Statistical Data Analysis
Stat 3: *p*-values, parameter estimation

London Postgraduate Lectures on Particle Physics;

University of London MSci course PH4515

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Course web page:

www.pp.rhul.ac.uk/~cowan/stat_course.html

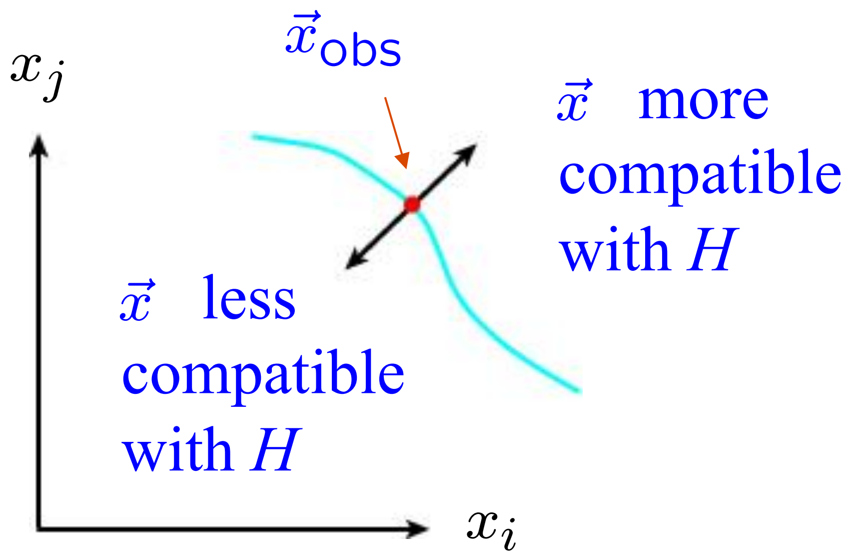
Testing significance / goodness-of-fit

Suppose hypothesis H predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\mathrm{obs}}$

What can we say about the validity of H in light of the data?

Decide what part of the data space represents less compatibility with H than does the point $\vec{x}_{\mathrm{obs}}$. (Not unique!)



p-values

Express 'goodness-of-fit' by giving the p-value for H:

p = probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

This is not the probability that H is true!

In frequentist statistics we don't talk about $P(H)$ (unless H represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for H.

For now stick with the frequentist approach;
result is p-value, regrettably easy to misinterpret as $P(H)$.

Distribution of the *p*-value

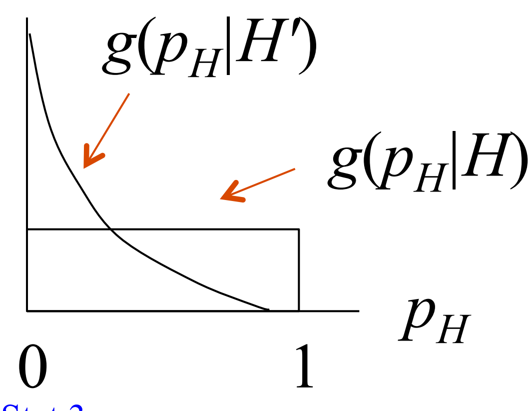
The *p*-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the *p*-value of *H* is found from a test statistic $t(\boldsymbol{x})$ as

$$p_H = \int_t^\infty f(t'|H)\,dt'$$

The pdf of p_H under assumption of *H* is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H/\partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \leq p_H \leq 1)$$

In general for continuous data, under assumption of *H*, $p_H \sim$ Uniform[0,1] and is concentrated toward zero for Some (broad) class of alternatives.



Using a p-value to define test of H_0

So the probability to find the p-value of H_0, p_0, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We started by defining critical region in the original data space ($\boldsymbol{x}$), then reformulated this in terms of a scalar test statistic $t(\boldsymbol{x})$.

We can take this one step further and define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.

p-value example: testing whether a coin is 'fair'

Probability to observe n heads in N coin tosses is binomial:

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Hypothesis H: the coin is fair ($p = 0.5$).

Suppose we toss the coin $N = 20$ times and get $n = 17$ heads.

Region of data space with equal or lesser compatibility with H relative to $n = 17$ is: $n = 17, 18, 19, 20, 0, 1, 2, 3$. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \; .$$

i.e. $p = 0.0026$ is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of H.

The Poisson counting experiment

Suppose we do a counting experiment and observe *n* events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s, e.g.,

test $s = 0$ (rejecting $H_0 \approx$ "discovery of signal process")

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

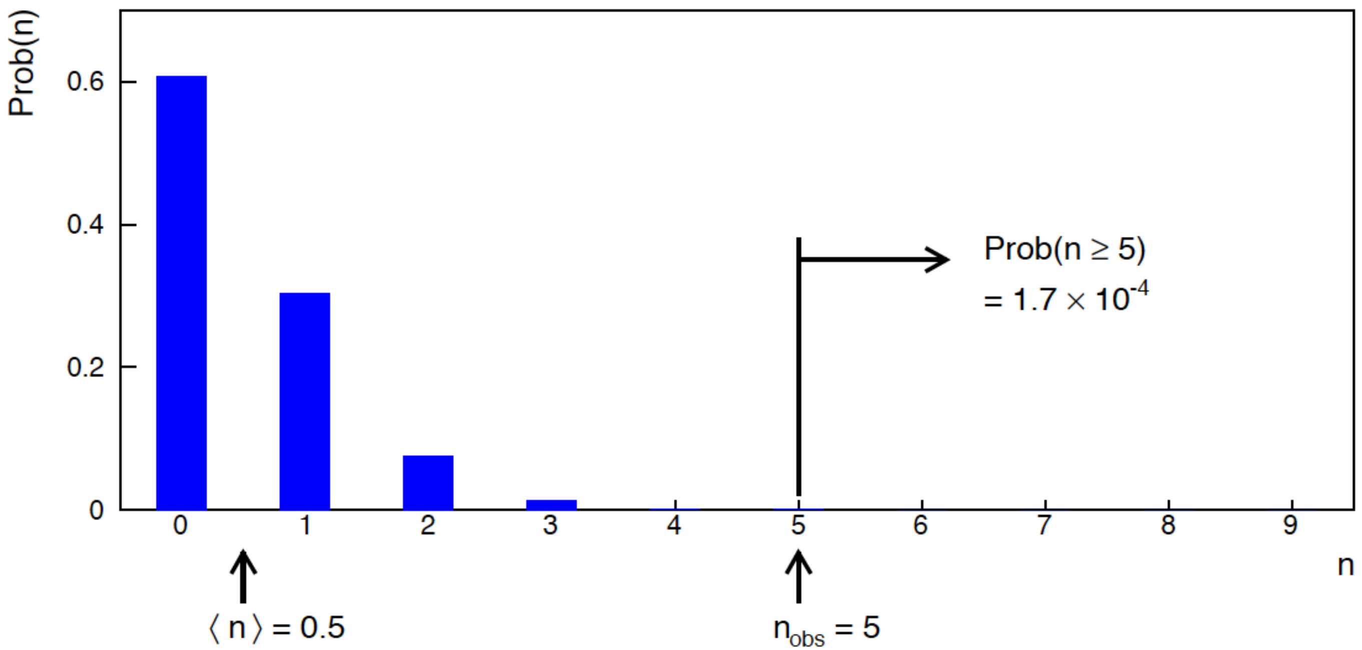
Poisson counting experiment: discovery p-value

Suppose $b = 0.5$ (known), and we observe $n_{obs} = 5$.

Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

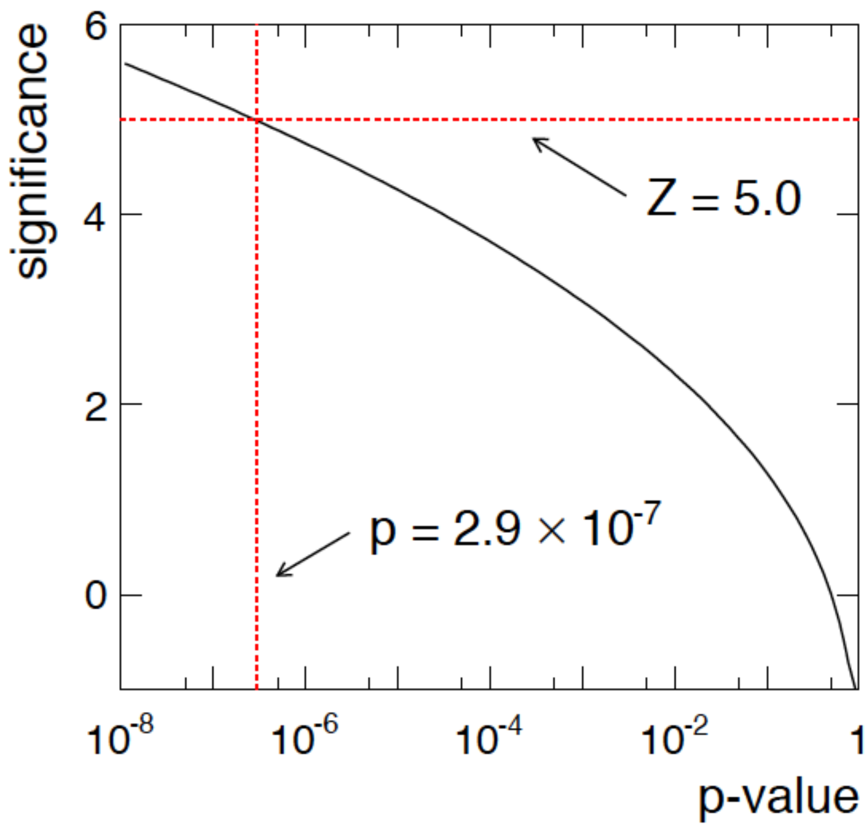
$$
\begin{aligned}
p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\
&= 1.7 \times 10^{-4} \neq P(s = 0)!
\end{aligned}
$$



Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1-p) = 3.6$

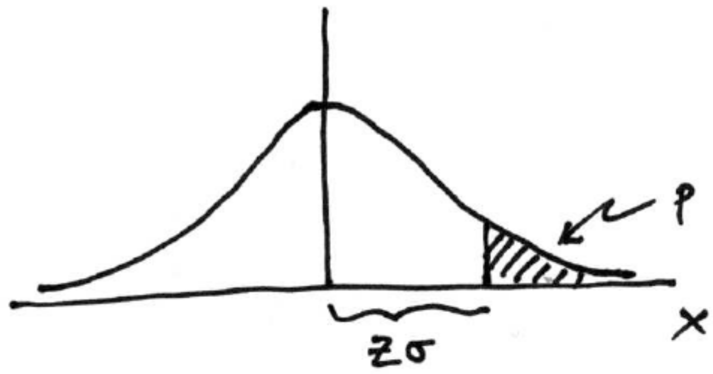
Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



In fact this tradition should be revisited: p-value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

Significance from p-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.

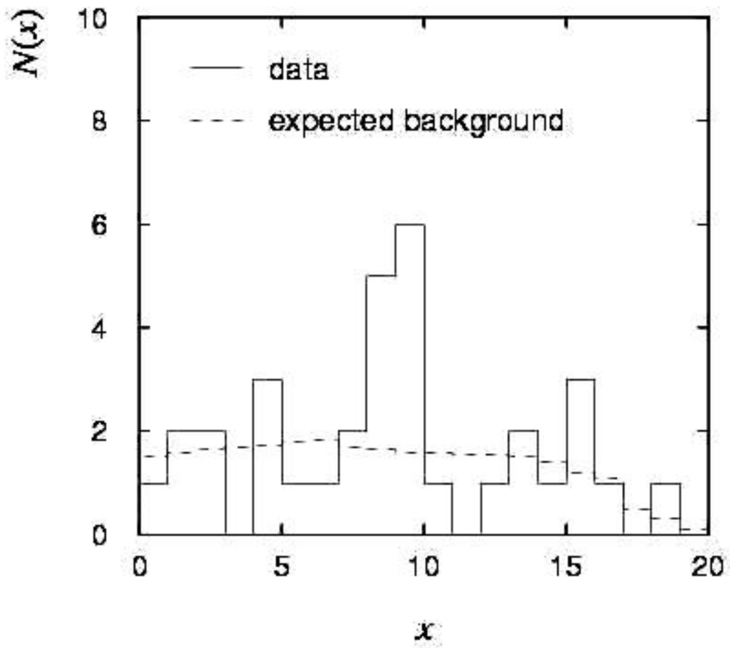


$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

The significance of a peak

Suppose we measure a value x for each event and find:



Each bin (observed) is a Poisson r.v., means are given by dashed lines.

In the two bins with the peak, 11 entries found with $b = 3.2$. The p-value for the $s = 0$ hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ "look-elsewhere effect";
want probability to find peak at least as significant as the one seen anywhere in the histogram.

How many bins × distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Is the observed width consistent with the expected x resolution?

→ take x window several times the expected resolution

Did we adjust the cuts to 'enhance' the peak?

→ freeze cuts, repeat analysis with new data

Should we publish????

When to publish (why 5 sigma?)

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = \Phi^{-1}(1 - p) = 5$ (a 5σ effect).

There a number of reasons why one may want to require such a high threshold for discovery:

The "cost" of announcing a false discovery is high.

Unsure about systematic uncertainties in model.

Unsure about look-elsewhere effect (multiple testing).

The implied signal may be a priori highly improbable (e.g., violation of Lorentz invariance).

But one should also consider the degree to which the data are compatible with the new phenomenon, not only the level of disagreement with the null hypothesis; p-value is only first step!

Why 5 sigma (cont.)?

The primary role of the *p*-value is to quantify the probability that the background-only model gives a statistical fluctuation as big as the one seen or bigger.

It is not intended as a means to protect against hidden systematics or the high standard required for a claim of an important discovery.

In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an “effect”, and the focus shifts to whether this is new physics or a systematic.

Providing LEE is dealt with, that threshold is probably closer to 3σ than 5σ.

Pearson's χ^2 statistic

Test statistic for comparing observed data $\vec{n} = (n_1, \ldots, n_N)$ (n_i independent) to predicted mean values $\vec{\nu} = (\nu_1, \ldots, \nu_N)$:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2}, \quad \text{where } \sigma_i^2 = V[n_i] \,.$$

(Pearson's χ^2 statistic)

χ^2 = sum of squares of the deviations of the ith measurement from the ith prediction, using σ_i as the 'yardstick' for the comparison.

For $n_i \sim$ Poisson(ν_i) we have $V[n_i] = \nu_i$, so this becomes

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} \,.$$

Pearson's χ^2 test

If n_i are Gaussian with mean ν_i and std. dev. σ_i, i.e., $n_i \sim N(\nu_i, \sigma_i^2)$, then Pearson's χ^2 will follow the χ^2 pdf (here for $\chi^2 = z$):

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

If the n_i are Poisson with $\nu_i >> 1$ (in practice OK for $\nu_i > 5$) then the Poisson dist. becomes Gaussian and therefore Pearson's χ^2 statistic here as well follows the χ^2 pdf.

The χ^2 value obtained from the data then gives the p-value:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z; N)\, dz \; .$$

The 'χ^2 per degree of freedom'

Recall that for the chi-square pdf for N degrees of freedom,

$$E[z] = N , \quad V[z] = 2N .$$

This makes sense: if the hypothesized n_i are right, the rms deviation of n_i from ν_i is σ_i, so each term in the sum contributes ~ 1.

One often sees χ^2/N reported as a measure of goodness-of-fit. But... better to give χ^2 and N separately. Consider, e.g.,

$$\begin{aligned} \chi^2 &= 15, \ N = 10 \rightarrow p - \text{value} = 0.13 , \\ \chi^2 &= 150, \ N = 100 \rightarrow p - \text{value} = 9.0 \times 10^{-4} . \end{aligned}$$

i.e. for N large, even a χ^2 per dof only a bit greater than one can imply a small p-value, i.e., poor goodness-of-fit.

Pearson's χ^2 with multinomial data

If $n_{\text{tot}} = \sum_{i=1}^{N}$ is fixed, then we might model $n_i \sim$ binomial

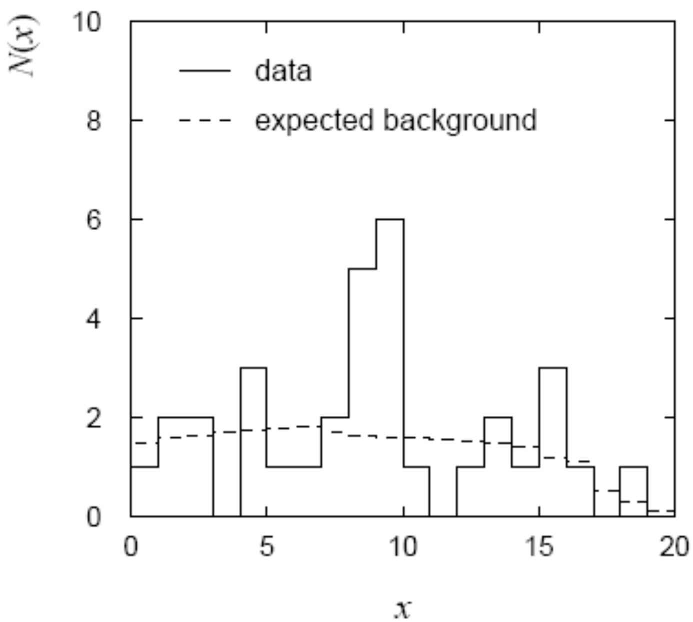
with $p_i = n_i / n_{\text{tot}}$. I.e. $\vec{n} = (n_1, \ldots, n_N) \sim$ multinomial.

In this case we can take Pearson's χ^2 statistic to be

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}}$$

If all $p_i \, n_{\text{tot}} >> 1$ then this will follow the chi-square pdf for $N-1$ degrees of freedom.

Example of a χ^2 test



← This gives

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for $N = 20$ dof.

Now need to find p-value, but... many bins have few (or no) entries, so here we do not expect χ^2 to follow the chi-square pdf.

Using MC to find distribution of χ^2 statistic

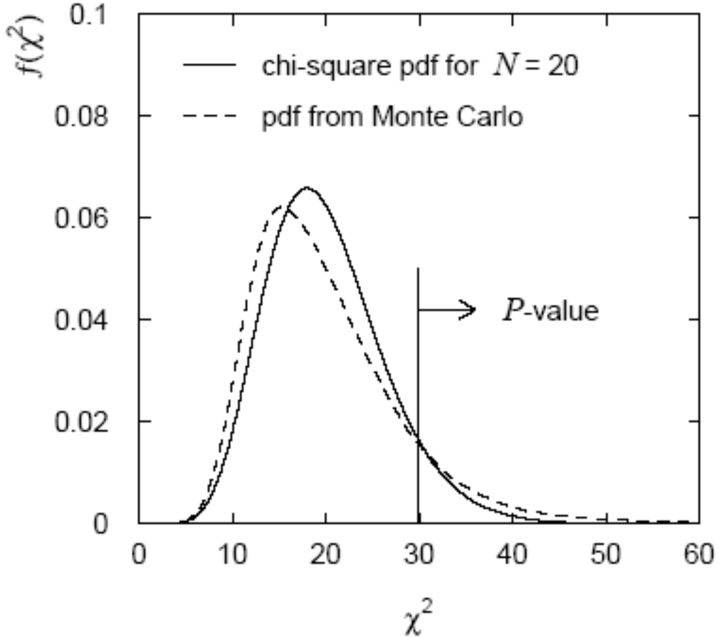
The Pearson χ^2 statistic still reflects the level of agreement between data and prediction, i.e., it is still a 'valid' test statistic.

To find its sampling distribution, simulate the data with a Monte Carlo program: $n_i \sim \text{Poisson}(\nu_i)\,, \; i = 1, N\,.$

Here data sample simulated 10^6 times. The fraction of times we find $\chi^2 > 29.8$ gives the p-value:

$$p = 0.11$$

If we had used the chi-square pdf we would find $p = 0.073$.



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: $\vec{x} = (x_1, \ldots, x_n)$

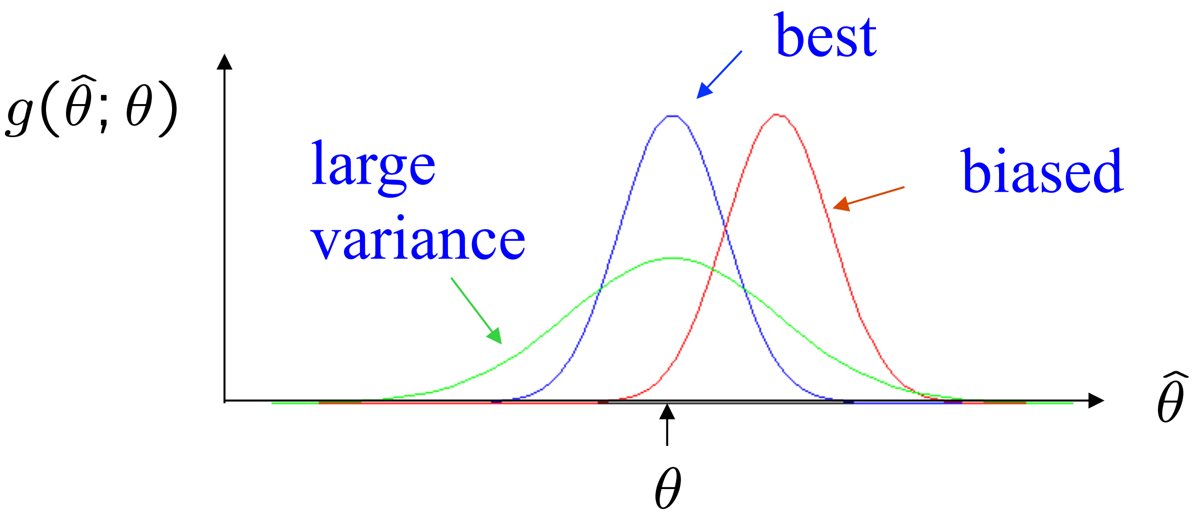
We want to find some function of the data to estimate the parameter(s):

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

Sometimes we say 'estimator' for the function of $x_1, \ldots, x_n$;
'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): $b = E[\widehat{\theta}] - \theta$

$\rightarrow$ average of repeated measurements should tend to true value.

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

$\rightarrow$ small bias & variance are in general conflicting criteria

An estimator for the mean (expectation value)

Parameter: $\mu = E[x]$

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

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

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

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 entire result of an experiment (set of measurements) is a collection of numbers ***x***, and suppose the joint pdf for the data ***x*** is a function that depends on a set of parameters ***θ***:

$$f(\vec{x};\vec{\theta})$$

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

$$\boxed{L(\vec{\theta}) = f(\vec{x};\vec{\theta})}$$

(***x*** constant)

The likelihood function for i.i.d.*. data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: x_1, ..., x_n, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

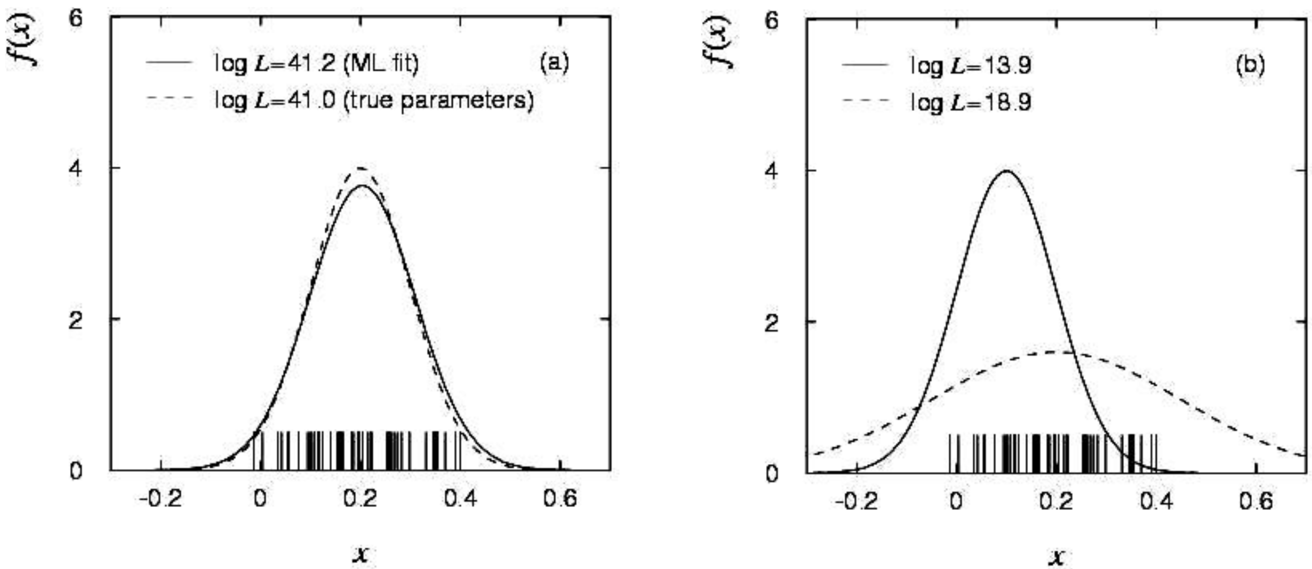
$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In this case the likelihood function is

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

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 i.i.d. 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):

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

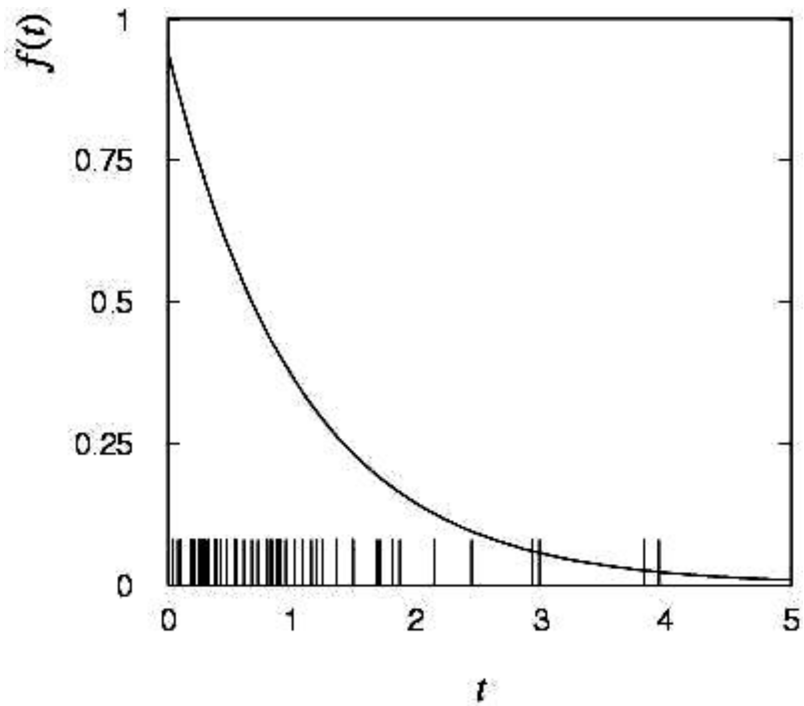
ML example: parameter of exponential pdf (2)

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:

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

Functions of ML estimators

Suppose we had written the exponential pdf as $f(t;\lambda) = \lambda e^{-\lambda t}$, i.e., we use $\lambda = 1/\tau$. What is the ML estimator for λ?

For a function $\alpha(\theta)$ of a parameter θ, it doesn't matter whether we express L as a function of α or θ.

The ML estimator of a function $\alpha(\theta)$ is simply $\hat{\alpha} = \alpha(\hat{\theta})$.

So for the decay constant we have $\hat{\lambda} = \frac{1}{\hat{\tau}} = \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right)^{-1}$.

Caveat: $\hat{\lambda}$ is biased, even though $\hat{\tau}$ is unbiased.

Can show $E[\hat{\lambda}] = \lambda \frac{n}{n-1}$. (bias $\to 0$ for $n \to \infty$)

Example of ML: parameters of Gaussian pdf

Consider independent $x_1, \ldots, x_n$, with $x_i \sim$ Gaussian (μ, σ^2)

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

The log-likelihood function is

$$\begin{aligned}
\ln L(\mu,\sigma^2) &= \sum_{i=1}^{n} \ln f(x_i;\mu,\sigma^2) \\
&= \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \ln \frac{1}{\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^2} \right) .
\end{aligned}$$

Example of ML: parameters of Gaussian pdf (2)

Set derivatives with respect to μ, σ^2 to zero and solve,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2 \,.$$

We already know that the estimator for μ is unbiased.

But we find, however, $E[\widehat{\sigma^2}] = \frac{n-1}{n}\sigma^2$, so ML estimator for σ^2 has a bias, but $b \to 0$ for $n \to \infty$. Recall, however, that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

is an unbiased estimator for σ^2.

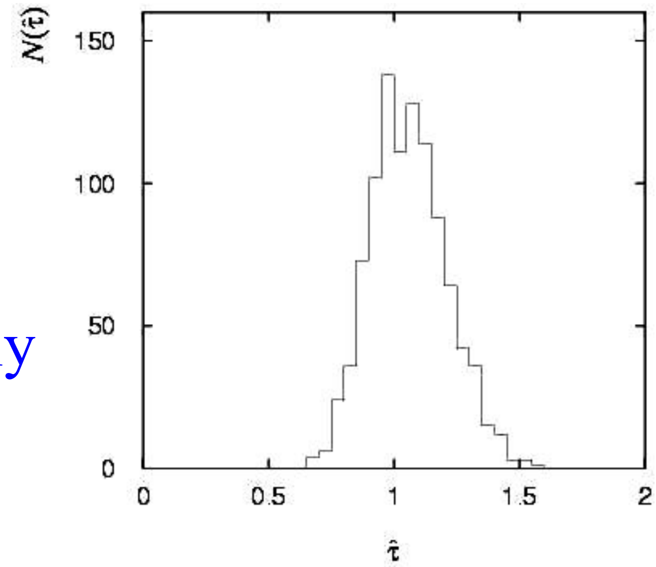
Variance of estimators: Monte Carlo method

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:

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



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):

Minimum Variance Bound (MVB)

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

$$(b = E[\widehat{\theta}] - \theta)$$

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

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

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

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

Variance of estimators: graphical method

Expand $\ln L(\theta)$ 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):

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

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

→ 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

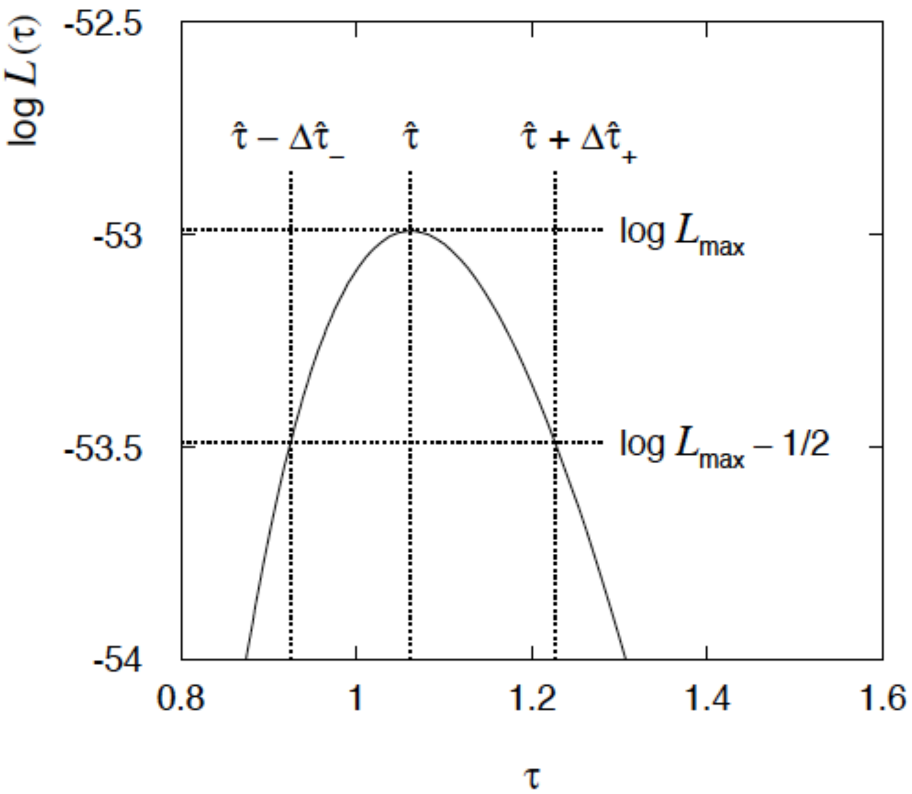
ML example with exponential:

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

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

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

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



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

Information inequality for N parameters

Suppose we have estimated N parameters $\vec{\theta} = (\theta_1, \ldots, \theta_N)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right] = -n \int f(x; \vec{\theta}) \frac{\partial^2 \ln f(x; \vec{\theta})}{\partial \theta_i \partial \theta_j} \, dx$$

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

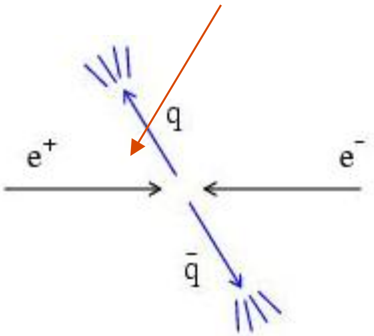
$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L.

Example of ML with 2 parameters

Consider a scattering angle distribution with $x = \cos\theta$,

$$f(x;\alpha,\beta) = \frac{1 + \alpha x + \beta x^2}{2 + 2\beta/3}$$



or if $x_{\min} < x < x_{\max}$, need always to normalize so that

$$\int_{x_{\min}}^{x_{\max}} f(x;\alpha,\beta)\, dx = 1 \ .$$

Example: $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$, generate $n = 2000$ events with Monte Carlo.

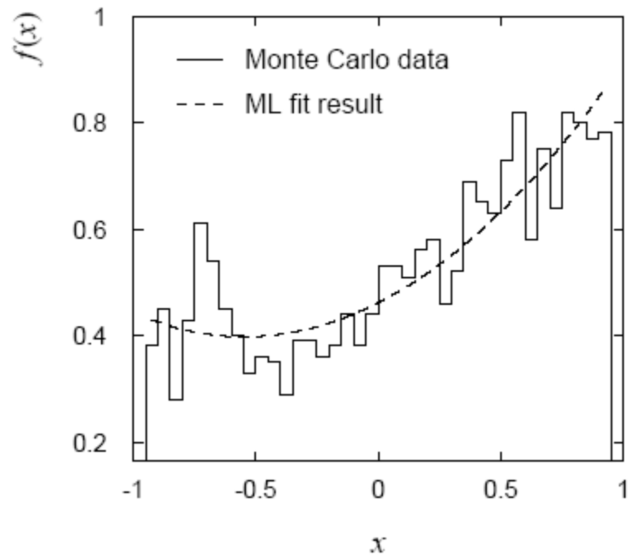
Example of ML with 2 parameters: fit result

Finding maximum of ln $L(\alpha, \beta)$ numerically (**MINUIT**) gives

$$\hat{\alpha} = 0.508$$

$$\hat{\beta} = 0.47$$

N.B. No binning of data for fit, but can compare to histogram for goodness-of-fit (e.g. 'visual' or χ^2).



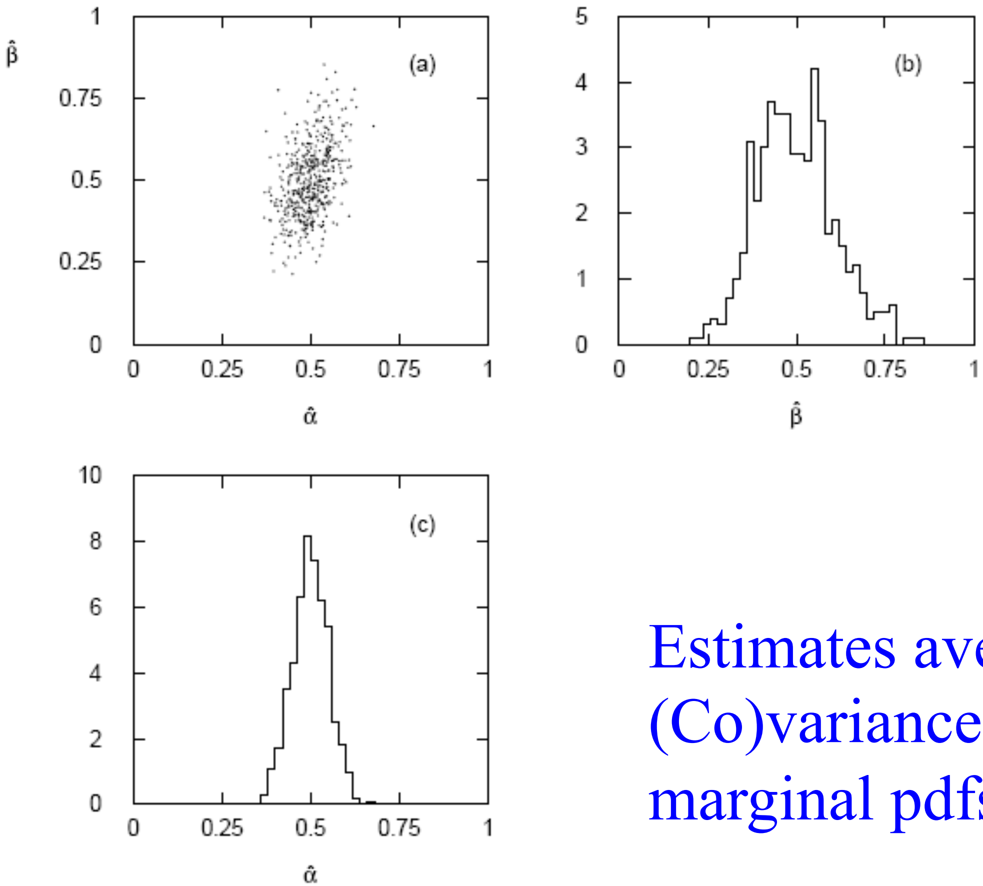
(Co)variances from $\widehat{(V^{-1})}_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta} = \hat{\vec{\theta}}}$ (**MINUIT** routine **HESSE**)

$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

$$\hat{\sigma}_{\hat{\beta}} = 0.11 \qquad r = 0.46$$

Two-parameter fit: MC study

Repeat ML fit with 500 experiments, all with n = 2000 events:



$$\overline{\hat{\alpha}} = 0.499$$
$$s_{\hat{\alpha}} = 0.051$$
$$\overline{\hat{\beta}} = 0.498$$
$$s_{\hat{\beta}} = 0.111$$
$$\widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0024$$
$$r = 0.42$$

Estimates average to ~ true values;
(Co)variances close to previous estimates;
marginal pdfs approximately Gaussian.

The ln $L_{\max}$ − 1/2 contour

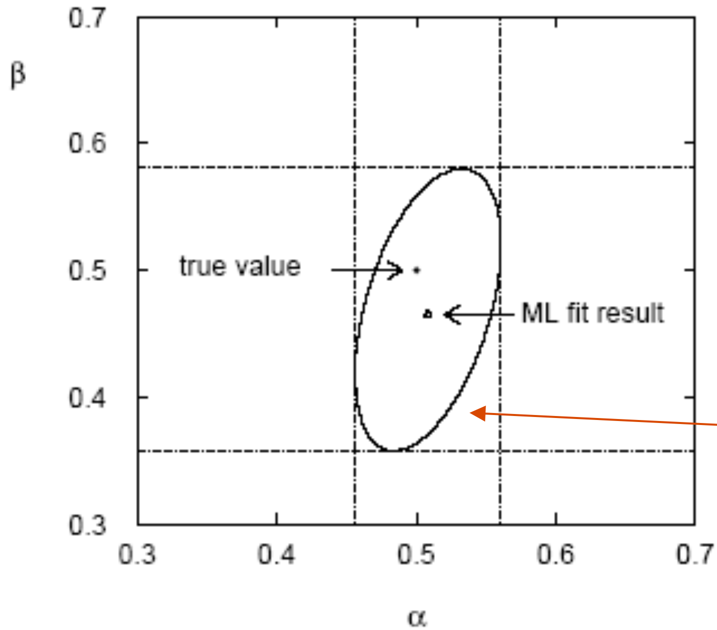
For large n, ln L takes on quadratic form near maximum:

$$\ln L(\alpha,\beta) \approx \ln L_{\max} - \frac{1}{2(1-\rho^2)}\left[\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2 + \left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right]$$

The contour $\ln L(\alpha,\beta) = \ln L_{\max} - 1/2$ is an ellipse:

$$\frac{1}{(1-\rho^2)}\left[\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)^2 + \left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha-\widehat{\alpha}}{\sigma_{\widehat{\alpha}}}\right)\left(\frac{\beta-\widehat{\beta}}{\sigma_{\widehat{\beta}}}\right)\right] = 1$$

(Co)variances from ln *L* contour



The α, β plane for the first MC data set

$$\ln L(\alpha, \beta) = \ln L_{\mathrm{max}} - 1/2$$

→ Tangent lines to contours give standard deviations.

→ Angle of ellipse ϕ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma^2_{\hat{\alpha}} - \sigma\hat{\beta}^2}$

Correlations between estimators result in an increase in their standard deviations (statistical errors).

Information inequality for *n* parameters

Suppose we have estimated *n* parameters $\vec{\theta} = (\theta_1, \ldots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -n \int f(x;\vec{\theta}) \frac{\partial^2 \ln f(x;\vec{\theta})}{\partial\theta_i \partial\theta_j}\, dx$$

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L.

Extended ML

Sometimes regard n not as fixed, but as a Poisson r.v., mean ν.

Result of experiment defined as: $n, x_1, \ldots, x_n$.

The (extended) likelihood function is:

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i; \vec{\theta})$$

Suppose theory gives $\nu = \nu(\boldsymbol{\theta})$, then the log-likelihood is

$$\ln L(\vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^{n} \ln(\nu(\vec{\theta}) f(x_i; \vec{\theta})) + C$$

where C represents terms not depending on $\boldsymbol{\theta}$.

Extended ML (2)

Example: expected number of events $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L \, dt$

where the total cross section $\sigma(\boldsymbol{\theta})$ is predicted as a function of the parameters of a theory, as is the distribution of a variable x.

Extended ML uses more info → smaller errors for $\hat{\vec{\theta}}$

Important e.g. for anomalous couplings in $e^+e^- \to W^+W^-$

If n does not depend on $\boldsymbol{\theta}$ but remains a free parameter, extended ML gives:

$$\begin{aligned} \hat{\nu} &= n \\ \hat{\theta} &= \text{same as ML} \end{aligned}$$

Extended ML example

Consider two types of events (e.g., signal and background) each of which predict a given pdf for the variable x: $f_s(x)$ and $f_b(x)$.

We observe a mixture of the two event types, signal fraction = θ, expected total number = ν, observed total number = n.

Let $\mu_{\mathrm{s}} = \theta\nu$, $\mu_{\mathrm{b}} = (1-\theta)\nu$, goal is to estimate μ_s, μ_b.

$$f(x;\mu_{\mathrm{s}},\mu_{\mathrm{b}}) = \frac{\mu_{\mathrm{s}}}{\mu_{\mathrm{s}}+\mu_{\mathrm{b}}} f_{\mathrm{s}}(x) + \frac{\mu_{\mathrm{b}}}{\mu_{\mathrm{s}}+\mu_{\mathrm{b}}} f_{\mathrm{b}}(x)$$

$$P(n;\mu_{\mathrm{s}},\mu_{\mathrm{b}}) = \frac{(\mu_{\mathrm{s}}+\mu_{\mathrm{b}})^n}{n!} e^{-(\mu_{\mathrm{s}}+\mu_{\mathrm{b}})}$$

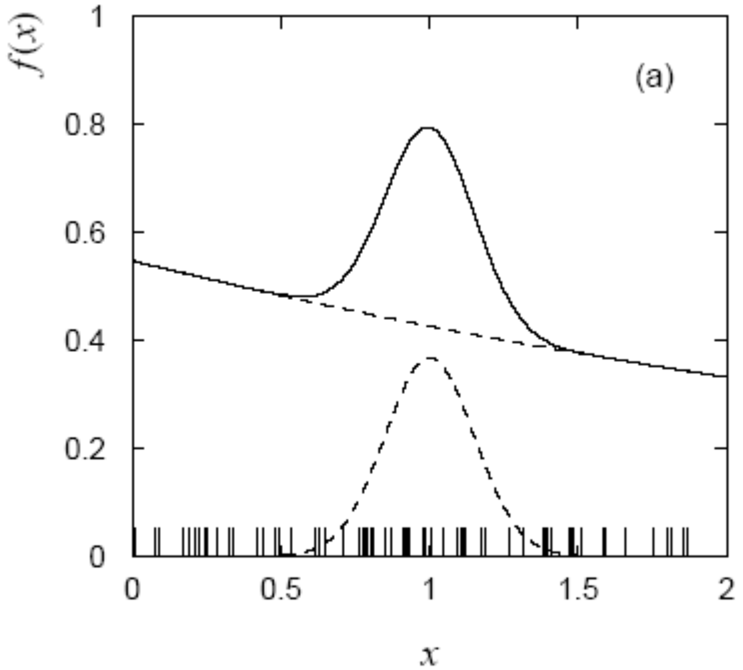
$$\rightarrow \ln L(\mu_{\mathrm{s}},\mu_{\mathrm{b}}) = -(\mu_{\mathrm{s}}+\mu_{\mathrm{b}}) + \sum_{i=1}^{n} \ln\left[(\mu_{\mathrm{s}}+\mu_{\mathrm{b}}) f(x_i;\mu_{\mathrm{s}},\mu_{\mathrm{b}})\right]$$

Extended ML example (2)

Monte Carlo example
with combination of
exponential and Gaussian:

$$\mu_{\rm s} = 6$$

$$\mu_{\rm b} = 60$$



Maximize log-likelihood in terms of μ_s and μ_b:

$$\hat{\mu}_{\rm s} = 8.7 \pm 5.5$$

$$\hat{\mu}_{\rm b} = 54.3 \pm 8.8$$

Here errors reflect total Poisson fluctuation as well as that in proportion of signal/background.

ML with binned data

Often put data into a histogram: $\vec{n} = (n_1, \ldots, n_N), \; n_{\rm tot} = \sum_{i=1}^{N} n_i$

Hypothesis is $\vec{\nu} = (\nu_1, \ldots, \nu_N), \; \nu_{\rm tot} = \sum_{i=1}^{N} \nu_i$ where

$$\nu_i(\vec{\theta}) = \nu_{\rm tot} \int_{{\rm bin}\, i} f(x; \vec{\theta})\, dx$$

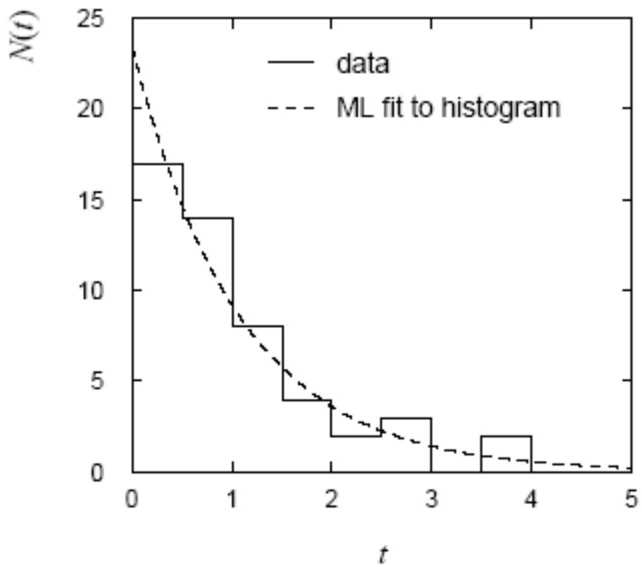
If we model the data as multinomial ($n_{\rm tot}$ constant),

$$f(\vec{n}; \vec{\nu}) = \frac{n_{\rm tot}!}{n_1! \ldots n_N!} \left(\frac{\nu_1}{n_{\rm tot}} \right)^{n_1} \cdots \left(\frac{\nu_N}{n_{\rm tot}} \right)^{n_N}$$

then the log-likelihood function is: $\ln L(\vec{\theta}) = \sum_{i=1}^{N} n_i \ln \nu_i(\vec{\theta}) + C$

ML example with binned data

Previous example with exponential, now put data into histogram:



$\hat{\tau} = 1.07 \pm 0.17$

(1.06 ± 0.15 for unbinned

ML with same sample)

Limit of zero bin width → usual unbinned ML.

If n_i treated as Poisson, we get extended log-likelihood:

$$\ln L(\nu_{\text{tot}}, \vec{\theta}) = -\nu_{\text{tot}} + \sum_{i=1}^{N} n_i \ln \nu_i(\nu_{\text{tot}}, \vec{\theta}) + C$$

Relationship between ML and Bayesian estimators

In Bayesian statistics, both $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are random variables:

$$L(\theta) = L(\vec{x}|\theta) = f_{\mathsf{joint}}(\vec{x}|\theta)$$

Recall the Bayesian method:

Use subjective probability for hypotheses (θ);

before experiment, knowledge summarized by prior pdf $\pi(\theta)$;

use Bayes' theorem to update prior in light of data:

$$p(\theta|\vec{x}) = \frac{L(\vec{x}|\theta)\pi(\theta)}{\int L(\vec{x}|\theta')\pi(\theta')\,d\theta'}$$

Posterior pdf (conditional pdf for $\boldsymbol{\theta}$ given $\boldsymbol{x}$)

ML and Bayesian estimators (2)

Purist Bayesian: $p(\theta|x)$ contains all knowledge about θ.

Pragmatist Bayesian: $p(\theta|x)$ could be a complicated function,

→ summarize using an estimator $\hat{\theta}_{\mathrm{Bayes}}$

Take mode of $p(\theta|x)$, (could also use e.g. expectation value)

What do we use for $\pi(\theta)$? No golden rule (subjective!), often represent 'prior ignorance' by $\pi(\theta)$ = constant, in which case

$$\hat{\theta}_{\mathrm{Bayes}} = \hat{\theta}_{\mathrm{ML}}$$

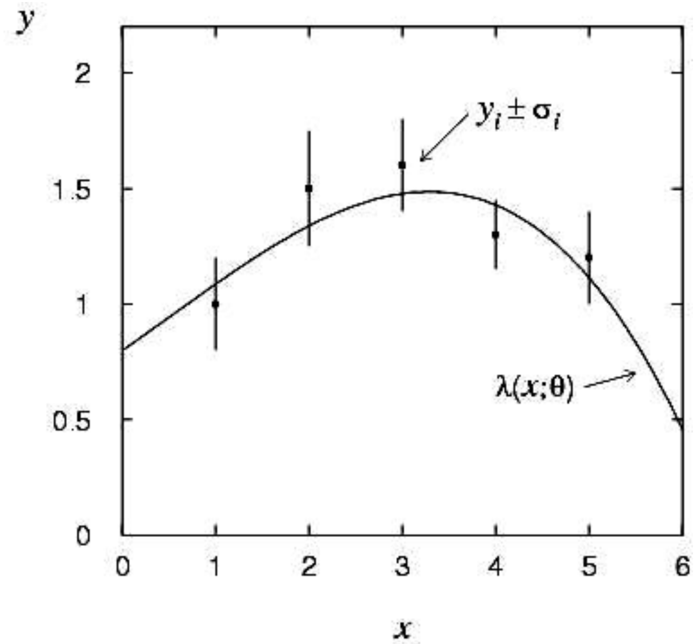
But... we could have used a different parameter, e.g., $\lambda = 1/\theta$, and if prior $\pi_\theta(\theta)$ is constant, then $\pi_\lambda(\lambda)$ is not!

'Complete prior ignorance' is not well defined.

The method of least squares

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

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



Assume known values of the control variable $x_1, \ldots, x_N$ 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 defines the least squares (LS) estimator $\hat{\theta}$.

Very often measurement errors are ~Gaussian and so ML and LS are essentially the same.

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

LS references

C.F. Gauss, Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambentium, Hamburgi Sumtibus Frid. Perthes et H.Besser Liber II, Sectio II (1809);

C.F. Gauss, Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, pars prior (15.2.1821) et pars posterior (2.2.1823), Commentationes Societatis Regiae Scientiarium Gottingensis Recectiores Vol. V (MDCCCXXIII).

LS with correlated measurements

If the y_i follow a multivariate Gaussian, covariance matrix V,

$$g(\vec{y}, \vec{\lambda}, V) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda})\right]$$

Then maximizing the likelihood is equivalent to minimizing

$$\chi^2(\vec{\theta}) = \sum_{i,j=1}^{N} (y_i - \lambda(x_i; \vec{\theta}))(V^{-1})_{ij}(y_j - \lambda(x_j; \vec{\theta}))$$

Linear LS problem

LS has particularly simple properties if $\lambda(x; \vec{\theta})$ linear in $\vec{\theta}$:

$$\lambda(x; \vec{\theta}) = \sum_{j=1}^{m} a_j(x)\theta_j$$

where $a_j(x)$ are any linearly independent functions of x.

Matrix notation: let $A_{ij} = a_j(x_i)$,

$$\chi^2(\vec{\theta}) = (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda})$$

$$= (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta})$$

Linear LS problem (2)

Set derivitives with respect to θ_i to zero,

$$\nabla \chi^2 = -2(A^T V^{-1} \vec{y} - A^T V^{-1} A \vec{\theta}) = 0$$

Solve to get the LS estimators,

$$\hat{\vec{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{y} \equiv B \vec{y}$$

N.B. estimators $\hat{\theta}_i$ are linear functions of the measurements y_i.

Linear LS problem (3)

Error propagation (exact for linear problem) for $U_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$:

$$U = B\,V\,B^T = (A^T V^{-1} A)^{-1}$$

Equivalently, use

$$(U^{-1})_{ij} = \frac{1}{2}\left[\frac{\partial^2 \chi^2}{\partial\theta_i \partial\theta_j}\right]_{\vec{\theta}=\hat{\vec{\theta}}}$$

Equals MVB if y_i Gaussian)

Linear LS problem (4)

For $\lambda(x;\vec{\theta})$ linear in the parameters, $\chi^2(\vec{\theta})$ is quadratic,

$$\chi^2(\vec{\theta}) = \chi^2(\hat{\vec{\theta}}) + \frac{1}{2}\sum_{i,j=1}^{m}\left[\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\right]_{\vec{\theta}=\hat{\vec{\theta}}}(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

$\rightarrow$ variances from tangent planes to (hyper)ellipse,

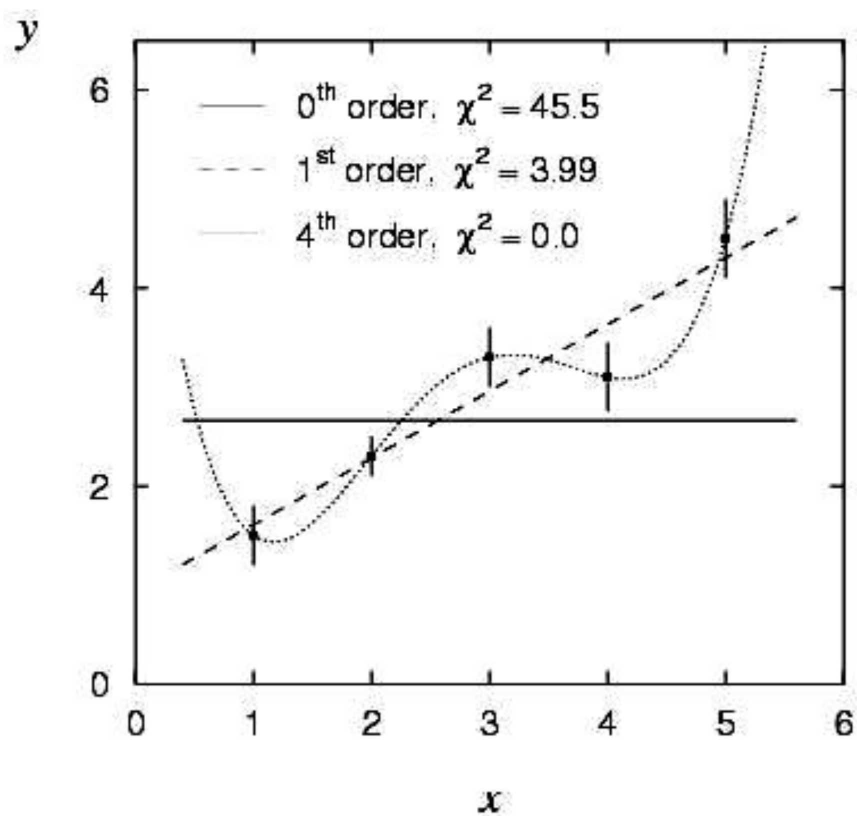
$$\chi^2(\vec{\theta}) = \chi^2(\hat{\vec{\theta}}) + 1 = \chi^2_{\min} + 1$$

If $\lambda(x;\vec{\theta})$ not linear in $\vec{\theta}$, then expressions above not exact

(but may still be good approximations).

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

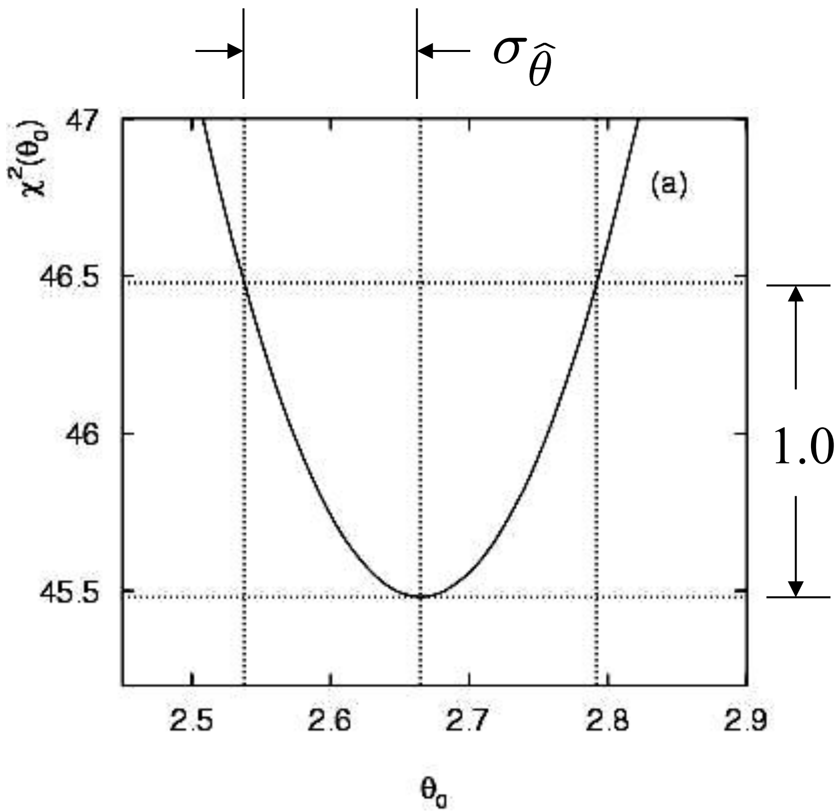
$$\chi^2(\theta) = -2 \ln L(\theta)$$

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

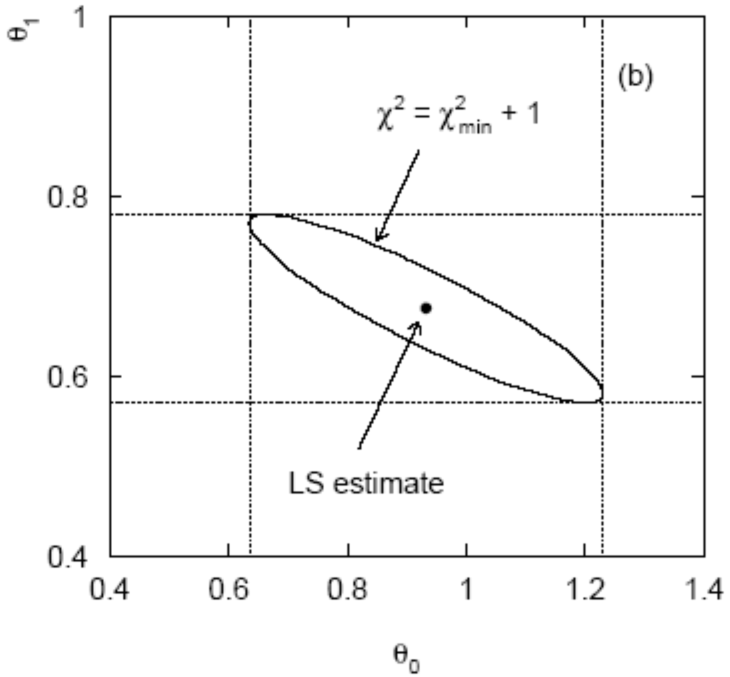
$$\chi^2(\theta) = \chi^2_{\min} + 1$$



Two-parameter LS fit

2-parameter case (line with nonzero slope):

$$\hat{\theta}_0 = 0.93 \pm 0.30,$$
$$\hat{\theta}_1 = 0.68 \pm 0.10$$
$$\widehat{\mathrm{cov}}[\hat{\theta}_0, \hat{\theta}_1] = -0.028$$
$$r = -0.90$$
$$\chi^2 = 3.99$$



Tangent lines $\rightarrow$ $\sigma_{\hat{\theta}_0}$, $\sigma_{\hat{\theta}_1}$.

Angle of ellipse $\rightarrow$ correlation (same as for ML)

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:

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

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 $t = \chi^2_{\min}$ follows the chi-square pdf,

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

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 $\chi^2_{\min} \approx n_{\mathrm{d}}$ the fit is 'good'.

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

This is the probability of obtaining a $\chi^2_{\min}$ 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),

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

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

$$\chi^2_{\min} = 45.5\,, \qquad n_{\mathrm{d}} = 5-1 = 4\,, \qquad p = 3.1\times 10^{-9}$$

Goodness-of-fit vs. statistical errors

Small statistical error does not mean a good fit (nor vice versa).

Curvature of χ^2 near its minimum $\rightarrow$ statistical errors ($\sigma_{\hat{\theta}}$)

Value of $\chi^2_{\rm min}$ $\rightarrow$ goodness-of-fit

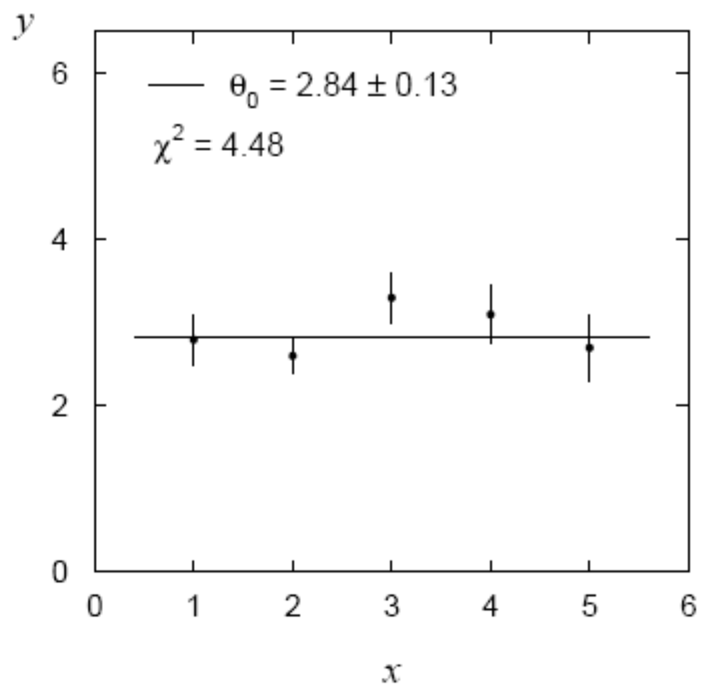
Horizontal line fit, move the data points, keep errors on points same:

$$\hat{\theta}_0 = 2.84 \pm 0.13$$

$$\chi^2_{\rm min} = 4.48$$

Variance same as before,

now $\chi^2_{\rm min}$ 'good'.



Goodness-of-fit vs. stat. errors (2)

$\rightarrow \chi^2(\theta_0)$ shifted down, same curvature as before.

Variance of estimator (statistical error) tells us:

if experiment repeated many times, how wide is the distribution of the estimates $\hat{\theta}$. (Doesn't tell us whether hypothesis correct.)

P-value tells us:

if hypothesis is correct and experiment repeated many times, what fraction will give equal or worse agreement between data and hypothesis according to the statistic $\chi^2_{\min}$.

Low P-value $\rightarrow$ hypothesis may be wrong $\rightarrow$ systematic error.

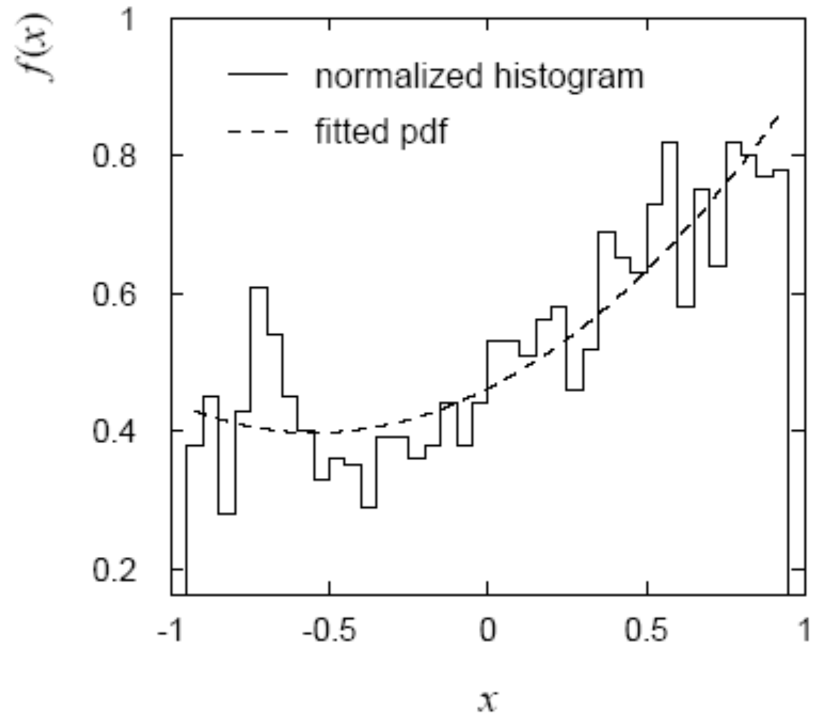
LS with binned data

Histogram:

N bins, n entries.

Hypothesized pdf:

$f(x;\vec{\theta})$



We have

$$y_i = \text{number of entries in bin } i,$$

$$\lambda_i(\vec{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x;\vec{\theta})dx = np_i(\vec{\theta})$$

LS with binned data (2)

LS fit: minimize

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$$

where $\sigma_i^2 = V[y_i]$, here not known a priori.

Treat the y_i as Poisson r.v.s, in place of true variance take either

$\sigma_i^2 = \lambda_i(\vec{\theta})$ (LS method)

$\sigma_i^2 = y_i$ (Modified LS method)

MLS sometimes easier computationally, but χ^2_{min} no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.

LS with binned data — normalization

Do not 'fit the normalization':

$$\lambda_i(\vec{\theta}, \nu) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = \nu p_i(\vec{\theta})$$

i.e. introduce adjustable ν, fit along with $\vec{\theta}$.

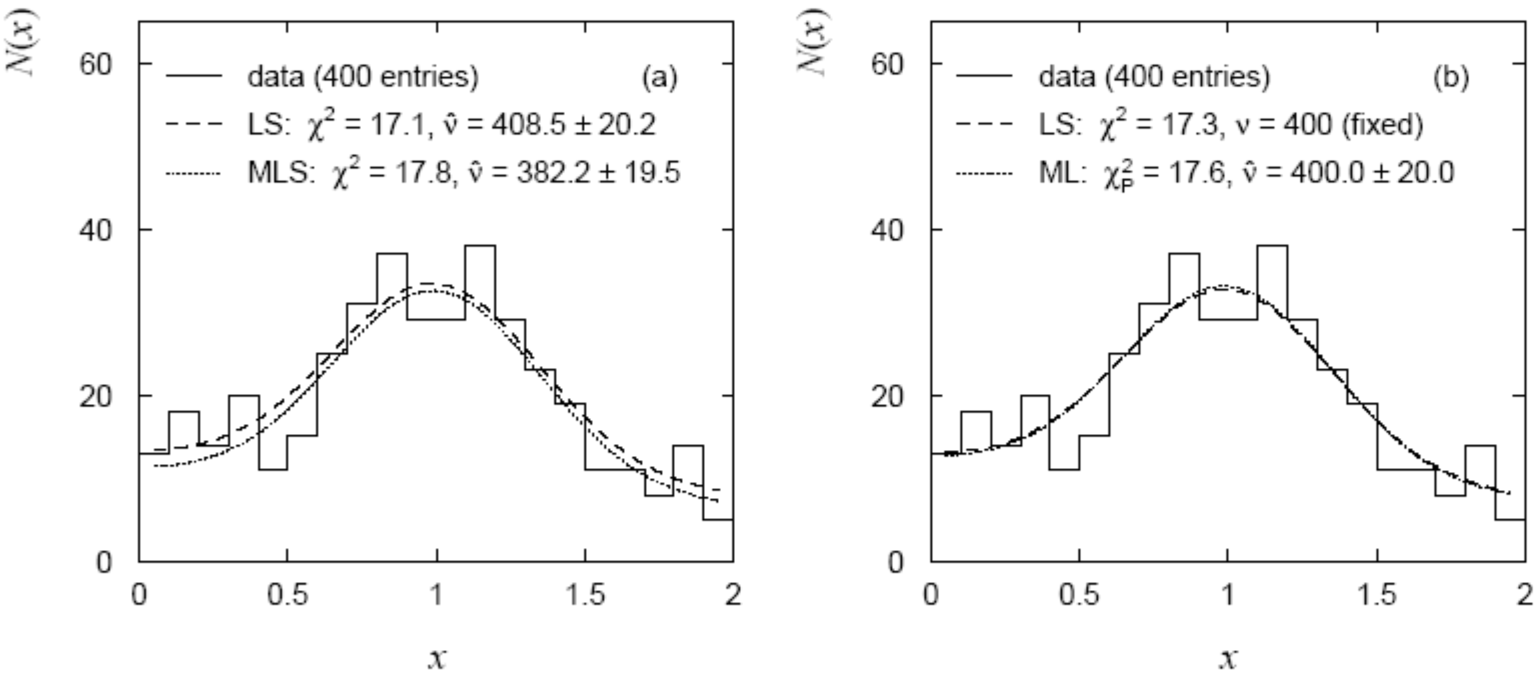
$\hat{\nu}$ is a bad estimator for n (which we know, anyway!)

$$\hat{\nu}_{\mathrm{LS}} = n + \frac{\chi^2_{\min}}{2}$$

$$\hat{\nu}_{\mathrm{MLS}} = n - \chi^2_{\min}$$

LS normalization example

Example with $n = 400$ entries, $N = 20$ bins:



Expect $\chi^2_{\min}$ around $N - m$,

$\rightarrow$ relative error in $\hat{\nu}$ large when N large, n small

Either get n directly from data for LS (or better, use ML).

Goodness of fit from the likelihood ratio

Suppose we model data using a likelihood $L(\boldsymbol{\mu})$ that depends on N parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$. Define the statistic

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})}$$

where $\hat{\boldsymbol{\mu}}$ is the ML estimator for $\boldsymbol{\mu}$. Value of $t_{\boldsymbol{\mu}}$ reflects agreement between hypothesized $\boldsymbol{\mu}$ and the data.

Good agreement means $\hat{\boldsymbol{\mu}} \approx \boldsymbol{\mu}$, so $t_{\boldsymbol{\mu}}$ is small;

Larger $t_{\boldsymbol{\mu}}$ means less compatibility between data and $\boldsymbol{\mu}$.

Quantify "goodness of fit" with p-value: $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) \, dt_{\boldsymbol{\mu}}$

need this pdf

Likelihood ratio (2)

Now suppose the parameters $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$ can be determined by another set of parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$, with $M < N$.

E.g. in LS fit, use $\mu_i = \mu(x_i; \boldsymbol{\theta})$ where x is a control variable.

Define the statistic

fit M parameters

$$q_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}$$

fit N parameters

Use $q_{\boldsymbol{\mu}}$ to test hypothesized functional form of $\mu(x; \boldsymbol{\theta})$.

To get p-value, need pdf $f(q_{\boldsymbol{\mu}}|\mu)$.

Wilks' Theorem (1938)

Wilks' Theorem: if the hypothesized parameters $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)$ are true then in the large sample limit (and provided certain conditions are satisfied) $t_{\boldsymbol{\mu}}$ and $q_{\boldsymbol{\mu}}$ follow chi-square distributions.

For case with $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)$ fixed in numerator:

$$t_{\boldsymbol{\mu}} = -2\ln\frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})} \qquad f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) \sim \chi^2_N$$

Or if M parameters adjusted in numerator,

degrees of freedom

$$q_{\boldsymbol{\mu}} = -2\ln\frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})} \qquad f(q_{\boldsymbol{\mu}}|\boldsymbol{\mu}) \sim \chi^2_{N-M}$$

S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

Goodness of fit with Gaussian data

Suppose the data are N independent Gaussian distributed values:

$$y_i \sim \text{Gauss}(\mu_i, \sigma_i) \,, \qquad i = 1, \ldots, N$$

want to estimate

known

Likelihood: $$L(\boldsymbol{\mu}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2}$$

Log-likelihood: $$\ln L(\boldsymbol{\mu}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C$$

ML estimators: $$\hat{\mu}_i = y_i \qquad i = 1, \ldots, N$$

Likelihood ratios for Gaussian data

The goodness-of-fit statistics become

$$t_\mu = -2\ln\frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})} = \sum_{i=1}^{N}\frac{(y_i - \mu_i)^2}{\sigma_i^2} \qquad f(t_\mu|\boldsymbol{\mu}) \sim \chi^2_N$$

$$q_\mu = -2\ln\frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})} = \sum_{i=1}^{N}\frac{(y_i - \mu_i(\hat{\boldsymbol{\theta}}))^2}{\sigma_i^2} \qquad f(q_\mu|\boldsymbol{\mu}) \sim \chi^2_{N-M}$$

So Wilks' theorem formally states the well-known property of the minimized chi-squared from an LS fit.

Likelihood ratio for Poisson data

Suppose the data are a set of values $\boldsymbol{n} = (n_1,\ldots, n_N)$, e.g., the numbers of events in a histogram with N bins.

Assume $n_i \sim \text{Poisson}(\nu_i)$, $i = 1,\ldots, N$, all independent.

Goal is to estimate $\boldsymbol{\nu} = (\nu_1,\ldots, \nu_N)$.

Likelihood:
$$L(\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Log-likelihood:
$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} [n_i \ln \nu_i - \nu_i] + C$$

ML estimators:
$$\hat{\nu}_i = n_i \;, \qquad i = 1, \ldots, N$$

Goodness of fit with Poisson data

The likelihood ratio statistic (all parameters fixed in numerator):

$$
\begin{aligned}
t_{\boldsymbol{\nu}} &= -2 \ln \frac{L(\boldsymbol{\nu})}{L(\hat{\boldsymbol{\nu}})} \\
&= -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i}{\hat{\nu}_i} - \nu_i + \hat{\nu}_i \right] \\
&= -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i}{n_i} - \nu_i + n_i \right]
\end{aligned}
$$

Wilks' theorem: $f(t_{\boldsymbol{\nu}} | \boldsymbol{\nu}) \sim \chi^2_N$

Goodness of fit with Poisson data (2)

Or with M fitted parameters in numerator:

$$q_{\nu} = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} - \nu_i(\hat{\boldsymbol{\theta}}) + n_i \right]$$

Wilks' theorem: $f(q_{\nu}|\boldsymbol{\nu}) \sim \chi^2_{N-M}$

Use t_{μ}, q_{μ} to quantify goodness of fit (p-value).

Sampling distribution from Wilks' theorem (chi-square).

Exact in large sample limit; in practice good approximation for surprisingly small n_i (~several).

Goodness of fit with multinomial data

Similar if data $\boldsymbol{n} = (n_1,\ldots, n_N)$ follow multinomial distribution:

$$P(\mathbf{n}|\mathbf{p}, n_{\rm tot}) = \frac{n_{\rm tot}!}{n_1! n_2! \ldots n_N!} p_1^{n_1} p_2^{n_2} \ldots p_N^{n_N}$$

E.g. histogram with N bins but fix: $$n_{\rm tot} = \sum_{i=1}^{N} n_i$$

Log-likelihood: $$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_{\rm tot}} + C \qquad (\nu_i = p_i n_{\rm tot})$$

ML estimators: $\hat{\nu}_i = n_i$ (Only $N-1$ independent; one is $n_{\rm tot}$ minus sum of rest.)

Goodness of fit with multinomial data (2)

The likelihood ratio statistics become:

$$t_{\nu} = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_i} \qquad f(t_{\nu}|\boldsymbol{\nu}) \sim \chi^2_{N-1}$$

$$q_{\nu} = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} \qquad f(q_{\nu}|\boldsymbol{\nu}) \sim \chi^2_{N-M-1}$$

One less degree of freedom than in Poisson case because effectively only $N-1$ parameters fitted in denominator.

Estimators and g.o.f. all at once

Evaluate numerators with $\boldsymbol{\theta}$ (not its estimator):

$$\chi^2_{\mathrm{P}}(\boldsymbol{\theta}) = -2 \sum_{i=1}^{N} \left[n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} - \nu_i(\boldsymbol{\theta}) + n_i \right] \qquad \text{(Poisson)}$$

$$\chi^2_{\mathrm{M}}(\boldsymbol{\theta}) = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i} \qquad \text{(Multinomial)}$$

These are equal to the corresponding $-2 \ln L(\boldsymbol{\theta})$ plus terms not depending on $\boldsymbol{\theta}$, so minimizing them gives the usual ML estimators for $\boldsymbol{\theta}$.

The minimized value gives the statistic q_μ, so we get goodness-of-fit for free.

Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.

Using LS to combine measurements

Use LS to obtain weighted average of N measurements of λ:

$$y_i = \text{result of measurement } i,\ i = 1, \ldots, N;$$

$$\sigma_i^2 = V[y_i], \text{ assume known};$$

$$\lambda = \text{true value (plays role of } \theta).$$

For uncorrelated y_i, minimize

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

Set $\frac{\partial \chi^2}{\partial \lambda} = 0$ and solve,

$$\rightarrow \quad \hat{\lambda} = \frac{\Sigma_{i=1}^N y_i/\sigma_i^2}{\Sigma_{j=1}^N 1/\sigma_j^2} \qquad V[\hat{\lambda}] = \frac{1}{\Sigma_{i=1}^N 1/\sigma_i^2}$$

Combining correlated measurements with LS

If $\mathrm{cov}[y_i, y_j] = V_{ij}$, minimize

$$\chi^2(\lambda) = \sum_{i,j=1}^{N} (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda) \,,$$

$$\rightarrow \quad \hat{\lambda} = \sum_{i=1}^{N} w_i y_i, \qquad w_i = \frac{\sum_{j=1}^{N} (V^{-1})_{ij}}{\sum_{k,l=1}^{N} (V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^{N} w_i V_{ij} w_j$$

LS $\hat{\lambda}$ has zero bias, minimum variance (Gauss–Markov theorem).

Example: averaging two correlated measurements

Suppose we have y_1, y_2, and $V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

$$\rightarrow \quad \hat{\lambda} = w y_1 + (1-w) y_2, \quad w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$V[\hat{\lambda}] = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \sigma^2$$

The increase in inverse variance due to 2nd measurement is

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} = \frac{1}{1-\rho^2}\left(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2}\right)^2 > 0$$

$\rightarrow$ 2nd measurement can only help.

Negative weights in LS average

If $\rho > \sigma_1/\sigma_2$, $\rightarrow$ $w < 0$,

$\rightarrow$ weighted average is not between y_1 and y_2 (!?)

Cannot happen if correlation due to common data, but possible for shared random effect; very unreliable if e.g. ρ, σ_1, σ_2 incorrect.

See example in SDA Section 7.6.1 with two measurements at same temperature using two rulers, different thermal expansion coefficients:

average is outside the two measurements; used to improve estimate of temperature.