Introduction to Statistics – Day 4

Lecture 1
- Probability
- Random variables, probability densities, etc.

Lecture 2
- Brief catalogue of probability densities
- The Monte Carlo method.

Lecture 3
- Statistical tests
- Fisher discriminants, neural networks, etc
- Significance and goodness-of-fit tests

→ Lecture 4
- Parameter estimation
- Maximum likelihood and least squares
- Interval estimation (setting limits)

Parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v. parameter

Suppose we have a sample of observed values: $\vec{x} = (x_1, \ldots, x_n)$

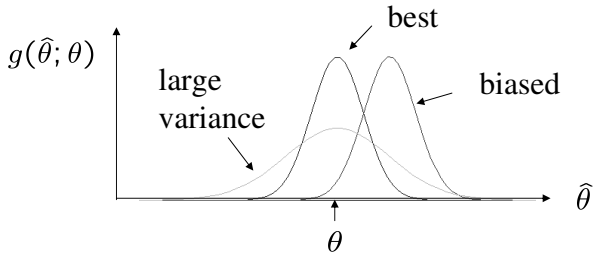
We want to find some function of the data to estimate the parameter(s):

$\boxed{\hat{\theta}(\vec{x})}$ ← estimator written with a hat

Sometimes we say 'estimator' for the function of x_1, ..., x_n; 'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

An estimator for the mean (expectation value)

Parameter: $\mu = E[x]$

Estimator: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \overline{x}$ ('sample mean')

We find: $b = E[\hat{\mu}] - \mu = 0$

$$V[\hat{\mu}] = \frac{\sigma^2}{n} \qquad \left(\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$

An estimator for the variance

Parameter: $\sigma^2 = V[x]$

Estimator: $\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \equiv s^2$ ('sample variance')

We find:

$b = E[\widehat{\sigma^2}] - \sigma^2 = 0$ (factor of n–1 makes this so)

$V[\widehat{\sigma^2}] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2\right)$, where

$\mu_k = \int (x-\mu)^k f(x)\,dx$

The likelihood function

Suppose the outcome of an experiment is: x_1, ..., x_n, which is modeled as a sample from a joint pdf with parameter(s) θ:

$$f(x_1, \ldots, x_n; \theta)$$

Now evaluate this with the data sample obtained and regard it as a function of the parameter(s). This is the likelihood function:

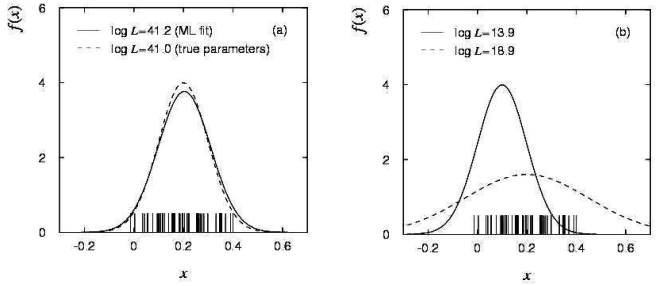
$$\boxed{L(\theta) = f(x_1, \ldots, x_n; \theta)} \qquad (x_i \text{ constant})$$

If the x_i are independent observations of $x \sim f(x;\theta)$, then,

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Maximum likelihood estimators

If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

ML estimators not guaranteed to have any 'optimal' properties, (but in practice they're very good).

ML example: parameter of exponential pdf

Consider exponential pdf, $f(t;\tau) = \frac{1}{\tau}e^{-t/\tau}$

and suppose we have data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \prod_{i=1}^{n}\frac{1}{\tau}e^{-t_i/\tau}$

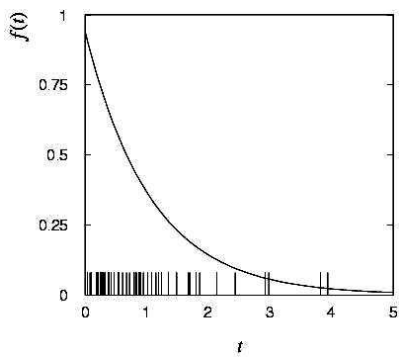
The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n}\left(\ln\frac{1}{\tau} - \frac{t_i}{\tau}\right)$$

ML example: parameter of exponential pdf (2)

Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$



Monte Carlo test:
generate 50 values
using $\tau = 1$:

We find the ML estimate:

$$\hat{\tau} = 1.062$$

Variance of estimators: Monte Carlo method

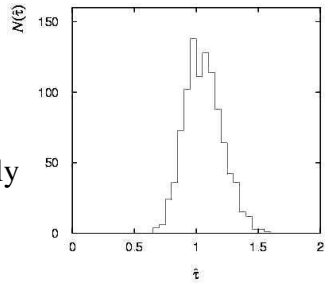
Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\hat{\sigma}_{\hat{\tau}} = 0.151$$

Note distribution of estimates is roughly Gaussian – (almost) always true for ML in large sample limit.



Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] \qquad (b = E[\hat{\theta}] - \theta)$$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\hat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of ln L at its maximum:

$$\hat{V}[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

Variance of estimators: graphical method

Expand ln L (θ) about its maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta})^2 + \ldots$$

First term is ln $L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta-\hat{\theta})^2}{2\hat{\sigma}^2_{\hat{\theta}}}$$

i.e., $\quad \ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \dfrac{1}{2}$

$\rightarrow$ to get $\hat{\sigma}_{\hat{\theta}}$, change θ away from $\hat{\theta}$ until ln L decreases by 1/2.

Example of variance by graphical method

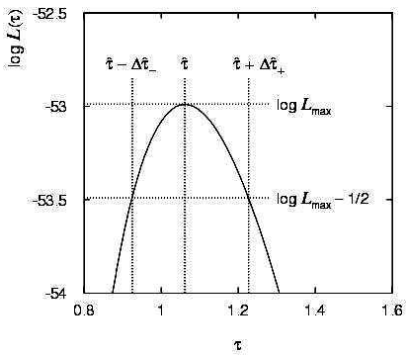
ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$



Not quite parabolic ln L since finite sample size (n = 50).

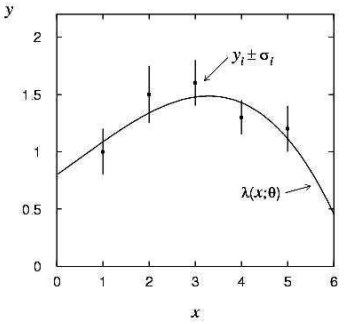
The method of least squares

Suppose we measure N values, y_1, ..., y_N, assumed to be independent Gaussian r.v.s with

$$E[y_i] = \lambda(x_i;\theta) \ .$$

Assume known values of the control variable x_1, ..., x_N and known variances

$$V[y_i] = \sigma_i^2 \ .$$



We want to estimate θ, i.e., fit the curve to the data points.

The likelihood function is

$$L(\theta) = \prod_{i=1}^{N} f(y_i;\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - \lambda(x_i;\theta))^2}{2\sigma_i^2}\right]$$

The method of least squares (2)

The log-likelihood function is therefore

$$\ln L(\theta) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2} + \text{terms not depending on } \theta$$

So maximizing the likelihood is equivalent to minimizing

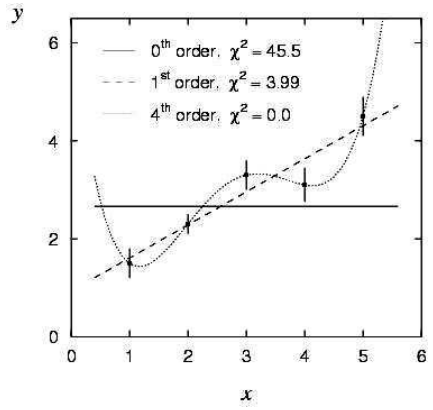
$$\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

Minimum of this quantity defines the least squares estimator $\hat{\theta}$.

Often minimize χ^2 numerically (e.g. program **MINUIT**).

Example of least squares fit

Fit a polynomial of order p: $\lambda(x;\theta_0,\ldots,\theta_p) = \sum_{n=0}^{p} \theta_n x^n$



Variance of LS estimators

In most cases of interest we obtain the variance in a manner similar to ML. E.g. for data ~ Gaussian we have

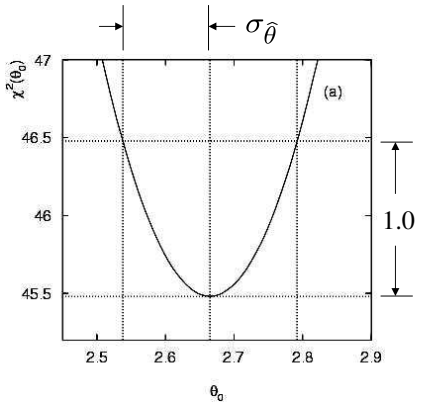
$$\chi^2(\theta) = -2 \ln L(\theta)$$

and so

$$\widehat{\sigma^2}_{\hat{\theta}} \approx 2 \left[\frac{\partial^2 \chi^2}{\partial \theta^2} \right]^{-1}_{\theta = \hat{\theta}}$$

or for the graphical method we take the values of θ where

$$\chi^2(\theta) = \chi^2_{\text{min}} + 1$$



Goodness-of-fit with least squares

The value of the χ^2 at its minimum is a measure of the level of agreement between the data and fitted curve:

$$\chi^2_{\text{min}} = \sum_{i=1}^{N} \frac{(y_i - \lambda(x_i; \hat{\theta}))^2}{\sigma_i^2}$$

It can therefore be employed as a goodness-of-fit statistic to test the hypothesized functional form $\lambda(x; \theta)$.

We can show that if the hypothesis is correct, then the statistic $t = \chi^2_{\text{min}}$ follows the chi-square pdf,

$$f(t; n_{\text{d}}) = \frac{1}{2^{n_{\text{d}}/2} \Gamma(n_{\text{d}}/2)} t^{n_{\text{d}}/2 - 1} e^{-t/2}$$

where the number of degrees of freedom is

n_{d} = number of data points − number of fitted parameters

Goodness-of-fit with least squares (2)

The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if $\chi^2_{\text{min}} \approx n_{\text{d}}$ the fit is 'good'.

More generally, find the p-value: $p = \int_{\chi^2_{\text{min}}}^{\infty} f(t; n_{\text{d}})\, dt$

This is the probability of obtaining a χ^2_{min} as high as the one we got, or higher, if the hypothesis is correct.

E.g. for the previous example with 1st order polynomial (line),

$$\chi^2_{\text{min}} = 3.99\,, \qquad n_{\text{d}} = 5-2 = 3\,, \qquad p = 0.263$$

whereas for the 0th order polynomial (horizontal line),

$$\chi^2_{\text{min}} = 45.5\,, \qquad n_{\text{d}} = 5-1 = 4\,, \qquad p = 3.1\times10^{-9}$$

Setting limits

Consider again the case of finding $n = n_{\text{s}} + n_{\text{b}}$ events where

n_{b} events from known processes (background)
n_{s} events from a new process (signal)

are Poisson r.v.s with means s, b, and thus $n = n_{\text{s}} + n_{\text{b}}$ is also Poisson with mean = $s + b$. Assume b is known.

Suppose we are searching for evidence of the signal process, but the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{\text{obs}} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

→ set upper limit on the parameter s.

Example of an upper limit

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\text{obs}}; s, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\text{up}}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\text{obs}} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\text{up}} = -\ln\gamma \approx 3.00$$

The interval [0, s_{up}] is an example of a confidence interval, designed to cover the true value of s with a probability $1 - \gamma$.

Calculating Poisson parameter limits

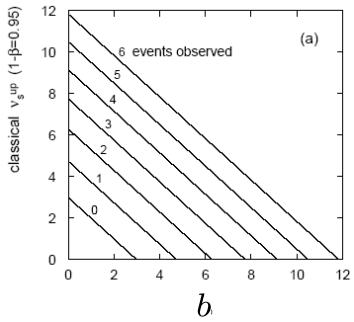
To solve for s_{lo}, s_{up}, can exploit relation to χ^2 distribution:

$$s_{\text{lo}} = \frac{1}{2} F_{\chi^2}^{-1}(\alpha; 2n) - b$$

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1-\beta; 2(n+1)) - b$$

Quantile of χ^2 distribution
`TMath::ChisquareQuantile`

For low fluctuation of n this can give negative result for s_{up}; i.e. confidence interval is empty.



Many subtle issues here – see e.g. CERN (2000) and Fermilab (2001) confidence limit workshops and PHYSTAT conferences.

Wrapping up lecture 4

We've seen some main ideas about parameter estimation,

ML and LS,
how to obtain/interpret stat. errors from a fit,

and what to do if you don't find the effect you're looking for,

setting limits.

In four days we've only looked at some basic ideas and tools, skipping entirely many important topics. Keep an eye out for new methods, especially multivariate, machine learning, Bayesian methods, etc.

Extra slides

Setting limits

Frequentist intervals (limits) for a parameter s can be found by defining a test of the hypothesized value s (do this for all s):

Specify values of the data n that are 'disfavoured' by s (critical region) such that $P(n \text{ in critical region}) \leq \gamma$ for a prespecified γ, e.g., 0.05 or 0.1.

(Because of discrete data, need inequality here.)

If n is observed in the critical region, reject the value s.

Now invert the test to define a confidence interval as:

set of s values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

The interval will cover the true value of s with probability $\geq 1 - \gamma$.

Setting limits: 'classical method'

E.g. for upper limit on s, take critical region to be low values of n, limit s_{up} at confidence level $1 - \beta$ thus found from

$$\beta = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)} .$$

Similarly for lower limit at confidence level $1 - \alpha$,

$$\alpha = P(n \geq n_{\text{obs}}; s_{\text{lo}}, b) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{(s_{\text{lo}} + b)^n}{n!} e^{-(s_{\text{lo}}+b)} .$$

Sometimes choose $\alpha = \beta = \gamma/2$ $\rightarrow$ central confidence interval.

Likelihood ratio limits (Feldman-Cousins)

Define likelihood ratio for hypothesized parameter value s:

$$l(s) = \frac{L(n|s,b)}{L(n|\hat{s},b)} \quad \text{where} \quad \hat{s} = \begin{cases} n - b & n \geq b, \\ 0 & \text{otherwise} \end{cases}$$

Here $\hat{s}$ is the ML estimator, note $0 \leq l(s) \leq 1$.

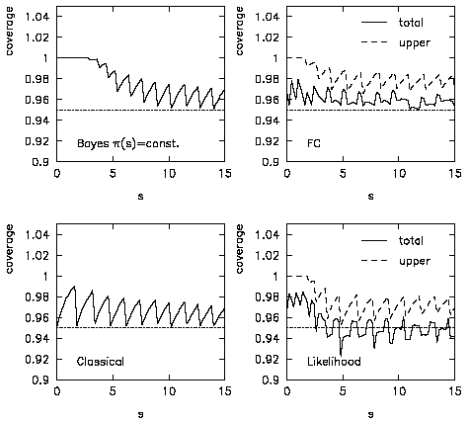
Critical region defined by low values of likelihood ratio.

Resulting intervals can be one- or two-sided (depending on n).

(Re)discovered for HEP by Feldman and Cousins, Phys. Rev. D 57 (1998) 3873.

Coverage probability of confidence intervals

Because of discreteness of Poisson data, probability for interval to include true value in general > confidence level ('over-coverage')



More on intervals from LR test (Feldman-Cousins)

Caveat with coverage: suppose we find $n >> b$.
Usually one then quotes a measurement: $\hat{s} = n - b\,, \quad \hat{\sigma}_{\hat{s}} = \sqrt{n}$

If, however, n isn't large enough to claim discovery, one sets a limit on s.

FC pointed out that if this decision is made based on n, then the actual coverage probability of the interval can be less than the stated confidence level ('flip-flopping').

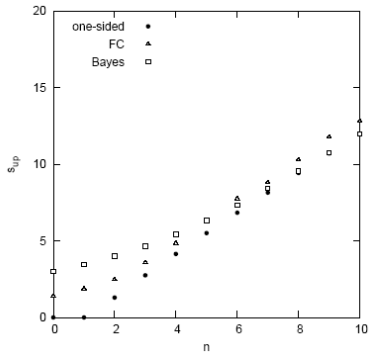
FC intervals remove this, providing a smooth transition from 1- to 2-sided intervals, depending on n.

But, suppose FC gives e.g. $0.1 < s < 5$ at 90% CL,
p-value of s=0 still substantial. Part of upper-limit 'wasted'?

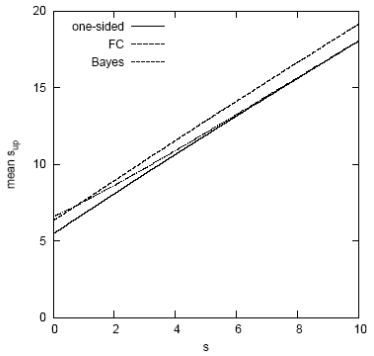
Properties of upper limits

Example: take $b = 5.0$, $1 - \gamma = 0.95$

Upper limit s_{up} vs. n

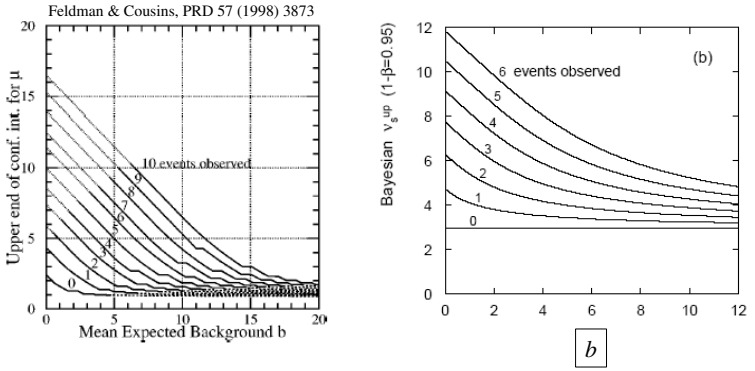


Mean upper limit vs. s



Upper limit versus b

Feldman & Cousins, PRD 57 (1998) 3873



If $n = 0$ observed, should upper limit depend on b?

Classical: yes
Bayesian: no
FC: yes