Parameter estimation and information

V. Blobel
Universität Hamburg

Zeuthen, May 9 – 11,2001

- Estimates
- Samples
- Information
- Method of moments
- Robust estimates

Estimates

In this chapter the amount of *information* about parameters given by a sample $x_1, \ldots, x_i, \ldots, x_n$ (or short $\{\boldsymbol{x}\}$) of measured values from a distribution $p(x|a)$ is expressed and related to the variance of the estimated parameter value.

$$\text{Estimate} \qquad \widehat{a} = \widehat{a}(x_1, x_2 \ldots x_n)$$

is a function of the random variables $x_1, \ldots, x_i, \ldots, x_n$ and is itself a random variable.

$$\text{for example} \qquad \widehat{a} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{mean value}$$

$$\text{Estimate} \qquad V(\widehat{a}) = V[\widehat{a}(x_1, x_2 \ldots x_n)]$$

Samples

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

- 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

$$\text{Sample mean} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = x_e + \frac{1}{n} S_x$$

$$\text{Sample variance} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(S_{xx} - \frac{1}{n}S_x^2\right)$$

The latter formulas use some approximate mean value x_e and

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

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:

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

with normalization

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

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 $\widehat{a}$.

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* $\text{eff}(\widehat{a})$ by the ratio of the mininal to the actual variance:

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

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

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

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

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 $a_j, j = 1, \dots 4$

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 $x_i, i = 1, \dots 4$ 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 $\sigma < 0.01\, a$.

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 $B_n(\widehat{a}) \propto 1/n$.

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.

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" (standard deviation) around the mean be measured?

Estimates for the scatter

Possible formulas for scatter:

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

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

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

A value of γ must be chosen, for example $\gamma = 0.2$

g = value of γn 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 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}

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