Statistics for HEP

Lecture 1: Introduction and basic formalism

`http://indico.cern.ch/conferenceDisplay.py?confId=173727`

Academic Training Lectures

CERN, 2–5 April, 2012

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

Outline

→ Lecture 1: Introduction and basic formalism

Probability, statistical tests, parameter estimation.

Lecture 2: Discovery and Limits

Quantifying discovery significance and sensitivity

Frequentist and Bayesian intervals/limits

Lecture 3: More on discovery and limits

The Look-Elsewhere Effect

Dealing with nuisance parameters

Lecture 4: Unfolding (deconvolution)

Correcting distributions for effects of smearing

Quick review of probablility

Frequentist (A = outcome of repeatable observation):

$$P(A) = \lim_{n\to\infty} \frac{\text{outcome is } A}{n}$$

Subjective (A = hypothesis):

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\Sigma_i P(B|A_i)P(A_i)}$$

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
$P\,(0.117 < \alpha_s < 0.121)$,

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian Statistics – general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:

probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

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

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayesian methods can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors ("if-then" character of Bayes' thm.)

Hypothesis testing

A hypothesis H specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x.

x could be uni-/multivariate, continuous or discrete.

E.g. write $x \sim f(x|H)$.

x could represent e.g. observation of a single particle, a single event, or an entire "experiment".

Possible values of x form the sample space S (or "data space").

Simple (or "point") hypothesis: $f(x|H)$ completely specified.

Composite hypothesis: H contains unspecified parameter(s).

The probability for x given H is also called the likelihood of the hypothesis, written $L(x|H)$.

Definition of a (frequentist) hypothesis test

Consider e.g. a simple hypothesis H_0 and alternative H_1.

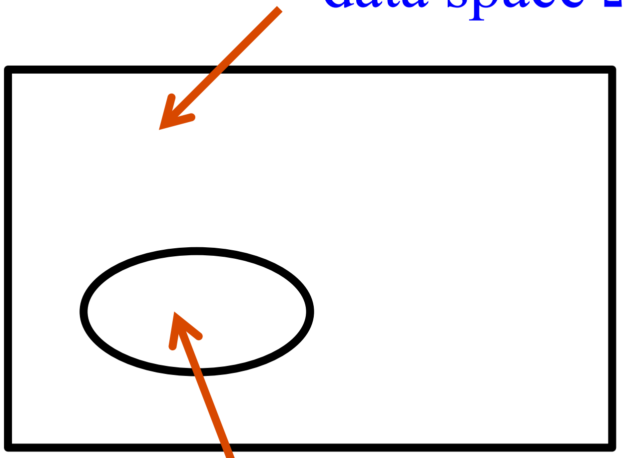
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

α is called the size or significance level of the test.

If x is observed in the critical region, reject H_0.

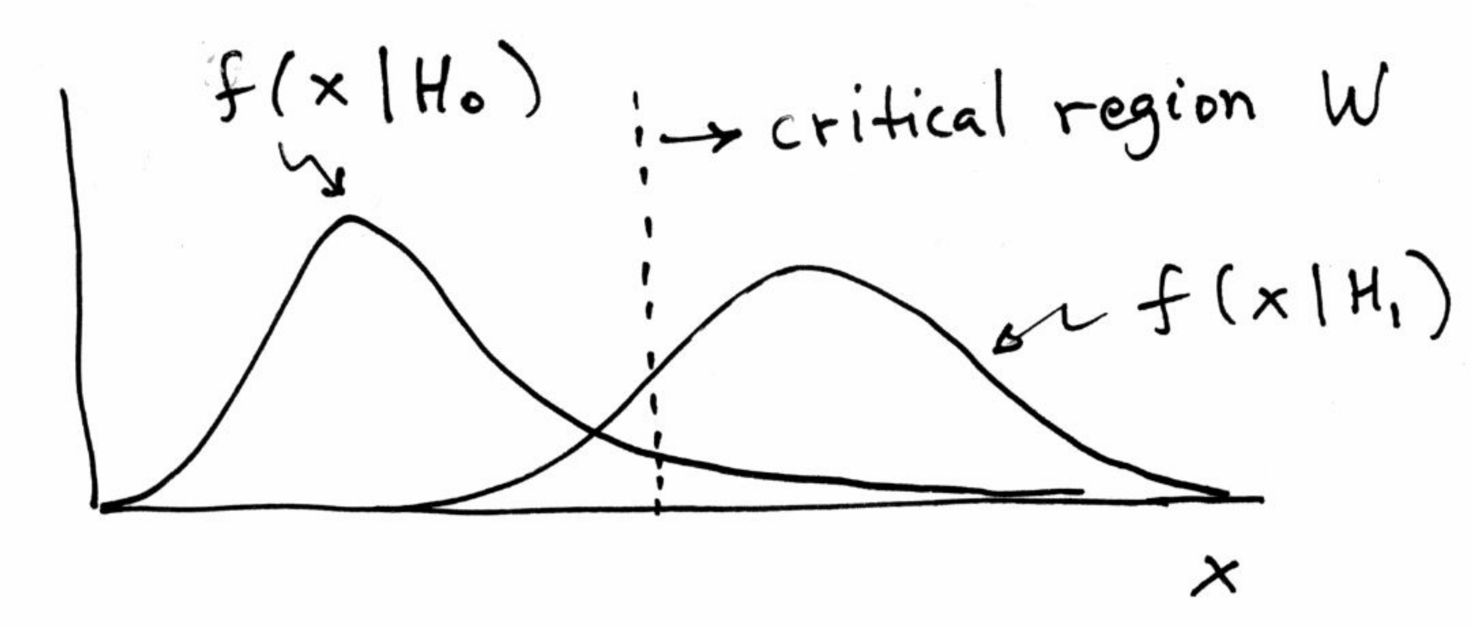

data space Ω

critical region w

Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same significance level α.

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if the alternative H_1 is true:



Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in w \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in S - w \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

Choosing a critical region

To construct a test of a hypothesis H_0, we can ask what are the relevant alternatives for which one would like to have a high power.

Maximize power wrt H_1 = maximize probability to reject H_0 if H_1 is true.

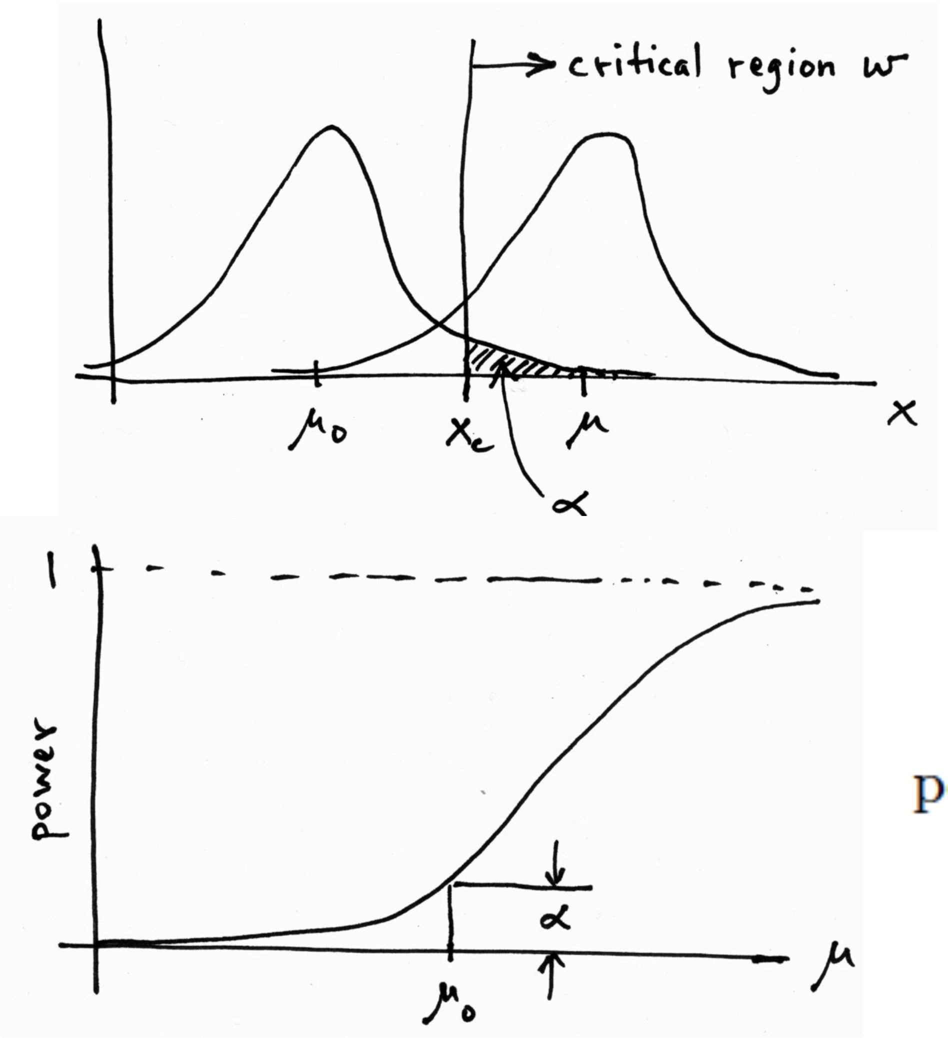
Often such a test has a high power not only with respect to a specific point alternative but for a class of alternatives. E.g., using a measurement $x \sim$ Gauss (μ, σ) we may test

$$H_0 : \mu = \mu_0 \text{ versus the composite alternative } H_1 : \mu > \mu_0$$

We get the highest power with respect to any $\mu > \mu_0$ by taking the critical region $x \geq x_\mathrm{c}$ where the cut-off x_c is determined by the significance level such that

$$\alpha = P(x \geq x_\mathrm{c} | \mu_0).$$

Test of $\mu = \mu_0$ vs. $\mu > \mu_0$ with $x \sim \mathrm{Gauss}(\mu,\sigma)$



Standard Gaussian cumulative distribution

$$\alpha = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma}\right)$$

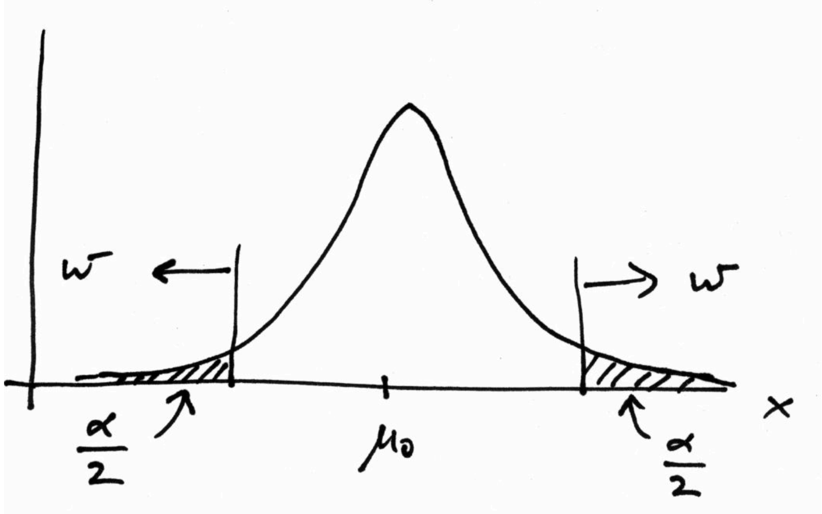
$$x_c = \mu_0 + \sigma\Phi^{-1}(1-\alpha)$$

Standard Gaussian quantile

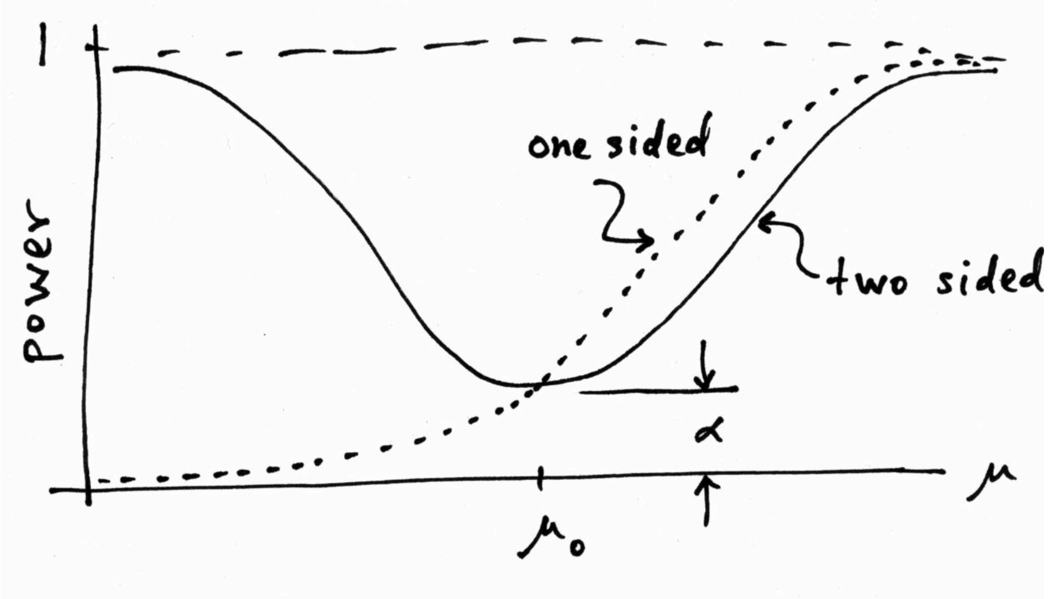
$$\text{power} = 1 - \beta = P(x > x_c|\mu) =$$

$$1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma} + \Phi^{-1}(1-\alpha)\right)$$

Choice of critical region based on power (3)



But we might consider $\mu < \mu_0$ as well as $\mu > \mu_0$ to be viable alternatives, and choose the critical region to contain both high and low x (a two-sided test).



New critical region now gives reasonable power for $\mu < \mu_0$, but less power for $\mu > \mu_0$ than the original one-sided test.

No such thing as a model-independent test

In general we cannot find a single critical region that gives the maximum power for all possible alternatives (no "Uniformly Most Powerful" test).

In HEP we often try to construct a test of

$$H_0 : \text{Standard Model (or ``background only'', etc.)}$$

such that we have a well specified "false discovery rate",

$$\alpha = \text{Probability to reject } H_0 \text{ if it is true,}$$

and high power with respect to some interesting alternative,

$$H_1 : \text{SUSY, } Z', \text{ etc.}$$

But there is no such thing as a "model independent" test. Any statistical test will inevitably have high power with respect to some alternatives and less power with respect to others.

Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true. In frequentist statistics only associate probability with outcomes of repeatable observations (the data).

In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

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

which depends on the prior probability $\pi(H)$.

What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.

Test statistics

The boundary of the critical region for an n-dimensional data space $\boldsymbol{x} = (x_1,\ldots, x_n)$ can be defined by an equation of the form

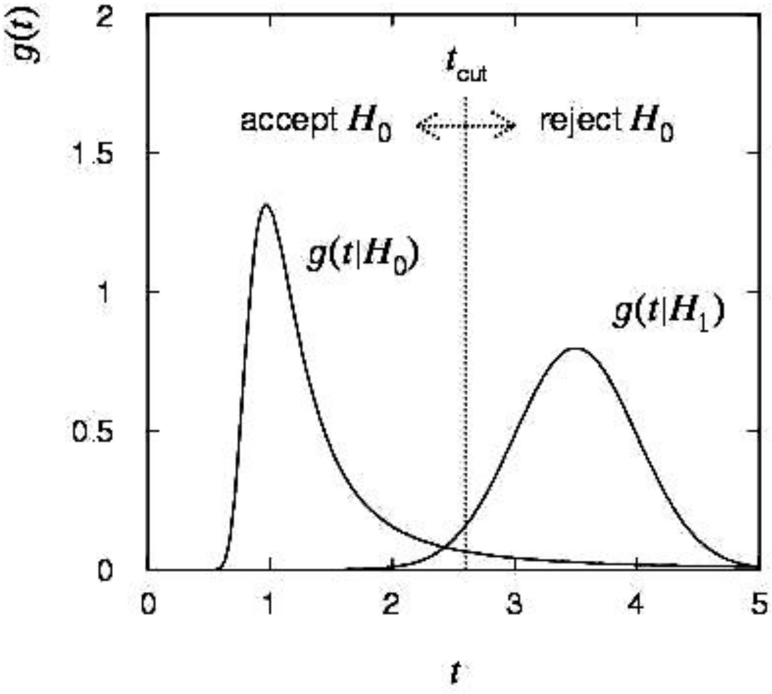
$$t(x_1,\ldots,x_n) = t_{\text{cut}}$$

where $t(x_1,\ldots, x_n)$ is a scalar test statistic.

We can work out the pdfs $g(t|H_0),\ g(t|H_1),\ \ldots$

Decision boundary is now a single ‘cut’ on t, defining the critical region.

So for an n-dimensional problem we have a corresponding 1-d problem.



Constructing a test statistic

How can we choose a test's critical region in an 'optimal way'?

Neyman-Pearson lemma states:

For a test of size α of the simple hypothesis H_0, to obtain the highest power with respect to the simple alternative H_1, choose the critical region w such that the likelihood ratio satisfies

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq k$$

everywhere in w and is less than k elsewhere, where k is a constant chosen such that the test has size α.

Equivalently, optimal scalar test statistic is

$$\boxed{t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}}$$

N.B. any monotonic function of this is leads to the same test.

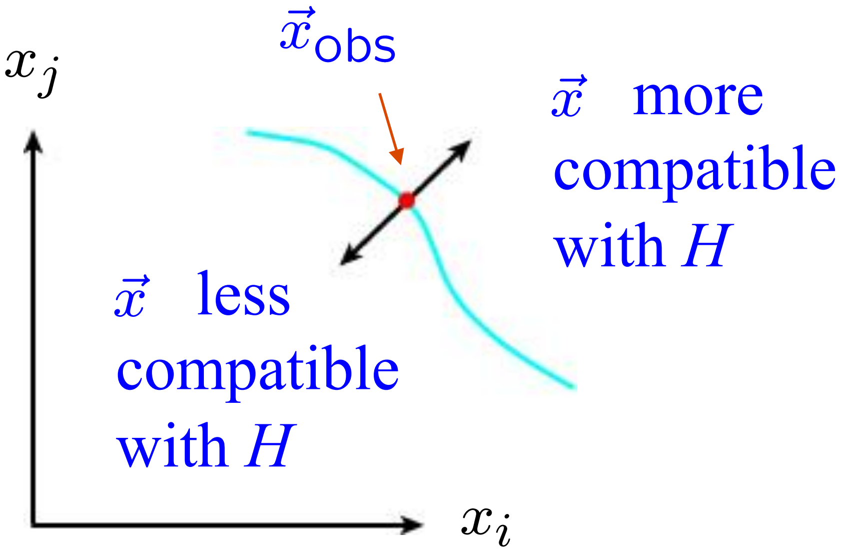
Testing significance / goodness-of-fit

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!)



$\vec{x}$ more compatible with *H*

$\vec{x}$ less compatible with *H*

p-values

Express level of agreement between data and *H* with *p*-value:

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

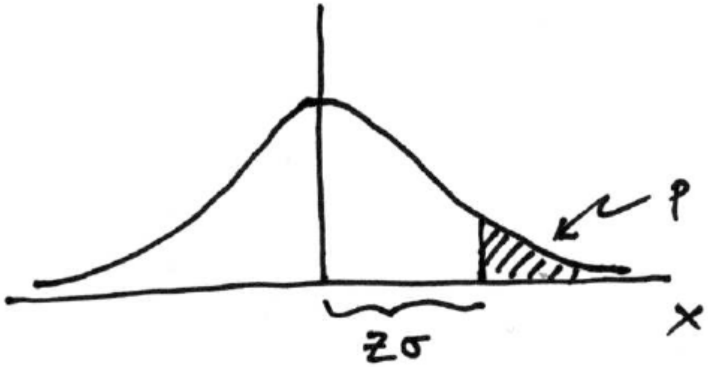
$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

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*).

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`

E.g. $Z = 5$ (a "5 sigma effect") corresponds to $p = 2.9 \times 10^{-7}$.

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b, then $n = n_\mathrm{s} + n_\mathrm{b}$ is also Poisson, mean = $s + b$:

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

Suppose $b = 0.5$, and we observe $n_\mathrm{obs} = 5$. Should we claim evidence for a new discovery?

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

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

Distribution of the *p*-value

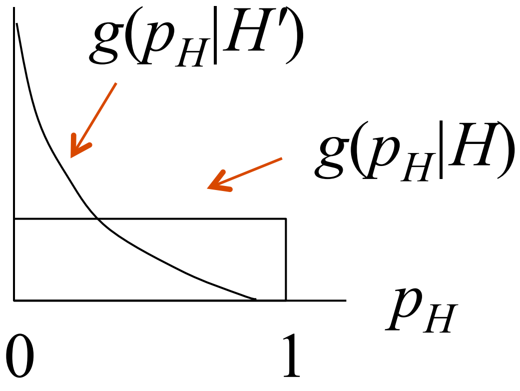
The *p*-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the *p*-value of *H* is found from a test statistic $t(\boldsymbol{x})$ as

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

The pdf of p_H under assumption of *H* is

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

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

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

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 $p_0 \le \alpha$.

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.

Quick review of parameter estimation

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

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

random variable

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):

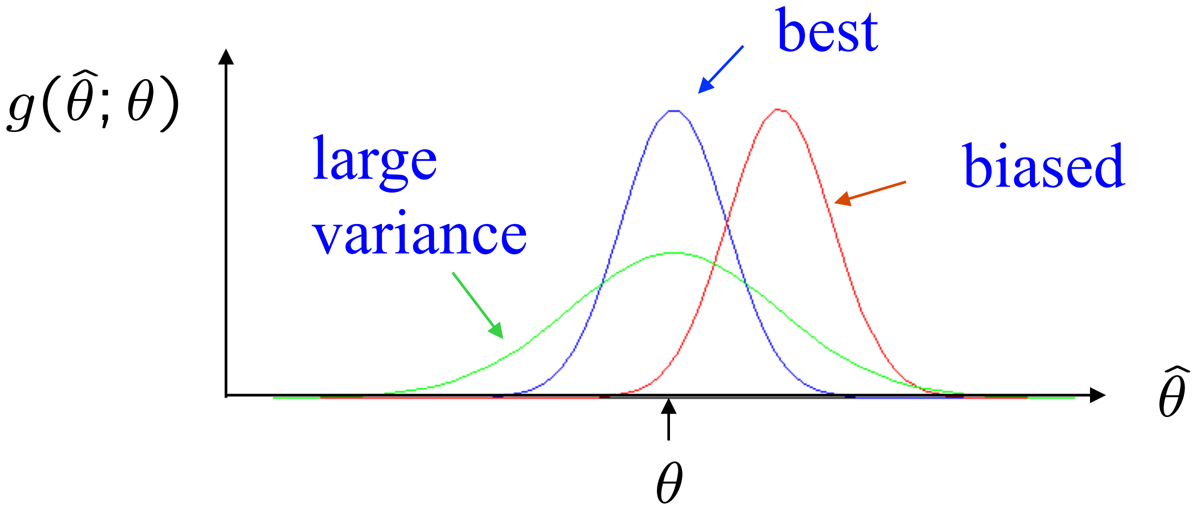
$$\boxed{\widehat{\theta}(\vec{x})}$$

← 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): $b = E[\hat{\theta}] - \theta$

→ 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

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:

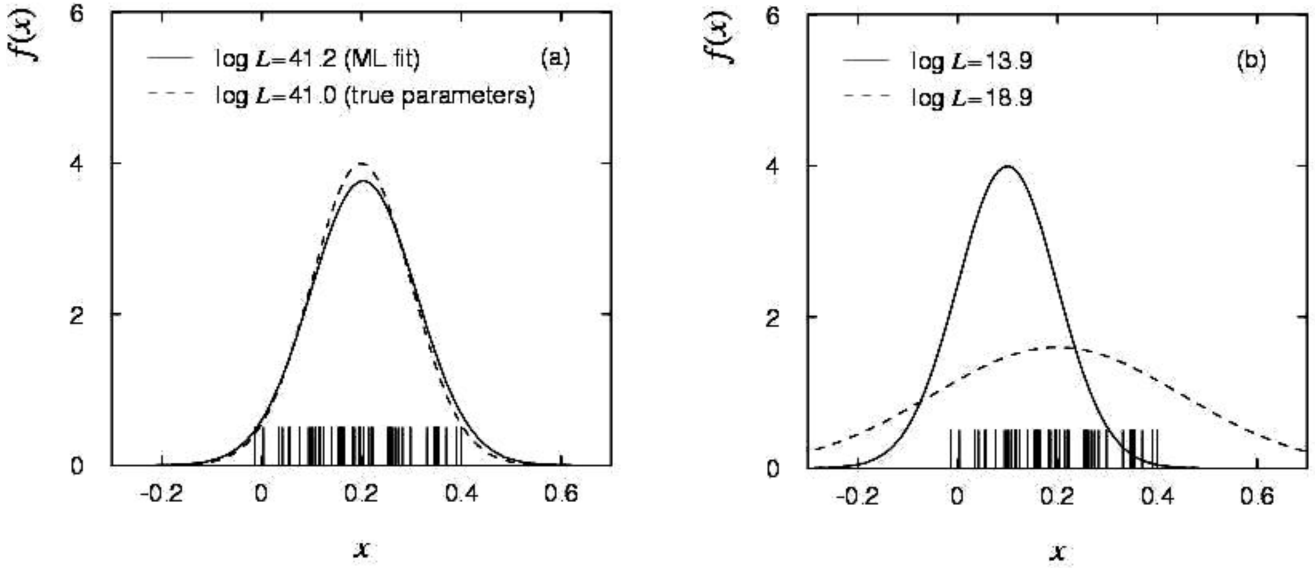
$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

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).

Example: fitting a straight line

Data: $(x_i, y_i, \sigma_i)\ , i = 1, \ldots, n\ .$

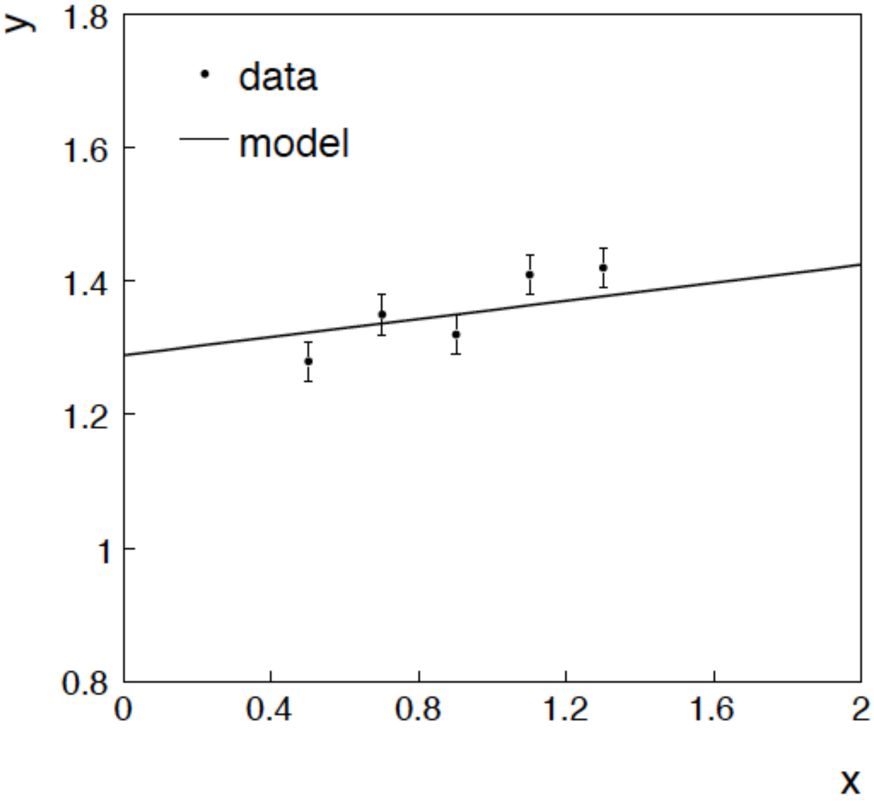
Model: y_i independent and all follow $y_i \sim \mathrm{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x\ ,$$

assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a "nuisance parameter")



Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}\right] ,$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

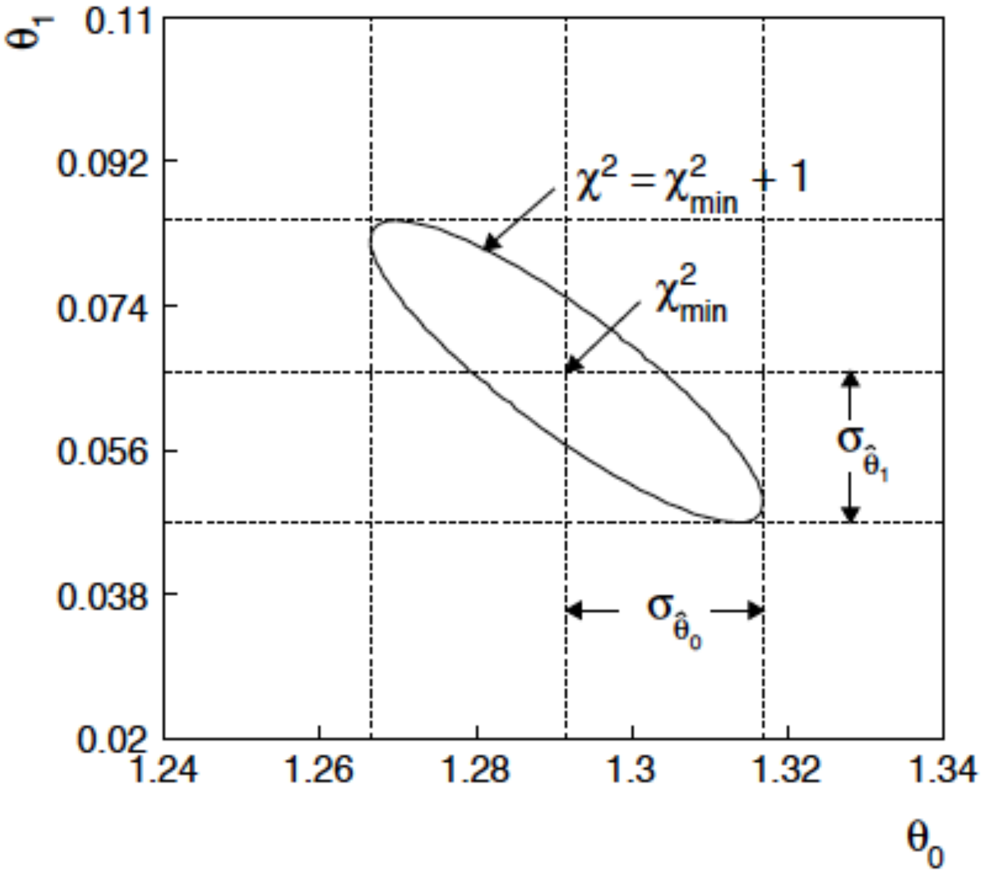
ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from tangent lines to contour

$$\chi^2 = \chi^2_{\text{min}} + 1 .$$

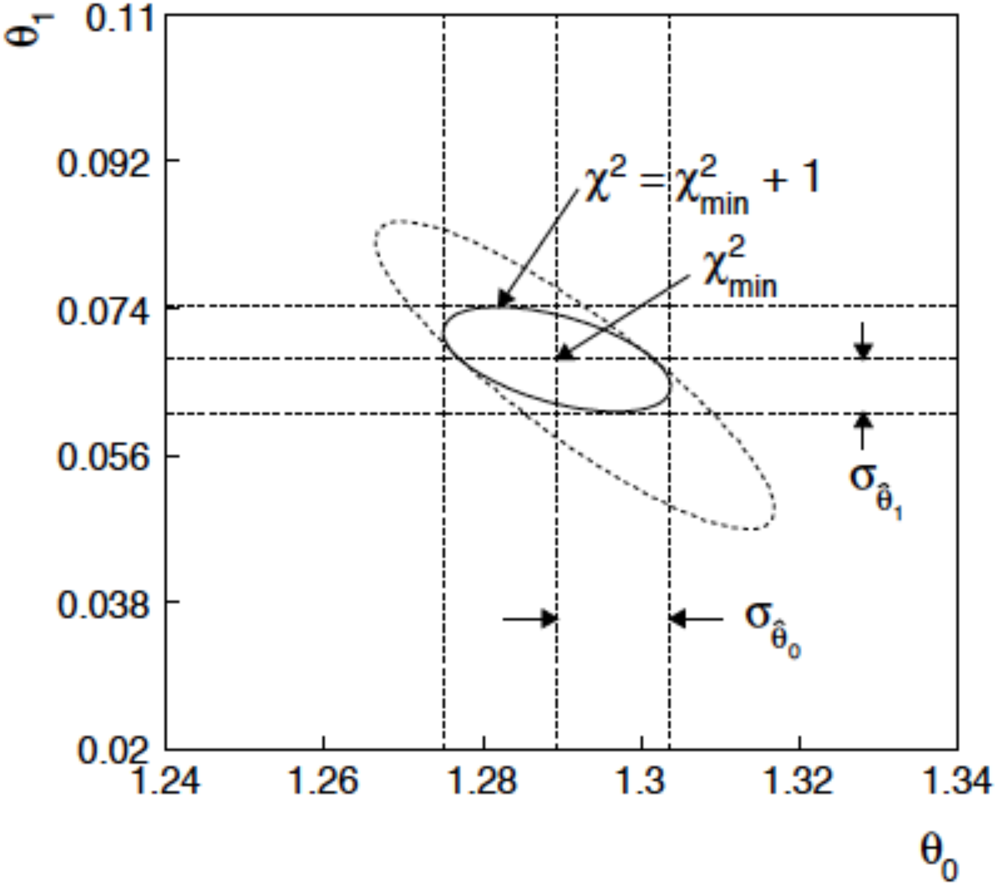
Correlation between $\hat{\theta}_0, \hat{\theta}_1$ causes errors to increase.



If we have a measurement $t_1 \sim$ Gauss (θ_1, σ_{t_1})

$$\chi^2(\theta_0,\theta_1) = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} + \frac{(\theta_1 - t_1)^2}{\sigma_{t_1}^2} \, .$$

The information on θ_1 improves accuracy of $\hat{\theta}_0$.



Bayesian method

We need to associate prior probabilities with θ_0 and θ_1, e.g.,

$$\begin{aligned}\pi(\theta_0,\theta_1) &= \pi_0(\theta_0)\,\pi_1(\theta_1)\\ \pi_0(\theta_0) &= \text{const.}\\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1-t_1)^2/2\sigma_{t_1}^2}\end{aligned}$$

← 'non-informative', in any case much broader than $L(\theta_0)$

← based on previous measurement

Putting this into Bayes' theorem gives:

$$p(\theta_0,\theta_1|\vec{y}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i-\mu(x_i;\theta_0,\theta_1))^2/2\sigma_i^2}\ \pi_0\ \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1-t_1)^2/2\sigma_{t_1}^2}$$

posterior ∝ likelihood × prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 \,|\, x)$ to find $p(\theta_0 \,|\, x)$:

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

In this example we can do the integral (rare). We find

$$\begin{aligned} p(\theta_0|x) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0}^2} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \ \text{(same as before)} \end{aligned}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x)\, d\theta_1 \ .$$

often high dimensionality and impossible in closed form, also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized Bayesian computation.

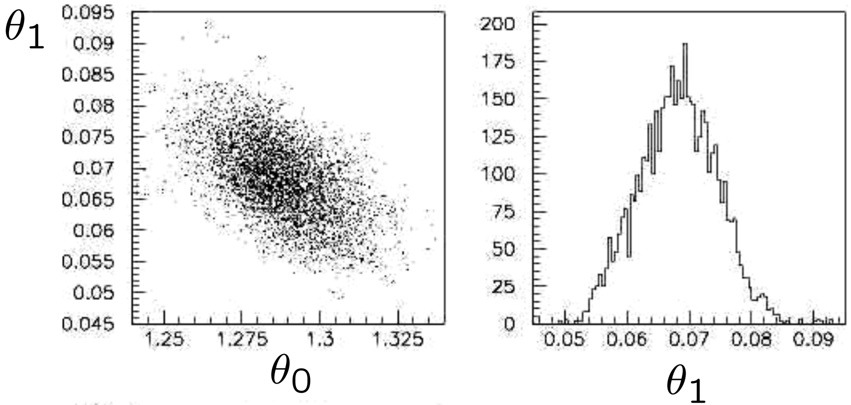
MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

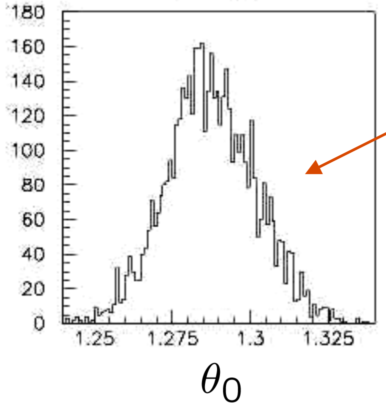
cannot use for many applications, e.g., detector MC;
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:





Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

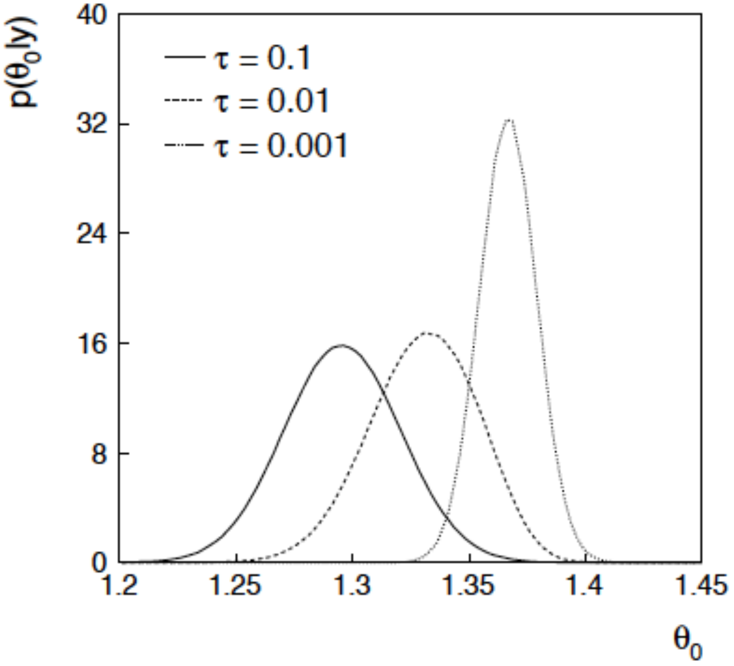
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau} \,, \quad \theta_1 \geq 0 \,, \quad \tau = 0.1 \,.$$

From this we obtain (numerically) the posterior pdf for θ_0:



This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

Wrapping up lecture 1

General idea of a statistical test:

Divide data spaced into two regions; depending on where data are then observed, accept or reject hypothesis.

Properties:

significance level or size (rate of Type-I error)
power (one minus rate of Type-II error)

Significance tests (also for goodness-of-fit):

p-value = probability to see level of incompatibility between data and hypothesis equal to or greater than level found with the actual data.

Parameter estimation

Maximize likelihood function → ML estimator.
Bayesian estimator based on posterior pdf.

Extra slides

Proof of Neyman-Pearson lemma

We want to determine the critical region *W* that maximizes the power

$$1 - \beta = \int_W P(x|H_1)\, dx$$

subject to the constraint

$$\alpha = \int_W P(x|H_0)\, dx$$

First, include in *W* all points where $P(x|H_0) = 0$, as they contribute nothing to the size, but potentially increase the power.

Proof of Neyman-Pearson lemma (2)

For $P(x|H_0) \neq 0$ we can write the power as

$$1 - \beta = \int_W \frac{P(x|H_1)}{P(x|H_0)} P(x|H_0)\, dx$$

The ratio of $1 - \beta$ to α is therefore

$$\frac{1-\beta}{\alpha} = \frac{\int_W \frac{P(x|H_1)}{P(x|H_0)} P(x|H_0)\, dx}{\int_W P(x|H_0)\, dx}$$

which is the average of the likelihood ratio $\mathrm{P}(x|H_1)$ / $\mathrm{P}(x|H_0)$ over the critical region W, assuming H_0.

$(1 - \beta) / \alpha$ is thus maximized if W contains the part of the sample space with the largest values of the likelihood ratio.

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.

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:

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

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.

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):

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

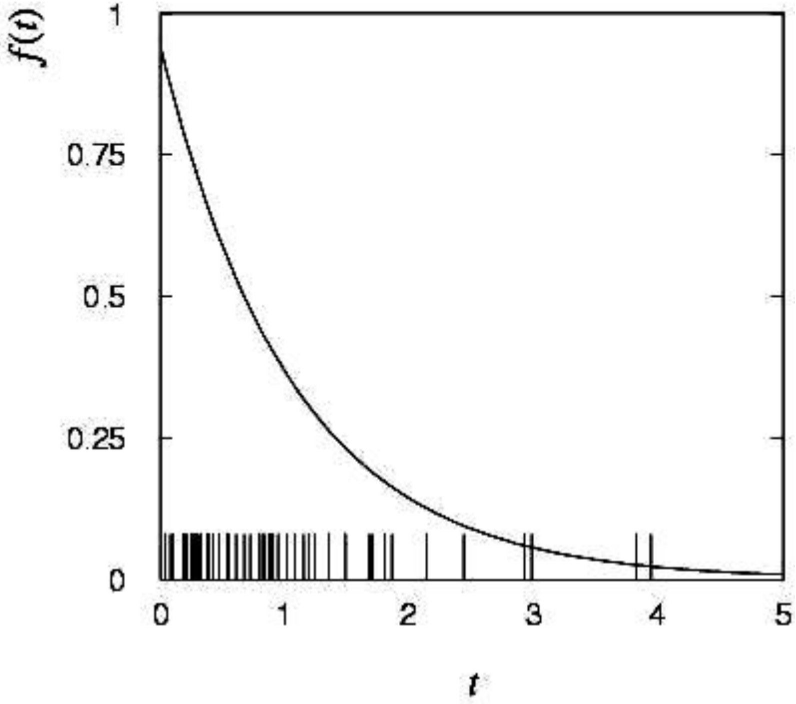
ML example: parameter of exponential pdf (2)

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

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

Monte Carlo test:

generate 50 values
using $\tau = 1$:



We find the ML estimate:

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

θ_1 known a priori

$$L(\theta_0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right] .$$

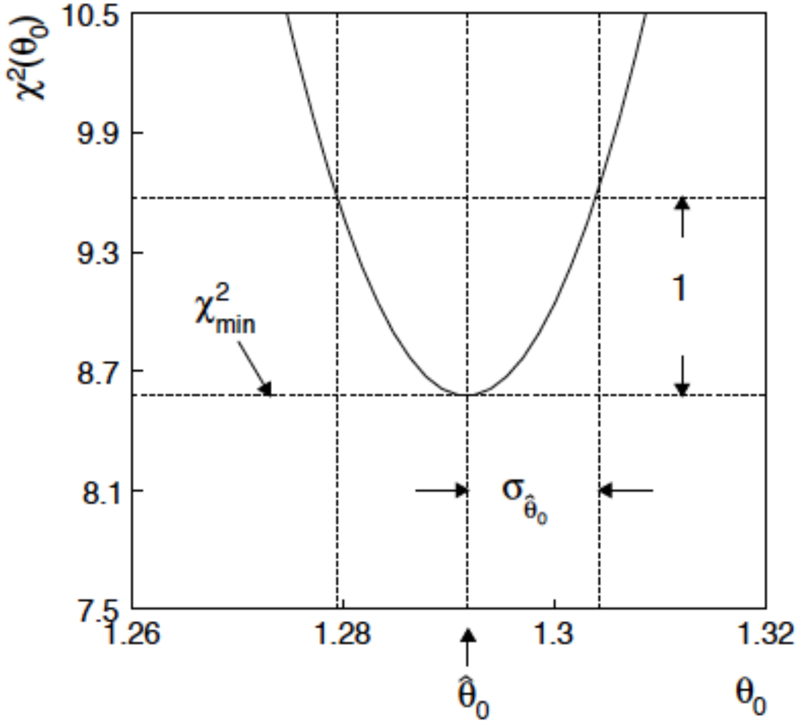
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from χ^2_{min}

to find $\sigma_{\hat{\theta}_0}$.



MCMC basics: Metropolis-Hastings algorithm

Goal: given an *n*-dimensional pdf $p(\vec{\theta})$,

generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$ e.g. Gaussian centred about $\vec{\theta}_0$

3) Form Hastings test ratio $\alpha = \min\left[1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

4) Generate $u \sim \mathrm{Uniform}[0, 1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ ← old point repeated

6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than naive $\sqrt{n}$.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

Test ratio is (*Metropolis*-Hastings): $\alpha = \min\left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)}\right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$. If proposed step rejected, hop in place.