G. Parameter estimation and information

V. Blobel
Inst. für Experimentalphysik, Hamburg

Sommersemester 2002

1. Samples

Given a sample $x_1, \ldots, x_i, \ldots, x_n$ (or short $\{\boldsymbol{x}\}$) of measured values from some distribution.

Examples for tasks:

- determine simple measures for the location, the variance, and the scale.
  Examples: sample mean, sample variance.
- given several values of different precision, form the best average of the data.
- if the data are from a distribution $p(x|a)$ (i.e. normalized density), what is the information in the sample about the parameter(s) a. Obtain best estimates of the parameter values.

Sample mean and variance

Given a sample $x_1, \ldots, x_i, \ldots, x_n$ (or short $\{\boldsymbol{x}\}$) of measured values from some distribution.

Define some estimate x_e for the mean and determine

$$S_x = \sum_{i=1}^{n} (x_i - x_e) \qquad S_{xx} = \sum_{i=1}^{n} (x_i - x_e)^2$$

$$\begin{aligned}
\text{sample mean} \qquad \bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i = x_e + \frac{1}{n} S_x \\
\text{sample variance} \qquad s^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left(S_{xx} - \frac{1}{n} S_x^2 \right) \\
\text{variance of mean} \qquad &= s^2/n
\end{aligned}$$

Note: $(n-1)$ instead of n removes bias of s^2.

t test for mean of distribution

Test: Is the observed mean $\bar{x}$ from a sample of size n consistent with a theoretical value μ?

$$\boxed{\text{test quantity} \quad t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}$$ follows the $N(0,1)$ distribution

If s replaces the unknown true standard deviation σ, then the normal distribution *underestimates* tails of distribution, especially for small n. $\rightarrow$ t-distribution

$$\boxed{\text{test quantity} \quad t = \frac{\bar{x} - \mu}{s/\sqrt{n}}}$$

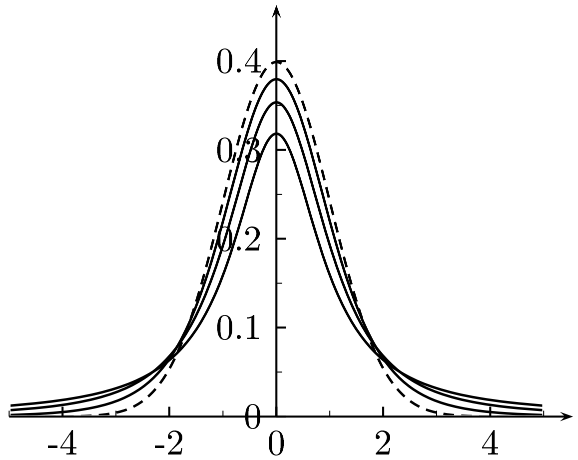
follows Students t-distribution with $(n-1)$ degrees of freedom.

Students t distribution

If x and $x_1, x_2, \ldots x_n$ are independent random variables following the standardized normal distribution, the quantity t defined by

$$t = x \Big/ \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

is distributed according to a Students t distribution with n degrees of freedom.



$n = 1, 2$ und 5

(dashed curve $= N(0,1)$).

Probability density

$$f_n(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \qquad -\infty < t < +\infty$$

Students t distribution contnd.

Mean $\quad E[t] = 0 \qquad$ Variance $\quad V[t] = \dfrac{n}{n-2} \qquad$ for $\quad n > 2$

Properties of the t distribution:

- Moments of order $> n$ do not exist.
- For $n = 1$ Students t distribution and the Cauchy distribution are identical (variance etc. divergent).
- For $n \to \infty$ it approaches the standard normal distribution.

Averaging data of different precision

Formulas used by the PDG for averaging:

$$\text{define weight factors} \qquad w_i = 1/\sigma_i^2$$

Averages from independent measurements x_i with standard deviations σ_i are formed by weighted means

$$\overline{x} \pm \delta\overline{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \pm \left(\sum_{i=1}^n w_i \right)^{-1/2}$$

Are the measurements consistent?

Averaging tests

The measurements may be incompatible or inconsistent; for a test wether the data are consistent with each other the test quantity

$$\chi^2 = \sum_{i=1}^n w_i \left(\overline{x} - x_i\right)^2 \qquad (n-1) \quad \text{degrees of freedom}$$

should be calculated. Expected value is $E\left[\chi^2\right] = n - 1$.

Depending on the value of χ^2 the following decisions can be taken:

$\chi^2/(n-1) \leq 1$ the result can be accepted;

$\chi^2/(n-1) \gg 1$ **(very large)** the data contradict each other. Either averaging should not be done, or the calculated average may be quoted, but with an conservative estimate of the uncertainty, taking into account the problems;

$\chi^2/(n-1) \gtrsim 1$ (not much greater than 1) averaging can still be done, but the uncertainties δx_i appear to be underestimated on at least one of the measurements.

Averaging tests contnd.

The latter case is of great practical importance.

One can assume that all errors are underestimated by the factor

$$S = \left[\chi^2/(n-1)\right]^{1/2}$$

This factor S can be used as a scale factor used to increase the calculated uncertainty $\delta\overline{x}$.

This is equivalent to scaling up *all* input errors δx_i before taking the average.

2. Information

Given a sample $x_1, \ldots, x_i, \ldots, x_n$ (or short $\{\boldsymbol{x}\}$) of measured values from a distribution $p(x; a)$ (i.e. normalized density).

What is the information in the sample about the parameter(s) a?

What is the accuracy that can be achieved from the sample about the parameter(s) a?

The likelihood function as *joint density* of the observed values of the random variable x:

$$\boxed{\text{Likelihood function} \qquad \mathcal{L}(a) = \prod_{i=1}^{n} p(x_i; a)}$$

with normalization

$$\int_\Omega \mathcal{L}(a)\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 1$$

Information inequality

$$\boxed{\text{Information} \quad I(a) = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)^2\right] = \int_\Omega \left(\frac{\partial \ln \mathcal{L}}{\partial a}\right)^2 \mathcal{L}\, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n}$$

This is the definition of *information*, where $\mathcal{L}$ is the joint density of the n observed values of the random variable x.

$$\boxed{\text{Information inequality} \qquad V[\widehat{a}] \geq \frac{1}{I}}$$

The *inverse* of the information $I_n(a)$, or short I, is the lower limit of the variance of the parameter estimate $\widehat{a}$ – minimum variance bound MVB.

The inequality is also called Rao-Cramér-Frechet inequality, and is valid in this form for any unbiased estimate $\widehat{a} = \widehat{a}(x)$.

Efficiency and bias

Definition of the *efficiency* of an estimator:

For an unbiased estimator $\hat{a}$ one can define the *efficiency* eff$(\widehat{a})$ by the ratio of the mininal to the actual variance:

$$\boxed{\text{Efficiency} \qquad \mathrm{eff}(\widehat{a}) = \frac{I^{-1}}{V[\widehat{a}]} \qquad 0 \leq \mathrm{eff}(\widehat{a}) \leq 1}$$

The actual efficiency of an estimator depends on the specific problem and method.

Variance limit in case of a bias $B_n(\widehat{a}) = E[\widehat{a}] - a_{\text{true}} \neq 0$:

$$\boxed{\text{Variance} \qquad V[\widehat{a}] \geq \frac{(1 + \partial B/\partial a)^2}{I}}$$

Proof of the information inequality

The proof[a] ist based on the two following equations:

$$\begin{aligned} E[\widehat{a}] = \int_\Omega \widehat{a}\,\mathcal{L}(a)\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n &= a_{\text{true}} & (1)\\ \int_\Omega \mathcal{L}(a)\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n &= 1 & (2)\end{aligned}$$

First expression (1) valid for an unbiased estimate $\widehat{a} = \widehat{a}(x)$.

Second expression (2) follows from the normalization of the probability density $p(x;a)$.

[a] In the proof sufficient regularity has to be assumed to allow differentiation with respect to a under the integral sign. This assumption may be wrong if the density function $p(x,a)$ does not fall off rapidly enough to ensure existence of the integral of the differentiated integrand. The assumption requires in addition that the effective range (where $p(x,a)$ is non-zero) of the x-integration does not depend on the value of the parameter a.

Proof of the information inequality (contnd.)

First step: take the derivative of the expression (1) with respect to the parameter a.

$$\frac{\partial}{\partial a}\int_\Omega \widehat{a}\,\mathcal{L}(a)\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n = 1\,.$$

or (note that $\widehat{a}$ is not a function of a)

$$\int_\Omega \widehat{a}(x)\frac{\partial\mathcal{L}(a)}{\partial a}\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n = \int_\Omega \widehat{a}(x)\frac{\partial\ln\mathcal{L}(a)}{\partial a}\,\mathcal{L}\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n = 1$$

Second step: take the derivative of the second expression (2) with repect to the parameter a:

$$\int_\Omega \frac{\partial\mathcal{L}}{\partial a}\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n = \int_\Omega \frac{\partial\ln\mathcal{L}}{\partial a}\,\mathcal{L}\,\mathrm{d}x_1\mathrm{d}x_2\ldots\mathrm{d}x_n = E\left[\left(\frac{\partial\ln\mathcal{L}}{\partial a}\right)\right] = 0$$

Proof of the information inequality (contnd.)

Multiplying last equation by a und subtracting it from the previous equation:

$$\int_\Omega (\widehat{a} - a) \left(\frac{\partial \ln \mathcal{L}}{\partial a} \right) \mathcal{L} \, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 1$$

Invoking the Schwarz inequality with $f \equiv (\widehat{a} - a) \sqrt{\mathcal{L}}$ and $g \equiv (\partial \ln \mathcal{L} / \partial a) \sqrt{\mathcal{L}}$:

$$\int f^2 \, \mathrm{d}x \cdot \int g^2 \, \mathrm{d}x \geq \left(\int f \, g \, \mathrm{d}x \right)^2$$

$$\boxed{\underbrace{\left(\int_\Omega (\widehat{a} - a)^2 \, \mathcal{L} \, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \right)}_{V[\widehat{a}]} \underbrace{\left(\int_\Omega \left(\frac{\partial \ln \mathcal{L}}{\partial a} \right)^2 \mathcal{L} \, \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n \right)}_{I(a)} \geq 1}$$

q.e.d.

Alternative expression of information I

Differentiating equation (2) a second time with respect to the parameter a:

$$\int_\Omega \left(\frac{\partial \ln \mathcal{L}}{\partial a} \frac{\partial \mathcal{L}}{\partial a} + \frac{\partial^2 \ln \mathcal{L}}{\partial a^2} \mathcal{L} \right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n = 0 \, ,$$

Rewritten in terms of expectation values:

$$I(a) = E \left[\left(\frac{\partial \ln \mathcal{L}}{\partial a} \right)^2 \right] = -E \left[\frac{\partial^2 \ln \mathcal{L}}{\partial a^2} \right]$$

i.e. either square of first derivative or negative second derivative.

Case of several variables

Case of m variables $a_1, \ldots, a_j, \ldots, a_m$: information I becomes a m-by-m symmetric matrix $\boldsymbol{I}$ with elements

$$I_{jk} = E\left[\frac{\partial \ln \mathcal{L}}{\partial a_j}\frac{\partial \ln \mathcal{L}}{\partial a_k}\right] = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k}\right]$$

The minimal variance $\boldsymbol{V}[\hat{\boldsymbol{a}}]$ of an estimate $\hat{\boldsymbol{a}}$ is given by the inverse of the information matrix $\boldsymbol{I}$:

$$\boxed{\text{minimal variance} \qquad \boldsymbol{V}[\hat{\boldsymbol{a}}] = \boldsymbol{I}^{-1}}$$

Example: Weighing four object with a balance

Weighing four objects $a_j, j = 1, \ldots 4$

Direct approach: weigh each object separately in four independent measurements.

Alternatively: try to improve the precision of four measurements by weighing suitable combinations of the objects:

$$x_i = \sum_{j=1}^{4} A_{ij} a_j \qquad \text{with} \qquad A_{ij} = -1, 0 \quad \text{or} \quad +1$$

Assumptions:

- measurements follow a normal distribution law, and
- standard deviation σ in each measurement is always the same.

Example: Weighing four object with a balance (2)

Likelihood function

$$\mathcal{L}(\boldsymbol{a}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^4 \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{4}\left(x_i - \sum_{j=1}^{4} A_{ij}a_j\right)^2\right]$$

in a combination of measured quantities $x_i, i = 1, \ldots 4$ with $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{a}$.

The elements of the information matrix are given by

$$I_{jk} = E\left[-\frac{\partial^2 \ln \mathcal{L}}{\partial a_j \partial a_k}\right] = \frac{1}{\sigma^2}\sum_{i=1}^{4} A_{ij}A_{ik} \qquad \text{or short} \quad \boldsymbol{I} = \frac{1}{\sigma^2}\boldsymbol{A}^T\boldsymbol{A}$$

Example: Weighing four object with a balance (3)

Direct approach: matrix $\boldsymbol{A}$ is the unit matrix ($A_{ij} = \delta_{ij}$)

$$\boldsymbol{I} = 1/\sigma^2\,\mathbf{1} \qquad \boldsymbol{V}[\widehat{a}] = \sigma^2\,\mathbf{1}$$

i.e. the objects a_j are measured with standard deviation σ.

Alternatively: linearly independent combination given by the matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad \boldsymbol{I} = \frac{4}{\sigma^2}\mathbf{1} \qquad \boldsymbol{V}[\widehat{a}] = \frac{\sigma^2}{4}\mathbf{1}$$

Alternative method is preferable: objects a_j are now measured with standard deviation $\sigma/2$.

In both cases the measured values are uncorrelated.

Experimental design – Example

Decay angular distribution of an elementary particle

The decay angular distribution is given by

$$p(x|a) = \frac{1}{2}(1 + ax) \qquad \text{with} \quad x \equiv \cos\vartheta$$

The value of the parameter a is expected to be around $a = -1/3$. The question is, how many decays have to be observed to determine the parameter a with a relative precision of 1 %.

The information is calculated for $n = 1$:

$$I_1 = \int_{-1}^{+1} \frac{1}{p(x,a)} \left(\frac{\partial p}{\partial a}\right)^2 \mathrm{d}x = \begin{cases} \frac{1}{2a^3}\left[\ln\frac{1+a}{1-a} - 2a\right] & \text{if} \quad a \neq 0 \\ \frac{1}{3} & \text{if} \quad a = 0 \end{cases}$$

The standard deviation depends on the value of the parameter. For $a = -1/3$ the value is $\sigma = \sqrt{2.797/n}$ and it follows that more than 2.52×10^5 decays have to be measured in order to obtain a standard deviation $\sigma < 0.01\,a$.

3. Method of moments

The method of moments is a general and simple method of parameter estimation – generally not optimal, but useful in many cases.

Given a sample x_i, $i = 1, 2, \ldots n$, assumed to come from a distribution with density $f(x;a)$, where x is the random variable and a is a parameter, to be estimated from the data.

Method of moments

The expectation of a function $h(x)$ of the random variable, given the density $f(x;a)$, is itself a function $g(a)$ of the parameter a:

$$E\left[h(x)\right] = \int_{-\infty}^{+\infty} h(x)\, f(x,a)\, \mathrm{d}x = g(a)$$

Compared to the average value of the function $h(x)$ for the data sample:

$$\overline{h(x)} = \frac{1}{n} \sum_{i=1}^{n} h(x_i) \; .$$

Principle of the method of moments: equate the theoretical expectation value and the empirical average of the function $h(x)$, to obtain the parameter estimate $\widehat{a}$.

Method of moments contnd.

$$g(\widehat{a}) = \overline{h(x)} \qquad\qquad \widehat{a} = g^{-1}\left(\overline{h(x)}\right)$$

which requires the function g^{-1} inverse to $g(a)$. According to the central limit theorem the mean value $\overline{h(x)}$ is asymptotically normal distributed with variance

$$V\left[\overline{h(x)}\right] = \frac{1}{n} V\left[h(x)\right]$$

The variance of the estimate is calculated using standard error propagation.

The approximate normality of the estimate is often destroyed by the inverse function $g^{-1}\left(\overline{h(x)}\right)$. The estimate has a potential bias $B_n(\widehat{a}) \propto 1/n$.

Example – angular distribution

The angular distribution $f(x) = 1/2\,(1+ax)$ with $x \equiv \cos\vartheta$ is assumed. In order to get an estimate $\widehat{a}$ the function $h(x) \equiv x$ is used.

$$
\begin{aligned}
E[x] &= \int_{-1}^{+1} x\frac{1}{2}(1+ax)\,\mathrm{d}x = \frac{1}{3}a \qquad \Rightarrow \qquad \boxed{\widehat{a} = 3\,\widehat{x}} \\
V[x] &= E\left[\left(x-\frac{1}{3}a\right)^2\right] = \int_{-1}^{+1}\left(x-\frac{1}{3}a\right)^2 \frac{1}{2}(1+ax)\,\mathrm{d}x = \frac{1}{3}-\frac{1}{9}a^2
\end{aligned}
$$

For the variance of the average $\overline{x}$ and the variance of the estimate $\widehat{a}$ follows:

$$
V[\overline{x}] = \frac{1}{n}\left(\frac{1}{3}-\frac{1}{9}a^2\right) \qquad V[\widehat{a}] = \frac{1}{n}(3-a^2)
$$

4. Robust estimates

Many estimates are based on the normal distribution. The normal distribution is the most important distribution in all statistics.

> Under random sampling, as the sample size gets large, the limiting distribution of the sample mean is normal." (Central limit theorem)

> "Everyone believes in the normal law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact" (Poincaré)

In the normal distribution only two numbers are required to tell us everything about the probabilities associated with a random variable:

the mean μ and the variance σ^2

Departures from normality

But: **Distributions are never normal!**

Departures from normality or from any other assumed distrinution are rather common in practice. Random samples …

- often have outliers: unusual large or small values among a sample,
- can have heavy tails (thicker than a normal distribution),
- can be highly skewed.

Estimates of statistical parameters can be sensitively influenced by such deviations.

Robustness signifies insensivity to small deviations from the assumptions.

Basic parameters of a sample

A standard task in data analysis of a sample of a random variable is the determination of

- the location, i.e. mean or center of the random variable, and
- the dispersion, i.e. the variance or standard deviation of the random variable.

How to define robust estimates for the location and the dispersion of a random variable?

Estimation of the centre of a symmetric distribution

The estimator $\bar{x}_p$ is the L_p estimate of location if it minimizes the quantity

$$L_p = \sum_{i=1}^{n} |x_i - \bar{x}_p|^p \ .$$

The L_p estimate ...

(for $p = 1$) ...is the sample median;

(for $p = 2$) ...is the sample mean;

(for $p = \infty$) (Tchebycheff norm) ...is equivalent to the midrange $= 0.5(x_{\text{min}} + x_{\text{max}})$, when the possible x-values are limited to the range $[x_{\text{min}}, x_{\text{max}}]$.

The Tchebycheff norm is optimal in the special case of a uniform distribution within the finite limits.

Estimation of the centre of a symmetric distribution contnd.

If a "Gaussian" sample contains background or some fraction of outliers, i.e. if some elements are from a different distribution, then the efficiency of the *mean* will not be optimal.

Models for symmetric distributions with longer tails are

- the Cauchy (Breit-Wigner) distribution, and the
- double-exponential distribution.

Distribution	Sample median	Sample mean	Sample midrange
Normal (Gaussian)	0.64	1.00	0.00
Cauchy	0.82	0.00	0.00
Double-exponential	1.00	0.50	0.00
Uniform	0.00	0.00	1.00

Properties of location estimators: asymptotic efficiencies.

Asymptotic variances

Properties of location estimators: the n-dependence of the asymptotic variances are:

Distribution	L_1 median	L_2 mean	L_∞ midrange
Normal (Gaussian)	$\pi/2n$	$1/n$	$\pi^2/12\ \ln n$
Cauchy	$\pi^2/4n$	∞	∞
Double-exponential	$1/2n$	$2/n$	$\pi^2/12$
Uniform	$1/4n$	$1/12n$	$1/2n^2$

The trimmed mean

The trimmed mean depends on the parameter r and is (after sorting)

- the mean of the $2rn$ central values;
- the $(1-2r)n$ lowest and
- the $(1-2r)n$ largest values are not used.

Special cases of the value of r are

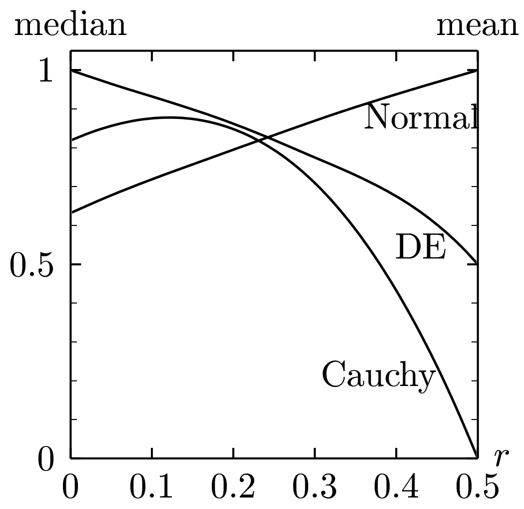
case $r = 0.5$**:** the mean;

case $r = 0.23$**:** a trimmed mean, which in a certain sense is optimal;

case $r = 0$**:** the median.

Asymptotic efficiency of the trimmed mean

...as function of the parameter r for different symmetric distribution: the Gaussian distribution, the Cauchy distribution and the double-exponential distribution (DE):



For $r = 0.23$ the efficiency is at least 82 % for either distribution, and is thus a robust estimate for a general symmetric distribution.

Estimation of the mean of a non-symmetric distribution

The situation is more difficult, if the location parameter of an non-symmetric distribution has to be determined.

If and only if the parametrization of the distribution is well known, the ML estimate is optimal.

For some distributions a nonlinear transformation $y = y(x)$ from a distribution

$$f(x)\mathrm{d}x \quad \text{to a distribution} \quad g(y)\mathrm{d}y \quad \text{with} \quad g(y) = f(x)/y'(x)$$

to an *almost* symmetric distribution may be useful, which allows to apply the robust and stable estimates of the mean.

The transformation of the mean and standard deviation is approximately given by

$$\bar{y} = y(\bar{x}) + \frac{1}{2}\sigma_x^2 \left.\frac{\mathrm{d}^2 y}{\mathrm{d}x^2}\right|_{x=\mu_x} + \ldots \qquad \sigma_y^2 = \sigma_x^2 \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 + \ldots .$$

Example for transformation: shower energy

The energy of a particle or a jet of particles can be measured in an (electromagnetic or hadronic) calorimeter. The resolution of sampling calorimeters is usually dominated by sampling fluctuations, leading to a fractional resolution

$$\frac{\sigma_E}{E} = \frac{\sigma_0}{\sqrt{E}} \qquad \text{or} \qquad \sigma_E = \sigma_0 \cdot \sqrt{E}$$

with typical values $\sigma_0 = 10$ % $(\mathrm{GeV})^{1/2}$ (electromagnetic calorimeter) and $\sigma_0 = 50$ % $(\mathrm{GeV})^{1/2}$ (hadronic calorimeter).

The quantity $h = \sqrt{E}$ has a constant standard deviation of

$$\sigma_h = 1/2\, \sigma_0$$

(by error propagation).

For a sample of fixed true energy values the h-distribution will be more Gaussian-like than the E-distribution.

Example: contaminated normal distribution

A large, randomly mixed batch of n "good" and "bad" observations of the same quantity μ is given.

- A "good" observation x_i from $N(\mu, \sigma^2)$ occurs with probability $(1-\varepsilon)$ (for small ε);
- a "bad" observation x_i from $N(\mu, K^2\sigma^2)$ occurs with probability ε with e.g. $K = 3$.

i.e. the x_i are independent, identically distributed with the underlying distribution

$$f(x) = (1-\varepsilon) \cdot n\left(\frac{x-\mu}{\sigma}\right) + \varepsilon \cdot n\left(\frac{x-\mu}{K\sigma}\right) \qquad \text{(contaminated normal)}$$

($n(x)$ = standard normal density).

How should the scatter (dispersion) around the mean be measured?

Estimates for the dispersion

Possible measures for the dispersion:

$$s_n = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} \qquad \text{mean square deviation}$$

$$d_n = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \qquad \text{mean absolute deviation}$$

Conventional answer: for normal observations s_n is about 12 % more efficient than d_n.

$$\lim_{n\to\infty} s_n = \sigma \qquad \lim_{n\to\infty} d_n = \sqrt{2/\pi}\,\sigma \approx 0.80\,\sigma$$

Efficiency for contaminated normal

But the distribution is not normal, it is the contaminated normal distribution!

Asymptotic relative efficiency of d_n relative to s_n (for $K = 3$):

$$A(\varepsilon) = \lim_{n\to\infty} \frac{V[s_n]/\left(E[s_n]\right)^2}{V[d_n]/\left(E[d_n]\right)^2} = \frac{1}{4} \frac{\dfrac{3(1+80\varepsilon)}{(1+8\varepsilon)^2} - 1}{\dfrac{\pi(1+8\varepsilon)}{2(1+2\varepsilon} - 1}$$

ε	0	0.001	0.002	0.005	0.01	0.02	0.05	0.10	0.25
$A(\varepsilon)$	0.876	0.948	1.016	1.198	1.439	1.752	2.035	1.903	1.371

Just 2 bad observations in 1000 are sufficient to get a better efficiency for the mean absolute deviation, and for $\varepsilon = 5\%$ the efficiency is better by a factor of > 2.

Method for mean value I: Trimmed mean

The random sample is $x_1, x_2, \dots x_i \dots x_n$. The observations are written in ascending order: $x_{(1)}, x_{(2)}, \dots x_{(i)} \dots x_{(n)}$.

A value of r must be chosen, for example $r = 0.23$

g = value of $r \cdot n$ rounded down to the nearest integer.

The sample trimmed mean is computed by removing the g smallest and g largest observations:

$$\hat{x}_t = \frac{x_{(g+1)} + \cdots + x_{(n-g)}}{n - 2g}$$

Method for mean value and for LSQ II: Winsorized mean

Winzorization of a random sample x_i: the g smallest values are pulled in and set equal to x_{g+1} and the g largest values are pulled in and set equal to x_{n-g}

$$z_i = \begin{cases} x_{g+1} & \text{if} \quad x_i \le x_{g+1} \\ x_i & \text{if} \quad x_{g+1} < x_i < x_{n-g} \\ x_{n-g} & \text{if} \quad x_i \ge < x_{n-g} \end{cases} \qquad \bar{x}_w = \frac{1}{n} \sum_{i=1}^{n} z_i$$

Algorithm can be used also for Least Squares (LSQ) Fits in iterations to reduce effect of outliers: very large residuals (expectation value 0) are "moved in" (e.g. limit to 3 standard deviations).

Fitting a line by minimizing absolute deviation

Fit of the line $y(x;\, a,\, b) = a + bx$ by minimizing

$$S(a,b) = \sum_{i=1}^{n} |y_i - a - bx_i| \ . \qquad (*)$$

The median c_{median} minimizes the sum $\sum |c_i - c_{\text{median}}|$ of absolute deviations.
For fixed slope b the sum $(*)$ is minimized by the intercept a, given as a function of the fixed b

$$a(b) = \text{median}\, \{y_i - bx_i\} \ .$$

Fitted value of b from derivative of the sum $(*)$ w.r.t. slope b

$$\frac{\partial S}{\partial b} = F(b) = \sum_{i=1}^{n} \text{sign}\ (y_i - a - bx_i) \cdot x_i \qquad \text{and} \quad F(\hat{b}) = 0$$

(with $\text{sign}\,(0) = 0$) with a replaced by $a(b)$. Function $F(b)$ has discontinuities and has to be solved for $\hat{b}$ by e.g. bracketing and bisection.