Statistics course outline

Lecture 1

1. **Probability**
2. **Random variables, probability densities, etc.**
3. **Brief catalogue of probability densities**
4. **The Monte Carlo method**

Lecture 2

1. **Statistical tests**
2. **Fisher discriminants, neural networks, etc.**
3. **Goodness-of-fit tests**
4. **The significance of a signal**
5. **Introduction to parameter estimation**

Lecture 3

1. **The method of maximum likelihood (ML)**
2. **Variance of ML estimators**
3. **The method of least squares (LS)**
4. **Interval estimation, setting limits**

The likelihood function

Consider data sample $\vec{x} = (x_1, \ldots, x_n)$ where x follows $f(x;\theta)$.

Goal: estimate θ (or in general $\vec{\theta} = (\theta_1, \ldots, \theta_m)$).

If $f(x;\theta)$ is true, then

$$P(\text{all } x_i \text{ found in } [x_i, x_i + dx_i]) = \prod_{i=1}^{n} f(x_i, \theta)\, dx_i$$

If hypothesis (including value of θ) is true,

$\rightarrow$ expect high probability for the data we actually got.

If hypothesized θ far away from true value,

$\rightarrow$ low probability to have observed what we did.

$\Rightarrow$ true θ should give high value for

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) \qquad \text{(the likelihood function)}$$

N.B. $L(\theta) = f_{\text{sample}}(\vec{x};\theta)$, but $L(\theta)$ regarded as function of θ,

$\vec{x}$ treated as constant (experiment is over).

N.B. In classical statistics, $L(\theta)$ is not a 'pdf' for θ.

$\rightarrow$ θ is not a random variable ($\hat{\theta}$ is).

In Bayesian statistics, treat $L(\theta) = L(\vec{x}|\theta)$ as pdf for $\vec{x}$ given θ, then use Bayes' theorem to get posterior pdf $p(\theta|\vec{x})$.

Maximum likelihood estimators

Define ML estimator $\hat{\theta}$ as the value of θ that maximizes $L(\theta)$.

Write estimators with hat ($\hat{\theta}$) to distinguish from true value θ, which may forever remain unknown.

For m parameters, usually find solution $\hat{\theta}_1, \ldots, \hat{\theta}_m$ by solving

$$\frac{\partial L}{\partial \theta_i} = 0 \quad i = 1, \ldots, m.$$

Sometimes $L(\theta)$ has more than one local maximum,

$\rightarrow$ take highest one.

N.B. no binning of data ('all information used').

N.B. the definition of ML estimators does not guarantee that they are in any way 'optimal'.

$\rightarrow$ investigate properties such as bias, variance.

For many cases of interest and for sufficiently large sample, ML turns out to be about as good as we can do.

Not always optimal for small n, but still usually best practical solution.

Example of ML estimator: parameter of exponential pdf

Consider the exponential pdf,

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

and suppose we have a data sample $t_1, \ldots, t_n$.

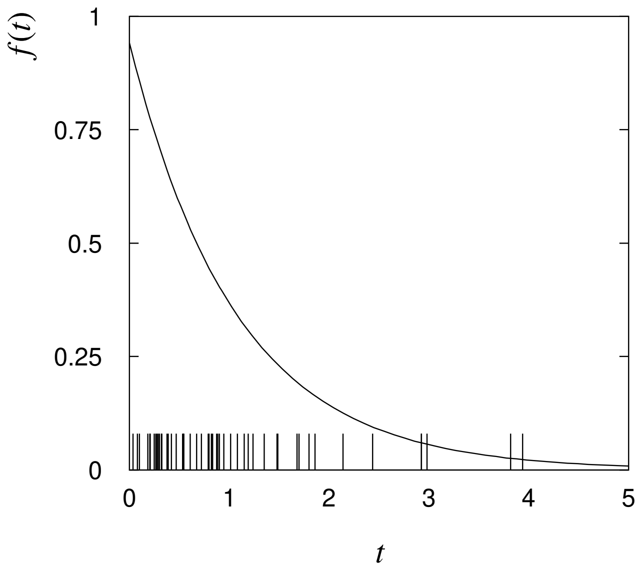
Usually use log-likelihood (maximum at same value of parameter),

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

Set $\dfrac{\partial \log L}{\partial \tau} = 0$ and solve for τ,

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

Example: generate 50 values of t with MC using $\tau = 1$,



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

Bias of ML estimator $\hat{\tau}$

Is $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ an unbiased estimator for τ?

The hard way to check:

find pdf $g(\hat{\tau}; \tau)$, compute $b = E[\hat{\tau}] - \tau$

Or use an easier way to compute $E[\hat{\tau}]$,

$$E[\hat{\tau}(t_1, \ldots, t_n)] = \int \ldots \int \hat{\tau}(\vec{t}) \, f_{\text{joint}}(\vec{t}; \tau) \, dt_1 \ldots dt_n$$

$$= \int \ldots \int \left(\frac{1}{n} \sum_{i=1}^{n} t_i \right) \frac{1}{\tau} e^{-t_1/\tau} \ldots \frac{1}{\tau} e^{-t_n/\tau} dt_1 \ldots dt_n$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\int t_i \frac{1}{\tau} e^{-t_i/\tau} dt_i \prod_{j \neq i} \int \frac{1}{\tau} e^{-t_j/\tau} dt_j \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \tau = \tau$$

$\rightarrow$ $\hat{\tau}$ is an unbiased estimator for τ.

The really easy way:

We already showed that the sample mean $\overline{t}$ is an unbiased estimator for $E[t]$, and for the exponential pdf, $E[t] = \tau$.

Variance of estimator: analyitic method

Recall estimator for mean of exponential: $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$.

How wide is the pdf $g(\hat{\tau}; \tau, n)$?

$$V[\hat{\tau}] = E[\hat{\tau}^2] - (E[\hat{\tau}])^2$$

$$= \int \ldots \int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^2 \frac{1}{\tau}e^{-t_1/\tau} \ldots \frac{1}{\tau}e^{-t_n/\tau} dt_1 \ldots dt_n$$

$$- \left(\int \ldots \int \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right) \frac{1}{\tau}e^{-t_1/\tau} \ldots \frac{1}{\tau}e^{-t_n/\tau} dt_1 \ldots dt_n\right)^2$$

$$= \frac{\tau^2}{n}.$$

$\rightarrow$ variance of $\hat{\tau}$ is n times smaller than variance of t.

(In fact we knew this already, since here $\hat{\tau} = \bar{t}$.)

N.B. $V[\hat{\tau}]$, $\sigma_{\hat{\tau}}$ functions of true (unknown!) τ. Estimate using

$$\hat{\sigma}_{\hat{\tau}} = \frac{\hat{\tau}}{\sqrt{n}}$$

Often given as 'statistical error', e.g. $\hat{\tau} \pm \hat{\sigma}_{\hat{\tau}} = 1.062 \pm 0.150$

This means: ML estimate for τ is 1.062.

ML estimate for σ of $g(\hat{\tau}; \tau, n)$ is 0.150.

If $g(\hat{\tau}; \tau, n)$ is Gaussian, $[\hat{\tau} - \hat{\sigma}_{\hat{\tau}}, \hat{\tau} + \hat{\sigma}_{\hat{\tau}}]$ same as

'68% confindence interval' (more on this later).

Variance of estimators: Monte Carlo method

Often form of $\hat{\theta}$, $g(\hat{\theta};\theta,n)$ not known explicitly,

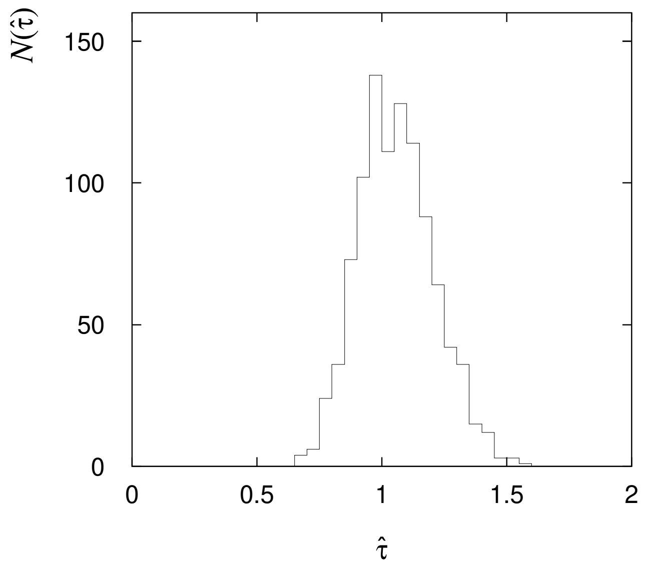
→ get $g(\hat{\theta};\theta,n)$ from Monte Carlo.

For example with exponential pdf we had $\hat{\tau} = 1.062$.

Use this as 'true' τ in MC,

generate samples of $n = 50$ values (1000 experiments),

compute $\hat{\tau}$ for each experiment and histogram:



Sample standard deviation from MC experiments gives

$$\hat{\sigma}_{\hat{\tau}} = \left[\frac{1}{N_{\mathrm{exp}} - 1} \sum_{i=1}^{N_{\mathrm{exp}}} (\hat{\tau}_i - \overline{\hat{\tau}})^2 \right]^{1/2} = 0.151$$

Similar to previous estimate $\dfrac{\hat{\tau}}{\sqrt{n}} = 0.150.$

N.B. $g(\hat{\tau};\tau,n)$ approximately Gaussian (cf. central limit theorem)

→ true in general for ML estimators in large sample limit.

The RCF bound (information inequality)

A lower bound on the variance of any estimator (not just ML) is

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

This is the Rao–Cramér–Frechet inequality (information inequality).

If equality is met, $\hat{\theta}$ is said to be efficient.

→ ML estimators are (almost always) efficient for large n,

often assume this to be true and use RCF bound to estimate $V[\hat{\theta}]$.

For the example with the exponential pdf, we obtain

$$\frac{\partial^2 \log L}{\partial \tau^2} = \frac{n}{\tau^2}\left(1 - \frac{2}{\tau}\frac{1}{n}\sum_{i=1}^{n} t_i\right) = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

and we know that $b = 0$, so

$$V[\hat{\tau}] \geq \frac{1}{E\left[-\frac{n}{\tau^2}(1 - \frac{2\hat{\tau}}{\tau})\right]} = \frac{1}{-\frac{n}{\tau^2}\left(1 - \frac{2E[\hat{\tau}]}{\tau}\right)} = \frac{\tau^2}{n}$$

This is equal to the true variance → ML $\hat{\tau}$ is efficient for any n.

For $\vec{\theta} = (\theta_1, \ldots, \theta_m)$ with efficient estimator and zero bias,

$$(V^{-1})_{ij} = E\left[-\frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}\right] = -n \int f(x;\vec{\theta})\, \frac{\partial^2 \log f(x;\vec{\theta})}{\partial \theta_i \partial \theta_j} dx$$

→ variance of efficient estimators $\propto \frac{1}{n}$.

The RCF bound (continued)

The expectation value of $\dfrac{\partial^2 \log L}{\partial \theta_i \partial \theta_j}$ in the RCF bound is a function of the true parameters.

$\rightarrow$ estimate by evaluating with the (single) ML estimate:

$$(\widehat{V^{-1}})_{ij} = - \left. \frac{\partial^2 \log L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta}=\hat{\vec{\theta}}}$$

For a single parameter one has

$$\widehat{\sigma^2}_{\hat{\theta}} = \left. \left(-1 \Big/ \frac{\partial^2 \log L}{\partial \theta^2} \right) \right|_{\theta=\hat{\theta}} .$$

Often maximize $\log L$ numerically, estimate matrix of 2nd derivatives (Hessian matrix) using finite differences.

$\rightarrow$ `MINUIT` routine `HESSE`.

Variance of ML estimators: graphical method

Consider single parameter θ, expand $\log L(\theta)$ about $\hat{\theta}$,

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

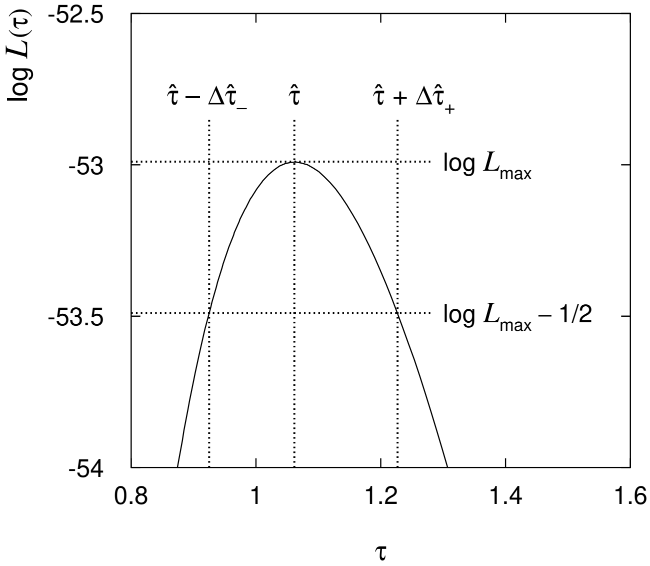
$\log L(\hat{\theta}) = \log L_{\max}$ and the second term is zero, therefore

$$\log L(\theta) = \log L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}},$$

that is,

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L_{\max} - \frac{1}{2}$$

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



Example of exponential:

$\hat{\tau} = 1.062$

$\Delta\hat{\tau}_{-} = 0.137$, $\Delta\hat{\tau}_{+} = 0.165$

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

Relationship between ML and Bayesian parameter estimation

In Bayesian statistics, both θ and $\vec{x}$ are r.v.s:

$$L(\theta) = L(\vec{x}|\theta) = f_{\text{joint}}(\vec{x}|\theta) \quad \text{(conditional pdf for } \vec{x} \text{ given } \theta)$$

The Bayesian Method:

Use subjective probability for hypotheses (θ),

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

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

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

$p(\theta|\vec{x})$ = posterior pdf (conditional pdf for θ given $\vec{x}$)

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

Pragmatist Bayesian: $p(\theta|\vec{x})$ is a complicated function,

$\rightarrow$ summarize by means of estimator $\hat{\theta}_{\text{Bayes}}$

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

What do we use for $\pi(\theta)$????

No golden rule (subjective!), often represent 'prior ignorance' by

$\pi(\theta) = \text{constant} \rightarrow \hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$

But . . . we could have used a different parameter, e.g. $\lambda = 1/\theta$.

If prior for $\pi_\theta(\theta)$ is constant, then $\pi_\lambda(\lambda)$ is not!

$\rightarrow$ 'complete prior ignorance' not well defined

The method of least squares (LS): connection with ML

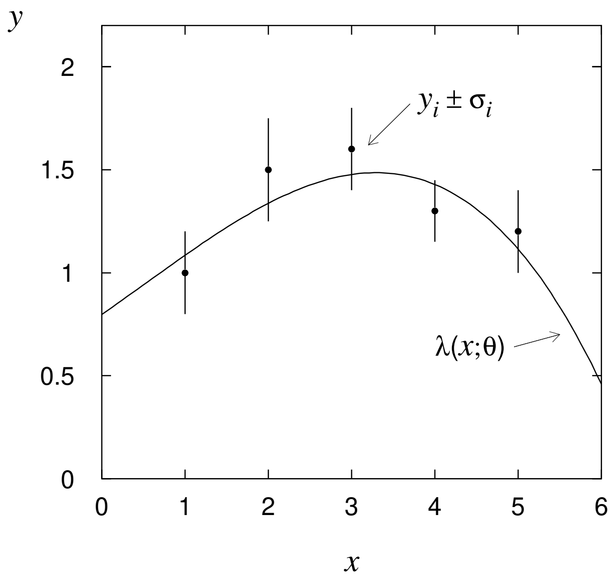
Suppose we have Gaussian r.v.s y_i, $i = 1, \ldots, N$

$E[y_i] = \lambda_i = \lambda(x_i; \vec{\theta})$,

where $x_1, \ldots, x_N$ and $V[y_i] = \sigma_i^2$ are known.

Goal: estimate parameters $\vec{\theta}$,

i.e. fit the curve through the points.



The joint pdf for independent Gaussian y_i is

$$g(\vec{y}; \vec{\lambda}, \vec{\sigma}^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{-(y_i - \lambda_i)^2}{2\sigma_i^2} \right)$$

i.e. the log-likelihood function is (drop terms not depending on $\vec{\theta}$),

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

→ maximizing $\log L(\vec{\theta})$ same as minimizing

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

Definition of least squares (LS) estimators

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 the log-likelihood is

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

i.e. we should minimize

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

Its minimum defines the least squares (LS) estimators $\hat{\vec{\theta}}$,

even when y_i not Gaussian. (In fact, y_i often Gaussian because central limit theorem leads to Gaussian measurement errors.)

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).

Linear least squares fit

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.

$\rightarrow$ $\hat{\vec{\theta}}$ have zero bias, minimum variance (Gauss–Markov theorem)

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

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

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

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.

Variance of LS estimators

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

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

$\rightarrow$ coincides with RCF bound if y_i are Gaussian.

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,

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

If $\lambda(x;\vec{\theta})$ not linear in $\vec{\theta}$, then expressions above not exact (but may still be good approximations).

Still interpret region $\chi^2(\vec{\theta}) \leq \chi^2_{\mathrm{min}} + 1$ as 'confidence region', having given probability of containing true $\vec{\theta}$ (more later).

N.B. formulae above don't depend on y_i being Gaussian, but in any case need $V_{ij} = \mathrm{cov}[y_i, y_j]$.

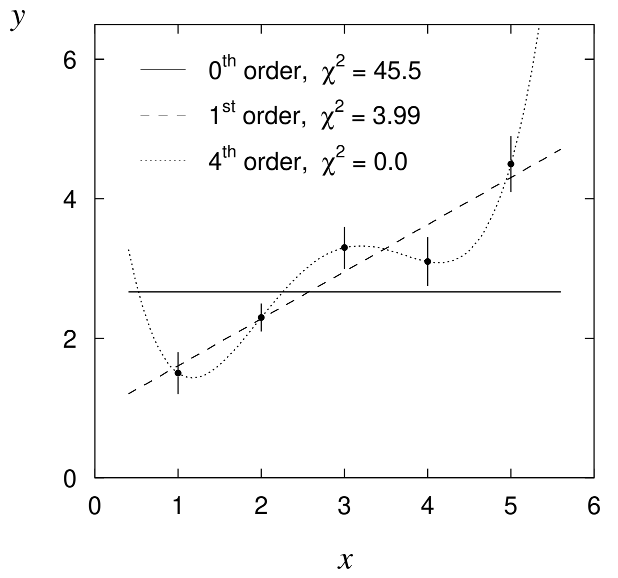
LS fit of a polynomial

Fit a polynomial: $\lambda(x;\theta_0,\ldots,\theta_m) = \sum_{j=0}^{m} \theta_j\, x^j$

$a_j(x) = x^j$

Examples:

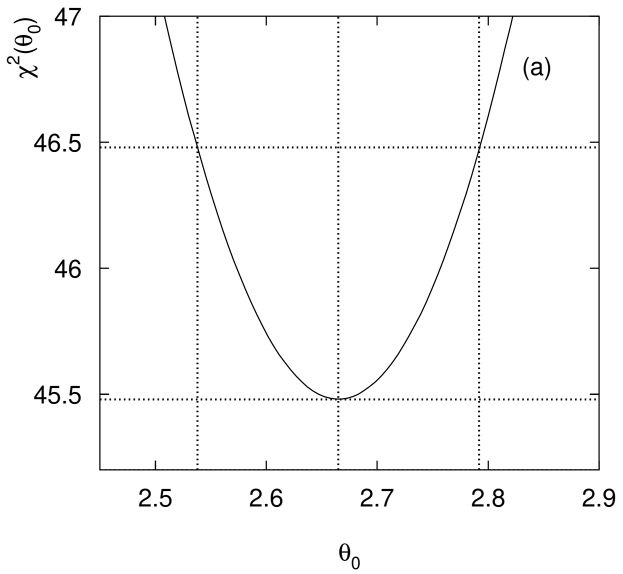
- 0th order (1 parameter)
- 1st order (2 parameters)
- 4th order (5 parameters)



1-parameter fit (i.e. horizontal line):

$$\hat{\theta}_0 = 2.66 \pm 0.13$$

$$\chi^2_{\min} = 45.5$$

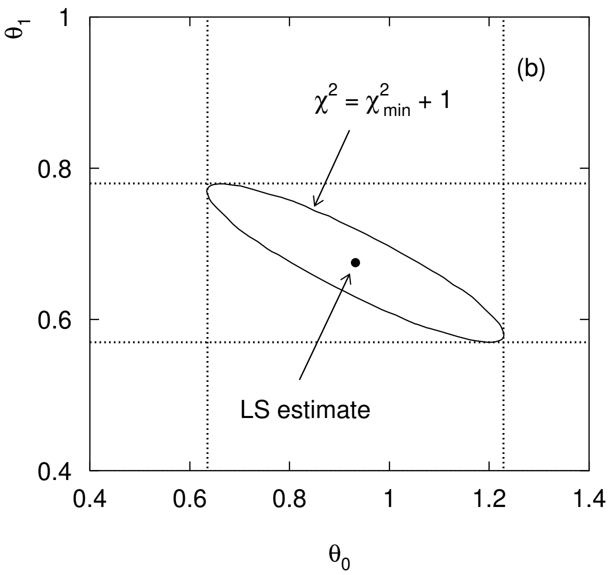


$\sigma_{\hat{\theta}_0}$ from $\chi^2(\hat{\theta}_0 \pm \sigma_{\hat{\theta}_0}) = \chi^2_{\min} + 1$.

Polynomial fit (continued)

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)

Could transform $(\hat{\theta}_0, \hat{\theta}_1) \rightarrow (\hat{\eta}_0, \hat{\eta}_1)$ such that $\mathrm{cov}[\hat{\eta}_0, \hat{\eta}_1] = 0$,

easier to work with uncorrelated estimators, but interpretation

of new parameters may not be obvious, cf. SDA Section 1.7.

5-parameter case:

curve goes through all points,

$\chi^2_{\mathrm{min}} = 0$,

(number of parameters = number of data points)

Value of χ^2_{min} reflects agreement between data and hypothesis,

$\rightarrow$ use as goodness-of-fit test statistic

Testing goodness-of-fit with LS

If: the y_i, $i = 1, \ldots, N$, are Gaussian (V_{ij} known),

the hypothesis $\lambda(x; \vec{\theta})$ is linear in θ_i, $i = 1, \ldots, m$, and

the form of the hypothesis $\lambda(x; \vec{\theta})$ is correct,

then $\chi^2_{\rm min}$ follows chi-square pdf for $N - m$ degrees of freedom.

From this compute P-value,

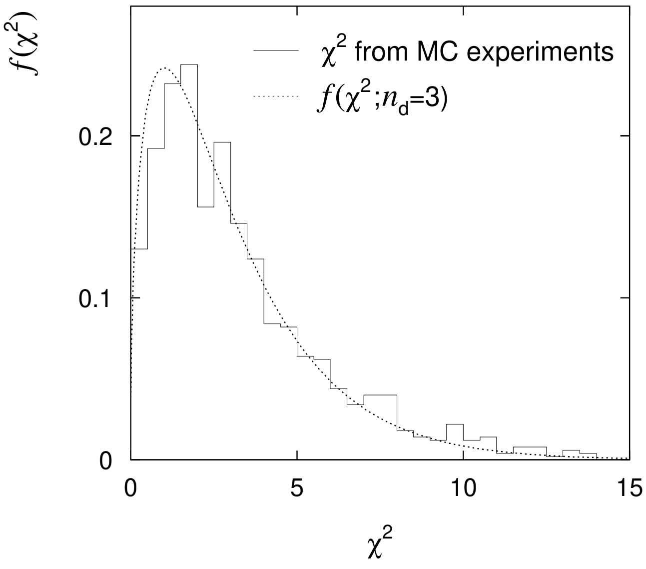
$$P = \int_{\chi^2_{\rm min}}^{\infty} f(z; n_{\rm d}) dz$$

Consider e.g. 2-parameter fit:

$$\chi^2_{\rm min} = 3.99, \; N - m = 3 \rightarrow P = 0.263$$

i.e. repeat experiment many times, 26.3% will have higher $\chi^2_{\rm min}$:

1000 MC experiments:



For the horizontal line fit, we had

$$\chi^2_{\rm min} = 45.5, \; N - m = 4 \rightarrow P = 3.1 \times 10^{-9}$$

Goodness-of-fit vs. smallness-of-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 $\chi^2_{\min}$ $\rightarrow$ goodness-of-fit

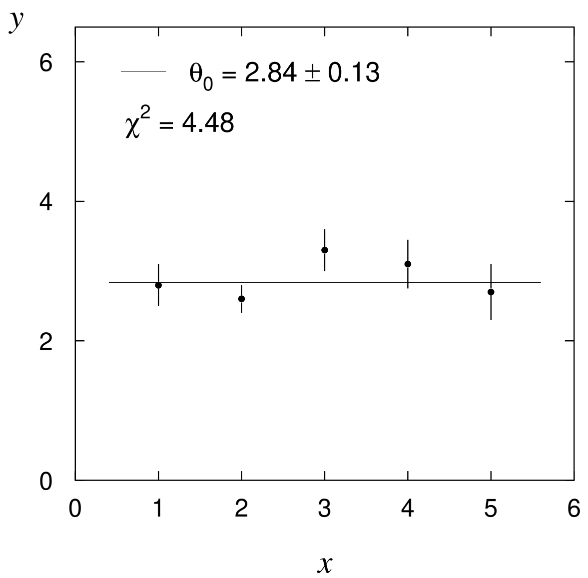
Horizontal line fit, move the data points, keep errors on points same:

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

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

Variance same as before,

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



$\rightarrow$ $\chi^2(\theta_0)$ 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.

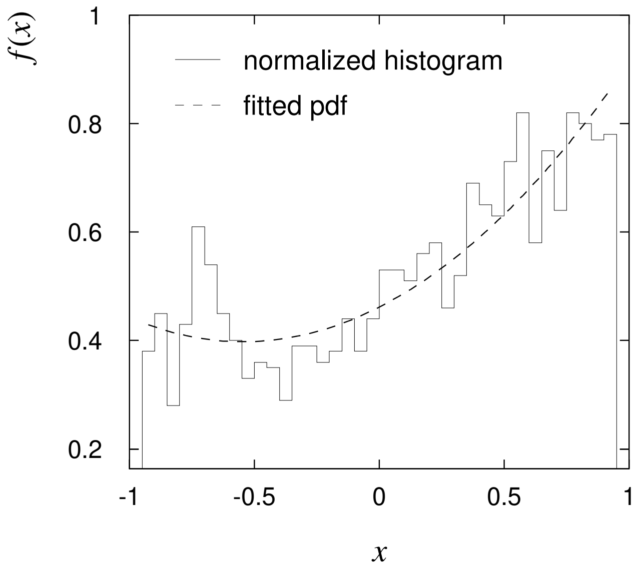
LS with binned data

Histogram:

N bins, n entries.

Hypothesized pdf:

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



We have

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

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

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 $\chi^2_{\min}$ no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.

Combining measurements with LS

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,

$$\rightarrow \quad \hat{\lambda} = \frac{\sum_{i=1}^{N} y_i / \sigma_i^2}{\sum_{j=1}^{N} 1/\sigma_j^2}$$

$$V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2}$$

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 of 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.

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.

The standard deviation as statistical error

My experiment: data $x_1, \ldots, x_n \rightarrow$ estimate $\hat{\theta}_{\mathrm{obs}}$

Also estimate variance of $\hat{\theta}$, $\widehat{\sigma^2}_{\hat{\theta}}$, often report something like

$$\hat{\theta}_{\mathrm{obs}} \pm \hat{\sigma}_{\hat{\theta}} = 5.73 \pm 0.21$$

What does this really mean?

We know $\hat{\theta}$ will follow some pdf $g(\hat{\theta};\theta)$,

estimate of θ is 5.73,

estimate of $\sigma_{\hat{\theta}}$ is 0.21 $\rightarrow$ $\sigma_{\hat{\theta}}$ measures width of $g(\hat{\theta};\theta)$

Often $g(\hat{\vec{\theta}};\vec{\theta})$ is multivariate Gaussian,

$\hat{\vec{\theta}}$, $\widehat{V} = \widehat{\mathrm{cov}}[\hat{\theta}_i, \hat{\theta}_j]$ summarize our (estimated) knowledge

about $g(\hat{\vec{\theta}};\vec{\theta})$, $\rightarrow$ input for error propagation, LS averaging, ...

We could stick with this as the convention for reporting errors,

regardless of the pdf of $g(\hat{\theta};\theta)$.

Sometimes we do (e.g. for PDG averaging), but ...

if $g(\hat{\theta};\theta)$ is Gaussian, then the interval

$$[\hat{\theta}_{\mathrm{obs}} - \hat{\sigma}_{\hat{\theta}}, \hat{\theta}_{\mathrm{obs}} + \hat{\sigma}_{\hat{\theta}}]$$

is a 68.3% central confidence interval (more later).

This is the more usual convention, and if $g(\hat{\theta};\theta)$ not Gaussian,

central confidence interval $\rightarrow$ asymmetric errors

Classical confidence intervals (1)

We have an estimator $\hat{\theta}$ for a parameter θ and an estimate $\hat{\theta}_{\rm obs}$,

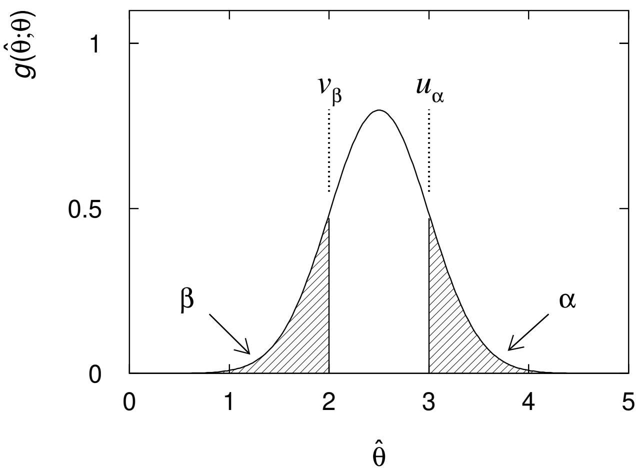
we also need $g(\hat{\theta};\theta)$ for all θ.

Specify 'upper and lower tail probabilities', e.g. $\alpha = \beta = 0.05$,

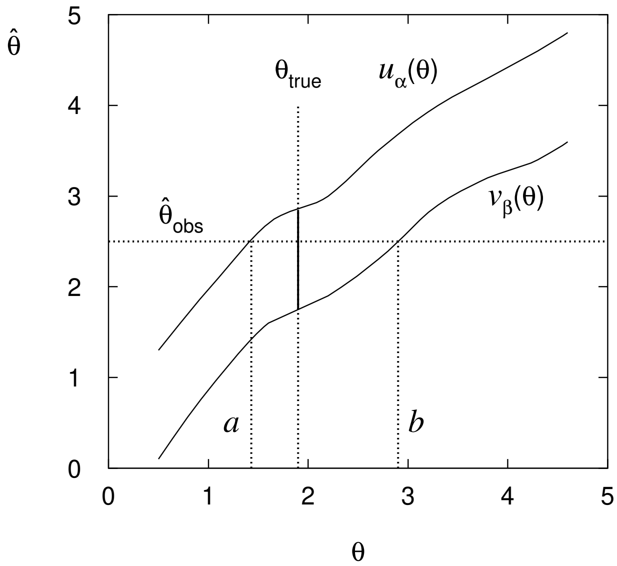
then, find functions $u_\alpha(\theta)$, $v_\beta(\theta)$ such that

$$\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)d\hat{\theta} = 1 - G(u_\alpha(\theta);\theta)\,,$$

$$\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)d\hat{\theta} = G(v_\beta(\theta);\theta)\,.$$



The region between $u_\alpha(\theta)$ and $v_\beta(\theta)$ is called the confidence belt.



Classical confidence intervals (2)

The probability to find $\hat{\theta}$ in the confidence belt, regardless of θ,

$$P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta.$$

Assume $u_\alpha(\theta)$, $v_\beta(\theta)$ monotonic, then

$$a(\hat{\theta}) \equiv u_\alpha^{-1}(\hat{\theta}) \;,$$

$$b(\hat{\theta}) \equiv v_\beta^{-1}(\hat{\theta}) \;.$$

The inequalities

$$\hat{\theta} \geq u_\alpha(\theta),$$

$$\hat{\theta} \leq v_\beta(\theta),$$

imply

$$a(\hat{\theta}) \geq \theta \;,$$

$$b(\hat{\theta}) \leq \theta \;.$$

$$\Rightarrow \quad P(a(\hat{\theta}) \geq \theta) = \alpha,$$

$$P(b(\hat{\theta}) \leq \theta) = \beta.$$

or together,

$$P(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = 1 - \alpha - \beta \;.$$

Classical confidence intervals (3)

The interval $[a(\hat{\theta}), b(\hat{\theta})]$ is called a confidence interval with confidence level or coverage probability $1 - \alpha - \beta$.

Its quintessential property:

probability to contain true parameter is $1 - \alpha - \beta$.

N.B. the interval is random, the true θ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover θ.

Sometimes only specify α or β, $\rightarrow$ one-sided interval (limit)

Often take $\alpha = \beta = \dfrac{\gamma}{2}$ $\rightarrow$ coverage probability $= 1 - \gamma$

$\rightarrow$ central confidence interval

N.B. ‘central’ confidence interval does not mean the interval is symmetric about $\hat{\theta}$, but only that $\alpha = \beta$.

The HEP error ‘convention’: 68.3% central confidence interval.

Classical confidence intervals (4)

Usually, we don't construct the confidence belt, but rather solve

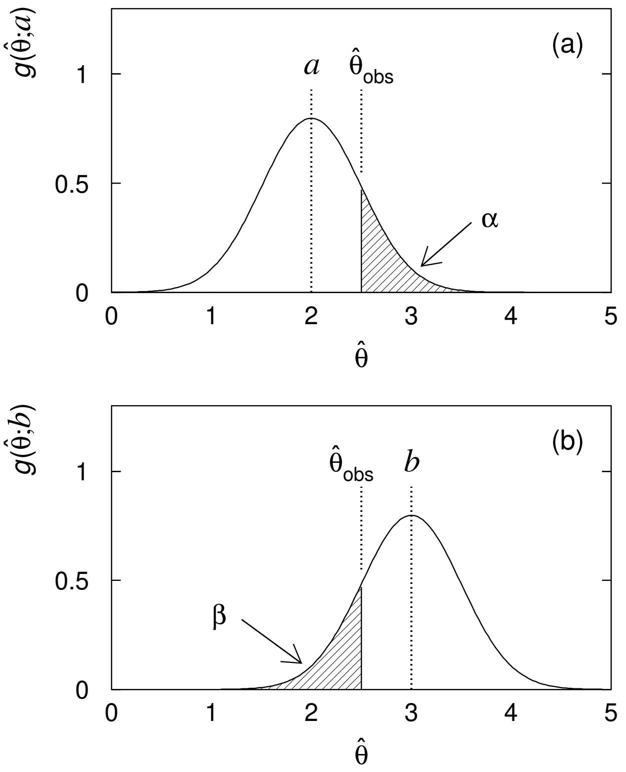
$$\alpha = \int_{\hat{\theta}_{\rm obs}}^{\infty} g(\hat{\theta}; a)\, d\hat{\theta} = 1 - G(\hat{\theta}_{\rm obs}; a)$$

$$\beta = \int_{-\infty}^{\hat{\theta}_{\rm obs}} g(\hat{\theta}; b)\, d\hat{\theta} = G(\hat{\theta}_{\rm obs}; b)$$

for interval limits a and b. (Gives same thing.)

→ a is hypothetical value of θ such that $P(\hat{\theta} > \hat{\theta}_{\rm obs}) = \alpha$;

b is hypothetical value of θ such that $P(\hat{\theta} < \hat{\theta}_{\rm obs}) = \beta$.



Confidence interval for Gaussian distributed estimator

Suppose we have $g(\hat{\theta};\theta) = \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(\frac{-(\hat{\theta}-\theta)^2}{2\sigma_{\hat{\theta}}^2}\right)$.

To find confidence interval for θ, solve

$$\alpha = 1 - G(\hat{\theta}_{\rm obs}; a, \sigma_{\hat{\theta}}) = 1 - \Phi\left(\frac{\hat{\theta}_{\rm obs} - a}{\sigma_{\hat{\theta}}}\right),$$

$$\beta = G(\hat{\theta}_{\rm obs}; b, \sigma_{\hat{\theta}}) = \Phi\left(\frac{\hat{\theta}_{\rm obs} - b}{\sigma_{\hat{\theta}}}\right),$$

for a, b, where G is cumulative distribution for $\hat{\theta}$ and

$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x'^2/2}\, dx'$ is cumulative of standard Gaussian.

$$\rightarrow \quad a = \hat{\theta}_{\rm obs} - \sigma_{\hat{\theta}}\, \Phi^{-1}(1-\alpha),$$

$$b = \hat{\theta}_{\rm obs} + \sigma_{\hat{\theta}}\, \Phi^{-1}(1-\beta).$$

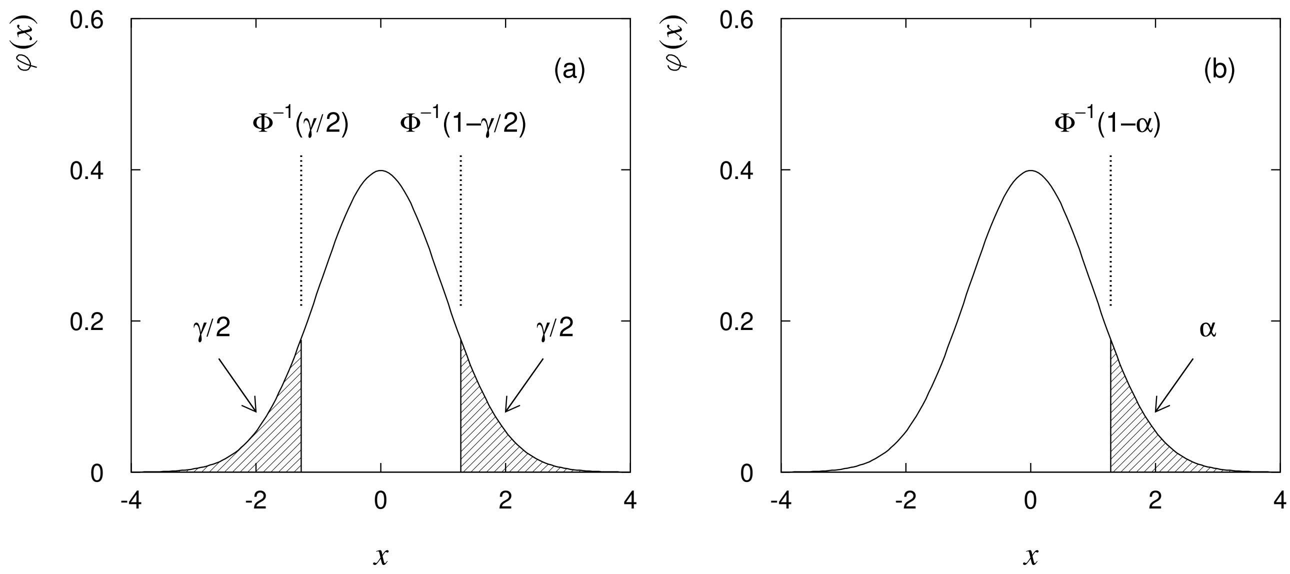
Φ^{-1} = quantile of standard Gaussian

(inverse of cumulative distribution, CERNLIB routine **GAUSIN**).

$\rightarrow$ $\Phi^{-1}(1-\alpha)$, $\Phi^{-1}(1-\beta)$ give how many standard deviations a, b are from $\hat{\theta}$.

Quantiles of the standard Gaussian

To find the confidence interval for a parameter with a Gaussian, estimator we need the following quantiles:



Usually take a round number for the quantile . . .

central		one-sided	
$\Phi^{-1}(1-\gamma/2)$	$1-\gamma$	$\Phi^{-1}(1-\alpha)$	$1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987

Sometimes take a round number for the coverage probability . . .

central		one-sided	
$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

Confidence interval for mean of Poisson distribution

Suppose n is Poisson, $\hat{\nu} = n$, estimate is $\hat{\nu}_{\text{obs}} = n_{\text{obs}}$,

$$P(n;\nu) = \frac{\nu^n}{n!} e^{-\nu}, \quad n = 0, 1, \ldots$$

Minor problem: for fixed α, β, confidence belt doesn't exist for all ν. No matter. Just solve

$$\alpha = P(\hat{\nu} \geq \hat{\nu}_{\text{obs}}; a) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{a^n}{n!} e^{-a},$$

$$\beta = P(\hat{\nu} \leq \hat{\nu}_{\text{obs}}; b) = \sum_{n=0}^{n_{\text{obs}}} \frac{b^n}{n!} e^{-b},$$

for a, b. Use trick:

$$\sum_{n=0}^{m} \frac{\nu^n}{n!} e^{-\nu} = 1 - F_{\chi^2}(2\nu; n_{\text{d}} = 2(m+1))$$

where F_{χ^2} is cumulative chi-square distribution for n_{d} dof,

$$a = \tfrac{1}{2} F_{\chi^2}^{-1}(\alpha; n_{\text{d}} = 2n_{\text{obs}}),$$

$$b = \tfrac{1}{2} F_{\chi^2}^{-1}(1-\beta; n_{\text{d}} = 2(n_{\text{obs}}+1)),$$

where $F_{\chi^2}^{-1}$ is the quantile of the chi-square distribution (CERNLIB routine `CHISIN`).

Interval for Poisson mean (continued)

Important special case: $n_{\mathrm{obs}} = 0$,

$$\rightarrow \beta = \sum_{n=0}^{0} \frac{b^n e^{-b}}{n!} = e^{-b}, \quad \rightarrow \quad b = -\log \beta.$$

For upper limit at confidence level $1 - \beta = 95\%$,

$$b = -\log(0.05) = 2.996 \approx 3.$$

Some more useful numbers...

n_{obs}	lower limit a			upper limit b		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\beta = 0.1$	$\beta = 0.05$	$\beta = 0.01$
0	–	–	–	2.30	3.00	4.61
1	0.105	0.051	0.010	3.89	4.74	6.64
2	0.532	0.355	0.149	5.32	6.30	8.41
3	1.10	0.818	0.436	6.68	7.75	10.04
4	1.74	1.37	0.823	7.99	9.15	11.60
5	2.43	1.97	1.28	9.27	10.51	13.11
6	3.15	2.61	1.79	10.53	11.84	14.57
7	3.89	3.29	2.33	11.77	13.15	16.00
8	4.66	3.98	2.91	12.99	14.43	17.40
9	5.43	4.70	3.51	14.21	15.71	18.78
10	6.22	5.43	4.13	15.41	16.96	20.14

Approximate confidence intervals from $\log L$ or χ^2

Recall trick for estimating $\sigma_{\hat{\theta}}$ if $\log L(\theta)$ parabolic:

$$\log L(\hat{\theta} \pm N\sigma_{\hat{\theta}}) = \log L_{\max} - \frac{N^2}{2}.$$

Claim: this still works even if $\log L$ not parabolic as an approximation for the confidence interval, i.e. use

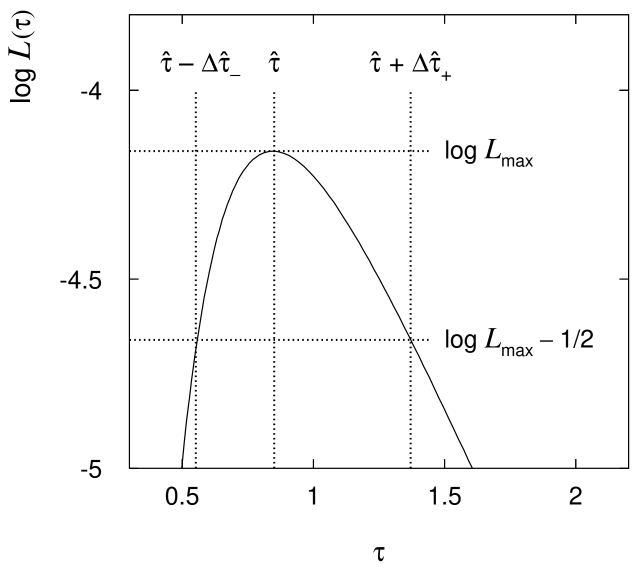
$$\log L(\hat{\theta}^{+d}_{-c}) = \log L_{\max} - \frac{N^2}{2},$$

$$\chi^2(\hat{\theta}^{+d}_{-c}) = \chi^2_{\min} + N^2,$$

where $N = \Phi^{-1}(1 - \gamma/2)$ is the quantile of the standard Gaussian corresponding to the confidence level $1 - \gamma$, e.g.

$$N = 1 \rightarrow 1 - \gamma = 0.683.$$

Our exponential example, now with $n = 5$ observations:



$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Lecture 3 summary

1. **The likelihood function, ML estimators:**

   $L(\theta)$ from joint pdf for the data, evaluate with the data we got.

   ML estimator $\hat{\theta}$ at maximum of L (or $\ln L$).

2. **Variance of ML estimators:**

   Analytic method: best when possible

   Monte Carlo method: useful but can be time consuming

   The information inequality: equality (approx.) for large sample.

   Graphical method: move θ from $\hat{\theta}$ until $\ln L \to \ln L_{\max} - 1/2$.

3. **The method of least squares:**

   ML and LS same if data Gaussian.

   χ^2 at minimum can be used for goodness-of-fit.

   LS can be used with binned data and for combining (averaging) measurements.

   Recipes for variances of LS estimators same as for ML with $\chi^2 \to -2 \ln L$.

4. **Interval estimation:**

   Often sufficient to give $\pm\sigma$ as 68.3% confidence interval.

   Neyman construction for confidence intervals: coverage probability independent of parameter's true value.

   Often quote one-sided interval as upper or lower limit (also 'unified intervals', c.f. Feldman & Cousins).

   Approximate confidence intervals from likelihood function.