) for the previous example (Poisson-dist, observed values k_j=(5,10)). The maximum of the function is located at $\tilde{\lambda}$=7.5

Another example

Assume that a quantity is measured repeatedly, but with Gaussian errors of different variance, where the mean is unknown. The likelihood function for measurement $x^{(j)}$ is thus

$$L(\mu \mid x^{(j)}) = f(x^{(j)};\mu) = \frac{1}{\sqrt{2\pi}\sigma_j}\exp\left(-\frac{(x^{(j)}-\mu)^2}{2\sigma_j^2}\right).$$

The combined likelihood for N measurements is the product of the individual likelihoods, and the log-likelihood function becomes

$$\ln L(\mu) = -\frac{1}{2}\sum_{j=1}^{N}\frac{(x^{(j)}-\mu)^2}{\sigma_j^2} + f(\sigma_{j,j=1,N}) \text{ with } f(\sigma_{j,j=1,N}) = -\frac{N}{2}\ln(2\pi) - \sum_{j=1}^{N}\ln(\sigma_j)$$

The solution of the likelihood equation, $\left.\frac{d\ln L}{d\mu}\right|_{\tilde{\mu}} = 0,$ results in $\tilde{\mu} = \frac{\sum_{j=1}^{N}\frac{x^{(j)}}{\sigma_j^2}}{\sum_{j=1}^{N}\frac{1}{\sigma_j^2}},$

which is just the "addition theorem" for calculating means from weighted measurements (see Chap. 6).

Conjugate priors

Remember (once more) Bayes theorem:

$$p(\lambda \,|\, \text{data},\text{model}) = \frac{p(\text{data}|\lambda, \text{model})\, p(\lambda)}{p(\text{data}|\text{model})},$$ where data are the measured quantities and

$p(\text{data}|\lambda, \text{model}) = L(\lambda \,|\, \text{data}, \text{model})$ is the likelihood function

For **conjugate** priors (conjugate to the specific likelihood function), the posterior follows the same distribution as the prior.

Conjugate priors play an important role for analytic solutions, parameter studies and test cases (e.g., for MCMC simulations, see Sect. 9).

Example 1: the prior from page 170 is NOT conjugate.

Example 2 for a conjugate prior :

normal likelihood with known σ + normal prior for μ $\rightarrow$ normally distributed posterior, $(\mu \,|\, \text{data}, \sigma)$

In particular: likelihood for a dataset distributed according to $N(\mu, \sigma)$, i.e., $L \propto \prod_{i=1}^{n} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$

normal prior with M and S, i.e., $p(\mu) \propto \exp\left(-\frac{(\mu - M)^2}{2S^2}\right)$

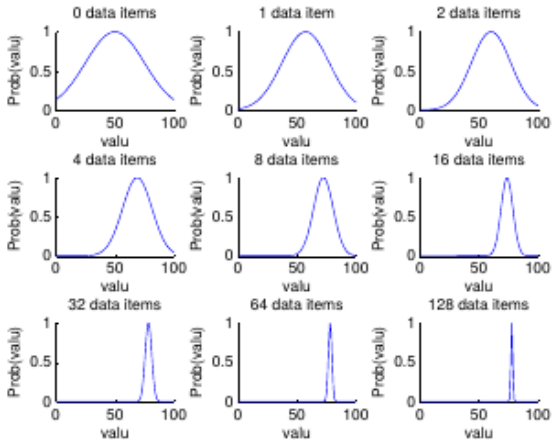
$\Rightarrow$ posterior $p(\mu \,|\, x, \sigma) = N(\mu_{tot}, \sigma_{tot})$ with $\mu_{tot} = \frac{\frac{1}{S^2}}{\frac{1}{S^2} + \frac{n}{\sigma^2}} M + \frac{\frac{n}{\sigma^2}}{\frac{1}{S^2} + \frac{n}{\sigma^2}} \bar{x}$ and $\sigma_{tot}^2 = \frac{\frac{\sigma^2}{n}}{S^2 + \frac{\sigma^2}{n}} S^2$

Conjugate priors

posterior $p(\mu \mid x, \sigma) = N(\mu_{tot}, \sigma_{tot})$ with $\mu_{tot} = \dfrac{\frac{1}{S^2}}{\frac{1}{S^2} + \frac{n}{\sigma^2}} M + \dfrac{\frac{n}{\sigma^2}}{\frac{1}{S^2} + \frac{n}{\sigma^2}} \bar{x}$ and $\sigma^2_{tot} = \dfrac{\frac{\sigma^2}{n}}{S^2 + \frac{\sigma^2}{n}} S^2$

Note:

- the mean is a weighted average of the prior mean M and the data mean $\bar{x}$. The weight on the prior mean is inversely proportional to the variance of the prior mean ($\propto 1/S^2$), and the weight on the data mean is inversely proportion to the variance of the data mean ($\propto n/\sigma^2$).
- if the prior mean is very precise relative to the data mean (low n), then it is highly weighted. Alternatively, if the data mean is more precise (large n), then it is assigned a larger weight. Thus, μ_{tot} varies from M to $\bar{x}$ for increasing n.
- the variance of $\mu|x,\sigma$ is smaller than the variance of the prior mean (S^2) and smaller than the variance of the data mean (σ^2/n), but varying between both limits if $n = 0 \rightarrow \infty$



Example on the left: Posterior distribution for the following scenario: data sampled from a normal distribution with a mean of 80 and a standard deviation of σ=10. The sample size n varies from 0 to 128. The prior distribution is assumed to be normal, with a mean M = 50, and a standard deviation, S = 25. Obviously, the posterior distribution for μ varies from the prior one (n=0) over some "compromise" solution for small n, to the distribution estimated from the sample, with mean $\bar{x}$, and variance of the mean σ^2/n. Note that the pdfs have been renormalized to a max. of unity.

Example taken from Robert Jacobs, lecture notes on Bayesian Statistics: Normal-Normal Model, based on Lynch, S. M. (2007), Introduction to Applied Bayesian Statistics and Estimation for Social Scientists. New York: Springer

Information inequality and minimum variance bound (Schranke minimaler Varianz)

Having defined the concept of likelihoods, we reconsider the problem of constructing estimators S with desirable properties. A "good" estimator for the parameter λ should be unbiased,

$$B(\lambda) = E(\mathrm{S}) - \lambda = 0,$$

and should have a variance $\sigma^2(\mathrm{S})$ as small as possible.

We will now show that there exists a relation between both quantities, the so-called information inequality, such that frequently a compromise between the requirement of minimum bias and variance has to be found.
(E.g., an estimator with $\sigma^2(\mathrm{S}) = 0$ can be always constructed, by choosing $\mathrm{S} = \text{const}$, but this estimator will be strongly biased).

We consider an estimator $\mathrm{S}(x^{(1)}, x^{(2)}, ..., x^{(N)})$, with a joint pdf for the sample

$$f(x^{(1)}, x^{(2)}, ..., x^{(N)}; \lambda) = f(x^{(1)}; \lambda) f(x^{(2)}; \lambda) \cdots f(x^{(N)}; \lambda).$$

Then,

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

$$E(\mathrm{S}) = \int S(x^{(1)}, x^{(2)}, ..., x^{(N)})\, f(x^{(1)};\lambda)\, f(x^{(2)};\lambda) \cdots f(x^{(N)};\lambda) \mathrm{d}x^{(1)} \mathrm{d}x^{(2)} \cdots \mathrm{d}x^{(N)} = B(\lambda) + \lambda. \quad \text{(Eq. 7.9)}$$

We now assume that we can differentiate under the integral w.r.t. λ (S is independent of λ), and obtain

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

$$1 + B'(\lambda) = E\left\{ \mathrm{S} \left(\sum_{j=1}^{N} \frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)} \right) \right\} = E\left\{ \mathrm{S} \frac{d \ln L(\lambda)}{d\lambda} \right\}.$$

The normalization of $f(x^{(1)}, x^{(2)}, ..., x^{(N)};\lambda)$ reads

$$\int f(x^{(1)};\lambda)\, f(x^{(2)};\lambda) \cdots f(x^{(N)};\lambda) \mathrm{d}x^{(1)} \mathrm{d}x^{(2)} \cdots \mathrm{d}x^{(N)} = 1, \quad \text{and the derivative with respect to } \lambda$$

$$\int \left(\sum_{j=1}^{N} \frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)} \right) f(x^{(1)};\lambda)\, f(x^{(2)};\lambda) \cdots f(x^{(N)};\lambda) \mathrm{d}x^{(1)} \mathrm{d}x^{(2)} \cdots \mathrm{d}x^{(N)} = 0 = E\left\{ \frac{d \ln L(\lambda)}{d\lambda} \right\}. \quad \text{(Eq. 7.10)}$$

Multiplying Eq. (7.10) with $E(\mathrm{S})$ (still $= 0$) and subtracting from Eq. (7.9), we obtain

$$1 + B'(\lambda) = E\left\{ \mathrm{S} \frac{d \ln L(\lambda)}{d\lambda} \right\} - E(\mathrm{S}) E\left\{ \frac{d \ln L(\lambda)}{d\lambda} \right\}.$$

$$1 + B'(\lambda) = E\left\{\mathrm{S}\frac{d\ln L(\lambda)}{d\lambda}\right\} - E(\mathrm{S})E\left\{\frac{d\ln L(\lambda)}{d\lambda}\right\} = E\left\{[\mathrm{S} - E(\mathrm{S})]\frac{d\ln L(\lambda)}{d\lambda}\right\}, \quad \text{(Eq. 7.11)}$$

which is of the type $E(\mathrm{xy})$.

To proceed further, we invoke the Schwarz inequality, $\left(E(\mathrm{xy})\right)^2 \le E(\mathrm{x}^2)E(\mathrm{y}^2)$ (Eq. 7.12)

Proof:

$E\left\{(a\mathrm{x} + \mathrm{y})^2\right\} = a^2E(\mathrm{x}^2) + 2aE(\mathrm{xy}) + E(\mathrm{y}^2) \ge 0$ is a non-negative number for *all* values of $a \in \mathbb{R}$

To fullfil this condition, the discriminant (regarding the solution $a_{1,2}$ of $E\left\{(a\mathrm{x} + \mathrm{y})^2\right\} = 0$),

$D = \left(2E(xy)\right)^2 - 4E(x^2)E(y^2)$,

(i) must be either $D = 0$ (unique solution, since only for $ax + y = 0$ we have $E\left\{(a\mathrm{x} + \mathrm{y})^2\right\} = 0$,

(ii) or $D < 0$ (no real solution for $E\left\{(a\mathrm{x} + \mathrm{y})^2\right\} = 0$)

$\Rightarrow$ together: $D \le 0$, i.e.,

$4\left(E(\mathrm{xy})\right)^2 - 4E(\mathrm{x}^2)E(\mathrm{y}^2) \le 0$, which proves the inequality.

$$\left(1+B'(\lambda)\right)^2 = \left(E\left\{\left[\mathrm{S}-E(\mathrm{S})\right]\frac{d\ln L(\lambda)}{d\lambda}\right\}\right)^2 \overset{\text{Schwarz inequality}}{\leq} E\left\{\left[\mathrm{S}-E(\mathrm{S})\right]^2\right\} E\left\{\left(\frac{d\ln L(\lambda)}{d\lambda}\right)^2\right\}.$$

$$\Rightarrow E\left\{\left[\mathrm{S}-E(\mathrm{S})\right]^2\right\} = Var(\mathrm{S}) \geq \frac{\left(1+B'(\lambda)\right)^2}{E\left\{\left(d\ln L(\lambda)/d\lambda\right)^2\right\}}$$

The quantity in the denominator is called the information of the sample w.r.t. λ, which is a non-negative number that vanishes if the likelihood-function does not depend on λ, and the inequality is called the Cramer-Rao- or Frechet- or information inequality. The r.h.s. is called the minimum variance bound, MVB.

Since

$$E\left\{\frac{d\ln L(\lambda)}{d\lambda}\right\} = \int\left(\sum_{j=1}^{N}\frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)}\right) f(x^{(1)};\lambda) f(x^{(2)};\lambda)\cdots f(x^{(N)};\lambda)\mathrm{d}x^{(1)}\mathrm{d}x^{(2)}\cdots\mathrm{d}x^{(N)} = 0,$$

further differentiation w.r.t. λ yields (in obvious notation: $dX = \mathrm{d}x^{(1)}\mathrm{d}x^{(2)}\cdots\mathrm{d}x^{(N)}$)

$$\frac{d}{d\lambda}\int\left(\sum_{j=1}^{N}\frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)}\right) f(x^{(1)};\lambda) f(x^{(2)};\lambda)\cdots f(x^{(N)};\lambda)\mathrm{d}x^{(1)}\mathrm{d}x^{(2)}\cdots\mathrm{d}x^{(N)} = \frac{d}{d\lambda}\int\frac{d\ln L}{d\lambda}L\,dX = 0$$

$$\Rightarrow \int\left(\frac{d^2\ln L}{d\lambda^2}L + \frac{d\ln L}{d\lambda}\frac{dL}{d\lambda}\right)dX = 0 = \int\left(\frac{d^2\ln L}{d\lambda^2} + \frac{d\ln L}{d\lambda}\frac{dL}{Ld\lambda}\right)L\,dX$$

$$\int\left(\frac{d^2\ln L}{d\lambda^2}L+\frac{d\ln L}{d\lambda}\frac{dL}{d\lambda}\right)dX=0=\int\left(\frac{d^2\ln L}{d\lambda^2}+\frac{d\ln L}{d\lambda}\frac{dL}{Ld\lambda}\right)LdX$$

$$\Rightarrow E\left\{\frac{d^2\ln L}{d\lambda^2}\right\}=-E\left\{\left(\frac{d\ln L}{d\lambda}\right)^2\right\},$$ so that an alternative formulation for the MVB is given by

$$Var(S)\geq\frac{\left(1+B'(\lambda)\right)^2}{I(\lambda)}=\frac{\left(1+B'(\lambda)\right)^2}{E\left\{\left(d\ln L(\lambda)/d\lambda\right)^2\right\}}=\frac{-\left(1+B'(\lambda)\right)^2}{E\left\{d^2\ln L(\lambda)/d\lambda^2\right\}}\qquad\text{(Eq. 7.13)}$$

The larger the information in the sample, the smaller the variance of the estimate!

Note: the nominator of the above equation depends on *S*,
whilst the denominator depends on the pdf and *N*

The information $I(\lambda)=E\left\{\left(\frac{d\ln L}{d\lambda}\right)^2\right\}=-E\left\{\frac{d^2\ln L}{d\lambda^2}\right\}$ can be also written in terms of the individual, *independent* probabilities

$$I(\lambda)=NE\left\{\left(\frac{f'(\mathrm{x},\lambda)}{f(\mathrm{x},\lambda)}\right)^2\right\}=-NE\left\{\left(\frac{f'(\mathrm{x},\lambda)}{f(\mathrm{x},\lambda)}\right)'\right\},\qquad\text{(Eq. 7.13a)}$$

if the expectation value is defined in analogy, $E(\mathrm{x})=\int_{-\infty}^{\infty}xf(x,\lambda)dx.$

Example

Efficiency of the sample mean for a Gaussian distribution

For a sample resulting from a normal distribution (identical mean and variance), the individual likelihoods are

$$L(\mu \mid x^{(j)}) = f(x^{(j)};\mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x^{(j)}-\mu)^2}{2\sigma^2}\right),$$

and the combined log-likelihood for a sample of size N becomes

$$\ln L(\mu) = -\sum_{j=1}^{N}\frac{(x^{(j)}-\mu)^2}{2\sigma^2} - N\ln\left(\sigma\sqrt{2\pi}\right)$$

$$\Rightarrow \frac{d^2 \ln L}{d\mu^2} = -\frac{N}{\sigma^2} \;\Rightarrow\; \text{MVB} = \frac{-(1+B'(\mu))^2}{E(-N/\sigma^2)} = \frac{\sigma^2}{N}(1+B'(\mu))^2 \text{ with } B(\mu) = E(S) - \mu$$

On the other side (from previous considerations), we also know that the variance of the sample mean is

$$Var(\bar{x}) = \frac{\sigma^2}{N}$$

which shows that this estimator has a variance according to the MVB (remember that the sample mean is an unbiased estimator, $B(\bar{x}) = 0$).

Such estimators with $Var(S)$=MVB are called efficient, otherwise their efficiency is given by the ratio

$$\eta = \frac{\text{MVB}}{Var(S)}$$

We might now ask under which general conditions an estimator is efficient, i.e., when do we find the equal sign in the information inequality?

Minimum variance estimators

To obtain equality in the Schwarzschild inequality Eq. (7.12) , we must have

$ax + y = 0,$ because only then $E\left\{(ax + y)^2\right\} = 0.$

This implies that (cf. Eq. 7.11)

$$a(\mathrm{S} - E(\mathrm{S})) + \frac{d \ln L}{d\lambda} = 0, \quad \text{or generally}$$

$$\frac{d \ln L}{d\lambda} = A(\lambda)(\mathrm{S} - E(\mathrm{S})), \qquad \text{(Eq. 7.14)}$$

where A must not depend on the sample $x^{(1)},...,x^{(N)},$ though it might depend on $\lambda.$

By integration (between $\lambda_{\min}$ and λ) and noting that $E(\mathrm{S}) = B(\lambda) + \lambda = f(\lambda),$

$$\ln L(\lambda) = \int \left[A(\lambda)\mathrm{S} - A(\lambda) f(\lambda)\right] d\lambda = C(\lambda)\mathrm{S} + D(\lambda) + \text{const}$$

$$L(\lambda) = d \exp\left\{C(\lambda)\mathrm{S} + D(\lambda)\right\}, \qquad \text{(Eq. 7.15)}$$

where d does not depend on λ. Estimators accompanied with likelihood functions of this type attain the MVB, and are called minimum variance estimators. In case of unbiased minimum estimators, $B(\lambda) = 0$, we then have

$$Var(S) = \frac{1}{I(\lambda)} = \frac{1}{E\left\{\left(d \ln L(\lambda)/d\lambda\right)^2\right\}} = \frac{1}{A^2(\lambda) E\left\{(\mathrm{S} - E(\mathrm{S}))^2\right\}} = \frac{1}{A^2(\lambda) E\left\{(\mathrm{S} - \lambda)^2\right\}} = \frac{1}{A^2(\lambda) Var(\mathrm{S})} \Rightarrow$$

$$Var(S) = \frac{1}{A(\lambda)}$$

Sufficient estimators

If instead of $L = d \exp\{C(\lambda)\mathrm{S} + D(\lambda)\}$ only the weaker condition

$$L = g(\mathrm{S}, \lambda) \cdot c(x^{(1)}, ..., x^{(N)}) \qquad \text{(Eq. 7.16)}$$

holds, the estimator (statistic) is called sufficient. It can be shown that no other estimator can contribute knowledge to λ that is not already contained in S if this condition is fulfilled.

Example

For a sample drawn from a Poisson distribution, we had

$$\ln L(\lambda) = -N\lambda + \sum_{j=1}^{N} (k^{(j)} \ln \lambda - \ln k^{(j)}!), \quad \text{i.e.,}$$

$$\frac{d \ln L}{d\lambda} = -N + \sum_{j=1}^{N} \frac{k^{(j)}}{\lambda} = -N + \frac{N\bar{\mathrm{k}}}{\lambda} = \frac{N}{\lambda}\left(\bar{\mathrm{k}} - \lambda\right), \text{ which is of the form}$$

$$\frac{d \ln L}{d\lambda} = A(\lambda)(\mathrm{S} - E(\mathrm{S}))$$

when the estimator for the mean is the sample (=arithmetic) mean. Thus, the sample mean $\bar{\mathrm{k}}$ for a Poisson distributed sample is a minimum variance estimator, and, since it is unbiased, its variance is given by

$$Var(\bar{\mathrm{k}}) = \frac{1}{A(\lambda)} = \frac{\lambda}{N}$$

which we have already derived previously.

Bias for ML estimators

Example: ML estimators for the mean and the standard deviation of a normally distributed sample.

For a normally distributed sample, the likelihood equations for the ML estimators

$$\left.\frac{\partial \ln L}{\partial \mu}\right|_{\tilde{\mu},\tilde{\sigma}} = 0 \quad \Rightarrow \sum_{j=1}^{N}(x^{(j)} - \tilde{\mu}) = 0$$

$$\left.\frac{\partial \ln L}{\partial \sigma}\right|_{\tilde{\mu},\tilde{\sigma}} = 0 \quad \Rightarrow \left(\sum_{j=1}^{N}\frac{(x^{(j)} - \tilde{\mu})^2}{\tilde{\sigma}^3}\right) - \frac{N}{\tilde{\sigma}} = 0$$

result in the well known ML estimates for $\tilde{\mu} = \frac{1}{N}\sum_{j=1}^{N} x^{(j)}$, the sample mean, and

$$\tilde{\sigma} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x^{(j)} - \tilde{\mu})^2} = \text{s'}$$

From our previous considerations, we know that the latter estimator is biased, by a factor of $\sqrt{(N-1)/N}$.

Now, calculate the ML estimator for the variance, $S(\sigma^2)$ $\Rightarrow$ 2nd ML equation

$$\left.\frac{\partial \ln L}{\partial \sigma^2}\right|_{\tilde{\mu},\sigma^2} = 0 \quad \Rightarrow \left(\sum_{j=1}^{N}\frac{(x^{(j)} - \tilde{\mu})^2}{2\left(\sigma^2\right)^2}\right) - \frac{N}{2\sigma^2} = 0 \quad \Rightarrow \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x^{(j)} - \tilde{\mu})^2 = \tilde{\sigma}^2 \text{!!!}$$

ML: consistency, bias and invariance

- ML estimators are usually consistent (see below)
- but ML estimators are also biased (in most cases)
- bias becomes small for large samples (see below)
- bias is the price one has to pay for an advantageous property of ML estimators, namely that they are invariant under parameter transformations (see example above,
  ML estimator for σ^2 = (ML estimator for σ)^2

If the maximum of L occurs at some particular value $\tilde{\lambda}$, $\left.\frac{dL}{d\lambda}\right|_{\tilde{\lambda}} = 0,$

then the maximum of L w.r.t an alternative parameter $\alpha = f(\lambda)$

occurs at $\tilde{\alpha} = f(\tilde{\lambda})$, since $\left.\frac{dL}{d\alpha}\right|_{\alpha(\tilde{\lambda})} = 0$ because of $0 = \left.\frac{dL}{d\lambda}\right|_{\tilde{\lambda}} = \left.\frac{dL}{d\alpha}\frac{d\alpha}{d\lambda}\right|_{\tilde{\lambda}}$

- Thus, we have generally,

$$S_{ML}(f(\lambda) = \tilde{f} = f(\tilde{\lambda})$$

- Note: other, non-ML estimators preserve the difference
- Invariance under transformation is incompatible with lack of bias

Asymptotic properties of L and ML estimators

In the following, we investigate important properties for L and ML-estimators in the case of big samples, i.e., $N \to \infty$. The estimator $\tilde{\lambda}$ was found from the solution of the likelihood equation(s),

$$\left.\frac{d \ln L}{d\lambda}\right|_{\tilde{\lambda}} = \sum_{j=1}^{N} \frac{d}{d\lambda} \ln\left[f(x^{(j)};\lambda)\right]_{\tilde{\lambda}} = \sum_{j=1}^{N} \left(\frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)} \right)_{\tilde{\lambda}} = 0.$$

Let's develop the derivative of the likelihood function into a Taylor series about $\tilde{\lambda}$,

$$\frac{d \ln L}{d\lambda} = \left.\frac{d \ln L}{d\lambda}\right|_{\tilde{\lambda}} + (\lambda - \tilde{\lambda}) \left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}} + \ldots = (\lambda - \tilde{\lambda}) \left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}} + \ldots \quad \text{(first term vanishes, see above)}$$

$$\left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}} = \sum_{j=1}^{N} \left(\frac{f'(x^{(j)};\lambda)}{f(x^{(j)};\lambda)} \right)'_{\tilde{\lambda}}, \qquad \text{which has the form of a sample mean (times N).}$$

For large N, sample means can be replaced by the expectation value (since their variance decreases with $1/N$), and sums of random variables can be replaced by their expectation value, since their relative uncertainty, $\sigma(\Sigma t_i)/E(\Sigma t_i)$, decreases with $1/\sqrt{N}\,\sigma(t)/E(t)$

$$\left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}} \xrightarrow{N\to\infty} E\left\{ \left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}} \right\} = -E\left\{ \left(\left.\frac{d \ln L}{d\lambda}\right|_{\tilde{\lambda}} \right)^2 \right\} = -I(\tilde{\lambda}) =: -a$$

$$\left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}} \xrightarrow{N\to\infty} E\left\{\left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}}\right\} = -E\left\{\left(\left.\frac{d \ln L}{d\lambda}\right|_{\tilde{\lambda}}\right)^2\right\} = -I(\tilde{\lambda}) =: -a$$

NOTE

1. To be consistent with the fact that the likelihood (the sum) is evaluated at $\tilde{\lambda}$, also its expectation value needs to account for a distribution with $\tilde{\lambda}$,

$$E\left\{g\left(x^{(1)}...x^{(N)};\tilde{\lambda}\right)\right\} = \int g\left(x^{(1)}...x^{(N)};\tilde{\lambda}\right) f(x^{(1)};\tilde{\lambda})\, f(x^{(2)};\tilde{\lambda})\, f(x^{(N)};\tilde{\lambda}) dx^{(1)} dx^{(2)} \cdots dx^{(N)} = E\left\{g\left(x^{(1)}...x^{(N)};\lambda\right)\right\}_{\tilde{\lambda}},$$

i.e., $E\left\{\left.\frac{d^2 \ln L}{d\lambda^2}\right|_{\tilde{\lambda}}\right\} = E\left\{\frac{d^2 \ln L}{d\lambda^2}\right\}_{\tilde{\lambda}}$ and $E\left\{\left(\left.\frac{d \ln L}{d\lambda}\right|_{\tilde{\lambda}}\right)^2\right\} = E\left\{\left(\frac{d \ln L}{d\lambda}\right)^2\right\}_{\tilde{\lambda}}$

2. When calculating $E\left\{\left(\left.\frac{d \ln L}{d\lambda}\right|_{\tilde{\lambda}}\right)^2\right\}$, the argument is NOT generally zero

since $\left.\frac{d \ln L}{d\lambda}\right|_{\tilde{\lambda}} = 0$ results from the specific, "observed" sample-values $x^{(1)}, x^{(2)}, \cdots x^{(N)}$, and the corresponding estimator derived from these values,

whereas for the expectation value we need to integrate over all possible variates, $x^{(1)}, x^{(2)}, \cdots x^{(N)}$.

In other words, we have replaced the 2nd derivative of the likelihood function (which depends on the specific $x^{(1)}...x^{(N)}$) by a (negative) number $-a$ $(I(\tilde{\lambda}) \geq 0)$, which depends "only" on the pdf f and the estimator $\tilde{\lambda}$. (Note that only here we require the asymptotic limit $N \to \infty$).

To first order, we thus have for the Taylor expansion $\frac{d \ln L}{d\lambda} = -a(\lambda - \tilde{\lambda})$, which by integration results in

$$\ln L = -\frac{a}{2}(\lambda - \tilde{\lambda})^2 + \text{const} \;\Rightarrow\; L(\lambda) = k \exp\left\{-\frac{a}{2}(\lambda - \tilde{\lambda})^2\right\} \quad \text{[with constant } k\text{]}$$

Thus, for large N the likelihood function $L(\lambda)$ has the form of a normal distribution, with mean $\tilde{\lambda}$ and variance $1/a$ (standard deviation $\sqrt{1/a}$).

Now, since $\tilde{\lambda} = \mathrm{S}$, we can rewrite the above equation as $\frac{d \ln L}{d\lambda} = a(\mathrm{S} - \lambda)$. With $E(d \ln L / d\lambda) = 0$ generally,

$$E\left(\frac{d \ln L}{d\lambda}\right) = a(E(\mathrm{S}) - \lambda) = 0 \quad \Rightarrow \quad E(\mathrm{S}) = E(\tilde{\lambda}) = \lambda,$$ the estimator is asymptotically unbiased, and

$$\frac{d \ln L}{d\lambda} = a(\mathrm{S} - \lambda) = a(\mathrm{S} - E(\mathrm{S}))$$

Comparing further with the expression (7.14)

$\frac{d \ln L}{d\lambda} = A(\lambda)(\mathrm{S} - E(\mathrm{S}))$ valid for minimum variance estimators,

$S = \tilde{\lambda}$ is a minimum variance estimator for large N, and, since it is asymptotically unbiased,

$$Var(\mathrm{S}) = Var(\tilde{\lambda}) \to \frac{1}{a}.$$

Thus, we conclude that the ML estimator $\tilde{\lambda}$ has the following properties

i) it is asymptotically unbiased, $E(\mathrm{S}) = E(\tilde{\lambda}) \to \lambda$.

ii) it is aymptotically efficient (i.e., a minimum variance estimator), with

$$Var(\tilde{\lambda}) \to \frac{1}{a} = \frac{1}{I(\tilde{\lambda})} =$$

$$= -\frac{1}{E\left\{\frac{d^2 \ln L}{d\lambda^2}\right\}_{\tilde{\lambda}}} = \frac{1}{E\left\{\left(\frac{d \ln L}{d\lambda}\right)^2\right\}_{\tilde{\lambda}}} = \frac{1}{NE\left\{\left(\frac{f'(\mathrm{x},\lambda)}{f(\mathrm{x},\lambda)}\right)^2\right\}_{\tilde{\lambda}}} = -\frac{1}{NE\left\{\left(\frac{f'(\mathrm{x},\lambda)}{f(\mathrm{x},\lambda)}\right)'\right\}_{\tilde{\lambda}}} \qquad \text{(Eq. 7.17)}$$

Note: The asymptotic variance corresponds to the MVB (Eq. 7.13), evaluated at $\lambda = \tilde{\lambda}$ and with $B' = 0$!

iii) since $Var(\tilde{\lambda}) \to 0$ for $N \to \infty$ and $\tilde{\lambda}$ is asymptotically unbiased, $\tilde{\lambda}$ is also consistent.

iv) the likelihood function is asymptotically normal, and the log-likelihood is a parabola.

Remember: The expression for $Var(\tilde{\lambda})$,

$$Var(\tilde{\lambda}) = -\frac{1}{E\left\{\frac{d^2 \ln L}{d\lambda^2}\right\}_{\tilde{\lambda}}}$$

is not only (generally) valid for large N, but also in those cases where we know (from explicit calculation) that the ML estimator is unbiased and efficient, e.g., for the sample mean from a Poisson distribution (see previous example).

Errors on the ML estimators

As we have argued at the begin of this section, the likelihood function

$$L(\lambda \mid x^{(1)} \ldots x^{(N)})$$

can be transformed into a conventional pdf for λ, if the prior $h(\lambda)$ [corresponding to the marginal distribution $\int f(x,\lambda)dx$] is known and the distribution can be normalized,

$$f(\lambda \mid x^{(j)}) = \frac{f(x^{(j)} \mid \lambda)h(\lambda)}{\int f(x^{(j)} \mid \lambda)h(\lambda)d\lambda} = \frac{L(\lambda \mid x^{(j)})h(\lambda)}{\int L(\lambda \mid x^{(j)})h(\lambda)d\lambda}.$$

Though the prior is usually unknown, the (normalized) likelihood function corresponds directly to a pdf for λ if the prior is constant (all parameter values equally probable),

$$L(\lambda) \xrightarrow{\text{constant prior, } L \text{ normalized}} f(\lambda)$$

which is assumed (and justified) in most cases.

From the previous considerations, we know the the likelihood function is asymptotically normal, with mean $\tilde{\lambda}$ [the ML estimator obtained from the solution of the likelihood equation(s)] and variance $1/I(\tilde{\lambda})$ [the information of the sample w.r.t. $\tilde{\lambda}$].

Thus, the standard approach to define the errors of the estimated parameter

$\lambda = \tilde{\lambda} \pm \Delta\lambda := \tilde{\lambda} \pm \sigma(\tilde{\lambda})$

refers to the usual confidence intervals for a Gaussian:

in the 1-σ interval, the probability has decreased by $e^{-1/2} \Rightarrow \ln L$ has decreased by 0.5

in the 2-σ interval, the probability has decreased by $e^{-4/2} \Rightarrow \ln L$ has decreased by 2.0

in the 3-σ interval, the probability has decreased by $e^{-9/2} \Rightarrow \ln L$ has decreased by 4.5

with respect to the maximum, i.e, $L(\tilde{\lambda})$ or $\ln L(\tilde{\lambda})$, respectively (independent of the dimension of λ)

When the large N limit has not been reached, L will not be a Gaussian, and $\ln L$ not a parabola. Presumably, however, there will be an alternative parameter λ' which transforms the shape to a parabola. For this parameter then, the corresponding 1-σ limits can be derived from the values of λ' where $\ln L(\lambda')$ has decreased by 0.5 below its maximum, and one can calculate the corresponding limits w.r.t. λ. These, by the invariance property of the likelihood, are just those values of λ where $\ln L(\lambda)$ has decreased by 0.5. Thus, we can completely skip the transformation to λ' and read off the n-σ limits from $\ln L(\lambda)$, both for finite as well as for large N.

For finite N, the L-distribution is asymmetric about $\tilde{\lambda}$, such that asymmtric errors have to be quoted, in the form of, e.g., $\lambda = 1.51^{+0.27}_{-0.15}$.

Note: for asymmetric distributions, the 2-σ limits are *not* the double of the 1-σ limits!

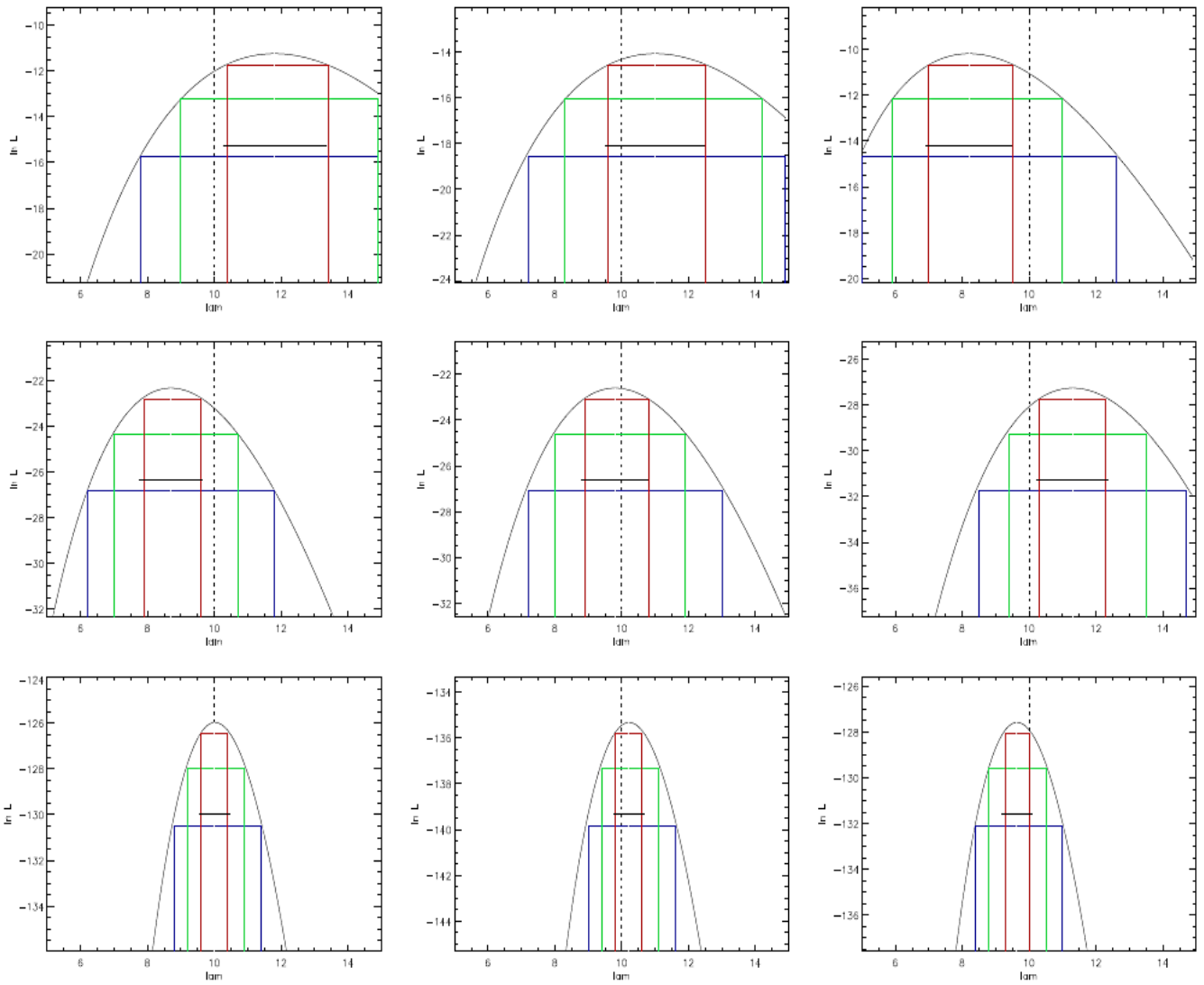
Examples

Log-likelihood functions ln L(λ) for the mean λ, calculated for 3 different samples with sizes N=5 (top),15 (middle) and 50 (bottom), drawn from a Poisson distribution with λ=10 (dashed).

The 1-,2- and 3-σ limits of the ML-estimator (corresponding to ln L values which are 0.5, 2.0 and 4.5 smaller than the maximum) are indicated in red, green and blue.

The corresponding estimate for the mean (including error), as calculated from the *sample mean,* is indicated by the horizontal black line.

Note that the shape of ln L becomes more and more symmetric (and converges to a parabola) when N increases, in parallel with decreasing errors on the estimated parameter.

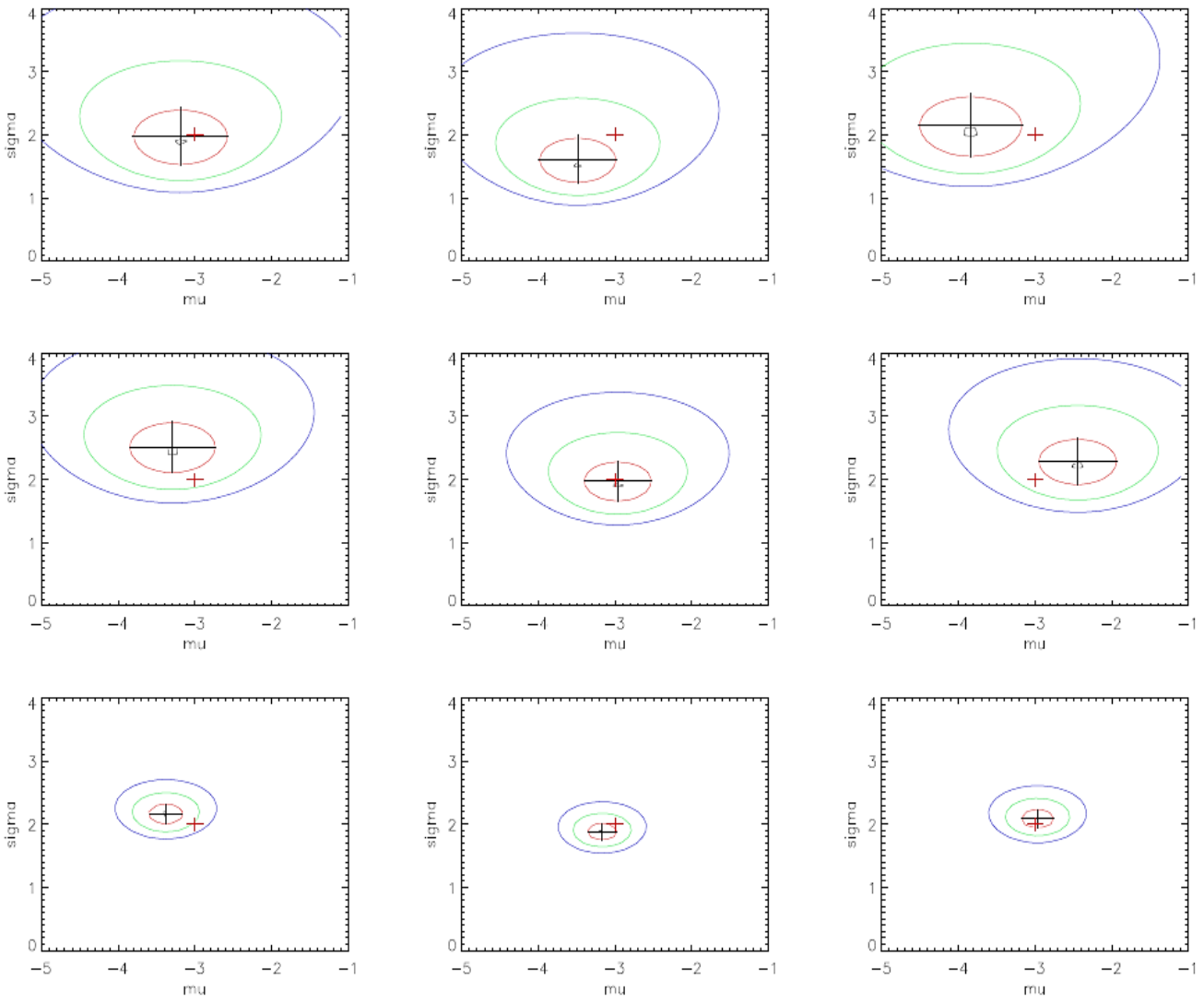


Iso-contours of the log-likelihood ln L(μ,σ) for the mean and standard deviation of 3 different samples with sizes N=10 (top), 20 (middle) and 100 (bottom), drawn from a normal distribution with μ=-3 and σ=2 (red plus).

The iso-contours denote the 1-, 2- and 3-σconfidence regions of the ML-estimators (corres-ponding to ln L values which are 0.5, 2.0 and 4.5 smaller that the maximum), indicated in red, green and blue.

The corresponding estimates for the mean and standard deviation (including errors), as calculated from the *sample mean* and variance, are indicated by the black crosses. The small difference between ML- and sample estimates is related to the bias in the ML-values.

Note that the shape of the ln L iso-contours becomes more and more symmetric (and converges to an error ellipse) when N increases.



Several parameters - covariances

To obtain a set of parameters λ_i, $i=1,M$, the set of likelihood equations has to be solved simultaneously. In the large N limit, the likelihood function becomes a Gaussian, and the variances of the ML-estimators are

$$Var(\tilde{\lambda}_i) = -\frac{1}{E\left\{\frac{\partial^2 \ln L}{\partial \lambda_i^{\,2}}\right\}_{\tilde{\lambda}_i}},\ i=1,M$$

From a Taylor expansion of the likelihood function, one then obtains the covariance matrix regarding the estimators

$$\mathbf{C}=\mathbf{B}^{-1}, \mathbf{B}=\left(\begin{array}{cccc} -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_1^{\,2}}\right\} & -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_1 \partial \lambda_2}\right\} & \ldots & -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_1 \partial \lambda_M}\right\} \\ -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_2 \partial \lambda_1}\right\} & -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_2^{\,2}}\right\} & \ldots & -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_2 \partial \lambda_M}\right\} \\ \ldots & \ldots & \ldots & \ldots \\ -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_M \partial \lambda_1}\right\} & -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_M \partial \lambda_2}\right\} & \ldots & -E\left\{\frac{\partial^2 \ln L}{\partial \lambda_M^{\,2}}\right\} \end{array}\right)_{\boldsymbol{\lambda}=\tilde{\boldsymbol{\lambda}}}$$

Fisher-matrix

i.e., $\mathbf{C}_{ij} = \left(\mathbf{B}^{-1}\right)_{ij}$, with $Var(\tilde{\lambda}_i) = \sigma^2(\tilde{\lambda}_i) = \mathbf{C}_{ii}$ and $\mathrm{cov}(\tilde{\lambda}_i, \tilde{\lambda}_j) = \mathbf{C}_{ij}$

Example

Calculate the variances and covariance of the mean and σ estimates for a Gaussian

$$\ln L(\mu,\sigma) = -\sum_{j=1}^{N} \frac{(x^{(j)}-\mu)^2}{2\sigma^2} - N\ln\left(\sigma\sqrt{2\pi}\right)$$

$$\tilde{\mu} = \overline{\mathrm{x}}, \quad \tilde{\sigma} = \sqrt{\overline{(\mathrm{x}-\tilde{\mu})^2}} = \sqrt{\overline{\mathrm{x}^2} - \tilde{\mu}^2} \quad (=\mathrm{s}', \text{ see previous example})$$

$$-E\left\{\frac{\partial^2 \ln L}{\partial \mu^2}\right\}_{\tilde{\mu},\tilde{\sigma}} = \frac{N}{\tilde{\sigma}^2};$$

$$-E\left\{\frac{\partial^2 \ln L}{\partial \sigma^2}\right\}_{\tilde{\mu},\tilde{\sigma}} = -E\left\{-3\sum_{j=1}^{N}\frac{(x^{(j)}-\mu)^2}{\sigma^4} + \frac{N}{\sigma^2}\right\}_{\tilde{\mu},\tilde{\sigma}} = -\left\{\frac{-3N\sigma^2}{\sigma^4} + \frac{N}{\sigma^2}\right\}_{\tilde{\mu},\tilde{\sigma}} = \frac{2N}{\tilde{\sigma}^2}$$

$$-E\left\{\frac{\partial^2 \ln L}{\partial \mu \partial \sigma}\right\}_{\tilde{\mu},\tilde{\sigma}} = -E\left\{-2\sum_{j=1}^{N}\frac{(x^{(j)}-\mu)}{\sigma^3}\right\}_{\tilde{\mu},\tilde{\sigma}} = 0$$

Thus, $\mathbf{B} = \begin{pmatrix} \frac{N}{\tilde{\sigma}^2} & 0 \\ 0 & \frac{2N}{\tilde{\sigma}^2} \end{pmatrix}$, and we obtain the estimates for the errors,

$$\mathrm{Var}(\tilde{\mu}) = \frac{\tilde{\sigma}^2}{N}, \ \mathrm{Var}(\tilde{\sigma}) = \frac{\tilde{\sigma}^2}{2N} \text{ and } \mathrm{cov}(\tilde{\mu},\tilde{\sigma}) = 0.$$

Estimates of mean and standard deviation are uncorrelated, which is also obvious from the position of the covariance ellipse in the previous figure.

Final comments on ML estimators

- For large samples, $\tilde{\lambda}$ is a consistent, unbiased and efficient estimator, so the best thing one can obtain
- for smaller samples, however, this is not true: in these cases, ML estimators *are* (often) biased
- advantages of ML:
  - easy to calculate estimators and corresponding errors
  - invariance under parameter transformation
  - very suitable if several parameters to be estimated in parallel
- major disadvantage:
  - one has to know the parent distribution. If the assumption on $f(x;\lambda)$ is wrong, there is no way of telling this from the results, since there is no quality factor or goodness of fit number
- minor problem:
  - to interpret $L(\lambda)$ as a pdf (which is required to estimate the errors of the estimates), one has to assume a uniform (constant) prior for the distribution of λ.
  - for non-uniform priors, one can use the MCMC method (Sect. 9) to obtain the (distribution of the) posteriors and thus the errors on the parameters
- Besides the basic estimators (sample mean and sample variance) and the ML estimators, there are other estimators as well, e.g., the method of moments (see literature) and the chi-squared minimization (next section).

8. Least squares

Suppose a data sample of pairs $(x^{(j)}, y^{(j)})$, where the $x^{(j)}$ are precisely known and the $y^{(j)}$ have been measured, with individual errors σ_j. A theory predicts that y should be a function $\varphi(x, \lambda)$, where the parameter(s) λ need to be estimated. The "ideal" y – values are assumed to be smeared out by measurement errors alone, and we further assume these errors to be normally distributed (invoking the CLT). Thus, the pdf to measure a certain $y^{(j)}$ for a given $x^{(j)}$ is given by

$$f(y^{(j)} \mid \lambda) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\left(\frac{\left[y^{(j)} - \varphi(x^{(j)}, \lambda) \right]^2}{2\sigma_j^2} \right) \right\} := L(\lambda \mid y^{(j)})$$

The combined log likelihood function is then

$$\ln L(\lambda) = -\frac{1}{2} \sum_{j=1}^{N} \left(\frac{y^{(j)} - \varphi(x^{(j)}, \lambda)}{\sigma_j} \right)^2 - \sum_{j=1}^{N} \ln\left(\sigma_j \sqrt{2\pi} \right)$$

and in order to maximize $\ln L(\lambda)$ one has to minimize the quantity

$$\chi^2 = \sum_{j=1}^{N} \left(\frac{y^{(j)} - \varphi(x^{(j)}, \lambda)}{\sigma_j} \right)^2,$$

i.e., one has to minimize the χ^2 (=χ^2 – minimization) or, in other words, the (weighted) sum of the squared differences (= least squares minimization).

- Though the method of least squares can be ‘derived’ from the ML principle, many people regard this method (and the corresponding estimator) to exist in its own rights, being obviously sensible and empirically tested.

- The predicted values are adjusted in such a way as to be close to the measurements; by squaring the differences, larger effort is spent on removing the larger deviations.

- Firstly published by Legendre in 1805 and by Gauss in 1809.The term “least squares” is from Legendre’s term, “moindres carrés”. However, Gauss claimed that he had known the method since 1795. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun.

- The minimization problem can be condensed in the form

$$\left.\frac{d\chi^2}{d\lambda}\right|_{\tilde{\lambda}} = 0 = \left(\sum_{j=1}^{N} \frac{1}{\sigma_j^2} \frac{d\varphi(x^{(j)};\lambda)}{d\lambda} \left[y_j - \varphi(x^{(j)};\lambda) \right] \right)_{\tilde{\lambda}}$$

- Since the estimator $\tilde{\lambda}$ is a function of the y_i, and the corresponding errors are known, the laws of error propagation can be used to calculate the error of the estimator (if we consider the least square method as being independent of the ML principle, we don’t have to worry about uniform priors here)

- If there are M parameter to estimate, a set of M simultaneous equations has to be solved.

Fitting to a straight line

- most common application of least square fitting
- $y=\varphi(x;m,b)=mx+b$
  with slope m and intercept (offset) b $=\varphi(x=0)$
- often denoted by “linear regression”
  - but note the difference between a straight line fit and regression:
  - regression is a statistical term related to “non perfect” laws, resulting in the formulation of a “trend” or correlation
  - firstly introduced by Francis Galton (1885), a cousin of Charles Darwin, to describe the biological phenomenon that the heights of descendants of tall ancestors tend to “regress down” towards a normal average
  - in the context considered here, we deal with the problem of parameter estimation, since we assume the law to be perfect (i.e., better measurements should lead to data very close or indistinguishable from a straight line)

Identical σ

Let's first assume that all measurement errors can be described by the same σ,

$\sigma := \sigma_j, \quad j = 1, N$

(which is frequently done, particularly if the errors are unknown). In this case, the "original" least squares problem needs to be solved, namely the sum

$$\sum_{j=1}^{N} \left(y_j - m x_j - b \right)^2$$

is to be minimized w.r.t. *m* and *b*. (Here and in the following, we abbreviate $y^{(j)}$ by y_j etc.).

Differentiating w.r.t. *m* and setting to zero yields

$$\sum_{j=1}^{N} -2 x_j \left(y_j - \tilde{m} x_j - \tilde{b} \right) = 0,$$

or, dividing by N and in terms of sample means $\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j, \bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j, \overline{xy} = \frac{1}{N}\sum_{j=1}^{N} x_j y_j$ etc.,

$$\overline{xy} - \tilde{m}\overline{x^2} - \tilde{b}\bar{x} = 0$$

Likewise, differentiating w.r.t. to *b*,

$$\sum_{j=1}^{N} -2 \left(y_j - \tilde{m} x_j - \tilde{b} \right) = 0$$

$$\bar{y} - \tilde{m}\bar{x} - \tilde{b} = 0.$$

Combining both equations, we obtain

$$\tilde{m} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \qquad \tilde{b} = \bar{y} - \tilde{m}\bar{x} = \frac{\overline{x^2}\bar{y} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2}.$$

The first expression for $\tilde{b}$ shows that the line goes through the center of gravity, $(\bar{x},\bar{y})$, since $y(\bar{x}) = \tilde{m}\bar{x} + \tilde{b} = \bar{y}$.

The errors on the parameters follow from the law of error propagation. Writing the expression for $\tilde{m}$ in the alternative way

$$\tilde{m} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \sum_j \frac{x_j - \bar{x}}{N\left(\overline{x^2} - \bar{x}^2\right)} y_j,$$

we immediately find (for identical errors in y_j)

$$Var(\tilde{m}) = \sum_j \left(\frac{x_j - \bar{x}}{N\left(\overline{x^2} - \bar{x}^2\right)} \right)^2 \sigma^2 = \frac{\sigma^2}{N\left(\overline{x^2} - \bar{x}^2\right)^2} \underbrace{\frac{1}{N}\sum_j \left(x_j - \bar{x}\right)^2}_{\overline{x^2} - \bar{x}^2} \Rightarrow$$

$$Var(\tilde{m}) = \frac{\sigma^2}{N\left(\overline{x^2} - \bar{x}^2\right)}$$

In case we have no info on σ^2, we approximate this quantity from the resulting χ^2 (see below)

Likewise

$$Var(\tilde{b}) = \sum_j \left(\frac{\overline{x^2} - \bar{x} x_j}{N\left(\overline{x^2} - \bar{x}^2\right)} \right)^2 \sigma^2 = \frac{\sigma^2 \overline{x^2}}{N\left(\overline{x^2} - \bar{x}^2\right)} = \overline{x^2} Var(\tilde{m}).$$

The general law of error propagation was

$$\mathrm{cov}(f_k, f_l) = \sum_{i,j=1}^{n} \left(\frac{\partial f_k}{\partial y_i} \right) \left(\frac{\partial f_l}{\partial y_j} \right) \mathrm{cov}(y_i, y_j).$$

Application to our case of *independent* y_j (i.e., $\mathrm{cov}(y_i, y_j) = \delta_{ij}\sigma^2$), we obtain

$$\mathrm{cov}(\tilde{m}, \tilde{b}) = \sum_{j=1}^{N} \left(\frac{\partial \tilde{m}}{\partial y_j} \right) \left(\frac{\partial \tilde{b}}{\partial y_j} \right) \sigma^2 = \sum_j \frac{\left(x_j - \bar{x}\right)\left(\overline{x^2} - \bar{x} x_j\right)}{N^2\left(\overline{x^2} - \bar{x}^2\right)^2} \sigma^2 = \frac{-\bar{x}\sigma^2}{N\left(\overline{x^2} - \bar{x}^2\right)} = -\bar{x} Var(\tilde{m}).$$

Before we give further comments on the error of the fit parameters, let's calculate the resulting (minimized) χ^2, which we will need later on to assess the fit quality.

$$\chi^2_{\min} = \sum_{j=1}^{N} \frac{\left(y_j - \tilde{m} x_j - \tilde{b}\right)^2}{\sigma^2} = N \frac{Var(y)}{\sigma^2}\left(1 - \rho^2_{x,y}\right),$$ if we use the abbreviations

$$Var(y) = \overline{y^2} - \bar{y}^2 \quad \text{and} \quad \rho^2_{x,y} = \frac{\left(\overline{xy} - \bar{x}\bar{y}\right)^2}{\left(\overline{x^2} - \bar{x}^2\right)\left(\overline{y^2} - \bar{y}^2\right)}.$$

Note that as long as the scatter in the y_j – values is on the order of σ, $\chi^2_{\min}$ is $O(N)$.

Variances and correlation

Summarizing the previous results, the error matrix for $(\tilde{m},\tilde{b})$ is given by

$$\mathbf{C} = \begin{pmatrix} Var(\tilde{m}) & \mathrm{cov}(\tilde{m},\tilde{b}) \\ \mathrm{cov}(\tilde{m},\tilde{b}) & Var(\tilde{b}) \end{pmatrix} = Var(\tilde{m}) \begin{pmatrix} 1 & -\bar{x} \\ -\bar{x} & \overline{x^2} \end{pmatrix}, \text{ with } Var(\tilde{m}) = \frac{\sigma^2}{N\left(\overline{x^2} - \bar{x}^2\right)},$$

and the correlation coefficient is

$$\rho_{\tilde{m},\tilde{b}} = \frac{\mathrm{cov}(\tilde{m},\tilde{b})}{\sqrt{Var(\tilde{m})Var(\tilde{b})}} = \frac{-\bar{x}}{\sqrt{\overline{x^2}}}$$

Obviously, the errors on the parameters and their correlation depend exclusively on the abscissae of the measurements. The larger the spread of these values, the smaller the variances and the correlation. This is reasonable, since measurements which cluster around a certain *x*-value allow for a lot of freedom in slope and intercept.

Interestingly, however, the covariance and the correlation is proportional to the sample mean of the x_j-values, and $\rho_{\tilde{m},\tilde{b}} < 0,\ > 0$ and $= 0$ for $\bar{x} > 0,\ <0$ and $= 0$ (see the example in Chap. 6).

In so far, the correlation can be simply avoided if one uses shifted *x*-coordinates $\hat{x} = x - \bar{x}$, and fits $y = m(x - \bar{x}) + b = m\hat{x} + b$ instead of $y = mx + b$.

With these new coordinates, $Var(\tilde{m}) = \sigma^2/(N\overline{\hat{x}^2}),\ Var(\tilde{b}) = \sigma^2/N,$ and $\rho_{\tilde{m},\tilde{b}} = 0.$

Inter-/Extrapolation – individual σ_j

If one inter-/extrapolates a value of Y for a given value X, in the conventional coordinates we find

$$Y(X) = \tilde{m}X + \tilde{b}, \qquad Var(Y) = X^2 Var(\tilde{m}) + Var(\tilde{b}) + 2X\,\mathrm{cov}(\tilde{m},\tilde{b}) \qquad \text{(see Chap. 6)}$$

whereas in the new coordinates $\hat{x} = x - \overline{\mathrm{x}}$ (with $\overline{\hat{\mathrm{x}}} = 0$ and $\mathrm{cov} = 0$) we obtain, after inserting the corresponding values for $Var(\tilde{m})$ and $Var(\tilde{b})$,

$$Var(Y) = \left(X - \overline{\mathrm{x}}\right)^2 \frac{\sigma^2}{N\overline{\hat{\mathrm{x}}^2}} + \frac{\sigma^2}{N} = \frac{\sigma^2}{N}\left(1 + \frac{\left(X - \overline{\mathrm{x}}\right)^2}{\overline{\hat{\mathrm{x}}^2}}\right) = \frac{\sigma^2}{N}\left(1 + \frac{\left(X - \overline{\mathrm{x}}\right)^2}{\overline{\mathrm{x}^2} - \overline{\mathrm{x}}^2}\right)$$

which of course has the same numerical value as the expression above.
Large errors are induced if X is far from $\overline{\mathrm{x}}$, whereas for $X \approx \overline{\mathrm{x}}$ (i.e., close to the center of gravity), the variance is reduced by a factor of 1/N compared to the intrinsic scatter.

Finally, we assume that the measurements have individual errors σ_j. Now, χ^2 has to be minimized, and the individual errors have to be accounted for. All derived equations remain valid, if we replace the sample means by the corresponding, weighted means,
and the quantity σ^2 occuring in the error matrix likewise, e.g.,

$$\overline{\mathrm{y}} = \frac{\sum_j y_j}{N} \rightarrow \frac{\sum_j \frac{y_j}{\sigma_j^2}}{\sum_j \frac{1}{\sigma_j^2}}, \qquad \sigma^2 \rightarrow \overline{\sigma^2} = \frac{\sum_j \frac{\sigma_j^2}{\sigma_j^2}}{\sum_j \frac{1}{\sigma_j^2}} = \frac{N}{\sum_j \frac{1}{\sigma_j^2}}$$

Example

Four slightly different samples (figures next page), all with $N = 7$

sample 1

x_j	1	2	3	4	5	6	7
y_j	1.2	1.9	3.1	4.2	2.0	6.5	6.8
σ_j	0.5	0.5	0.5	0.5	0.5	0.5	0.5

sample 2

x_j	1	2	3	4	5	6	7
y_j	1.2	1.9	3.1	4.2	2.0	6.5	6.8
σ_j	0.5	0.5	0.5	0.5	2.0	0.5	0.5

sample 3

x_j	–9.5	1	2	3	4	6	7
y_j	–5	1.2	1.9	3.1	4.2	6.5	6.8
σ_j	0.5	0.5	0.5	0.5	0.5	0.5	0.5

sample 4

x_j	–9.5	1	2	3	4	6	7
y_j	–5	1.2	1.9	3.1	4.2	6.5	6.8
σ_j	2.0	0.5	0.5	0.5	0.5	0.5	0.5

Sample 1/2 and 3/4 differ in one point at either $x = 5$ or $x = -9.5$. All following results have been calculated from the previous expressions.

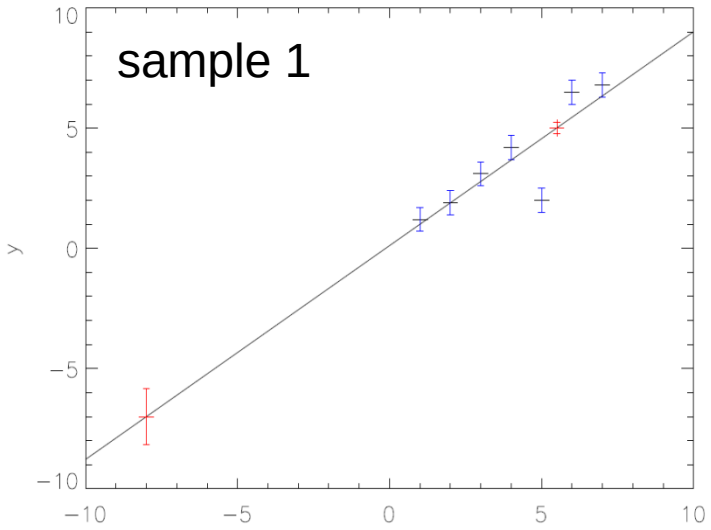
Results

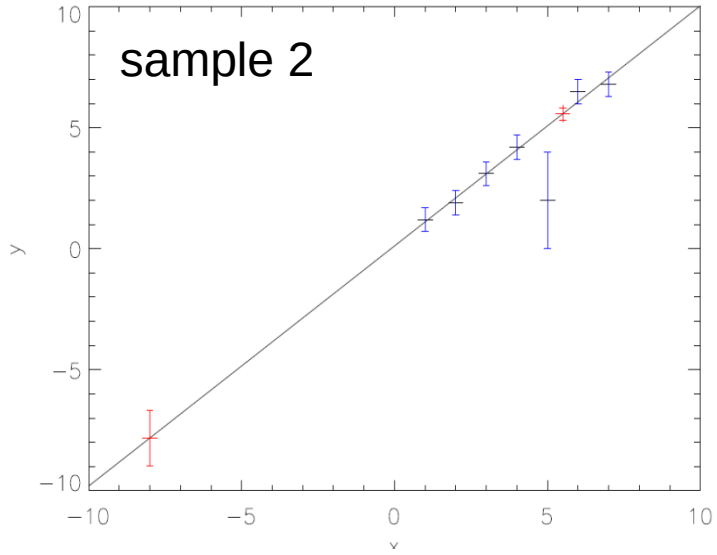
sample	$\tilde{m}$	$\tilde{b}$	$\mathrm{cov}(\tilde{m},\tilde{b})$	$\rho_{\tilde{m},\tilde{b}}$	$\chi^2_{\min}$	$Y(X=-8)$	$Y(X=5.5)$
1	0.889 ± 0.094	0.114 ± 0.423	-0.0357	-0.894	33.16	-7.00 ± 1.15	5.01 ± 0.24
2	0.992 ± 0.096	0.114 ± 0.423	-0.0357	-0.877	3.63	-7.82 ± 1.16	5.57 ± 0.26
3	0.721 ± 0.037	1.282 ± 0.202	-0.0027	-0.356	11.13	-4.48 ± 0.42	5.24 ± 0.23
4	0.905 ± 0.081	0.513 ± 0.363	-0.0244	-0.829	4.61	-6.73 ± 0.97	5.49 ± 0.25

The relevance of the obtained $\chi^2_{\min}$ will be discussed later on. $Y(X)$ are extra-/interpolated values.

Examples

red data: inter-/extrapolated values with error-bars



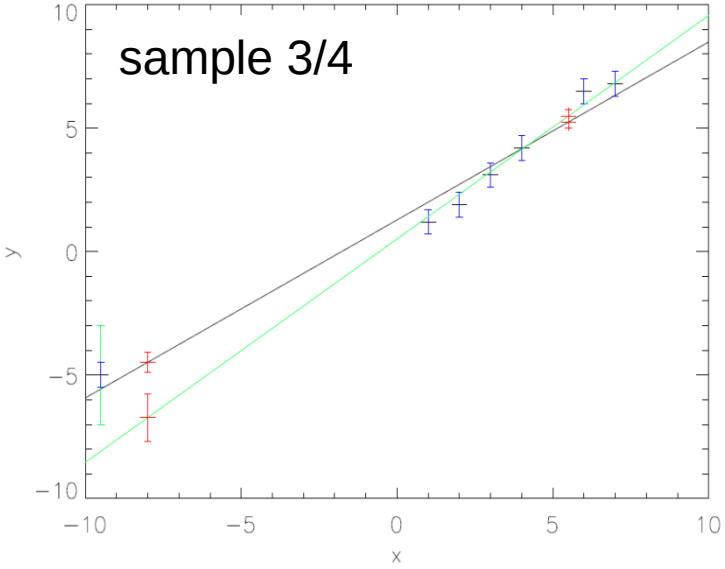


sample 1 vs. 2:

a) Note the influence of the measurement error σ for y(x=5)

b) extrapolated values have a larger error than values close to the center of gravity

c) Though the fit to sample 1 looks reasonable, the large χ^2 makes it rather unlikely (see later on)



sample 3/4 vs. 1/2

Note the strong influence of a data point far away from the center of gravity (and the influence of the corresponding error)

Binned data

- least squares fitting also possible for binned data (histograms)

- assume data can be binned (without significant loss of information)

- suppose there are N events, and the pdf is $f(x;\lambda)$

- the events are sorted into N_B bins, centered on point x_j and with width W_j (often uniform)

- the expected number of events in bin j is $\mu_j = NW_j f(x_j;\lambda)$

(here we use μ for the expectation value to avoid confusion with the parameter λ of the distribution)

- the actual, observed number is n_j, and can be described by a Poisson statistics (see page 96)

- thus, the error on the events in bin j is $\sigma_j = \sqrt{\mu_j}$

- the total χ^2 summed over all bins (sometimes called Pearson's χ^2) results in

$$\chi^2 = \sum_{j=1}^{N_B} \frac{(n_j - \mu_j)^2}{\sigma_j^2} = \sum_{j=1}^{N_B} \frac{(n_j - \mu_j)^2}{\mu_j}$$

- for discrete probabilities with individual probability $p_j(\lambda)$ for bin j, we likewise obtain

$$\chi^2 = \sum_{j=1}^{N_B} \frac{(n_j - Np_j)^2}{Np_j}$$

The parameter λ can be calculated from the usual minimization of χ^2

Note: (i) the bin width should be not too small (see (ii)) and not too large (such that $f(x;\lambda)$ does not vary too much over the bin j)

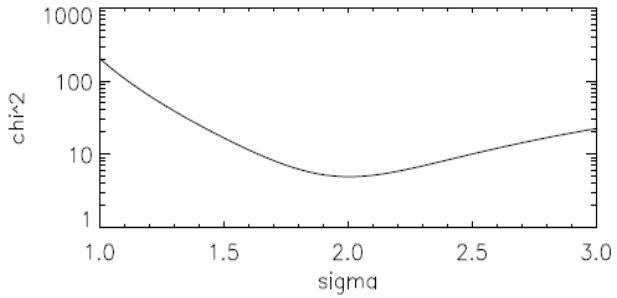
(ii) the expected number of events (μ_j or Np_j) should be at least 5 per bin. If this constraint is not fulfilled, the number of events N or the bin-width needs to be increased.

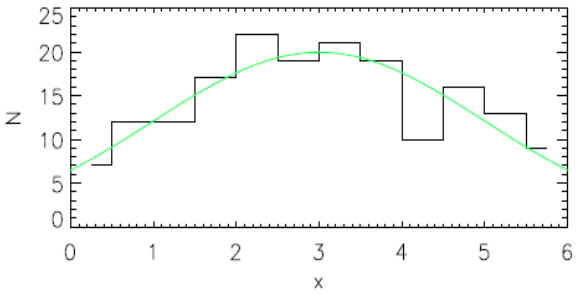
(iii) since the n_j are Poisson distributed, Pearson's χ^2 as defined above follows the actual χ^2 – distribution (based on a Gaussian) only for large N

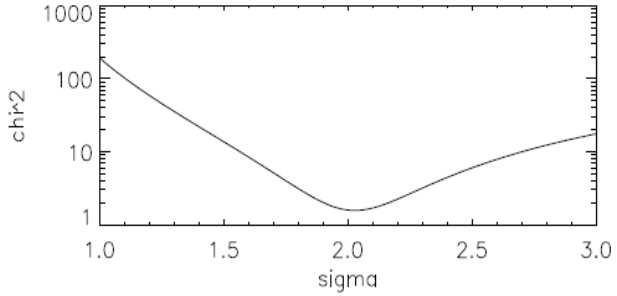
Example

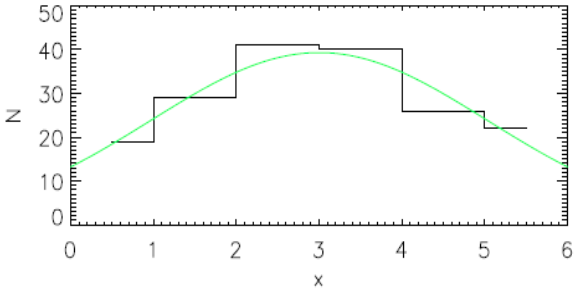
right panels: histograms with different bin-widths (W_j = 0.5, 1, 2) resulting from a normal distribution with μ=3 and σ=2, for N = 200

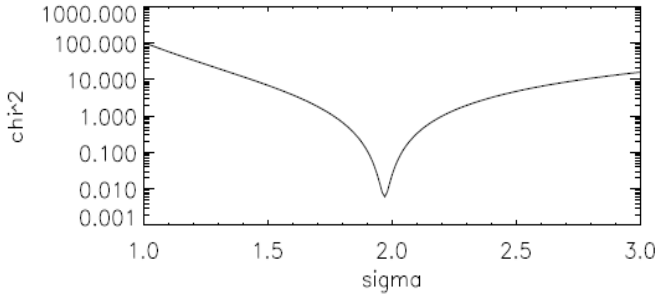
left panels: corresponding χ^2 as a function of σ. A fit to the binned data using σ(Min(χ^2)) is displayed in green on the right.

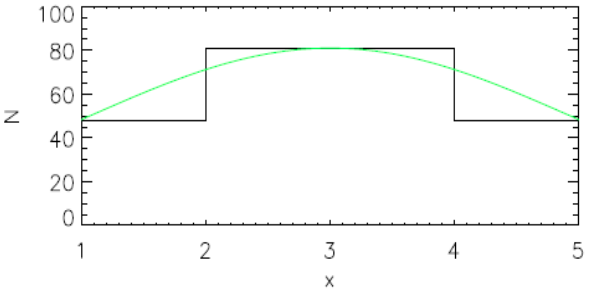












Goodness of fit (Fitgüte)

- So far, we have minimized χ^2 and obtained estimates (incl. errors) for the parameter(s) λ of the function $\varphi(x; \lambda)$ used to fit the measured data.
- From this approach, however, it is not clear whether the function itself (or the assumed errors) is/are reasonable.
- A suitable test can be found from the *goodness of fit*, which uses the properties of the χ^2 -distribution.
- Remember the basic assumption of the least squares method: the differences between theory ($\varphi(x_i; \lambda)$) and observations (y_i) are due to measurement errors alone.
- In so far, different samples (different series of measurements) will give different y_i and thus different minimized χ^2_{min} and different parameters λ.
- If we now assume that the measurement errors are Gaussian distributed (CLT), the (minimized) χ^2 is just a sum of squares of normally distributed reduced variables,

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} = \sum_{i=1}^{N} u_i^2 \qquad \text{with } \mu_i = \varphi(x_i; \lambda)$$

and thus should follow the corresponding χ^2 -distribution introduced in Chap. 5.

$P(\chi^2, f) = \frac{1}{\Gamma(f/2)2^{f/2}}(\chi^2)^{f/2-1}e^{-\chi^2/2}$, with Gamma-function Γ and

$E(\chi^2) = f$, $Var(\chi^2) = 2f$, where f denotes the so-called number of degrees of freedom.

For a sum of N *independent* r.v. u^2, the number of degrees of freedom is $f = N$ (Chap. 5).

If, however, χ^2 has been calculated after the minimization, not all terms in the sum are statistically independent, since they are subject to homogeneous linear constraints, at least for linear models, $y = \varphi(x;\boldsymbol{\lambda}) = \sum_{k=1}^{M} \lambda_k f_k(x)$

In this case, the number of degrees of freedom is reduced by the number of constraints (= number of parameters to be fitted),

$f = N - M$

Note:

(i) for a fit to a straight line, the linear model is $y = mx + b$, i.e.,

$M = 2$ with $\lambda_1 = b$, $f_1(x) = 1$, $\lambda_2 = m$, $f_2(x) = x$

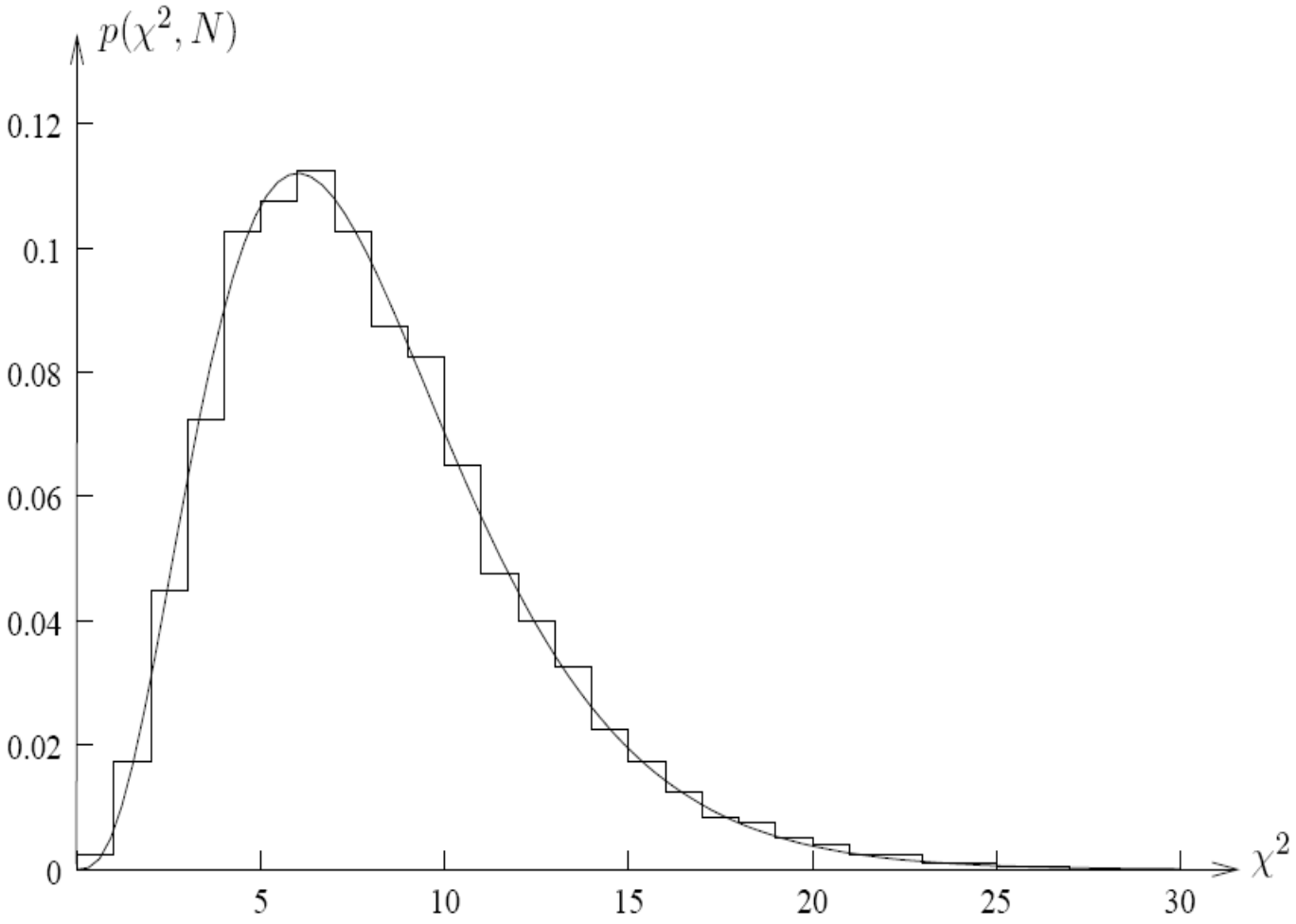
(ii) for non-linear models, e.g., $y = x^{\lambda_1} + x^{\lambda_2}$, this is no longer strictly valid, but might be applied if the non-linearity is not too drastic or if $N >> M$

Example

- consider a fit to a straight line with 10 measurements und two parameters, m, b:

  $y_i = b + mx_i$ + noise (Gaussian) (i = 1, 10);

- because of the Gaussian noise, all requirements to apply the least squares method are exactly fulfilled
- we calculate 10,000 different samples created in this way and calculate the corresponding minimized χ^2_{min}.
- the corresponding distribution is shown in the next plot, and compares very well with the theoretical χ^2 -distribution with f=10-2=8 degrees of freedom
- mean value and standard deviation of the experimental distribution, 8.07 and 4.03, respectively, are in good agreement with the theoretical prediction, $E(\chi^2)=f=8$ and $\sigma(\chi^2)=\sqrt{(2f)}= 4$



histogram of “measured” distribution of minimized χ^2_{min} (10,000 measurements); grey curve: theoretical pdf $p(\chi^2,f)$ for f=8

- Thus, a first simple test to decide whether all assumptions made (reasonable model, Gaussian errors of "correct" size) are fulfilled is to check the magnitude of the minimized χ^2_{min} . If it is in the range

  $\chi^2_{min} \approx f \pm \sqrt{(2f)}$,

  then there is no obvious problem.
- a better quantification is the calculation of the goodness of fit, Q
- Q describes the probability that any another $\chi^2 \geq \chi^2_{min}$ could have occurred by chance, i.e.,

  $Q(\chi^2_{min}, f) = P(\chi^2 \geq \chi^2_{min})$

- if, e.g., Q≈1, then the minimized χ^2_{min} is most likely too small (see below)

Q can be calculated from integrating the pdf of the χ^2-distribution:

$$Q(\chi^2_{\min}, f) = \int_{\chi^2_{\min}}^{\infty} P(\chi^2, f)\, d\chi^2 = \ldots = 1 - \frac{\gamma(f/2, \chi^2_{\min}/2)}{\Gamma(f/2)}$$ (Note the arguments!!!)

with incomplete Gamma-function $\gamma(a,x) = \int_0^x t^{a-1} e^{-t}\, dt$ and

(complete) Gamma-function $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\, dt$

- The calculation of the goodness of fit allows to determine whether the model and/or the assumed errors comply with the measurements.
- In dependence of the value of Q, we can make the following statements

  • if $0.05 \le Q \le 0.95$, the fit is OK $\left(\text{at least in principle}\right)$.

  In terms of hypothesis testing (Chap. 10),

  the fit cannot be rejected at a two-sided 10% significance level.

  • if $0.001 \lesssim Q \le 0.05$, the fit *maybe* OK, but one should perform an additional measurement series.

  • if $Q \lesssim 0.001$, $\chi^2_{\min}$ is too large (the probability to obtain any other, larger χ^2 is less than 10^{-3}), and the fit has to be rejected at a 0.1% significance level.

  Either, the model is "wrong", or certain errors have been assumed as too small, or the errors are not normally distributed

  • if Q is close to unity, $\chi^2_{\min}$ is too low! Either, the errors have been adopted as too large, or we have encountered a case of data-faking!

  There is, of course, a (quite) low probability that we were simply lucky. Again, an additional measurement series might be helpful.

Example

- continuation of example from page 214

The minimized $\chi^2_{\min}$ for sample 1,2,3,4 were

$\chi^2_{\min}$ =33.16, 3.63, 11.13, 4.61,

respectively. The number of data was $N = 7,$ and the number of fit-parameters was $M = 2$
$\Rightarrow f = 5$ (number of degrees of freedom)

Thus, we expect a minimized $\chi^2_{\min} = f \pm \sqrt{2f} = 5 \pm 3.16$

From this expectation, we conclude that the fits to sample 2 and 4 are OK. The fit to sample 1 can be discarded (obviously), whereas the fit to sample 3 lies within the 2-σ region and needs to be reinvestigated.

Quantitatively, we find goodness of fit parameters

$Q(\chi^2_{\min}, 5) = 3.52 \cdot 10^{-6}$, 0.604, 0.049, 0.465

which are consistent with our above expectations. Note that the value for sample 3 lies just slightly below the 5% margin and might be OK indeed, since there is still an almost 5% chance that other experiments would have yielded a higher $\chi^2_{\min}$!

Unknown errors

In order to calculate $\chi^2_{\min}$ and the goodness of fit, Q, but also the errors of the fitted parameters, one needs to know the individual errors σ_i (which may be identical for all data-pairs)

Sometimes, we do not know these errors. In this case, we may proceed as follows:

We assume that the errors are identical, i.e., $\sigma_i \equiv \sigma$, and determine the set of parameters from minimizing the quantity

$$\sum_{j=1}^{N}\left[y_j - \varphi(x^{(j)};\lambda)\right]^2 .$$ (Formally, we use the general algorithm and set $\sigma_i \equiv 1$)

Then, we assume that the fit is OK, and derive the value of σ from the corresponding expectation value of $\chi^2_{\min}$.

$$\Rightarrow \chi^2_{\min} := E(\chi^2_{\min}, f) = N - M = \sum_{j=1}^{N}\frac{\left[y_j - \varphi(x^{(j)};\tilde{\lambda})\right]^2}{\sigma^2}$$

$$\Rightarrow \quad \sigma = \sqrt{\sum_{j=1}^{N}\frac{\left[y_j - \varphi(x^{(j)};\tilde{\lambda})\right]}{N-M}}$$

Note that this approach does not allow to assess the quality of the fit, since we assume the fit to be perfect. However, the derived value of σ allows for a rough estimate on the errors of the parameters.

Errors on x and y

- Until now, we have assumed that the x-values are error-free.
- Often, one needs to perform a straight line fit to a data-set where both the x and the y data are affected by measurement errors. (Even worse, sometimes the x and y data have been calculated from other data and are correlated!)
- In this case, we have to minimize the quantity

$$\chi^2 = \sum_{j=1}^{N} \frac{\left[y_j - (mx_j + b) \right]^2}{\sigma_{tot,j}^2}$$

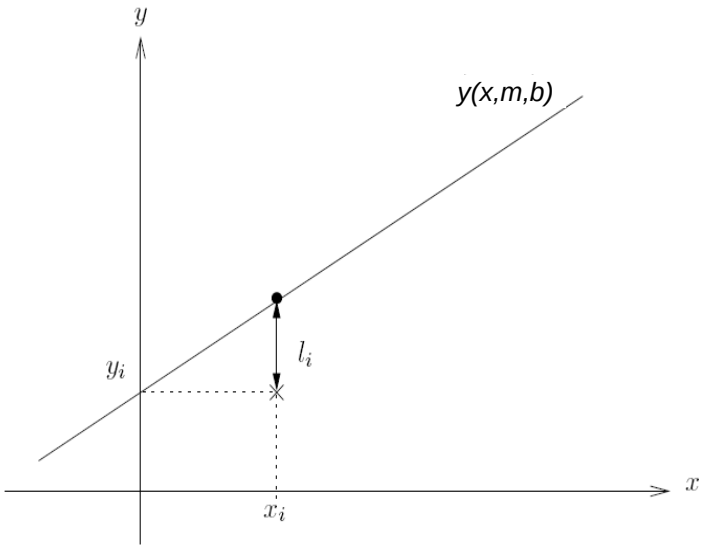
where $\sigma_{tot,j}$ is the total measurement error.

To calculate this error, we consider a certain residuum (see figure),

$l_j = y_j - f(x_j) = y_j - (mx_j + b)$

It is the error of this quantity which needs to be considered in the regression.

(Note that for the conventional case of $\sigma_{x_i} = 0$ we obtain $\sigma_j := \Delta l_j = \Delta y_j$ as usual).



For errors on both axes, we find from the law of error propagation

$$\sigma^2_{tot,j} = \Delta l_j^2 = \left[\left(\frac{\partial l_j}{\partial y_j} \right)^2 \Delta y_j^2 + \left(\frac{\partial l_j}{\partial x_j} \right)^2 \Delta x_j^2 \quad \left(+2 \frac{\partial l_j}{\partial y_j} \frac{\partial l_j}{\partial x_j} \mathrm{cov}(x_j, y_j) \right) \right] =$$

$$= \Delta y_j^2 + m^2 \Delta x_j^2 \left(-2m\rho(x_j, y_j) \Delta y_j \Delta x_j \right) \quad \text{(with correlation coeff. } \rho)$$

$$\Rightarrow \sigma_{tot,j}^2 = \sigma_{y_j}^2 \left[1 + m^2 \frac{\sigma_{x_j}^2}{\sigma_{y_j}^2} \left(-2m\rho(x_j, y_j) \frac{\sigma_{x_j}}{\sigma_{y_j}} \right) \right]$$

Obviously, errors in the x-data become of strong impact if the relation is steep.

The estimate for the intercept $\tilde{b}$ still follows from the equation $\overline{y} = \tilde{m}\overline{x} + \tilde{b}$
(because $\sigma_{tot,j}^2$ does not depend on b), i.e.,
the line goes through the center of gravity (w.r.t. weighted means) as usual.

$\tilde{m}$, on the other hand, has to be calculated by numerical minimization method
(or from an MCMC method, see next section),
unless the errors in x direction are identical, as well as the errors in y-direction, and the x and y data are uncorrelated. In that case, one obtains

$$\tilde{m} = \frac{\sigma_y}{\sigma_x} \left(A \pm \sqrt{A^2 + 1} \right) \quad \text{with} \quad A = \frac{\sigma_x^2 (\overline{y^2} - \overline{y}^2) - \sigma_y^2 (\overline{x^2} - \overline{x}^2)}{2\sigma_x \sigma_y (\overline{xy} - \overline{x}\,\overline{y})}.$$

The positive sign is taken if the denominator of A is positive, and vice versa.

Example

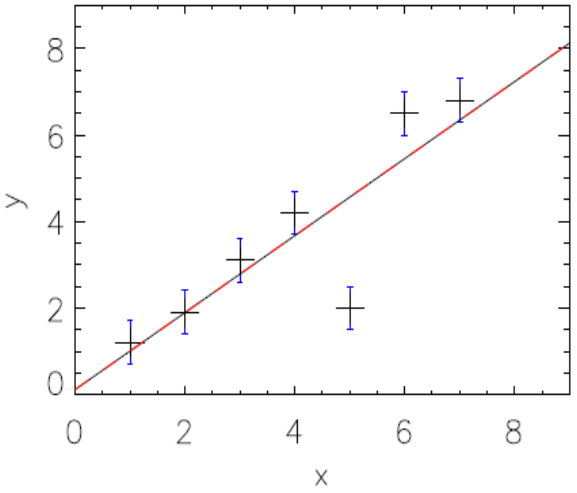
continuation of example from page 214/215

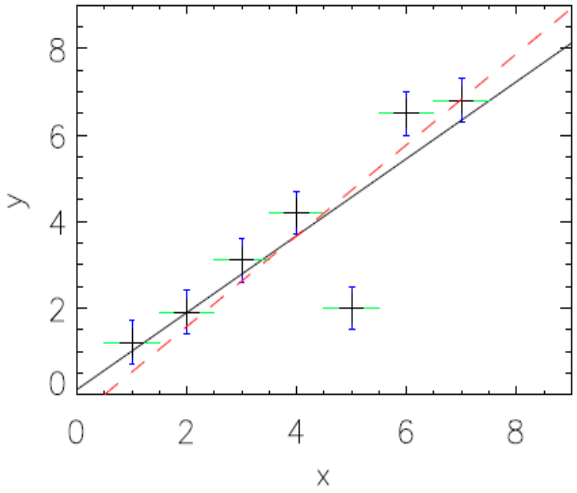
data and y-errors (σ_y =0.5, blue) as in sample 1; regression accounting for y-errors only in black:
m=0.89, b=0.09, χ^2_{min} = 33.16

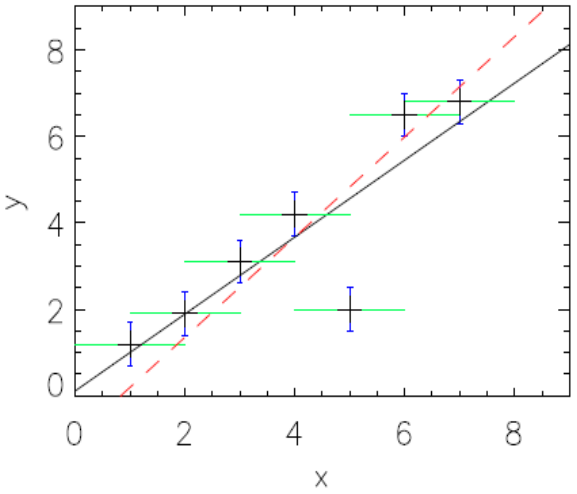
additional errors on the x-data (σ_x, green) and regression accounting for x- and y-errors in red

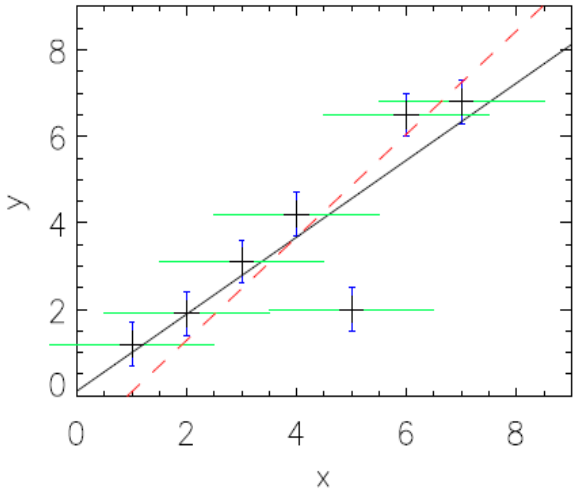
upper and lower panel, from left to right:
σ_x =0.01, m=0.89, b= 0.11, χ^2_{min} =33.15
σ_x =0.50, m=1.05, b=-0.53, χ^2_{min} =17.15
σ_x =1.00, m=1.16, b=-0.96, χ^2_{min} = 6.48
σ_x =1.50, m=1.19, b=-1.09, χ^2_{min} = 3.15

Formally (from χ^2_{min}), fit 3 and 4 are OK.









To be continued …

- Least squares fitting/ χ^2 minimization
  - arbitrary linear models: “normal equations”, orthogonal transformations
  - non-linear models: Gauss-Newton method, minimization methods, particularly Levenberg-Marquardt method and genetic algorithms

  → lecture notes “Numerik für Physiker” (in German), Chap. 5.5/5.6 (http://www.usm.uni-muenchen.de/people/puls/lessons/Numerik/Numerics.pdf)

9. Sampling the posterior – MCMC

- Often, the knowledge of the **distribution** of the parameters obtained from a “fit” is as (or even more) important as the knowledge of the parameters themselves.
- Well-known examples are the constraints on the cosmological parameters derived from various experiments (see next page)
- As we have seen in Sect. 7 (ML estimators), a direct interpretation of the likelihood as the distribution of the estimated parameters requires to assume that the priors are uniformly distributed, an assumption which often cannot be justified.
- Even if the assumption of uniform priors would be true, the calculation of errors on the derived parameters and their correlation (Fisher-matrix) becomes quite cumbersome if the model is complex and/or involves many parameters that need to be estimated in parallel.
- The major issue here is that the likelihood method provides us with a distribution of likelihoods (probability for the observed model given a set of parameters), but not with the distribution of the parameters themselves. This could be found – if at all, and then only for constant priors -- only by marginalization, i.e., integration over the complete set of all other parameters.
- Thus, a method which provides the distribution of the parameters themselves would be highly welcome, and indeed such a method exists, called **Markov Chain Monte Carlo (MCMC),** suggested (and used) first by Metropolis et al. (1953), and improved in a number of subsequent studies.
- In the astrophysical community, one particular implementation (in Python) is frequently used, since it’s easy to use, quite fast and reliable, namely *emcee*

MCMC: the basic idea

For simplicity, and in view of the example discussed below, we assume a dataset x, and a model with three parameters, m, b, f that should describe the dataset

Again, we make use of Bayes theorem, where the posterior, i.e., the distribution of m, b, f given the data x, is expressed via

$$p(m,b,f \mid x) = \frac{p(x \mid m,b,f)\, p(m,b,f)}{p(x)}$$

with "evidence" $p(x)$ [marginal likelihood for the data], $p(x \mid m,b,f)$ the likelihood of the "observed" dataset given a model with parameters m, b, f, and $p(m,b,f)$ the prior for the parameters, which are usually thought to be independent, i.e.

$$p(m,b,f) = p(b) \cdot p(m) \cdot p(f)$$

The procedure is now as follows:

(i) define the (log) likelihood for the observed dataset, in dependence of the (model) parameters

(ii) define the (log) prior (see below)

(iii) start the MCMC-"engine". The output from the engine is a (large) sample of the posterior (the parameter set given the data), enabling to quickly perform any kind of marginalization and correlation analysis.

MCMC: the basic idea

Inside the MCMC algorithm, the following steps are performed
[description partly taken from "**MCMC sampling for dummies - While My MCMC Gently Samples**"]

- initialize the parameter-set, here $\theta_1:=(m_1,b_1,f_1)$
- calculate the combined "probability" (not normalized), likelihood times prior, $p(x|\theta_1)p(\theta_1)$ (in this calculation, any factor that does not depend on the parameters can be omitted, since later on only relative probabilities will play a role)
- use a Markov process ("chain" if many such processes) to proceed from parameter θ_i to θ_{i+1}. Markov processes are stochastic processes where the r.v. θ_{i+1} depends only on θ_i, and **not** on the previous history i-1, i-2, ...
- The actual realization of this Markov step (direction, length) is the major difficulty in this method, since it must be warranted that the complete parameter space w.r.t. θ is *efficiently* sampled (→ literature)
- for the new "proposal" θ_{i+1}, calculate $p(x|\theta_{i+1})p(\theta_{i+1})$, and the probability ratio

$$r = \frac{p(x\,|\,\theta_{i+1})\,p(\theta_{i+1})}{p(x\,|\,\theta_i)\,p(\theta_i)} = \frac{\dfrac{p(x\,|\,\theta_{i+1})\,p(\theta_{i+1})}{p(x)}}{\dfrac{p(x\,|\,\theta_i)\,p(\theta_i)}{p(x)}} = \frac{p(\theta_{i+1}\,|\,x)}{p(\theta_i\,|\,x)}$$

- This so-called acceptance ratio is the ratio of the posteriors for the proposed θ_{i+1} and the current θ_i (since the evidence and any constant factor neglected above cancels out).
- By means of r, we check whether the proposed θ_{i+1} is a "good place" to jump to or not.

MCMC: the basic idea

- If r>1, the proposed parameter set θ_{i+1} is better supported by the data than the current θ_i, and θ_{i+1} is definitely a place to go.

- If we would only accept a jump when r>1, at some step we would reach the point of maximum posterior probability, from where no more moves would be possible. However, since we want to obtain the *distribution* of the posterior (and not only the optimum parameter set), we **sometimes** have also to accept proposals with r<1. E.g., if $p(\theta_i|x)$ is twice as large as $p(\theta_{i+1}|x)$, there is a 50% chance of moving towards θ_{i+1}.

- Thus, the typical MCMC strategy reads:

```
provide new proposal θ_{i+1} [Markov step]

calculate r(θ_{i+1}, θ_i)

    accept = random_uniform(seed) < r
#always true for r>1, sometimes true for r<1 (in relation to prob.)

if accept then  θ_i = θ_{i+1}   # update position

save θ_i  for the sample    # either the proposed or the current θ is saved
```

That way, we are visiting regions of high posterior probability relatively more often than those of low posterior probability, and this simple procedure gives us meaningful samples from the posterior. Note that the cumbersome quantity p(x) (many integrals → "nested sampling") does not need to be calculated. Note also that in the output chain certain parameter sets will sometimes repeat themselves (whenever r < random_uniform).

MCMC – additional notes

General note: when applying the MCMC algorithm, consult the literature, talk to experienced colleagues, and perform a couple of tests before relying on the results!

The MCMC method is also well-suited for (linear and non-linear) regressions involving *errors on x* and *correlations* between the x- and y-data (see Chap. 8)

Notes on the prior:
Often, it is a good idea to use either

- uniform ("flat") priors (if one has *only* an idea about the potential range of parameters, or
- normally distributed priors (if one has an idea about the typical value and it's uncertainty, or
- specific distributions resulting *from previous experience* (cf. Sect. 2, "The first night in paradise"
- when adequate, or for tests of the method, one might also use conjugate priors (see Sect. 7), which result in the same posterior distribution as the prior.

Note that uniform priors are sometimes not the most "non-informative" ones. E.g., using symmetry and scaling invariance arguments, it turns our that the most non-informative prior for the standard deviation in a model is $p(\sigma) \sim 1/\sigma$ (the so-called Jenkins-prior), whilst for a linear model with slope m and offset b the most uninformative choice is a uniform prior for b and $p(m) \sim (1+m^2)^{-3/2}$, corresponding to a uniform distribution in $\sin(\theta)$ where θ is the angle between the x-axis and the model line.

MCMC – additional notes

Notes on the sampled chain

- Because the initial samples are drawn from a specific start-point (e.g., from the prior, or from a Gaussian “ball” around the maximum likelihood), it can take time for them to converge on the desired sampling from the posterior distributions. One might try to estimate this first “*burn-in*” period (e.g., by eye-inspection of the “trace”, see example below, or by inspecting the auto-correlation time), and simply ignore all samples created during this phase.

- Because of the Markovian process, there is a certain (auto-)correlation between the drawn samples θ_i and θ_{i-N}, and it is worthwhile to estimate the autocorrelation time (e.g. by using the tool provided by *emcee,* `sampler.get_autocorr_time`).
  [briefly: test whether there is a significant correlation between θ_i and θ_{i-N}, N = 1,..., and provide the value for the maximum (auto-)correlation, N_{corr}].

- Then, the output-sample *might* be thinned by about half the autocorrelation time, i.e., only each $(N_{corr}/2)^{th}$ sample point would be finally considered. However, thinning is debated, and according to Link & Eaton (2011) “On thinning of chains in MCMC”, doi: 10.1111/j.2041-210X.2011.00131.x, thinning of chains is not usually appropriate when the goal is precise estimates from an MCMC sample. Anyhow, from the auto-correlation analysis, at least an estimate of the burn-in time might be obtained, namely a few times N_{corr}.

A worked out example: Fit to a straight line with unknown systematic error

- The following example bases (in part) on the tutorial for the *emcee*-sampler.
- for other samplers and a comparison, see, e.g., samplers-samplers-everywhere by Matt Pitkin.

"Observations" (data): 20 data points in x-y, errors only on y

model: linear, $y = mx + b$

y errors: (i) Gaussian, with known σ_{obs}

(ii) systematic, proportional to $|y|$;

$\rightarrow \sigma^2 = \sigma_{\text{obs}}^2 + \left(f \cdot y\right)^2$

Actually, we are only interested in the model parameters, m and b.
f is a so-called **nuisance** parameter, which *has* an influence on these values, but is of no further physical interest. I.e., the distributions of m an b have to be derived by marginalizing p(m,b,f | x) over f.

simulations: 20 x-values uniformly distributed in (0,10]: $x_i = 10 \cdot \text{Random\_uniform(seed)}$

$\rightarrow y_i = mx_i + b + yerrobs_i + yerrsys_i$

with $yerrobs_i = \sigma_{\text{obs},i} \cdot \text{Random\_normal(seed)}$ and $\sigma_{\text{obs},i} = (0.1 + 0.2 * \text{Random\_uniform(seed)})$

and $yerrsys_i = \text{abs}\,(f \cdot y_i) \cdot \text{Random\_normal(seed)}$

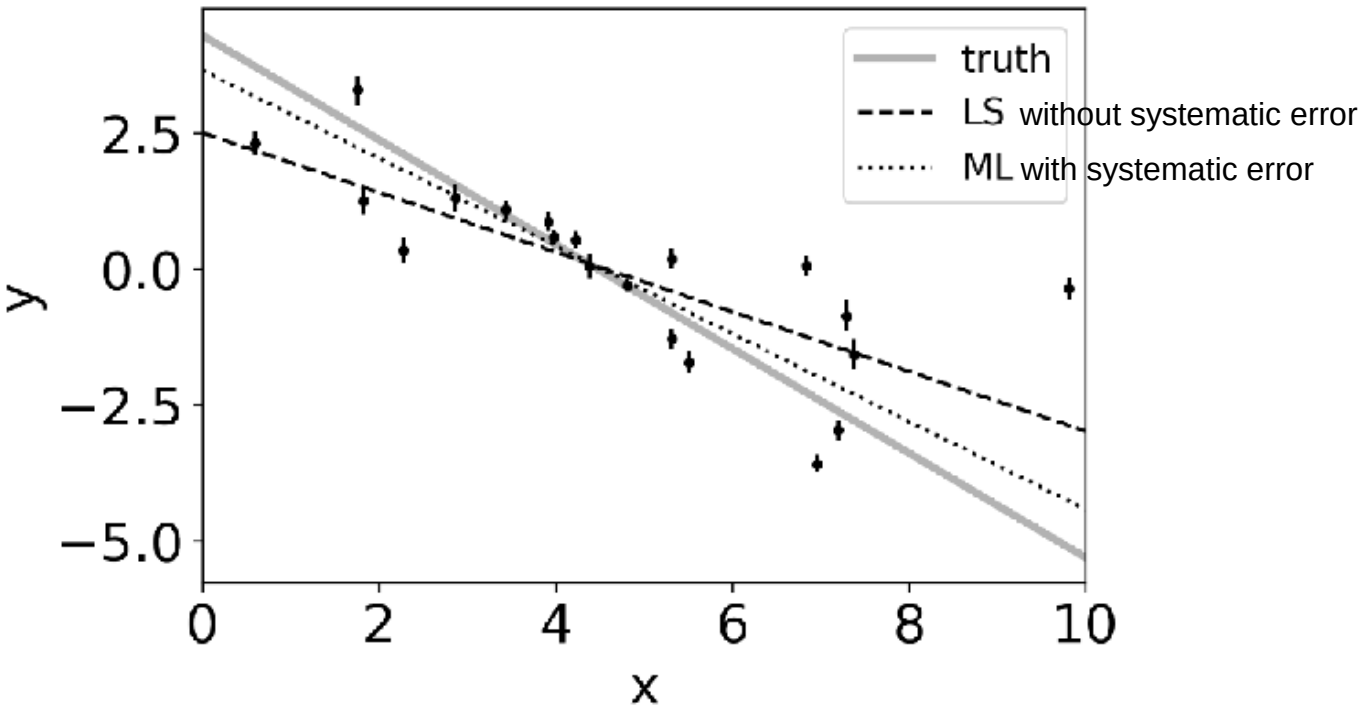
Example: fit to a straight line – Least squares and ML

m_{true}= -0.9594
b_{true} = 4.294
f_{true} = 0.534

from max. loglikelihood:

m_{ML}= -0.806
b_{ML} = 3.649
f_{ML} = 0.690

What about errors?



loglikelihood

$$\ln L(m,b,f \mid x^{(i)}) = -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{\left(y_i-(m\cdot x_i+b)\right)^2}{\sigma_i^2}+\ln\sigma_i^2\right] + \text{const}$$

cf. page 181

with $\sigma_i^2 = \sigma_{\text{obs},i}^2 + \left[f\cdot(m\cdot x_i+b)\right]^2$

Example: fit to a straight line – errors on the parameters from ML (flat priors)

priors -- in order to compare with the likelihood method, we use flat priors:

$$p(m) \propto \begin{cases} 1 \text{ for } m_{\text{true}} - 10 < m < m_{\text{true}} + 10 \\ 0 \text{ else} \end{cases}$$

$$p(b) \propto \begin{cases} 1 \text{ for } b_{\text{true}} - 10 < b < b_{\text{true}} + 10 \\ 0 \text{ else} \end{cases}$$

$$p(\ln f) \propto \begin{cases} 1 \text{ for } \ln f_{\text{true}} - 10 < \ln f < \ln f_{\text{true}} + 10 \\ 0 \text{ else} \end{cases}$$

(remember: likelihood invariant to parameter transformations)

Assuming flat priors, we can estimate the errors on the parameters from the condition that the 1-σ error is located at the (hyper-)surface for L_{max}-0.5

In this way, we find (by evaluating ln *L* on a dense grid and locating the iso-contours)

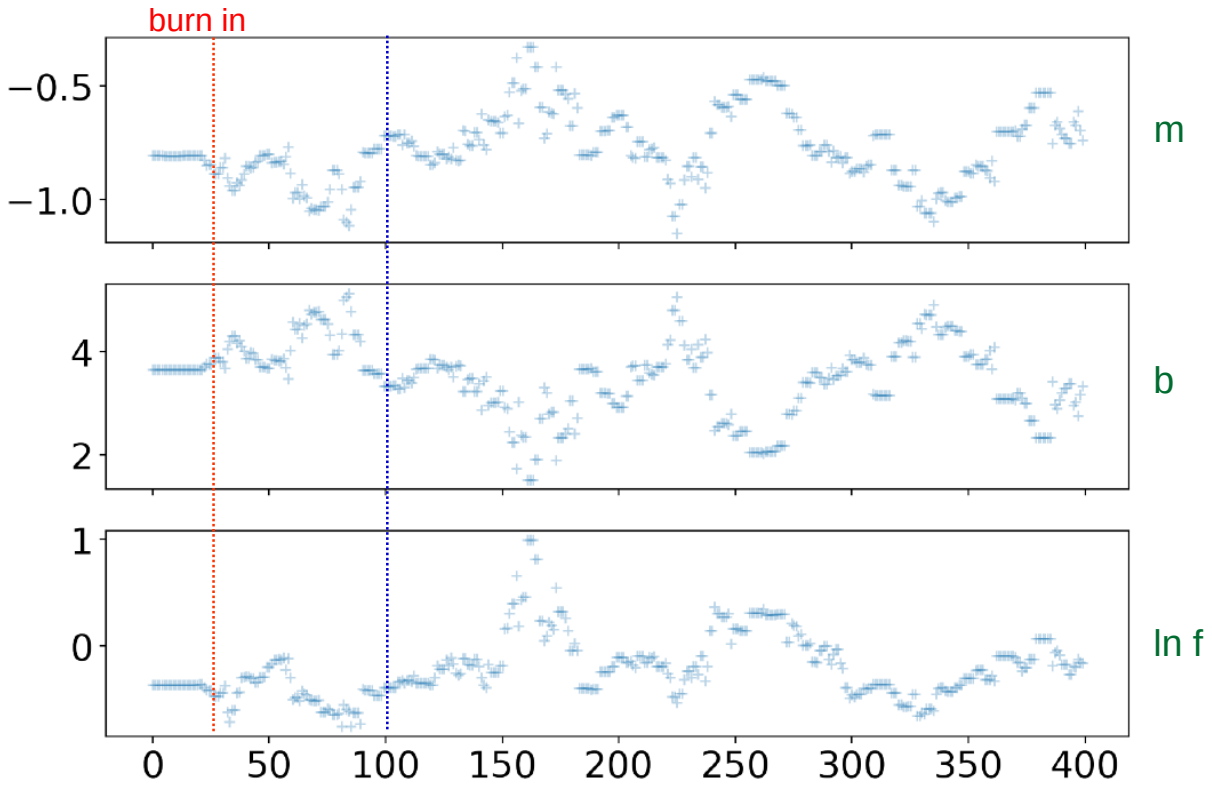
$$m_{ML} = -0.806^{+0.120}_{-0.140}$$

$$b_{ML} = 3.649^{+0.650}_{-0.550}$$

$$\ln f_{ML} = -0.371^{+0.200}_{-0.200}$$

Example: fit to a straight line – MCMC

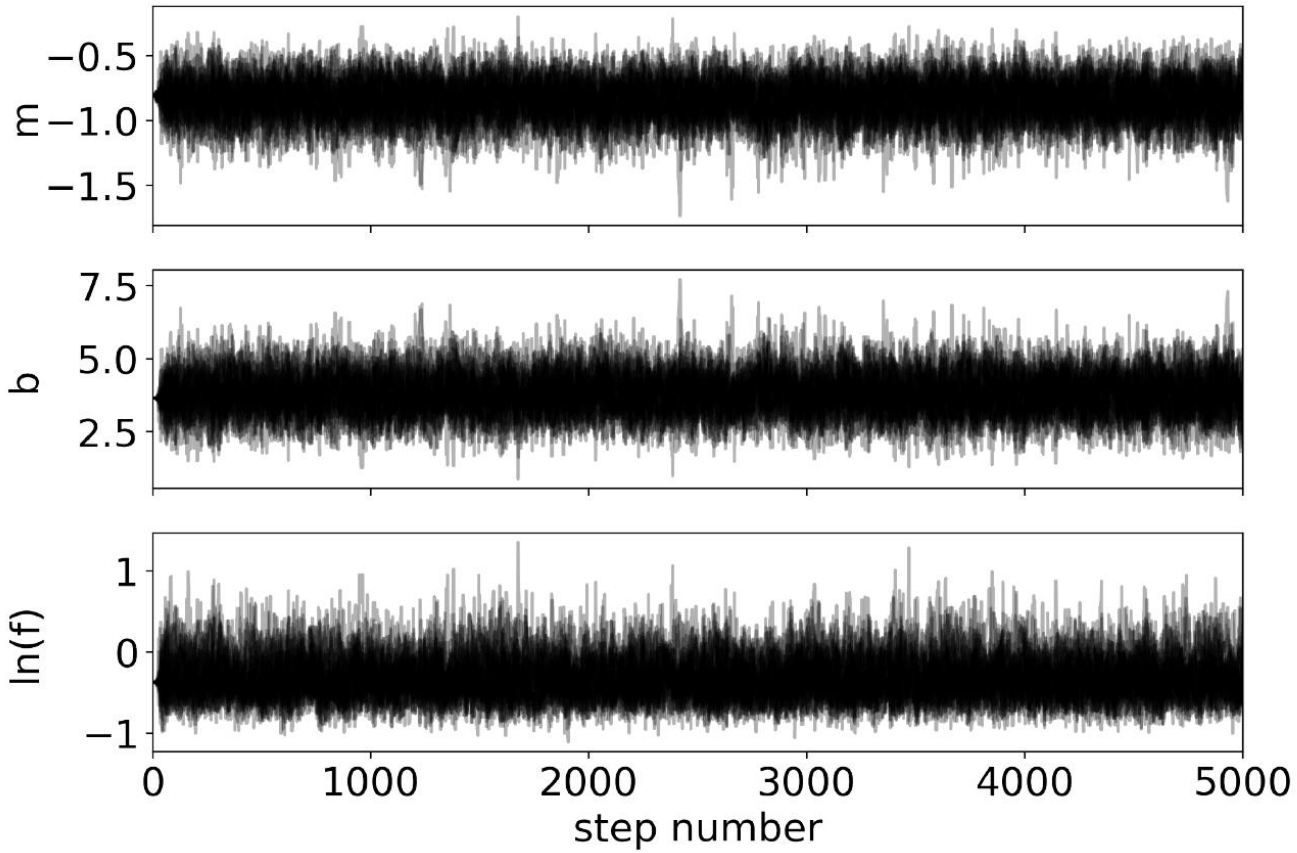
- Now MCMC (using emcee) with 32 “walkers” and 5000 steps per walker
  - *walkers:* in order to more efficiently sample the parameter space, many samplers run in parallel and periodically exchange states



First 400 steps (samples) from walker #0. Obviously, certain samples appear repeatedly (whenever r < random_uniform, e.g., around step 100, dotted blue line)

Example: fit to a straight line – MCMC

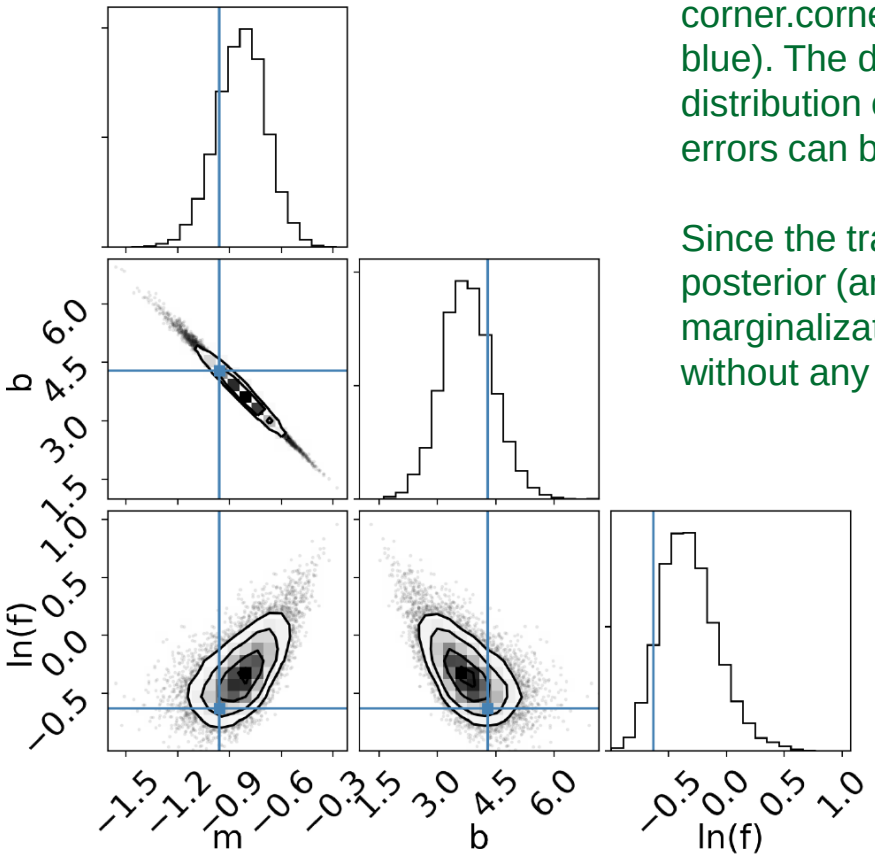
"trace" produced by *emcee*



All 5000 steps (samples) from all 32 walkers; the burn-in phase is still visible.

The following results have been produced by neglecting the first 100 steps, and by thinning the samples so that only each 15th is used.

Example: fit to a straight line – MCMC: distribution and errors



Marginalized distributions as obtained from the corner.corner Python-module ("true" values indicated in blue). The diagonal consists of the marginalized distribution of the individual parameters, and mean and errors can be easily calculated.

Since the trace (in case, thinned) is just a sample of the posterior (an array with dimension Nstep x Npara), marginalization and histogram-plotting can be performed without any effort.

$$m_{ML} = -0.806^{+0.120}_{-0.140}, \qquad m_{MCMC} = -0.829^{+0.136}_{-0.149}$$

$$b_{ML} = 3.649^{+0.650}_{-0.550}, \qquad b_{MCMC} = 3.754^{+0.675}_{-0.618}$$

$$\ln f_{ML} = -0.371^{+0.200}_{-0.200}, \qquad \ln f_{MCMC} = -0.341^{+0.263}_{-0.217}$$

Example: fit to a straight line – MCMC vs. marginalized likelihoods

In the following, we marginalize the likelihood (*L* as a function of m, b, ln f calculated on a dense grid) by "brute force" trapezoidal integration, assuming flat priors for m and b, and $p(f) \propto 1/f$ (consistent with a flat prior in ln f) . We don't care for pre-factors, since the normalization could be done in the end. What we want to show here is that the distribution from the marginalized likelihood is (and should be) quite similar (theoretically: identical) to the distribution from the MCMC, *when the priors are identical.*
Formally, we look for the marginalized distribution of the posteriors,

$$p(m, b \mid x) = p(m, b, x) / p(x),$$

where p(x) is the evidence, i.e., just a number -- though difficult to calculate if many parameters are present. Now

$$p(m, b, x) = \int p(m, b, f, x)\, df = \int p(x \mid m, b, f)\, p(m)\, p(b)\, p(f)\, df$$

and thus (for flat priors in m and b)

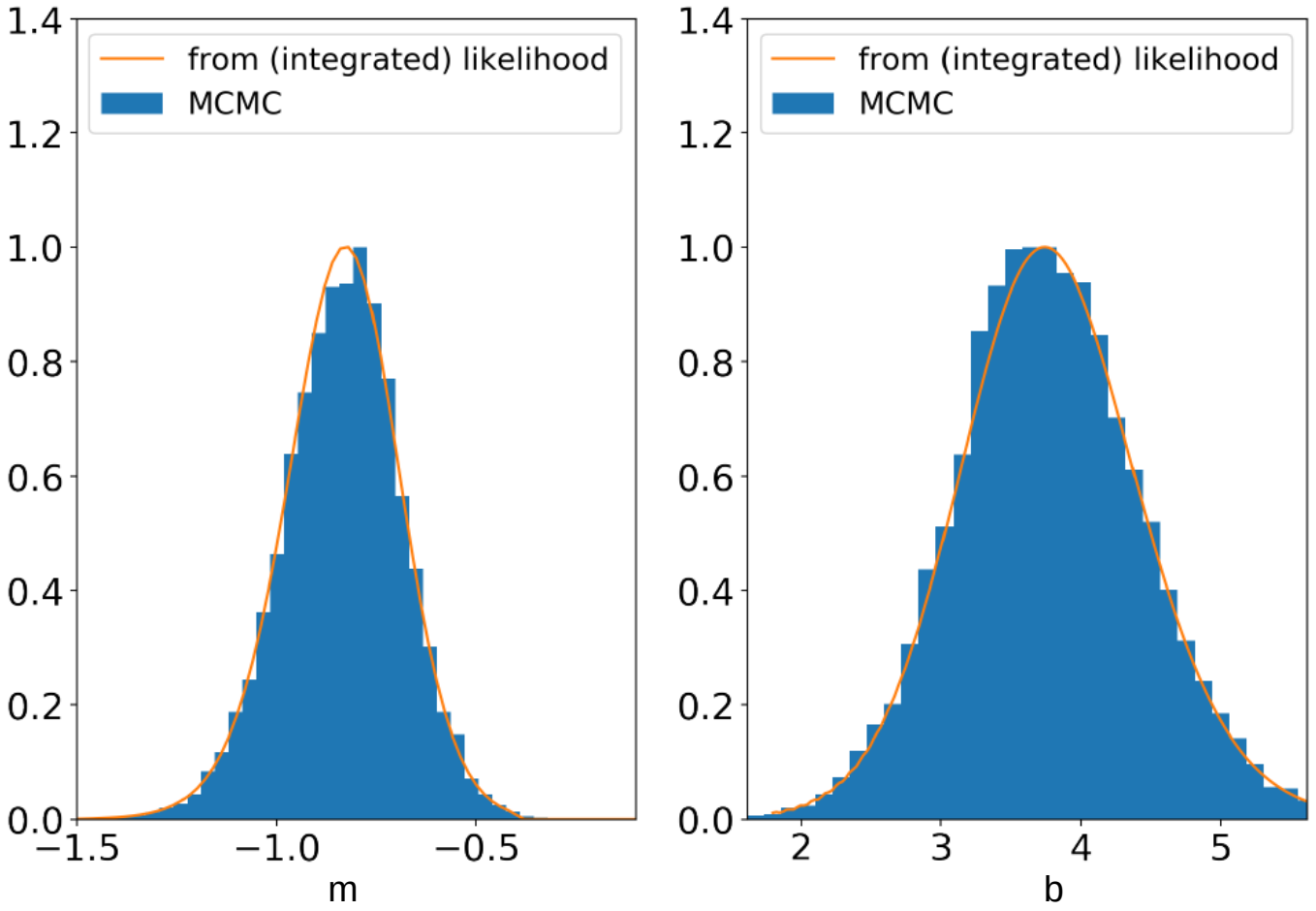
$$p(m, b \mid x) \propto \int p(x \mid m, b, f)\, p(f)\, df,$$

where $p(x \mid m, b, f)$ is just the likelihood *L* for the dataset. Hence, to obtain the marginalized distribution of the posterior, we just have to integrate the likelihood over $p(f)\, df \propto df / f = d\ln f$.
[For more complex problems, the corresponding integration is the major problem.]

Moreover, the marginalized posteriors for $p(m \mid x)$ and $p(b \mid x)$ are obtained (without any additional assumptions) via

$$p(m \mid x) = \int p(m, b \mid x)\, db \quad \text{and} \quad p(b \mid x) = \int p(m, b \mid x)\, dm.$$

Example: fit to a straight line – MCMC vs. marginalized likelihoods



Marginalized distribution for m (left) and b (right), normalized to a maximum of unity. Note that the distributions resulting from MCMC and the integrated likelihood are consistent, as required!

10. Confidence intervals and hypothesis testing

A few remarks on **confidence intervals**

- Measurements (in a statistical sense) are obtained from estimates of one or more parameters.
- The corresponding measurement error(s) are obtained from estimates of the confidence interval(s).
- not evident how to define these intervals. Various philosophies/methods are present, based on the “school” the author is belonging to.
- a small list of wishes for appropriate properties:
  - confidence interval(s) should contain the actual parameter(s) with a certain (pre-defined) probability (confidence level)
  - for a given confidence level, the interval should be small
  - independent of subjective assumptions
  - the definition should be consistent, i.e., observations with similar information content regarding the parameter should yield similar intervals.
  - method to derive intervals should be simple and transparent

- From Chap. 7, we know that it is impossible to obtain the pdf for a certain parameter from a sample alone, unless the prior's pdf is known. Thus, in most cases one should restrict oneself to the likelihood function.
  - if the prior is known with sufficient knowledge, the pdf of the parameter can be derived, and the corresponding confidence interval defined from the variance of this pdf.
  - for unknown prior, we use
    - the errors derived from the variation of the likelihood function around its maximum
    - in the large N limit (→ Gaussian likelihood function), the variances of the estimator(s) from

$$Var(\tilde{\lambda}_i) = -\frac{1}{E\left\{\frac{\partial^2 \ln L}{\partial \lambda_i^{\,2}}\right\}_{\tilde{\lambda}_i}},\ i = 1, M \ \text{ for } M \ \text{ parameters}$$

    - if possible, the limits as implied by a likelihood ratio (Chap. 7)
  - sometimes, one needs to integrate over the possible parameters. In this case, we assume constant priors (**maximum ignorance** in most cases). Remember, however, the inherent problem regarding parameter transformations

$$P(\lambda_1 < \lambda < \lambda_2) = \frac{\int_{\lambda_1}^{\lambda_2} L(\lambda)\mathrm{d}\lambda}{\int_{-\infty}^{\infty} L(\lambda)\mathrm{d}\lambda} = \frac{\int_{\alpha_1}^{\alpha_2} L(\alpha(\lambda))\frac{\partial\lambda}{\partial\alpha}\mathrm{d}\alpha}{\int_{-\infty}^{\infty} L(\lambda)\mathrm{d}\lambda} \neq \frac{\int_{\alpha_1}^{\alpha_2} L(\alpha)\mathrm{d}\alpha}{\int_{-\infty}^{\infty} L(\alpha)\mathrm{d}\alpha}$$

when $\alpha(\lambda)$ is a non-linear function.

The latter problem can be circumvented by using "Bartlett's S-function", see Brandt, Chap. 7.8

Classical confidence intervals

- "frequentist" interpretation, introduced by Pearson, difficult and sometimes problematic (not recommended, except for Gaussian case)

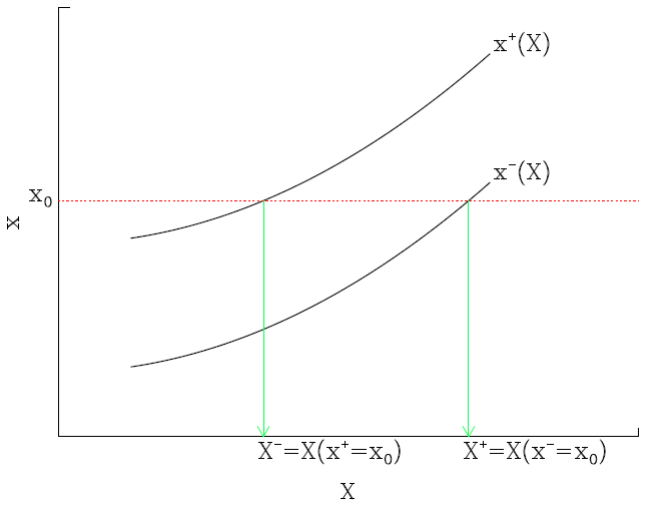
- **Definition**: The confidence interval with confidence level *p* contains **all** those values of the parameter *X*, for which the probability to obtain a measurement (estimate) x=S(*X*) within a certain interval is given by *p*.
- *Before* the measurement, one defines, for each value of X, an interval, which in the one-dimensional case is defined by

$$P(x^- < x < x^+ \mid X) = p$$

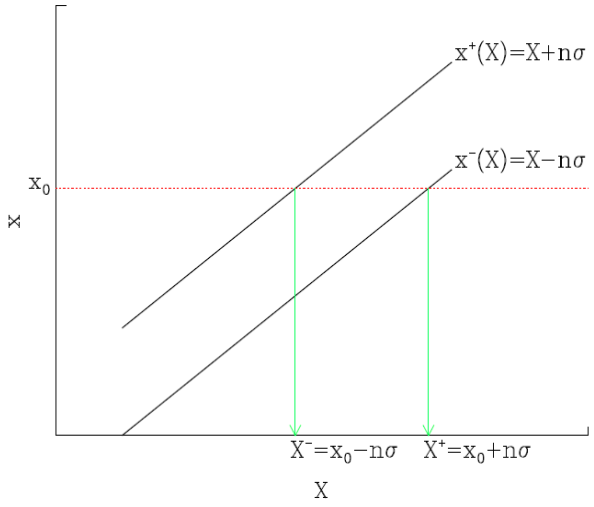
- For a given *X* and *p*, a measurement of x is thus expected within the interval $[x^-(X), x^+(X)]$.
- For an *actual* measurement of x_0, the confidence interval of *X* comprises all those *X*-values for which the above condition with $x = x_0$ is fulfilled.

- In the univariate case, usually central probability intervals are chosen, i.e.,

$$P(x < x^- \mid X) = P(x > x^+ \mid X) = (1 - p)/2$$



Classical confidence intervals, general case:
the confidence belt (i.e., the region in between x^- and x^+) is constructed from the condition $P(x^+ < x < x^- \mid X) = p$, as a function of the true parameter X. For an actual measurement (estimate) x_0, the confidence interval $[X^- \le X \le X^+]$ is derived from inverting the relations as indicated. Note that in general the classical confidence interval is not (directly) related to the errors of x_0.
Example: Poisson distribution, with mean X and measurement k_o. Here, the width of the distribution depends on X via $\sigma(k) = \sqrt{X}$. E.g., for large X, $x^{\pm} = X \pm n\sqrt{X}$, and the confidence belt diverges with increasing X (compare with right panel).



Classical confidence intervals, normal distribution:
Here, the confidence belt lies inside the region confined by $x^- = X - n\sigma$ and $x^+ = X + n\sigma$. Both relations are parallel to the one-to-one relation $x = X$. Thus, the confidence interval $[X^- \le X \le X^+]$ derived from an actual measurement x_0 is given by $X^{\pm} = x_0 \pm n\sigma$ (see figure), which justifies the "usual" approach of identifying the confidence region of X with the uncertainty of x_0 (which for the general case would be a misinterpretation and wrong.)

The discussed property for normal distributions can be generalized to such distributions that are symmetric with respect to parameter X, and their width etc. does not depend on X, i.e.

$f(x) = f\left(|x - X|\right)$.

Proof: $x^+(X)$ can be derived from the condition $\left(\text{if the central probability is } p\right)$

$$\int_{x^+(X)}^{\infty} f\left(|x' - X|\right) dx' = (1-p)/2$$

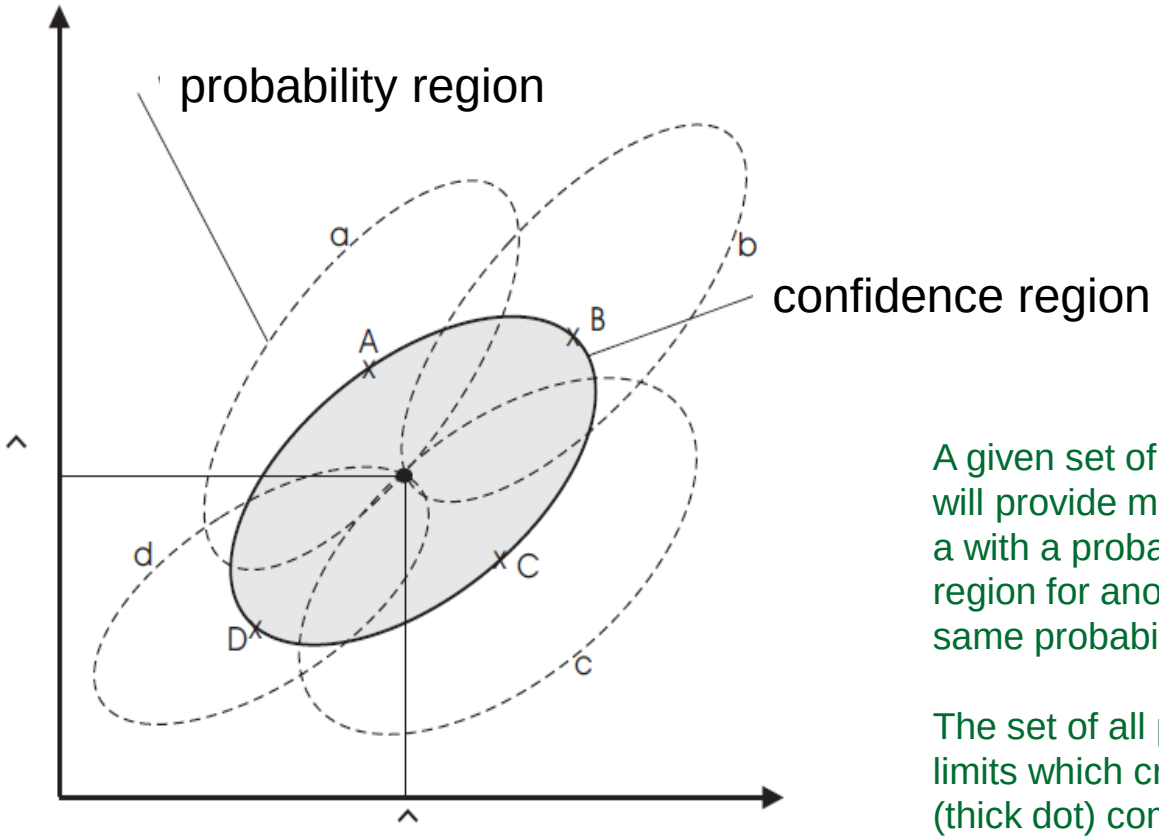
Now substitute $x' - X := x_o - y \quad \Rightarrow \; dx' = -dy$ and $y = -x' + x_0 + X$.

$$(1-p)/2 = -\int_{-x^+ + x_o + X}^{-\infty} f\left(|x_o - y|\right) dy = \int_{-\infty}^{X^-} f\left(|y - x_o|\right) dy = \int_{-\infty}^{X^-} f\left(|x' - x_o|\right) dx'$$

[Note: $-x^+ + x_o + X = X^-$, see previous sketch]

Thus, the role of X and x_o are interchanged, while the integrals have the same values. Consequently, the integration range must be the same, and because of the symmetric pdf we find that if $x^+ = X + \Delta$, then $X^- = x_o - \Delta$, q.e.d.

- in the multivariate case, probability regions are chosen which are enclosed by a curve or surface of equal probability.



A given set of parameters (here: 2), A, will provide measurements in the region a with a probability p. The probability region for another parameter set B (with same probability) is b, and so on.

The set of all parameters with probability limits which cross the measured value (thick dot) comprises the confidence region (grey).

Upper (lower) boundaries

- Sometimes there is the problem that a parameter cannot be measured accurate enough as to obtain a value which is significantly different from zero. In this case one might like to provide an upper boundary.
  Example: the life-time of a short-lived particle might be quoted as "with 90% confidence, the mean lifetime is smaller than …"
- In this example, the probability aspect is the dominating one, and one invokes a constant prior. In other words, one calculates the normalized likelihood for the considered parameter λ and interprets this as a pdf for the parameter (since we assume a constant prior, this cancels in the nominator and denominator, see Chap. 7) .
- The integral over the allowed range $[-\infty, \lambda_0]$ yields the confidence level C for the upper limit λ_0.

$$C(\lambda_0) = \frac{\int_{-\infty}^{\lambda_0} L(\lambda) d\lambda}{\int_{-\infty}^{\infty} L(\lambda) d\lambda}$$

Upper Poisson limit

- If, in a certain experiment, one is looking for a specific event (reaction, number of photons, particles etc.), but does not find any, it is appropriate to cite an upper limit instead of the measured value, again using a corresponding probability.
- In this case, the result might be quoted in the form “the number of events, photons … is smaller than λ_0 with p*100% confidence (p=0.90, 0.95 …)
- Again, the upper limit λ_0 is found by integration of the likelihood function assuming a constant prior.
- **Example:** We observe k events. Calculate the upper limit λ_0 for the expectation value with C = 90% confidence.
  - The normalization integral w.r.t. the Poisson distribution is unity, and thus we obtain

$$L(\lambda) = L(\lambda; k) = P(k \mid \lambda);$$

$$C = \int_0^{\lambda_0} P(k \mid \lambda)\mathrm{d}\lambda = \frac{1}{k!}\int_0^{\lambda_0} \exp(-\lambda)\lambda^k \mathrm{d}\lambda \underset{\text{partial integration}}{=} 1 - \sum_{j=0}^{k} \frac{\exp(-\lambda_0)\lambda_0{}^j}{j!} = 1 - \sum_{j=0}^{k} P(j \mid \lambda_0)$$

In particular, if no event has been found ($k = 0$), we obtain

$C = 1 - \exp(-\lambda_0)$ and thus

$\lambda_0 = -\ln(1 - C)$

For $C = 0.9$, the upper limit for the expectation value is λ_0=2.3 events

Unphysical parameters

- Sometimes, the allowed parameter range is restricted by physical laws (e.g., masses should be positive, though a specific experiment might yield a negative one).
- In this case, the prior is a step function,
  $P(\lambda)=0$ for $\lambda<0$ and constant for $\lambda>0$,
  and we have to adapt (restrict and to renormalize) the likelihood.

$$C(\lambda_0) = \frac{\int_{\lambda_{\min}}^{\lambda_0} L(\lambda) d\lambda}{\int_{\lambda_{\min}}^{\infty} L(\lambda) d\lambda}$$

- For confidence levels/regions for parameters derived from χ^2-minimization (in particular, parameter sets of higher dimension), see Numerical Recipes.

Tests of hypotheses

- So far, we have concentrated on the determination of unknown parameters from a sample ("measurement", estimation).
- Often, one has an expectation on the values of these parameters (e.g., from previous results or from models/theories)
- Then, the purpose of the sample is to test this **hypothesis** (see also Chap 7, section "likelihood", page 166 ff),

- **Example:**
  - In production control, one assumes that certain parameters are distributed normally about their nominal value. We consider the case of bread production. A certain type of bread should have a weight of 2kg, with a standard deviation of 20 g. From the daily production, we draw a random sample of ten breads, which display an *arithmetic mean* value of 1.99 kg, i.e., 0.01 kg below the nominal value.
  - If we assume the hypothesis to be true, then the mean difference (w.r.t. the nominal value) should be normally distributed, with mean 0 and standard deviation 20g/√10.
  - We now ask: What is the probability of "observing" a mean difference, $|\overline{\delta x}|$, ≥ 0.01 kg in our distribution? From Chap. 5, page 110, the corresponding probability is

$$P(\left|\overline{\delta \mathrm{x}}\right| \geq 0.01) = 1 - (2\psi_o(u) - 1) = 2(1 - \psi_o(u))$$

with $\psi_o(u)$ the cdf of the normal distribution and u the reduced random variable, $\mathrm{u} = (\mathrm{x} - \mu)/\sigma$.

Since we are dealing with mean values, the corresponding σ is a factor of $\frac{1}{\sqrt{\mathrm{N}}}$ smaller than the individual one, and we have $\mathrm{u} = 0.01/(0.02/\sqrt{10}) = \sqrt{10}/2 = 1.58$

Thus, $P(\left|\overline{\delta \mathrm{x}}\right| \geq 0.01) = 2(1 - \psi_o(1.58)) = 0.114$

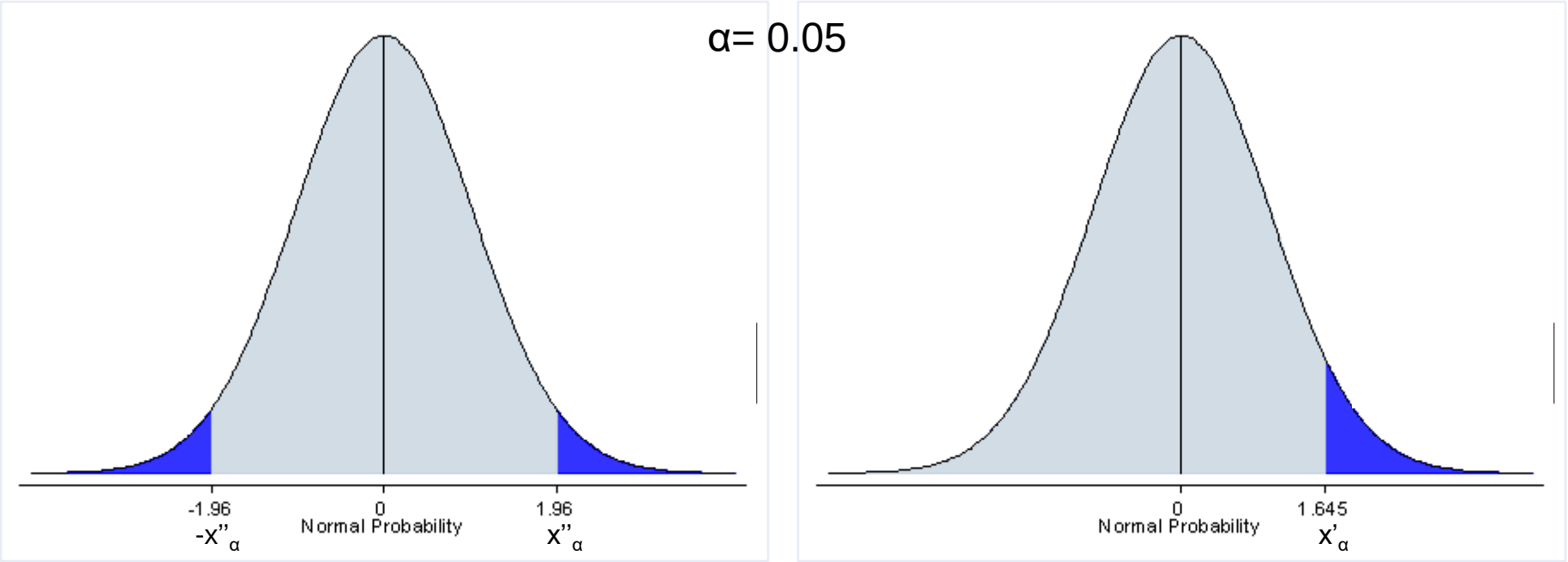
- Therefore, even is our hypothesis is true (i.e., that the weights are normally distributed around 2 kg with the given standard deviation), there is a probability of 11% that a sample of size 10 yields a mean that deviates by 0.01 kg or more from the nominal value (i.e., from the population mean)

- difficult to answer the simple question whether the hypothesis is true of false.
- but: all results in statistics are probability results. Therefore, we can improve the situation by introducing the concept of a **significance level.**
- We fix a certain (small) probability α (preferentially, before the sample has been analyzed).
- We then ask: Assuming the hypothesis to be true, is the probability of finding a sample with the observed properties larger or smaller than α? In the above example, our question would be whether $P(|\overline{\delta x}| \geq 0.01) < \alpha$
- If the probability is indeed smaller, we would conclude that it is unlikely for the assumed population to yield a sample as observed, and we would **reject** the hypothesis.
- Unfortunately, the reverse is **not** possible. If the probability exceeds α, one cannot say that the hypothesis is true, but only that it is not inconsistent with the result from the sampling and that there is no reason to reject it.

- Typical values of α are 10, 5,1 or 0.1%

One-tailed and two-tailed tests

- In some cases, the sign of the quantity in question is relevant. In our example, an underweight might be illegal and an overweight results in profit losses.
- Thus, we might test in one direction only and ask if $P(\bar{x} \geq x'_\alpha) < \alpha$
- This is called a one-tailed test, contrasted to the two-tailed test $P(|\bar{x}| \geq x''_\alpha) < \alpha$ from the example above



- Generally, quantities different from the mean might be tested.
- Then we define a corresponding test statistic T (i.e., a suitable function of the sample), and fix a significance level, α. For this level, we determine a sub-region, the co-called critical region S_c, within the complete definition region of T such that

$$P(\mathrm{T} \in S_c \mid H) = \alpha$$

- We draw a sample which yields a certain value for the test statistic. If this value falls inside S_c, we reject the hypothesis H (at a significance level α).

- more later on (page 270 ff)

F-test on equality of variances

- A quantity should be measured with two instruments which should have no systematic errors. We ask whether the measurements have the same variance (i.e., quality).
- We assume the populations to be normally distributed, and draw samples of size N_1 and N_2.
- Our hypothesis is that the variances are equal.
- We calculate the empirical variances (i.e., the bias-free estimators) s^2 for each sample, and consider the variance ratio.

$$F = \frac{s_1^2}{s_2^2}$$

If our hypothesis is true, $F \approx 1$. If the population is normally distributed, the quantities

$$X_i^2 = \sum_{j=1}^{N_i} \frac{(x_j - \bar{x})^2}{\sigma_i^2} = \frac{(N_i - 1)s_i^2}{\sigma_i^2} = \frac{f_i s_i^2}{\sigma_i^2}, \quad i = 1, 2$$

follow the χ^2 – distribution with $f_i = N_i - 1$ degrees of freedom. Assuming our hypothesis is true, $\sigma_1 = \sigma_2$, and

$$F = \frac{f_2}{f_1}\frac{X_1^2}{X_2^2} = \frac{s_1^2}{s_2^2}.$$

Using the distributions of X_i^2, the pdf of F can be calculated as

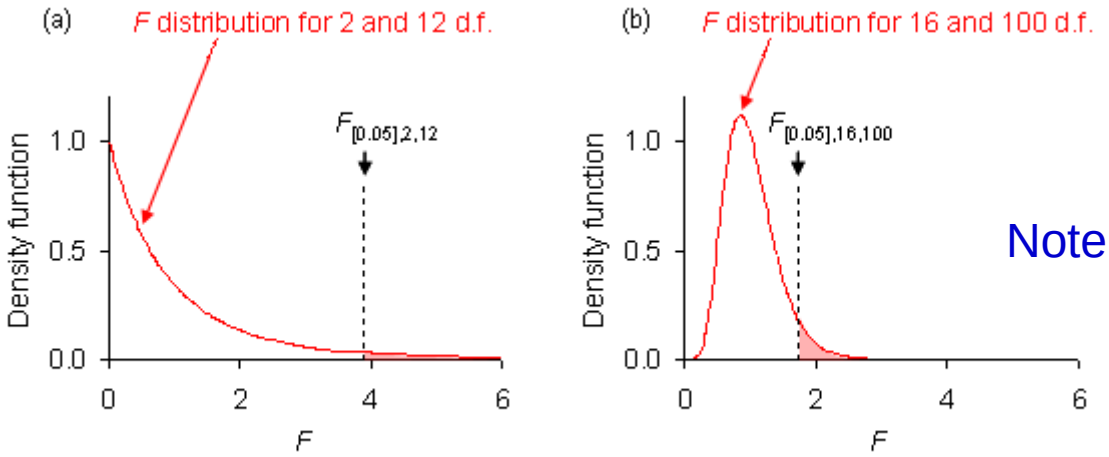
$$f(F) = \left(\frac{f_1}{f_2}\right)^{f_1/2} \frac{\Gamma((f_1 + f_2)/2)}{\Gamma(f_1/2)\Gamma(f_2/2)} F^{f_1/2-1} \left(1 + \frac{f_1}{f_2} F\right)^{-(f_1+f_2)/2}$$

The corresponding cdf

$$W(F) = P(\frac{s_1^2}{s_2^2} < F) = \int_0^F f(F')dF'$$

is called Fisher's F-distribution (sometime also Snedecor or variance-ratio distribution), and depends on F, f_1 and f_2.

Tables and programs to evaluate W(F) can be found in (almost) any textbook on statistics and within mathematical/graphical program packages, respectively.



Note that f(F) is asymmetric.

The shaded areas corresponds to a one-tailed significance level of 0.05, i.e., those variance ratios with $P(s_1^2/s_2^2 > F_{\alpha=0.05})=0.05$, where $F_{\alpha=0.05}$ is indicated by the arrows. By comparison with above, we have

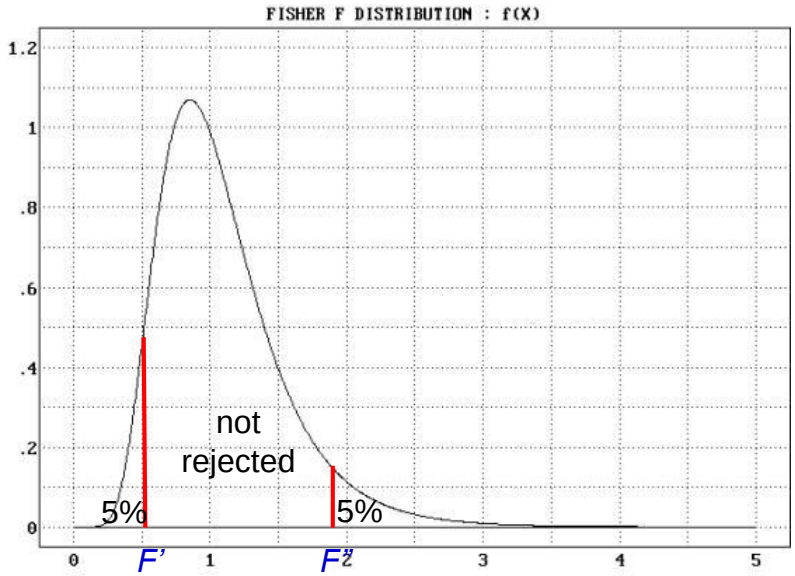
$$F_{\alpha=0.05} = F_{0.95} = F_{1-\alpha} \quad \text{and}$$
$$1-W(F_{1-\alpha})=\alpha \text{ or } W(F_{1-\alpha})=1-\alpha$$

Since we are testing for the equality of two test statistics, we have to perform a two-tailed test, i.e., we have (in principle) to calculate two cut-off values F' and F'' with

$$P\left(\frac{s_1^2}{s_2^2} < F'\right) = \alpha/2 \quad \text{and} \quad P\left(\frac{s_1^2}{s_2^2} > F''\right) = \alpha/2$$

We then would NOT reject the hypothesis of equal variances if $F' < F_{obs} < F''$.

Generally, $F' = F_{\alpha/2}(f_1, f_2)$ and $F'' = F_{1-\alpha/2}(f_1, f_2)$



$$F' = F_{\alpha/2}(f_1, f_2) = 0.511 \text{ and } F'' = F_{1-\alpha/2}(f_1, f_2) = 1.955 \text{ with } \alpha = 0.1,\ f_1 = 25,\ f_2 = 25$$

Because F is a ratio and because of certain properties of $W(F)$

[without proof : $F_{\alpha}(f_1, f_2) = 1/F_{1-\alpha}(f_2, f_1)$ and $F_{1-\alpha}(f_1, f_2) \geq 1$ for $\alpha \leq 0.1$],

for typical α's the two-tailed test can be performed in one rush, by testing whether

$$\frac{s_L^2}{s_S^2} > F_{1-\alpha/2}(f_L, f_S) \qquad (\alpha \leq 0.2) \quad ,$$

where L and S are the indices of the larger and smaller value of (s_1, s_2).

If the observed ratio is actually larger than the cutoff, then the hypothesis of equal variances needs to be rejected at a significance level of α, and we have actually performed a two-tailed test in the above sense, with $F'' = F_{1-\alpha/2}(f_L, f_S)$ and $F' = 1/F_{1-\alpha/2}(f_S, f_L)$

Example repeated measurements

Instrument 1	100	101	103	98	97	98	102	mean $\mu_1 = 99.86$	$s_1 = 2.268$	$f_1 = 7-1$
Instrument 2	97	102	103	96	100			mean $\mu_2 = 99.60$	$s_2 = 3.050$	$f_2 = 5-1$

$$F_{obs} = \frac{s_2^2}{s_1^2} = 1.81 \; < \; F_{0.95}(4,6) = 4.53 = F'' \qquad (F_{obs} > F' = F_{0.05}(4,6) = 1/F_{0.95}(6,4) = 0.16$$

simultaneously fulfilled)

Thus, the hypothesis that both samples have equal variances cannot be rejected at a 10% significance level.

We have no reason to believe that both instruments have a different quality.

Student's test – comparison of means

- Let's consider a normally distributed r.v. x.
- We draw a sample of size N with mean $\bar{x}$, which has a variance of $\sigma^2(\bar{x})= \sigma^2(x)/N$, which is unknown.
- The corresponding estimator is s^2/N.

We ask how much the reduced (random) variable

$$y = \frac{(\bar{x} - \mu)}{\sigma(\bar{x})}$$

deviates from the Gaussion when $\sigma(\bar{x})$ is replaced by the empirical standard deviation $s(\bar{x})=s(x)/\sqrt{N}$.

Let's shift coordinates to obtain $\mu=0$ (always possible). We consider the distribution of

$$t = \frac{\bar{x}}{s(\bar{x})} = \frac{\bar{x}\sqrt{N}}{s(x)}$$

Since (as in the previous section) $(N-1)s^2(x) = fs^2(x)$ is χ^2-distributed with $f = N-1$ d.o.f., we write

$t = \bar{x}\sqrt{\frac{Nf}{\chi^2}}$, with a distribution

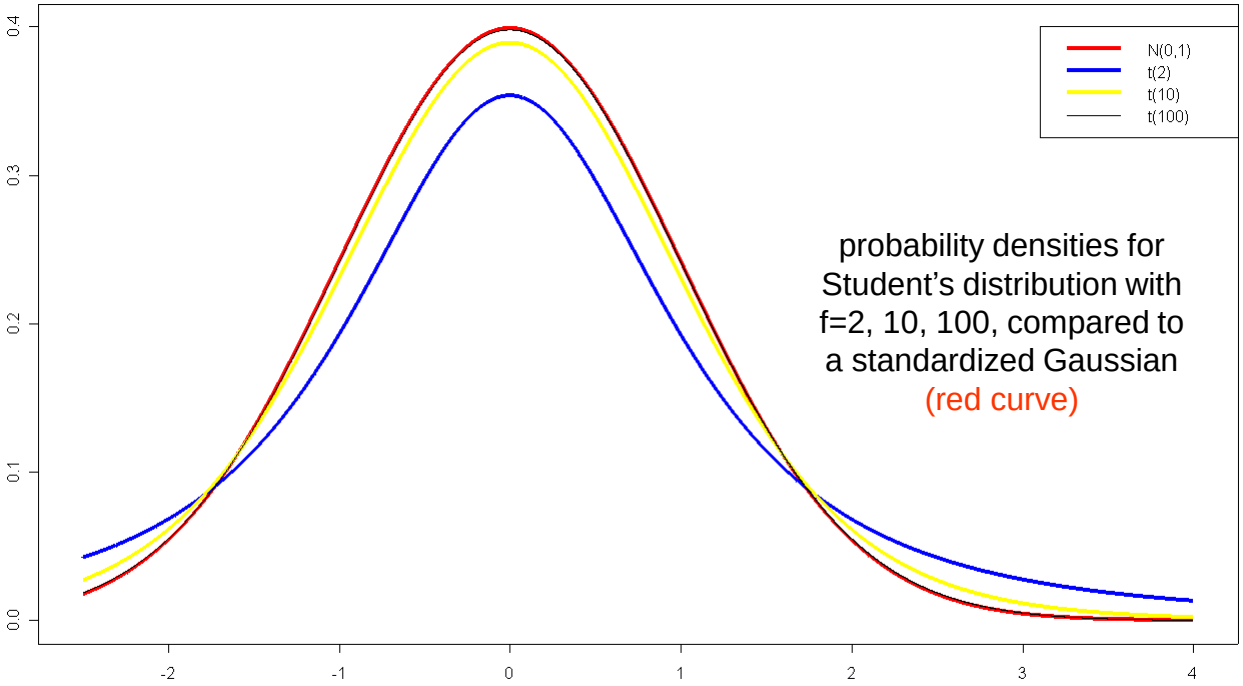
$$F(t) = P(t<t) = P(\bar{x}\sqrt{\frac{Nf}{\chi^2}} < t).$$

After some calculations, it turns out that

$$F(t) = \int_{-\infty}^{t} f(t)dt, \quad \text{with Student's distribution} \qquad \left(\text{published by W.S. Gosset under the pseudonym "Student"}\right)$$

$$f(t) = \frac{\Gamma((f+1)/2)}{\Gamma(f/2)\sqrt{\pi f}}\left(1+\frac{t^2}{f}\right)^{-(f+1)/2}$$

Student's distribution



- symmetric, bell-shaped, wider wings than Gaussian, core narrower
- *f(t)*→ Gaussian for d.o.f.→∞
- tabulated in all text-books etc.
- P(|t|≤*t*)=2F(*t*)-1, as for Gaussian

$$\text{bounds } \pm t'_\alpha \text{ from } \int_0^{t'_\alpha} f(t)dt = \frac{1}{2}(1-\alpha) \text{ for a given significance level } \alpha,$$

$$\text{i.e., } F(t'_\alpha) = \int_{-\infty}^{t'_\alpha} f(t)dt = \frac{1}{2} + \frac{1}{2}(1-\alpha) = 1 - \frac{\alpha}{2} \Rightarrow t'_\alpha = t_{1-\alpha/2}$$

Student's test: testing the mean

- A hypothesis predicts a certain expectation value μ.

A sample of size N is drawn, and $\bar{x}$ and s^2 calculated

If the inequality

$$|t| = \frac{|\bar{x} - \mu|\sqrt{N}}{s} > t'_\alpha = t_{1-\alpha/2}$$

is fulfilled, the hypothesis must be rejected at the α-level.

- **example** (continuation from page 262)
- hypothesis: the parent distribution has an expectation value μ=100 (use data from instrument 1)

$$t = \frac{|99.86 - 100|\sqrt{7}}{2.268} = 0.163 < 1.94 = t_{0.95} \quad \text{(for 6 d.o.f.)}$$

- The hypothesis cannot be rejected at the 10% level.
- If the hypothesis were μ=102, → t=2.50, and the hypothesis would need to be rejected.

Student's difference test: comparing two means

- From two populations samples of size N_1 and N_2 have been drawn. We like to test whether the population means are equal, i.e., whether <x> = <y>.

As before, the empirical variances are $s^2(\bar{x})=s^2(x)/N_1$ and $s^2(\bar{y})=s^2(y)/N_2$.

Because the sum of two normally distributed r.v.s is normally distributed as well (see exercises), also the difference $\Delta = \bar{x} - \bar{y}$ should be (nearly) normally distributed, with $s^2(\Delta) = s^2(\bar{x}) + s^2(\bar{y})$.

Our hypothesis of equal means implies $E(\Delta) = 0$, and we have to test for $t=\dfrac{(\Delta - 0)}{s(\Delta)}$.

Usually the hypothesis of equal means also implies that the samples have been drawn from the same parent population, thus $\sigma^2(x) = \sigma^2(y)$, and the best estimator for the common ("pooled") empirical variance, s_p^2 is the weighted average [with weights $\propto (N_1 - 1)$ and $(N_2 - 1)$],

$$s_p^2 = \frac{(N_1 - 1)\, s^2(x) + (N_2 - 1)s^2(y)}{(N_1 - 1) + (N_2 - 1)} \quad \Rightarrow \quad s(\Delta) = \sqrt{s^2(\bar{x}) + s^2(\bar{y})} = \sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}} = s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

The quotient

$$t = \frac{\Delta}{s(\Delta)} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}}$$

follows the Student distribution with $f = N_1 + N_2 - 2$ degrees of freedom. If the hypothesis of equal means needs to be rejected, one can assume that $E(x) > E(y)$ or $E(y) > E(x)$, depending on the sign of $\Delta = \bar{x} - \bar{y}$.

- **Note:** even if the distributions are not accurately Gaussian, Student's difference test works remarkably well, i.e., it is a **robust** test

- **example** (continuation from page 262)
- hypothesis: the two samples measured by instrument 1 and 2 are drawn from the same population.
  - equal variances could not be rejected (Fisher test)
  - test now equal means

$$s_p^2 = \frac{6 \cdot 2.268^2 + 4 \cdot 3.05^2}{6+4} = 6.8073, \quad s_p = 2.609$$

$$t = \frac{|99.86\text{-}99.60|}{2.609\sqrt{\frac{1}{7}+\frac{1}{5}}} = 0.17 < 1.81 = t_{0.95} \quad \text{for 10 d.o.f.}$$

The hypothesis cannot be rejected at a significance level of 10%.

For unequal variances, the unpooled t-test needs to be applied:

$$t = \frac{\Delta}{s(\Delta)} = \frac{\bar{x} - \bar{y}}{\sqrt{s^2(\bar{x}) + s^2(\bar{y})}} \quad \text{with} \quad f = \frac{\left(s^2(\bar{x}) + s^2(\bar{y})\right)^2}{\frac{\left(s^2(\bar{x})\right)^2}{N_1 - 1} + \frac{\left(s^2(\bar{y})\right)^2}{N_2 - 1}} \quad \text{d.o.f}$$

and $s^2(\bar{x}) = \frac{s^2(x)}{N_1}, \quad s^2(\bar{y}) = \frac{s^2(y)}{N_2}$

For the previous example (assuming now unequal variances), we obtain

$$t = \frac{|99.86 - 99.60|}{\sqrt{\frac{2.268^2}{7} + \frac{3.05^2}{5}}} = 0.161, \quad f = 7(.05), \; t_{0.95}(f = 7) = 1.89$$

Even if the variances were different, a hypothesis of equal means cannot be rejected at $\alpha = 0.1$.

General aspects – the null hypothesis

- Any hypothesis consists of an assumption on the pdf $f(x;\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p) = f(x; \boldsymbol{\lambda})$
- A hypothesis is called **simple** if the pdf f is completely specified (i.e., regarding all parameters λ_i).
- A hypothesis is called **composite** if the form of f is known but at least one of the parameters remains unspecified.
- The hypothesis to be tested is called the **null hypothesis,** H_0**.**
  - Remember that we can only *reject* a hypothesis (at a given significance). If we actually want to *confirm* a specific hypothesis, the null hypothesis must be the opposite and needs to be rejected by the test.
  - **example**: We want to show that $\lambda_1 \neq \lambda_2$. Thus, H_0 is $\lambda_1 = \lambda_2$. If we can reject H_0, we can "confirm" our actual hypothesis.
  - This approach has been/is criticized by different groups, particularly from the Bayesian school.
  - Also *publication bias:* journals and reviewers have developed a bias against articles that do *not* reject the null hypothesis. In connection with the *file drawer problem*, other researchers might waste their time by examining questions that have already been examined.
  - The *file drawer problem* exists due to the fact that academics tend not to publish results that indicate the null hypothesis could not be rejected. These results mostly end up unpublished, in file drawers.

Critical region

A reasonable H_0 should (but not necessarily needs to) be simple, i.e. $H_0(\boldsymbol{\lambda} = \boldsymbol{\lambda}_0)$.
Any other hypothesis is called alternative hypothesis, and can be simple
$H_1(\boldsymbol{\lambda} = \boldsymbol{\lambda}_1)$ or composite $\left(\text{e.g., } H_1(\boldsymbol{\lambda} \neq \boldsymbol{\lambda}_0)\right)$.

H_0 predicts the probability distribution in sample space, i.e., the probability of "observing" a point $X = (x^{(1)}, x^{(2)}, x^{(3)}, ..., x^{(N)})$ in any region of the possible sample space. We define a critical region S_c at a significance level α by requiring

$$P(X \in S_c \mid H_0) = \alpha,$$

i.e., that the probability of observing a point X inside S_c under the condition that H_0 is true is equal to α. If in a specific experiment X actually falls inside S_c, we reject H_0. Note that the above requirement does not uniquely determine S_c [simple example: two-tailed vs. single-tailed (upper, lower tail)].
For actual tests, we form a test statistic,

$$T = T(X) = T(x^{(1)}, x^{(2)}, x^{(3)}, ..., x^{(N)}),$$

and determine the corresponding critical region by mapping

$$X \to T(X), \quad S_c(X) \to U(X).$$

In this case, the null hypothesis is rejected if $T \in U$.

Errors of first and second kind

H_0 might be true even if the hypothesis was rejected because of $X \in S_c$. This is called an error of the first kind (or type I error), and its probability is equal to α by definition.

There is another possibilty to make a wrong decision, namely not to reject the hypothesis, because $X \notin S_c$, although H_0 is false and an alternative H_1 is true. This is called an error of the second kind (or type II error), with probability ("false negative rate")

$$P(X \notin S_c \mid H_1) = \beta,$$

which depends on the particular H_1. Thus, for a meaningful S_c, β should be small, or, the other way round, the power of the test,

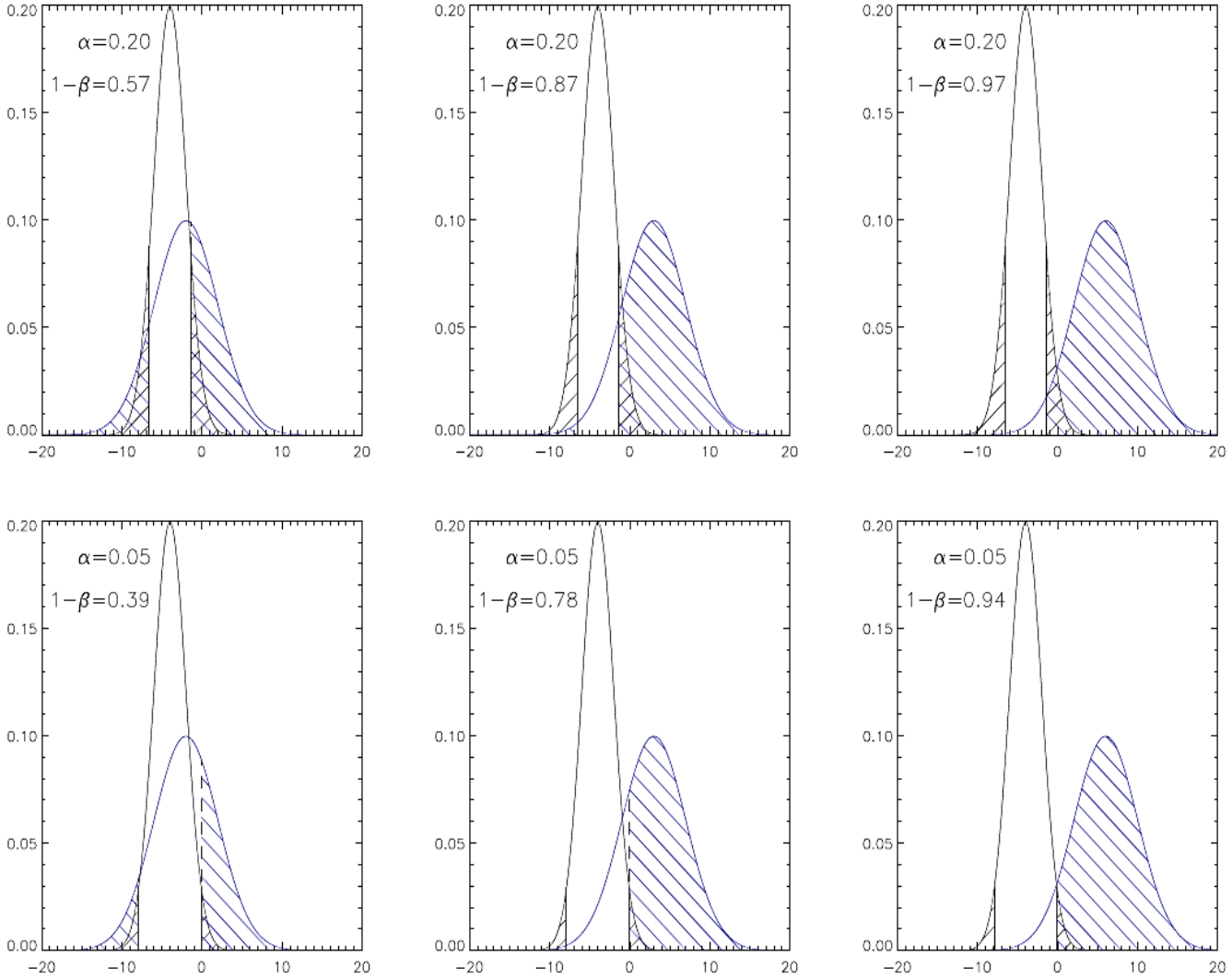
$$1 - \beta = P(X \in S_c \mid H_1),$$

should be large. If there is only one alternative hypothesis, the power is the probability of correctly rejecting H_0.

- A simple **example**
  An accused in front of a law court proclaims that he is innocent
  - H_0 : the accused is innocent
  - H_1 : the accused is guilty
  - Both the null and the alternative hypotheses are simple ones
- if the accused is innocent but wrongly convicted, this is a type I error
- if he is declared innocent but actually is guilty, this is a type II error

Significance and power

a simple null hypothesis and a simple alternative



H_0 = normal distribution,
with $\mu = -4, \sigma = 2$ (black)

H_1 = normal distribution,
$\mu_i = [-2, 3, 6], \sigma = 4$ (blue)
(from left to right)

significance level, two tailed
α=[0.2,0.05] (top and bottom)

indicated are the critical regions
$S_c = ([-\infty, x_{\alpha/2}] \bigcup [x_{1-\alpha/2}, \infty])$,
bounds displayed by vertical lines

with corresponding probability
$P(x \in S_c \mid H_0) = \alpha$:
significance, black shaded

and
$P(x \in S_c \mid H_1) = 1 - \beta$:
power, blue shaded

The false negative rate, β
(= prob. of type II error) corresponds to the area under the blue curve within the vertical lines

Summary of definitions

- **significance** (level) of a test, **α**: probability of type I errors (reject true H_0)
  - upper bound of probability for composite hypotheses
  - should be small, usually needs to be defined by the person who performs the test (typical value: 0.05)
  - a result is **significant** if the prob. that it could have arisen by chance from H_0 is small
- region of rejection / **critical region**, S_c
  - the set of values of the test statistic for which the null hypothesis is rejected (at a certain significance level)
- region of acceptance
  - the set of values of the test statistic for which we fail to reject the null hypothesis (at the above significance level)
- false negative rate, β: probability of (specified) alternative H_1 outside rejection region of H_0 = probability of type II error w.r.t. H_1 (accept H_0 though H_1 is true)
  - should be small
- **power of test**, 1-β: probability of alternative (H_1) in rejection region of H_0
  - probability of correctly rejecting H_0 (if only one alternative hypothesis exists)
  - should be large

Power function etc.

Once the critical region has been specified, one can calculate the probability of rejecting H_0 as a function of the "true" hypothesis or the parameters decribing it. This probability is called the power function of the test and is given by

$$M(S_c, \boldsymbol{\lambda}) = P(\mathrm{X} \in S_c \mid H) = P(\mathrm{X} \in S_c \mid \boldsymbol{\lambda}).$$

The complementary probability of "accepting" (i.e., not rejecting) H_0 as a function of the "true" hypothesis or its parameters is called the acceptance probability or the operational characteristic of the test,

$$L(S_c, \boldsymbol{\lambda}) = 1 - M(S_c, \boldsymbol{\lambda}).$$ Obviously,

$$M(S_c, \boldsymbol{\lambda}_0) = \alpha \qquad M(S_c, \boldsymbol{\lambda}_1) = 1 - \beta$$
$$L(S_c, \boldsymbol{\lambda}_0) = 1 - \alpha \qquad L(S_c, \boldsymbol{\lambda}_1) = \beta$$

Some possible test properties:

most powerful: $M(S_c, \boldsymbol{\lambda}_1) = 1 - \beta = \max$ (when testing a *simple* H_0 relative to a *simple* alternative H_1, implies an optimum S_c).

uniformly most powerful: the test is most powerful w.r.t. *any possible alternative* (particularly also composite ones).

unbiased: $M(S_c, \boldsymbol{\lambda}) \geq \alpha, \quad \boldsymbol{\lambda} \neq \boldsymbol{\lambda}_0$ reasonable since the probability of rejecting H_0 should be smallest when H_0 is true.

The Neyman-Pearson theorem

We define a conditional pdf

$$f(X \mid H_0)$$

where X is a "point" in sample space with coordinates $(x^{(1)},x^{(2)},x^{(3)},...,x^{(N)})$
(in the case of one random variable. Generalization to multi-variate processes is obvious).
Then we have

$$\int_{S_c} f(X \mid H_0)dX = P(\mathrm{X} \in S_c \mid H_0) = \alpha$$

Neyman-Pearson theorem:

A test of the *simple* hypothesis H_0 relative to the *simple* alternative H_1 is *most powerful* if the critical region S_c is such that

$$\frac{f(X \mid H_0)}{f(X \mid H_1)} \begin{cases} \leq c \text{ for each } \mathrm{X} \in S_c \\ \geq c \text{ for each } \mathrm{X} \notin S_c \end{cases}$$

where c is a (positive) constant depending on the significance level.

Proof: e.g., Brandt, Chap. 8.4.

Idea: show for arbitrary $S \neq S_c$ that $M(S_c,\boldsymbol{\lambda}_1) \geq M(S,\boldsymbol{\lambda}_1)$, using the above relations.

Example

- From Barlow, Chap. 8.2:
  The density of opal is 2.2 g/cm^3, and for quartz it is 2.6 g/cm^3. Various sites produce small quantities of crystals which could be either, since their density is measurable with a resolution of 0.2 g/cm^3 only. Which are worth the expense of further excavation?
- In formal terms, we have to perform a test on the hypothesis that a normal population of a given variance σ^2 has the mean $\mu=\mu_0$ instead of $\mu=\mu_1$, from a sample of size N=1 (see next figure).
- Both hypotheses are simple, so the Neyman-Pearson theorem can be applied

The conditional probabilities of the drawn sample point $X = x^{(1)}$ are

$$f(X \mid H_0) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x^{(1)} - \mu_0)^2}{2\sigma^2} \quad \text{for } H_0 \text{ and}$$

$$f(X \mid H_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x^{(1)} - \mu_1)^2}{2\sigma^2} \quad \text{for the alternative hypothesis with } \mu=\mu_1.$$

The quotient becomes

$$\frac{f(X \mid H_0)}{f(X \mid H_1)} = \exp\left(-\frac{1}{2\sigma^2}\left((\mu_0{}^2 - \mu_1{}^2) - 2x^{(1)}(\mu_0 - \mu_1)\right)\right) = k \exp\left(\frac{x^{(1)}(\mu_0 - \mu_1)}{\sigma^2}\right) \quad \text{with } k > 0$$

Note that this ratio increases monotonically with $x^{(1)}$ for $\mu_0 > \mu_1$ and decreases monot. for $\mu_0 < \mu_1$.
Thus, it is always possible to find a region where the ratio will remain below a certain threshold.
In particular, the Neyman-Pearson condition takes the form

$$k \exp\left(\frac{x^{(1)}(\mu_0 - \mu_1)}{\sigma^2}\right) \quad \begin{cases} \le c \text{ for } X \in S_c \\ \ge c \text{ for } X \notin S_c \end{cases}$$

which is equivalent to

$$x^{(1)}(\mu_0 - \mu_1) \begin{cases} \leq c' \text{ for } X \in S_c \\ \geq c' \text{ for } X \notin S_c \end{cases}$$

This places the required condition on S_c, and suggests that a suitable test statistic is $x^{(1)}$ itself.
Exercise: Show that for the same problem but a sample size $N > 1$ one obtaines a similar inequality, with $x^{(1)}$ replaced by the mean $\bar{x}$ (which then would serve as the test statistic).

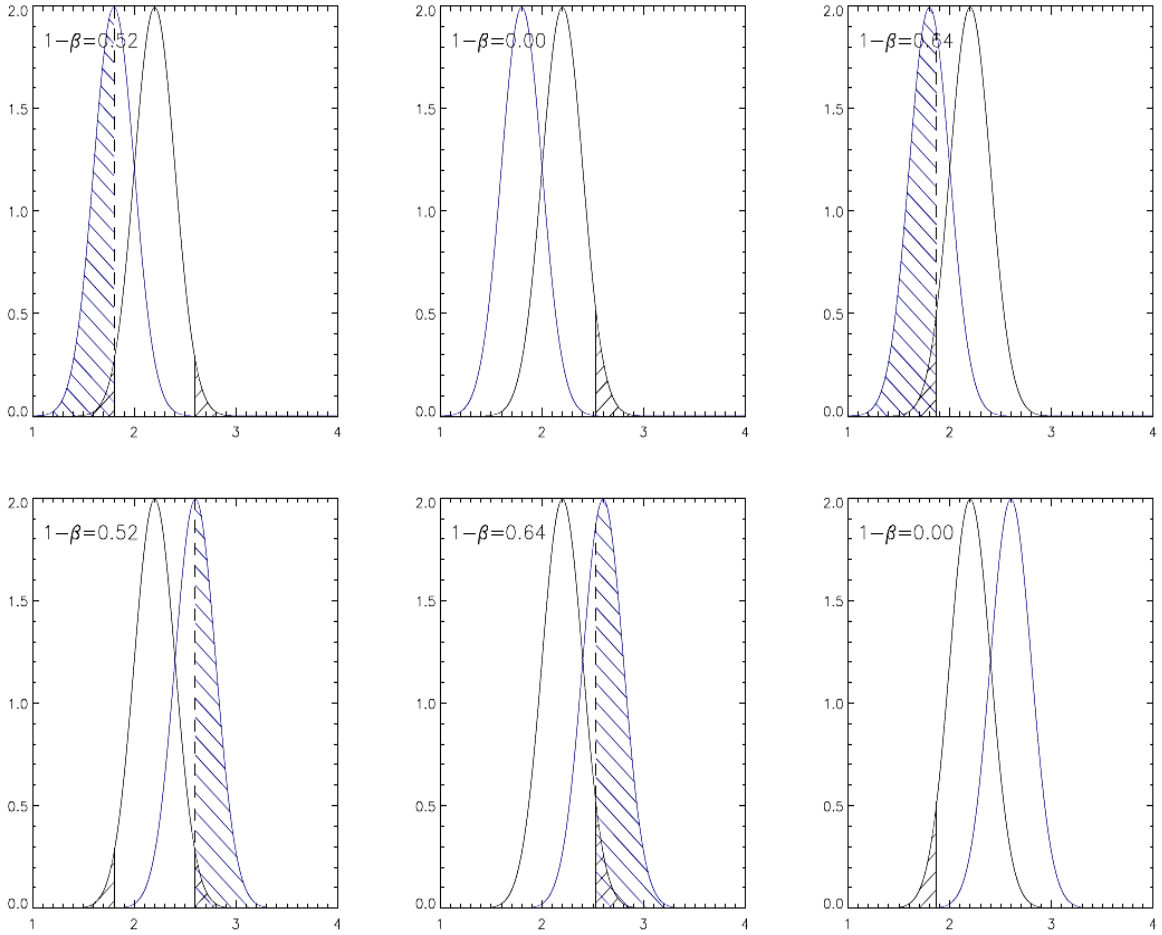
From above, we find that the critical region corresponding to a most powerful test depends on the sign of $(\mu_0 - \mu_1)$.
For $\mu_0 > \mu_1$, the critical region is "on the left", $x^{(1)} \leq c$" [quotient montonically increasing], and for $\mu_0 < \mu_1$, the critical region is "on the right", $x^{(1)} \geq c$" [quotient montonically decreasing].

Thus, the most powerful test on the above hypothesis is a one-tailed one, with S_c depending on wether $(\mu_0 - \mu_1) > 0$ or <0, respectively.
In so far, a uniformly most powerful test does not exist for this problem.

Note, however, that a two-tailed test can deal fairly well (i.e., with not significantly less power) with both possibilities (see figure next page).

Opal or quartz?

most powerful test according to the Neyman-Pearson theorem (see previous example)



upper panel: quotient $\frac{f(x\,|\,H_0)}{f(x\,|\,H_1)}$ monotonically increasing; lower panel: quotient monotonically decreasing.

H_0 : the sample is opal
$\Rightarrow$ normal distribution,
with $\mu = 2.2, \sigma = 0.2$ (black)

H_1 : normal distribution,
$\mu = 1.8, \sigma = 0.2$ (blue, upper panel, for comparison)
$\mu = 2.6, \sigma = 0.2$ (blue, lower panel, corresponding to quartz)
significance level α=0.05

from left to right: critical regions for
1. two-tailed test, $S_c = ([-\infty,\ x_{\alpha/2}] \bigcup [x_{1-\alpha/2}, \infty])$
2. one-tailed test, $S_c = [x_{1-\alpha}, \infty]$
3. one-tailed test, $S_c = [-\infty,\ x_{\alpha}]$

bounds displayed by vertical lines

significance, black shaded
power 1-β, blue shaded

For $\mu_0 > \mu_1$ (upper panel), the most powerful test comprises a critical region according to 3.
For $\mu_0 < \mu_1$ (lower panel), the most powerful test comprises a critical region according to 2.

- If we allow for a rejection of 5% of genuine opals (type I error), the most powerful test gives 1-β=0.64, i.e., a 36% probability of type II errors: 36% of the quartz deposits will be needlessly investigated.
- In this case, we reject all samples with a measured density of **larger** than 2.53 g/cm^3, which is 1.64σ above the opal density.
- In reality, one has to make a compromise regarding the importance of type I (ignoring valuable resources) vs. type II (needless expenses) errors.
- If, e.g., the costs for test excavations are very high, one might want to reduce β at the expense of discarding more genuine opal sites. E.g., for α=0.15 (corresponding to an upper bound of 2.41 g/cm^3 (= 1.04σ), β is reduced to 17%

Likelihood ratio test

- The Neyman-Pearson theorem gives a condition for a most powerful test when comparing two simple hypotheses.
- In general such a test does not exist if the alternative hypothesis is composite and comprises parameters which are both smaller and larger than the one of H_0 (no *uniformly* most powerful test possible).
- There is, however, a method which yields tests with desirable properties, based on a specific likelihood ratio.
  - Note that the quotient entering the Neyman-Pearson theorem is nothing else than the likelihood ratio Q defined in Chap. 7 (page 173) for the comparison of two simple hypotheses.

The null hypothesis shall be defined by a certain range of parameters, $\boldsymbol{\lambda} \in \Lambda_0$. The most general alternative is then described by that part of the total parameter space Λ which does not contain Λ_0, i.e., $\Lambda_1 = \Lambda - \Lambda_0$. The **likelihood ratio test** defines a statistic

$$T = \frac{f(x^{(1)},x^{(2)},x^{(3)},...,x^{(N)};\tilde{\boldsymbol{\lambda}}(\Lambda))}{f(x^{(1)},x^{(2)},x^{(3)},...,x^{(N)};\tilde{\boldsymbol{\lambda}}(\Lambda_0))}$$

where $\tilde{\boldsymbol{\lambda}}(\Lambda)$ and $\tilde{\boldsymbol{\lambda}}(\Lambda_0)$ are the maximum likelihood estimates (Chap. 7) for the parameter set $\boldsymbol{\lambda}$, evaluated for the parameter regions $\Lambda = \Lambda_0 + \Lambda_1$ and Λ_0, respectively.

Note 1: sometimes, T is defined by the inverse ratio.

Note 2: $f(x^{(1)},x^{(2)},x^{(3)},...,x^{(N)};\boldsymbol{\lambda})$ is the joint probability density of the sample $X = (x^{(1)},x^{(2)},x^{(3)},...,x^{(N)})$ *given* $\boldsymbol{\lambda}$, i.e., the likelihood function $L(\boldsymbol{\lambda} \mid x^{(1)}, x^{(2)}, x^{(3)},..., x^{(N)})$, and the values entering the quotient are the corresponding maximum likelihoods.

Thus, we can alternatively (and briefly) write

$$\mathrm{T} = \frac{L_{\max}(\boldsymbol{\lambda} \in \Lambda)}{L_{\max}(\boldsymbol{\lambda} \in \Lambda_0)} = \frac{L(\tilde{\boldsymbol{\lambda}})}{L(\tilde{\boldsymbol{\lambda}}_0)}$$

If, e.g., the overall maximum of L lies inside Λ_0, we have $\mathrm{T} = 1$. In this case, the null and the alternative hypothesis have an equal likelihood, and we cannot define a rejection region.
If, on the other hand, the null hypothesis is simple, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$, the parameter space Λ_0 degenerates into the point $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$, and the statistic becomes

$$\mathrm{T} = \frac{L_{\max}(\boldsymbol{\lambda} \in \Lambda)}{L(\boldsymbol{\lambda}_0)}.$$

Within the likelihood ratio test, we now reject H_0 if $\mathrm{T} > T_{1-\alpha}$, with

$$P(\mathrm{T} > T_{1-\alpha} \mid H_0) = \int_{T_{1-\alpha}}^{\infty} g(T \mid H_0)\mathrm{d}T$$

and $g(T \mid H_0)$ the conditional pdf of the statistic T.

Wilks' theorem (1938). For large N, the statistic $\mathrm{T}' = 2\ln\mathrm{T}$ follows a χ^2-distribution, with $f = \max(1, p - r)$ d.o.f., when p is the number of parameters and r out of those have been specified in H_0.

Usually, T is easily determined, and the only problem is to derive its distribution, if one does not rely on the asymptotic behaviour of $2\ln\mathrm{T}$. For further details, see Brandt, Chap. 8.5.

Example 1

Test the hypothesis that a normal population with variance σ^2 has the mean $\mu=\mu_0$ via the likelihood ratio test.

Parameters for H : $\tilde{\mu}(\Lambda) = \bar{x}$ (maximum likelihood estimator, Chap. 7) $\tilde{\sigma}^2(\Lambda) = \sigma^2$

Parameters for H_0 : $\tilde{\mu}(\Lambda_0)=\mu_0$, $\tilde{\sigma}^2(\Lambda_0)=\sigma^2$

$$\Rightarrow L(\tilde{\lambda}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{2\sigma^2}\right); \quad L(\tilde{\lambda}_0) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\sum_{i=1}^{N}\frac{(x_i-\mu_0)^2}{2\sigma^2}\right)$$

$$\mathrm{T} = \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{\sigma^2} - \frac{(x_i-\mu_0)^2}{\sigma^2}\right) = \exp\left(-\frac{1}{2}\frac{2(\mu_0-\bar{x})\sum_{i=1}^{N}x_i + N(\bar{x}^2-\mu_0^2)}{\sigma^2}\right) = \exp\left(\frac{1}{2}\frac{N(\mu_0-\bar{x})^2}{\sigma^2}\right)$$

$$2\ln \mathrm{T} = \frac{N(\mu_0-\bar{x})^2}{\sigma^2} = \frac{(\bar{x}-\mu_0)^2}{\sigma^2/N} = \mathrm{u}^2$$

Since the x_i are normally distributed, also the mean $\bar{x}$ is normally distributed, with $E(\bar{x})=\mu_0$ and $Var(\bar{x}) = \sigma^2/N$.

Thus, u is the square of a normally distributed reduced variable, and thus χ^2-distributed with $f = 1$ d.o.f.

In this case, Wilks' theorem ($2\ln \mathrm{T}$ is χ^2-distributed with $f = \max(1, 2-2) = 1$) is valid even in the non-asymptotic case, i.e., for all $N \geq 1$.

Example 2

Test the hypothesis that a normal population of unknown variance has the mean $\mu=\mu_0$ via the likelihood ratio test.

Parameters for H : $\tilde{\mu}(\Lambda)=\bar{x}$, $\quad \tilde{\sigma}^2(\Lambda)=s'^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$, maximum likelihood estimators (Chap. 7)

Parameters for H_0 : $\tilde{\mu}(\Lambda_0)=\mu_0$, $\quad \tilde{\sigma}^2(\Lambda_0)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_0)^2$

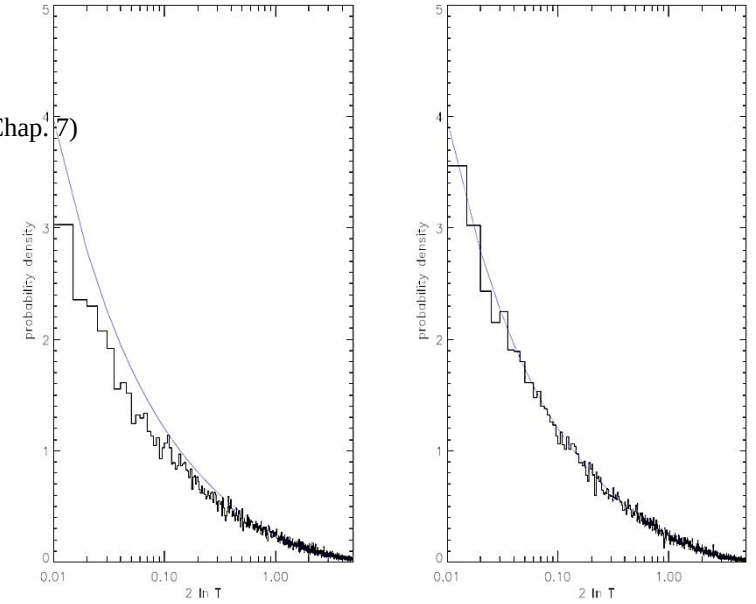
$$\Rightarrow L(\tilde{\lambda})=\left(\frac{N}{2\pi\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)^{N/2}\exp\left(-\sum_{i=1}^{N}\frac{(x_i-\bar{x})^2}{\frac{2}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)=\left(\frac{N}{2\pi\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)^{N/2}\exp\left(-\frac{N}{2}\right)$$

$$L(\tilde{\lambda}_0)=\left(\frac{N}{2\pi\sum_{i=1}^{N}(x_i-\mu_0)^2}\right)^{N/2}\exp\left(-\sum_{i=1}^{N}\frac{(x_i-\mu_0)^2}{\frac{2}{N}\sum_{i=1}^{N}(x_i-\mu_0)^2}\right)=\left(\frac{N}{2\pi\sum_{i=1}^{N}(x_i-\mu_0)^2}\right)^{N/2}\exp\left(-\frac{N}{2}\right)$$

$$T=\left(\frac{\sum_{i=1}^{N}(x_i-\mu_0)^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)^{N/2}=\left(\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2+N(\bar{x}-\mu_0)^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)^{N/2}=\left(1+\frac{t^2}{N-1}\right)^{N/2} \quad \text{with}$$

$t=\frac{(\bar{x}-\mu_0)\sqrt{N}}{s}$ the Student's test statistic. To test the null hypothesis, we can either (for large N) compare the sample's value of $2\ln T$ with the bound from the corresponding χ^2-distribution (left fig.). Alternatively, we can use t instead of T (since t(T) is a monotonic function of T, $T>0$), and reject the null hypothesis for a specific sample if $|t|>t_{1-\alpha/2}$.

In so far, the general likelihood ratio test is consistent with our former t-test on the mean of normal samples with unknown variance.



Wilks' theorem for example 2:
histogram of the distribution of 2lnT, for 50000 samples with sizes N=5 (left panel) and N=30 (right panel).
Overplotted in blue is the χ^2 distribution with (p-r)=2-1=1 degree of freedom. For large N, 2lnT becomes χ^2 distributed.

Exercise: Show that to test the hypothesis that a normal population of unknown mean has a variance σ_0^2, the appropriate test statistics is

$$T=\left(\frac{\sigma_0}{s'}\right)^N\exp\left(\frac{N}{2}\left(\frac{s'^2}{\sigma_0^2}-1\right)\right)$$ (*Note*: error in Brandt)

Goodness of fit (Fitgüte)

- So far, we have discussed **parameter tests**, i.e., tests on hypotheses which specify one or more parameters of a population.
  - Many more tests can be found in the literature, e.g.,
  - the analysis of variances (ANOVA): comparing parameters from several samples
  - "contingency table" analysis (particularly suited for non-numeric data)
- Null hypothesis is accepted or rejected on the basis of *one single number,* the test statistic, determined from the sample.
- Now: **tests of fit**. Here, we compare the distribution function of the sample directly with the (supposed) distribution of the population, or the distributions of two samples.
- The χ^2-test described in the following is a generalization of what has been already discussed in Chap 8 (Goodness of fit, binned data)

The χ^2-test for data pairs

- Assume that the data consist of a set of data x_i and y_i, where the x_i are exact and the y_i have measurement errors σ_i. The hypothetical function y=f(x) claims to predict the actual relation between x and y.

To test the hypothesis, we calculate the test statistic X^2,

$$X^2 = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i)\right)^2}{\sigma_i^2}$$ for independent measurement errors, and

$$X^2 = (\mathbf{y} - \mathbf{f})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{f})$$ for correlated errors, with $\mathbf{V}$ the corresponding covariance matrix (Chap. 6).

If the hypothesis is true, the differences in the nominators should be only due to measurement errors. Then, the quantity X^2 is the sum of N squares of reduced, normally distributed variables, and should be distributed according to the χ^2-distribution (Chap. 5 and 8). The hypothesis has to be rejected if

$$X^2 > \chi^2_{1-\alpha},$$

where $P(X^2 > \chi^2_{1-\alpha}) = \int_{\chi^2_{1-\alpha}}^{\infty} g(u, f)\,du = \alpha$

with $g(u, f)$ the probability density of the χ^2-distribution with $f = N$ d.o.f.

The so-called χ^2-probability $\chi^2_{1-\alpha}$ is tabulated in most text-books (as a function of f and α), and included in various program-packages.

If X^2 is too large to be accepted, there might be three reasons for this:

i) $f(x)$ is not the correct description for y, i.e., the "theory" is wrong
(this disproof is the original intention of the test)

ii) the errors σ_i might have been underestimated (check!)

iii) the errors are not Gaussian (seldom, but small samples!)

Additionally, one needs to check whether the obtained X^2 is not too small

(typically, when $P(\chi^2 < X^2) = \int_0^{X^2} g(u, f)\mathrm{d}u \leq 10^{-3}$ see Chap. 8).

In this case, there is the possibility that either the σ_i have been overestimated, or that the data have been faked. If possible, one should analyze a second sample.

If $f(x)$ depends on p parameters which have been *fitted* from the sample, these leads to constraints in the above sum of squares, and the number of d.o.f. has to be reduced by p, i.e., $f = N - p$.

It is recommended to always carry out such a χ^2-test after relations have been fitted from a sample, e.g., after a straight line fit has been performed (in the latter case, $f = N - 2$), and the variances are due to real measurement errors and not estimated from the fit itself (otherwise, the χ^2 will be always OK, see Chap. 8)

The χ^2-test for binned data

We like to test whether the population is distributed with a probability density $f(x)$.
The total range of the r.v. x is binned into r intervals $\xi_1, \xi_2, \xi_3, ..., \xi_k, ..., \xi_r$. Hypothetically, the probability of observing x in a specific ξ_k is given by

$$p_k = P(x \in \xi_k) = \int_{\xi_k} f(x)\mathrm{d}x, \qquad \text{with} \quad \sum_{k=1}^{r} p_k = 1$$

We draw a sample of size N, with n_k the number of elements of the sample which fall inside ξ_k.

Obviously, $\sum_{k=1}^{r} n_k = N$.

From the hypothetical probability density, we expect

$E(n_k) = Np_k$.

As a measure for the deviation of the sample distribution form the hypothetical one, we use

$$X^2 = \sum_{k=1}^{r} \frac{(n_k - E(n_k))^2}{\sigma^2(n_k)} = \sum_{k=1}^{r} \frac{(n_k - Np_k)^2}{Np_k}$$

The test statistic X^2 is asymptotically χ^2 – distributed, with $f = r - 1$ d.o.f.

idea of proof: The population probability of the individual channels can be either described by a multinomial process (Chap. 5), or by r independent variables following a Poisson distribution with mean *and* variance Np_k (Chap. 5). Both processes give identical population probabilities, independent from the sample size.

Now, by virtue of the CLT, in the asymptotic limit X^2 is just the sum of squares of reduced random variables which follow a normal distribution.
[Poisson distribution with 'large' mean $\rightarrow$ normal distribution (see exercise)]

Consequently, X^2 is asymptotically χ^2 – distributed, and the d.o.f. are $r-1$ because of the additional constraint $\sum n_k = N$.

If the function depends on p parameters fitted from the observed sample, there are additional p constraints, and the d.o.f are $r - p - 1$.

The distribution of the bins have to be carefully chosen (sometimes, one has to accept intervals which are not equidistant).

Rule of thumb to ensure that the Poisson distribution can be approximated by a normal:
In each bin, at least $E(n_k) = Np_k = 5$ events have to be expected (this gives a constraint on the size of the sample, N), and there should be also at least $n_k = 5$ events present in each bin.

The Kolmogorov-Smirnov (KS-)test

- provides a possibility to test the hypothesis that
  - a sample has been drawn from a given theoretical distribution
  - the distributions of two samples are equal
- is an alternative to the χ^2-test when the data sample is so small that binning becomes impossible.
- is applicable only to univariate distributions, which are fixed beforehand, i.e., not fitted to the data sample.
  - a generalization of the original KS-test for 2-D distributions can be found in Numerical Recipes, Chap. 14.7 and references therein

In the KS-test (one sample), we compare the hypothetical cdf

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

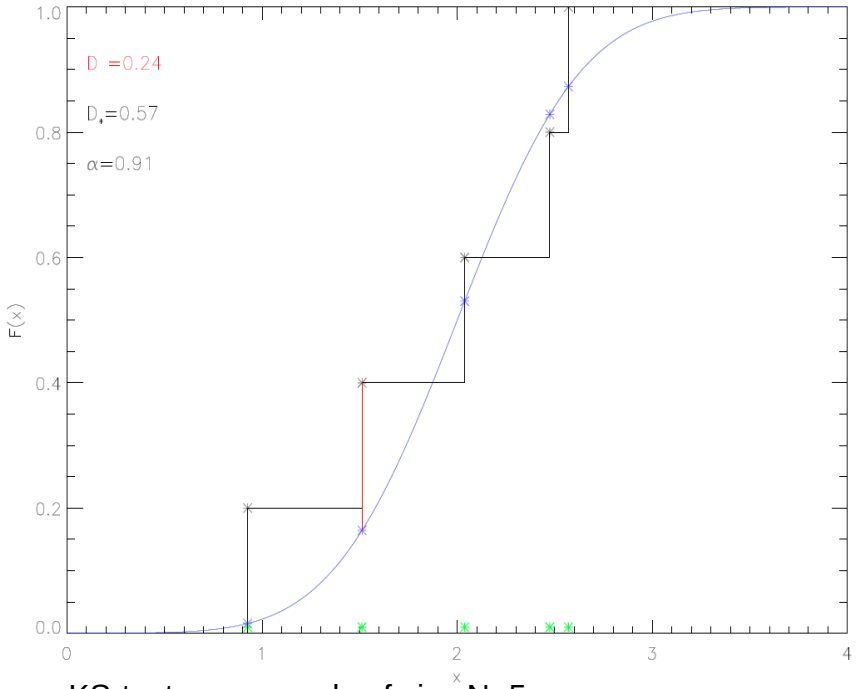
with the observed one,

$$S_N(x) = \frac{\text{number of events with } x_i < x}{\text{total number of events } N}.$$

If the events are located at x_i, then $S_N(x)$ gives the fraction of data points "to the left" of a given x. This function is constant between consecutive x_i (which have to be sorted into ascending order before the test can be performed), and jumps by the same constant $1/N$ at each x_i. See figure on the right. Thus,

$$S_N(x_i) = \frac{i}{N}, \quad i = 1, N$$

which is an unbiased estimator for the cdf of the distribution from which the sample was drawn.



KS-test on a sample of size N=5.
Theoretical distribution (cdf): normal, μ=2,σ=0.5
sample drawn from same population
green asterisks: sample data
step function: cdf for sample data, with steps of 1/N=0.2
red line: max deviation D
The significance of the test is α= 0.91.

The KS-test measures the maximum value of the absolute difference between two cdfs.

If we compare a sample's $S_N(x)$ with a given theoretical distribution, the KS *statistic* is

$$D = \max(|S_N(x) - F(x)|),$$

while for comparing two cdfs, the statistic is

$$D = \max(|S_{N_1}(x) - S_{N_2}(x)|)$$

Note that in both cases the differences have to be evalutated only at the data points x_i (see figure). The distribution of the KS-statistic given the null hypothesis (samples drawn from the same distribution) can be calculated, and gives the significance of any observed value of *D*. The central function entering the significance is the sum

$$Q_{\mathrm{KS}}(D_*) = 2\sum_{k=1}^{N}(-1)^{k-1}\exp(-2k^2 D_*^{\,2}),$$

which is a monotonic function with $Q_{\mathrm{KS}}(0) = 1$ and $Q_{\mathrm{KS}}(\infty) = 0,$ and the significance level of an observed value of *D* is

$$P(\mathrm{D}>D) = Q_{\mathrm{KS}}\left(D_*(D)\right).$$

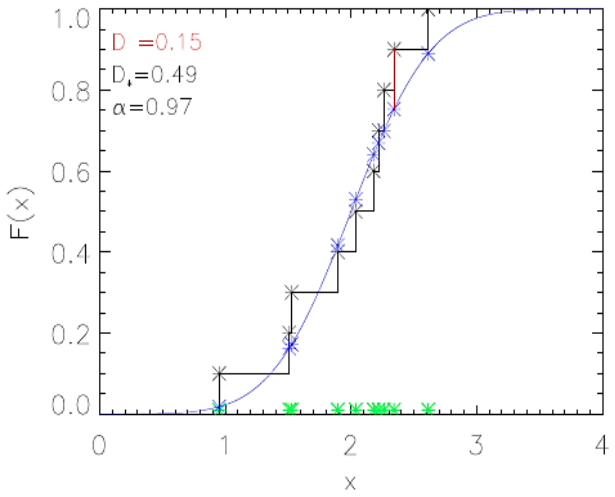
For sensible sample sizes $N \geq 4,$ one can approximate

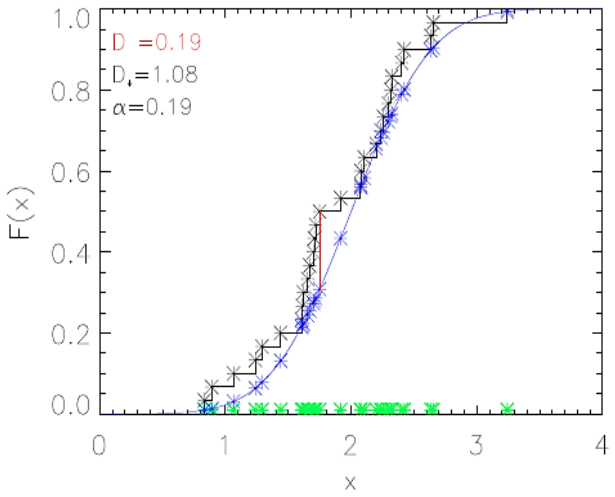
$$D_*(D) = D\left(\sqrt{N_{\mathrm{eff}}} + 0.12 + 0.11/\sqrt{N_{\mathrm{eff}}}\right) \qquad \text{with}$$

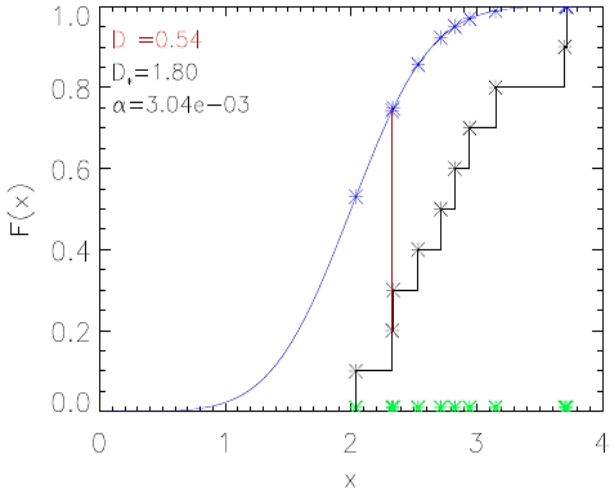
$$N_{\mathrm{eff}} = \begin{cases} N & \text{for the case of one distribution} \\ \dfrac{N_1 N_2}{N_1 + N_2} & \text{for the case of two distributions} \end{cases}$$

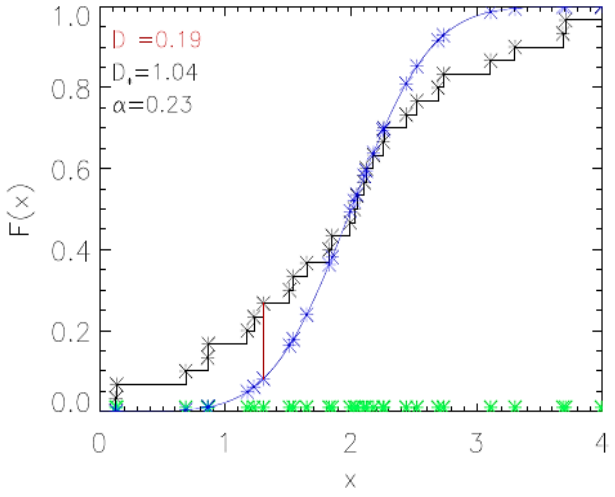
Critical value D_*	Significance $Q_{KS}(D_*)$
1.63	0.01
1.36	0.05
1.22	0.10
1.07	0.20

Examples









KS-test on samples of size N=10 (left panels) and N=30 (right panels).

Outline as in previous figure.

Theoretical distribution: normal, μ=2,σ=0.5

Upper panels: sample drawn from same population, hypothesis can be accepted.

Lower left panel: sample drawn from population with different mean, μ=3. Significance level very low, hypothesis needs to be rejected.

Lower right panel: sample drawn from population with larger variance, σ=1.0. Note that the significance is reasonable though the variances of theoretical distribution and sample are different, and N is quite large. *In this case we would accept a wrong hypothesis (type II error).*

Nonparametric tests

- Tests of hypothesis are usually classified into **parametric** and **nonparametric** methods.
  - Parametric methods make assumptions about the underlying distribution from which sample populations are selected.
  - Nonparametric methods make no assumptions about a sample population's distribution and are often based upon magnitude-based ranking, rather than actual measurement data.
- In many cases it is possible to replace a parametric test with a corresponding nonparametric test without significantly affecting the conclusion.

- In Chap. 7, we provided an unbiased estimator for the linear correlation coefficient, Pearson's r.
- Unfortunately, r is a rather poor statistic for deciding whether an observed correlation is statistically significant, or whether one observed correlation is stronger than another.
- Nonparametric or **rank correlation** can improve the situation.

Rank correlation

- We consider N data-pairs (x_i,y_i).
- We replace the values of x_i by the value of its rank among all other x_i's in the sample, that is by a number in between 1 and N.
- If there are some ties, they will be assigned a **midrank** which is the mean of the ranks they would have had if their values would have been slightly different.
- **Example** for *ordered* data:
  - **x**=[-1,4,4,8, 20, 50,50,50,100,100, 100]
  - if slightly different, we would obtain rank (**x**)= [1,**2,3**,4,5,**6,7,8**, **9,10,11**], where bold face ranks indicate ties.
  - averaging over ranks from ties results in rank (**x**)= [1,2.5,2.5,4,5,7,7,7,10,10,10].
  - due to this procedure, the **rank sum** will be just the sum of all integers between 1 and N, namely 1/2N(N+1) [=66 for our example].
- **example** for unordered data (same values, other order)
  - **x**=[100, 20,-1,50,8,50,4,100,100,50,4]
  - rank (**x**)=[10,5,1,7,4,7,2.5,10,10,7,2.5]
- The same ranking process is applied to the y_i-data, replacing each value by its rank among all other y_i's
- Finally, we have mapped (x_i,y_i) pairs to [rank (x_i), rank (y_i)] pairs.

- Now we can develop/use statistics for detecting correlations between uniform sets of integers/ half-integers in the range 1 to N.
- Due to ranking, there is some loss of information, …
- … but nonparametric correlation is (much) more robust than linear correlation: If a correlation has been detected by nonparametric methods (at a certain significance, of course), then it is really present.

Spearman rank-order correlation coefficient

Let R_i be the rank of x_i and S_i be the rank of y_i. The rank-order correlation coefficent is then the linear correlation coefficient w.r.t. the ranks,

$$r_S = \frac{\sum_{i=1}^{N}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \bar{R})^2 (S_i - \bar{S})^2}}$$

The significance of a nonzero value of r_S (i.e., that r_S deviates from zero) is tested by computing

$$t = r_S \sqrt{\frac{N-2}{1-r_S^2}}$$

which is approx. distributed as Stundent's distribution with $f = N - 2$ d.o.f.
Most importantly, this approximation does not depend on the original distribution of the x_i and y_i.

The Wilcoxon rank-sum test

- (also known as Mann-Whitney U-test) tests (loosely spoken) the hypothesis that two sample populations X and Y have the same mean. X and Y may be of different lengths, N_x and N_y.
- More precisely, the null hypothesis in this test is that the two samples are drawn from a single population, and therefore that their probability distributions are equal.
- somewhat similar to t-test on equal means. It does not require, however, assumptions about the form of the distribution of the parent population (contrasted to the t-test, which assumes normal distributions).
- to be used whenever the distributional assumptions that underlie the t-test cannot be satisfied.
- The Mann-Whitney statistics for X and Y are defined as

$$U_x = N_x N_y + \frac{1}{2} N_x (N_x + 1) - W_x$$

$$U_y = N_x N_y + \frac{1}{2} N_y (N_y + 1) - W_y$$

$$\left[U_x + U_y = N_x N_y \quad \text{(good test whether everything is OK)} \right]$$

  where W_x and W_y are the rank sums for X and Y *within the combined sample*, respectively.
- for small samples, compare Min(U_x, U_y) with critical values from tables, $U_{crit}(N_x, N_y, \alpha)$
- for larger samples, calculate the test statistic Z,

$$Z = \frac{U_x - N_x N_y / 2}{\sqrt{N_x N_y (N_x + N_y + 1)/12}}$$

  which follows closely a standardized normal distribution for sample sizes N_x, $N_y \geq 10$, and evaluate the significance.