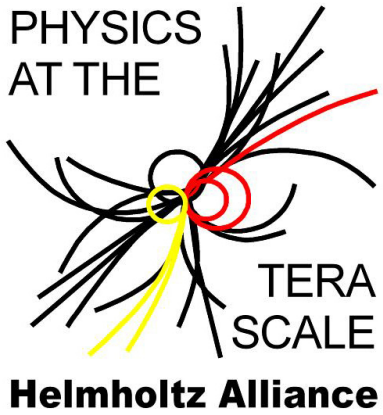
Statistical combinations, etc.

https://indico.desy.de/conferenceDisplay.py?confId=11244



Terascale Statistics School
DESY, Hamburg
March 23-27, 2015

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

Outline

1. Review of some formalism and analysis tasks
2. Broad view of combinations & review of parameter estimation
3. Combinations of parameter estimates.
4. Least-squares averages, including correlations
5. Comparison with Bayesian parameter estimation
6. Bayesian averages with outliers
7. PDG brief overview

Hypothesis, distribution, likelihood, model

Suppose the outcome of a measurement is x. (e.g., a number of events, a histogram, or some larger set of numbers).

A hypothesis H specifies the probability of the data $P(x|H)$.

Often H is labeled by parameter(s) $\theta \rightarrow P(x|\theta)$.

For the probability distribution $P(x|\theta)$, variable is x; θ is a constant.

If e.g. we evaluate $P(x|\theta)$ with the observed data and regard it as a function of the parameter(s), then this is the likelihood:

$$L(\theta) = P(x|\theta) \qquad \text{(Data } x \text{ fixed; treat } L \text{ as function of } \theta.\text{)}$$

Here use the term 'model' to refer to the full function $P(x|\theta)$ that contains the dependence both on x and θ.

(Sometimes write $L(x|\theta)$ for model or likelihood, depending on context.)

Bayesian use of the term 'likelihood'

We can write Bayes theorem as

prior

posterior

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

where $L(x|\theta)$ is the likelihood. It is the probability for x given θ, evaluated with the observed x, and viewed as a function of θ.

Bayes' theorem only needs $L(x|\theta)$ evaluated with a given data set (the 'likelihood principle').

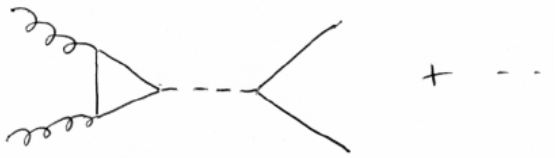
For frequentist methods, in general one needs the full model.

For some approximate frequentist methods, the likelihood is enough.

Theory ↔ Statistics ↔ Experiment

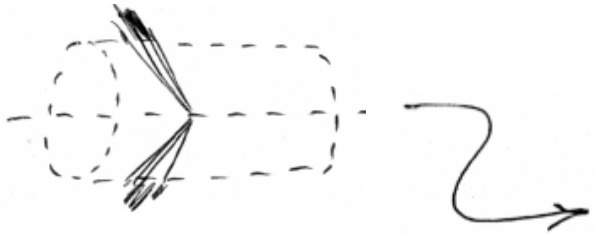
Theory (model, hypothesis):

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \not{D} \psi + \cdots$$

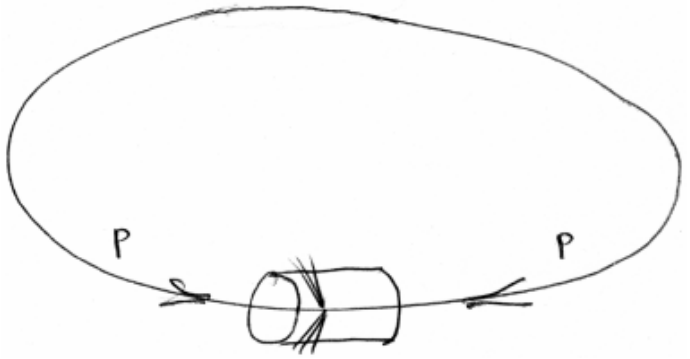


$$\sigma = \frac{G_F \alpha_s^2 m_H^2}{288 \sqrt{2\pi}} \times$$

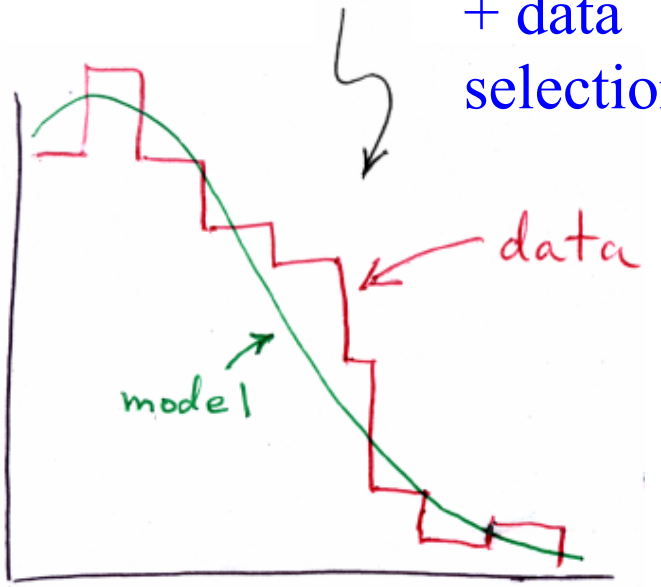
+ simulation of detector and cuts



Experiment:

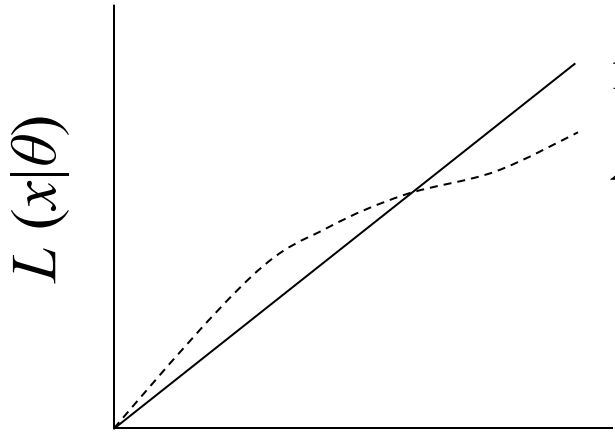


+ data selection



Nuisance parameters

In general our model of the data is not perfect:



model: $L(x|\theta) = \theta x$

truth: $L(x|\theta) = \theta x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$L(x|\theta) \to L(x|\theta, \nu)$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be "true".

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Data analysis in particle physics

Observe events (e.g., pp collisions) and for each, measure a set of characteristics:

particle momenta, number of muons, energy of jets,...

Compare observed distributions of these characteristics to predictions of theory. From this, we want to:

Estimate the free parameters of the theory: $m_H = 125.4$

Quantify the uncertainty in the estimates: ± 0.4 GeV

Assess how well a given theory stands in agreement with the observed data: 0^+ good, 2^+ bad

Test/exclude regions of the model's parameter space (→ limits)

Combinations

"Combination of results" can be taken to mean: how to construct a model that incorporates more data.

E.g. several experiments,

Experiment 1: data x, model $P(x|\theta) \rightarrow$ upper limit $\theta_{\text{up},1}$
Experiment 2: data y, model $P(y|\theta) \rightarrow$ upper limit $\theta_{\text{up},2}$

Or main experiment and control measurement(s).

The best way to do the combination is at the level of the data, e.g., (if x,y independent)

$$P(x,y|\theta) = P(x|\theta)\, P(y|\theta) \rightarrow \text{ "combined" limit } \theta_{\text{up,comb}}$$

If the data are not available but rather only the "results" (limits, parameter estimates, p-values) then possibilities are more limited.

Usually OK for parameter estimates, difficult/impossible for limits, p-values without additional assumptions & information loss.

Quick review of frequentist parameter estimation

Suppose we have a pdf characterized by one or more parameters:

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

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):

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

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

Maximum likelihood

The most important frequentist method for constructing estimators is to take the value of the parameter(s) that maximize the likelihood:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(x|\theta)$$

The resulting estimators are functions of the data and thus characterized by a sampling distribution with a given (co)variance:

$$V_{ij} = \operatorname{cov}[\hat{\theta}_i, \hat{\theta}_j]$$

In general they may have a nonzero bias:

$$b = E[\hat{\theta}] - \theta$$

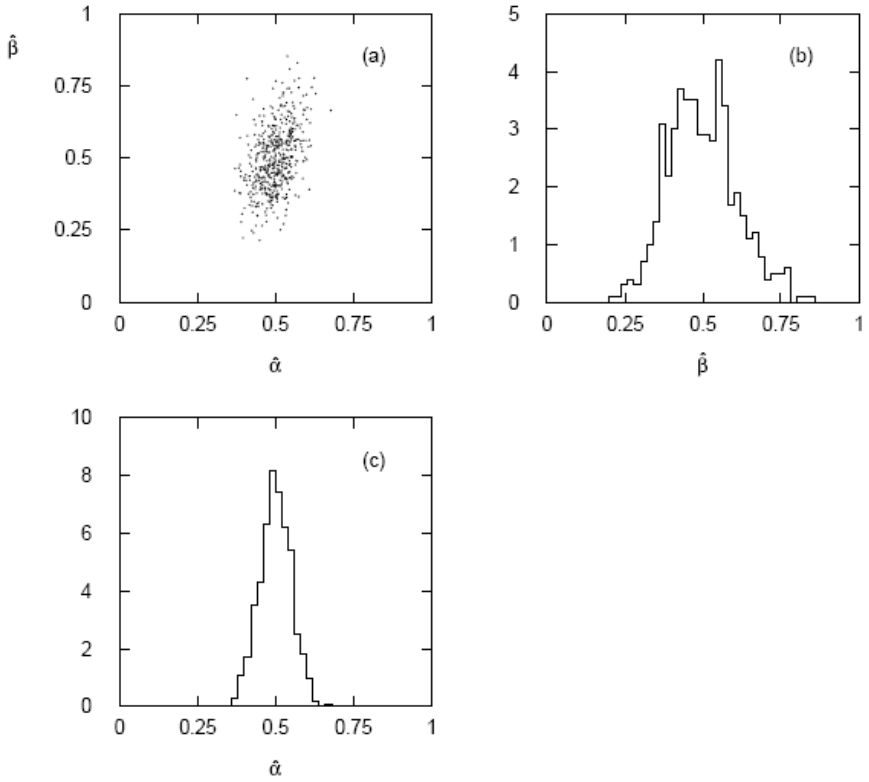
Under conditions usually satisfied in practice, bias of ML estimators is zero in the large sample limit, and the variance is as small as possible for unbiased estimators.

ML estimator may not in some cases be regarded as the optimal trade-off between these criteria (cf. regularized unfolding).

Ingredients for ML

To find the ML estimate itself one only needs the likelihood $L(\theta)$.

In principle to find the covariance of the estimators, one requires the full model $L(x|\theta)$. E.g., simulate many times independent data sets and look at distribution of the resulting estimates:

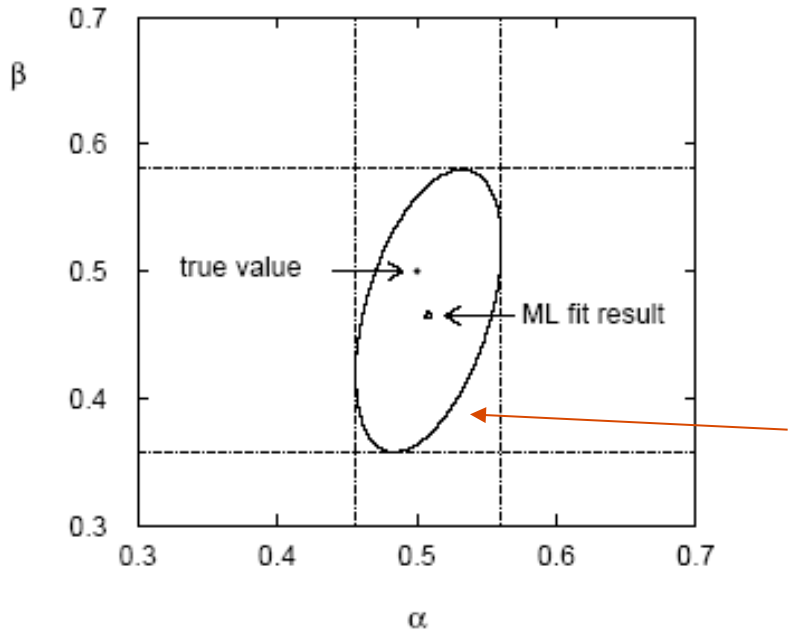


$$\overline{\hat{\alpha}} = 0.499$$
$$s_{\hat{\alpha}} = 0.051$$
$$\overline{\hat{\beta}} = 0.498$$
$$s_{\hat{\beta}} = 0.111$$
$$\widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0024$$
$$r = 0.42$$

Ingredients for ML (2)

Often (e.g., large sample case) one can approximate the covariances using only the likelihood $L(\theta)$:

$$\widehat{V}_{ij}^{-1} \approx -\left.\frac{\partial^2 \ln L}{\partial\theta_i\,\partial\theta_j}\right|_{\theta=\hat{\theta}}$$



This translates into a simple graphical recipe:

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

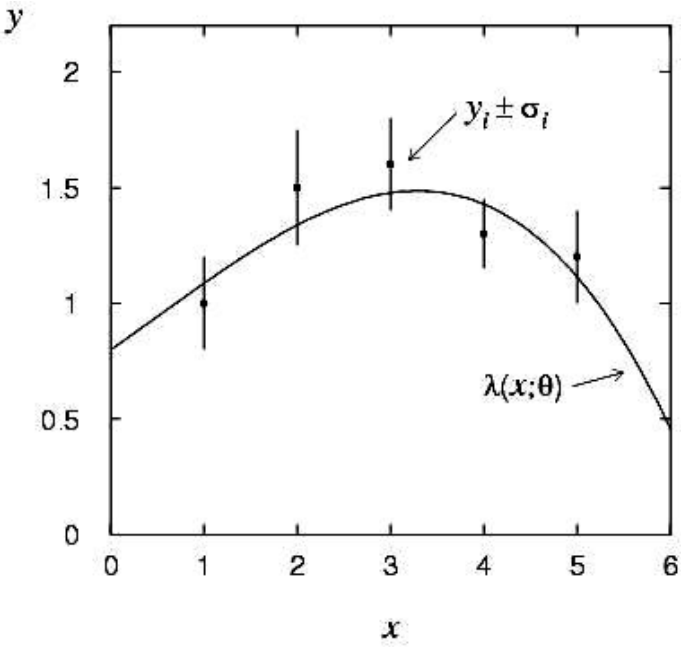
→ Tangent lines to contours give standard deviations.

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

The method of least squares

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

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



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

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

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

The likelihood function is

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

The method of least squares (2)

The log-likelihood function is therefore

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

So maximizing the likelihood is equivalent to minimizing

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

Minimum 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 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)

Goodness-of-fit with least squares

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

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

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

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

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

where the number of degrees of freedom is

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

Goodness-of-fit with least squares (2)

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

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

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

Using LS to combine measurements

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

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

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

λ = true value (plays role of θ).

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,

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

Combining correlated measurements with LS

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

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

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

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

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

That is, if we take the estimator to be a linear form $\Sigma_i\, w_i\, y_i$, and find the w_i that minimize its variance, we get the LS solution (= BLUE, Best Linear Unbiased Estimator).

Example: averaging two correlated measurements

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

$$\rightarrow \quad \hat{\lambda} = wy_1 + (1-w)y_2, \quad w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

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

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.

Covariance, correlation, etc.

For a pair of random variables x and y, the covariance and correlation are

$$\mathrm{cov}[x, y] = E[xy] - E[x]E[y] \qquad \rho_{xy} = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y}$$

One only talks about the correlation of two quantities to which one assigns probability (i.e., random variables).

So in frequentist statistics, estimators for parameters can be correlated, but not the parameters themselves.

In Bayesian statistics it does make sense to say that two parameters are correlated, e.g.,

$$\mathrm{cov}[\theta_i, \theta_j] = \int \theta_i \theta_j p(\boldsymbol{\theta}|x)\, d\boldsymbol{\theta} - \int \theta_i p(\boldsymbol{\theta}|x)\, d\boldsymbol{\theta} \int \theta_j p(\boldsymbol{\theta}|x)\, d\boldsymbol{\theta}$$

Example of "correlated systematics"

Suppose we carry out two independent measurements of the length of an object using two rulers with diferent thermal expansion properties.

Suppose the temperature is not known exactly but must be measured (but lengths measured together so T same for both),

$$T \sim \text{Gauss}(\tau, \sigma_T)$$

The expectation value of the measured length L_i ($i = 1, 2$) is related to true length λ at a reference temperature τ_0 by

$$E[L_i] = \lambda - \alpha_i(T - \tau_0), \qquad i = 1, 2$$

and the (uncorrected) length measurements are modeled as

$$L_i \sim \text{Gauss}(\lambda - \alpha_i(\tau - \tau_0), \sigma_i)$$

Two rulers (2)

The model thus treats the measurements T, L_1, L_2 as uncorrelated with standard deviations σ_T, σ_1, σ_2, respectively:

$$L(T, L_1, L_2|\lambda, \tau) = \frac{1}{\sqrt{2\pi}\sigma_T} e^{-(T-\tau)^2/2\sigma_T^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(L_i-\lambda+\alpha_i(\tau-T_0))^2/2\sigma_i^2}$$

Alternatively we could correct each raw measurement:

$$y_i = L_i + \alpha_i(T - \tau_0)$$

which introduces a correlation between y_1, y_2 and T

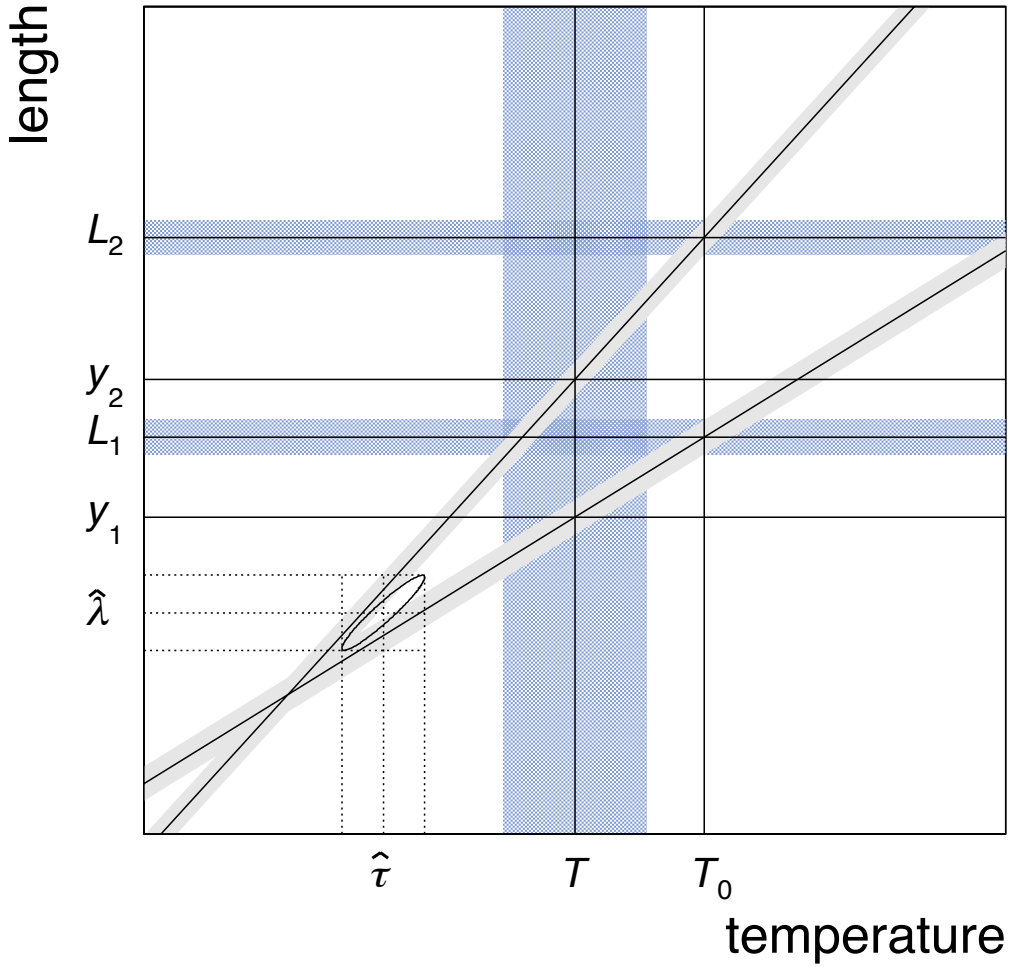
$$\mathrm{cov}[y_1, y_2] = \alpha_1\alpha_2\sigma_T^2 \qquad \mathrm{cov}[y_i, T] = \alpha_i\sigma_T^2$$

But the likelihood function (multivariate Gauss in T, y_1, y_2) is the same function of τ and λ as before (equivalent!).

Language of y_1, y_2: temperature gives correlated systematic.
Language of L_1, L_2: temperature gives "coherent" systematic.

Two rulers (3)

Outcome has some surprises:



Estimate of λ does not lie between y_1 and y_2.

Stat. error on new estimate of temperature substantially smaller than initial σ_T.

These are features, not bugs, that result from our model assumptions.

Two rulers (4)

We may re-examine the assumptions of our model and conclude that, say, the parameters α_1, α_2 and τ_0 were also uncertain.

We may treat their nominal values as measurements (need a model; Gaussian?) and regard α_1, α_2 and τ_0 as as nuisance parameters.

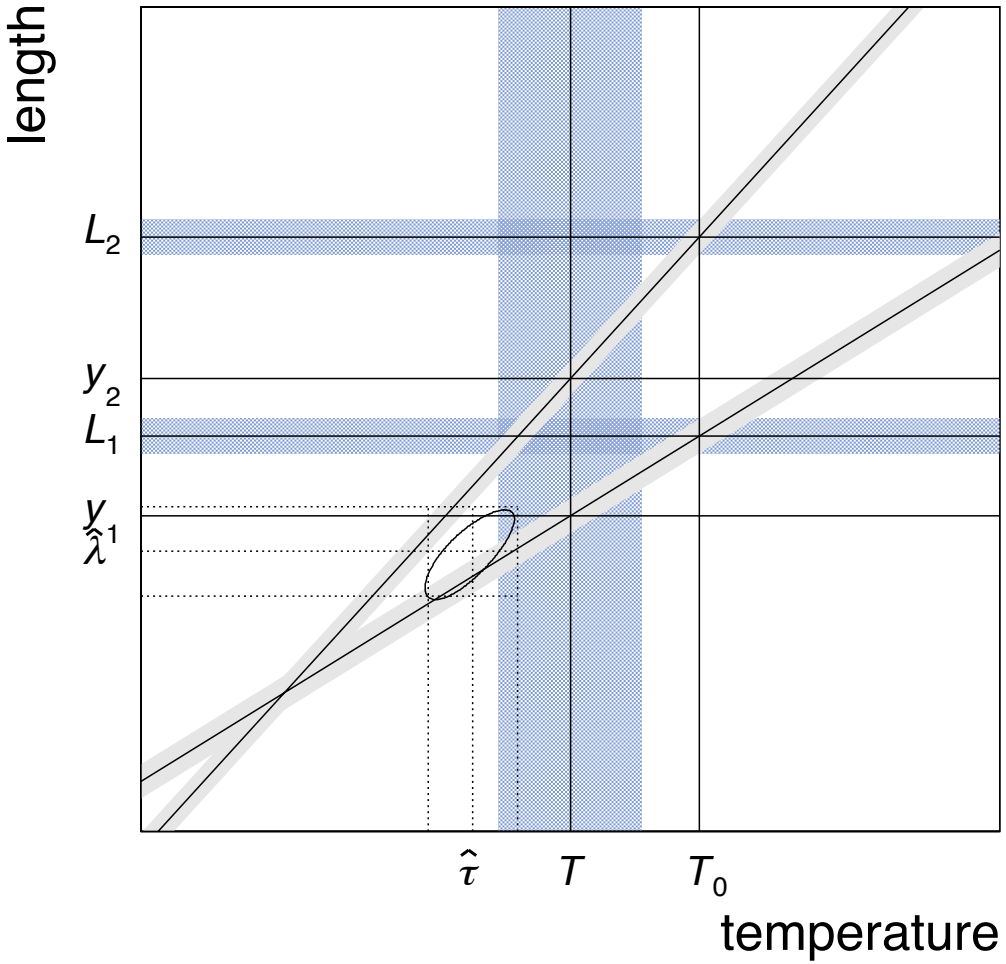
$$L(L_1, L_2, T, \tilde{\tau}_0, \tilde{\alpha}_1, \tilde{\alpha}_2 | \lambda, \tau, \tau_0, \alpha_1, \alpha_2) =$$

$$\frac{1}{\sqrt{2\pi}\sigma_T} e^{-(T-\tau)^2/2\sigma_T^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(L_i - \lambda + \alpha_i(\tau - \tau_0))^2/2\sigma_i^2}$$

$$\times \frac{1}{\sqrt{2\pi}\sigma_{\tilde{\tau}_0}} e^{-(\tilde{\tau}_0 - \tau_0)^2/2\sigma_{\tilde{\tau}_0}^2} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}\sigma_{\tilde{\alpha}_i}} e^{-(\tilde{\alpha}_i - \alpha_i)^2/2\sigma_{\tilde{\alpha}_i}^2}$$

Two rulers (5)

The outcome changes; some surprises may be "reduced".



"Related" parameters

Suppose the model for two independent measurements x and y contain the same parameter of interest μ and a common nuisance parameter, θ, such as the jet-energy scale.

To combine the measurements, construct the full likelihood:

$$L(\mu, \theta) = P(x|\mu, \theta)P(y|\mu, \theta)$$

Although one may think of θ as common to the two measurements, this could be a poor approximation (e.g., the two analyses use jets with different angles/energies, so a single jet-energy scale is not a good model).

Better model: suppose the parameter for x is θ, and for y it is

$$\theta' = \theta + \varepsilon$$

where ε is an additional nuisance parameter expected to be small.

Model with nuisance parameters

The additional nuisance parameters in the model may spoil our sensitivity to the parameter of interest μ, so we need to constrain them with control measurements.

Often we have no actual control measurements, but some "nominal values" for θ and ε, $\widetilde{\theta}$ and $\widetilde{\varepsilon}$, which we treat as if they were measurements, e.g., with a Gaussian model:

$$p(\tilde{\theta}, \tilde{\varepsilon}|\theta, \varepsilon) = \text{Gauss}(\tilde{\theta}|\theta, \sigma_{\tilde{\theta}})\text{Gauss}(\tilde{\varepsilon}|\varepsilon, \sigma_{\tilde{\varepsilon}})$$

We started by considering θ and θ' to be the same, so probably take $\widetilde{\varepsilon} = 0$.

So we now have an improved model

$$L(\mu, \theta, \varepsilon) = P(x|\mu, \theta)P(y|\mu, \theta, \varepsilon)p(\tilde{\theta}, \tilde{\varepsilon}|\theta, \varepsilon)$$

with which we can estimate μ, set limits, etc.

Example: fitting a straight line

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

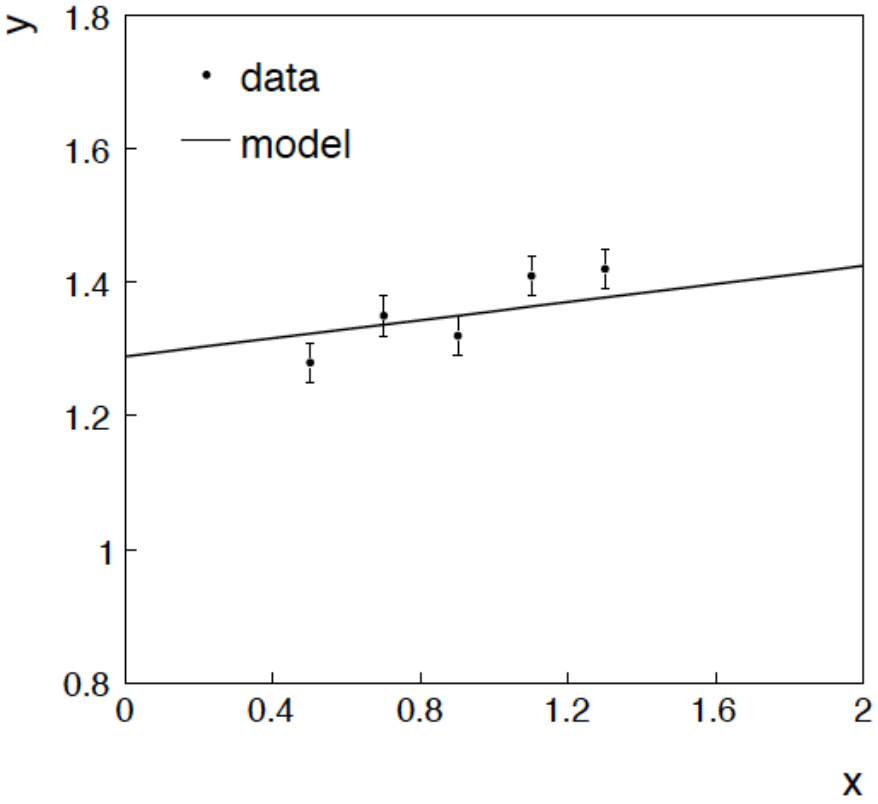
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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)

θ_1 known a priori

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

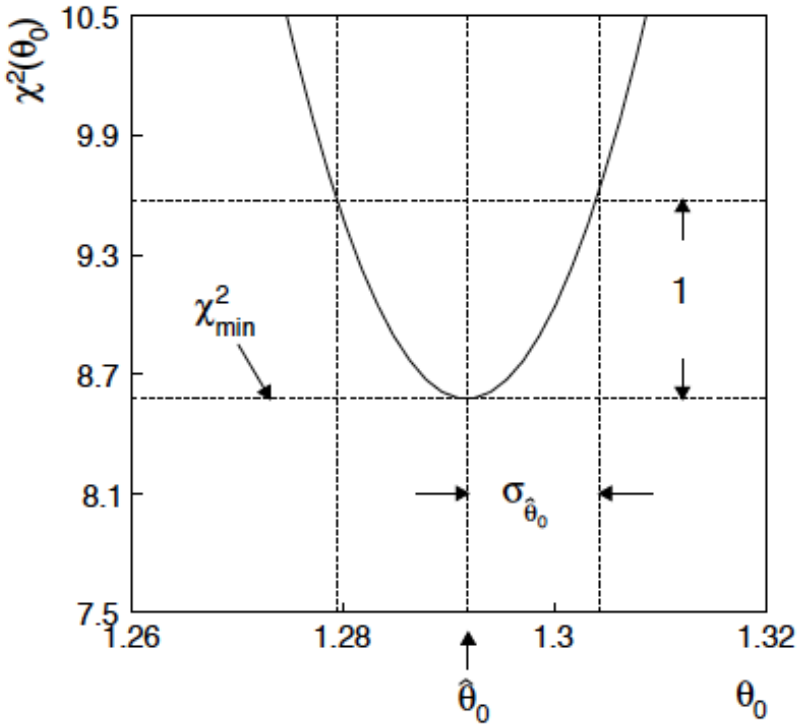
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

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



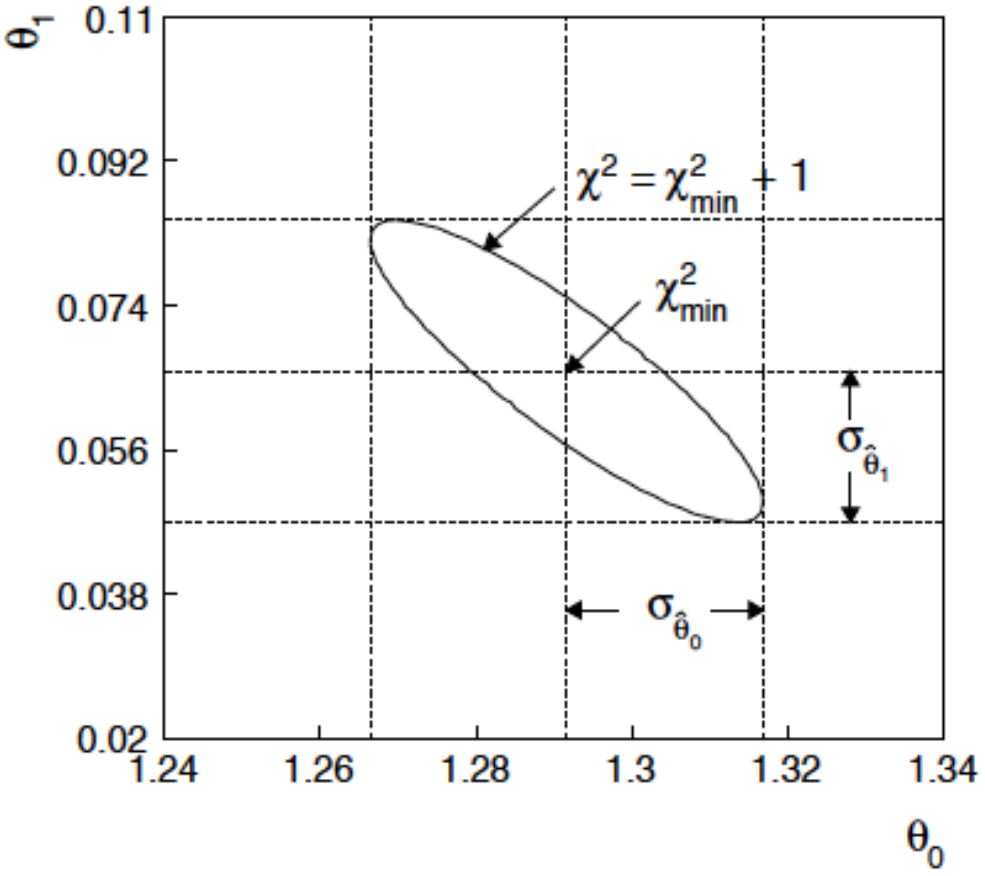
ML (or LS) fit of θ_0 and θ_1

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

Standard deviations from tangent lines to contour

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

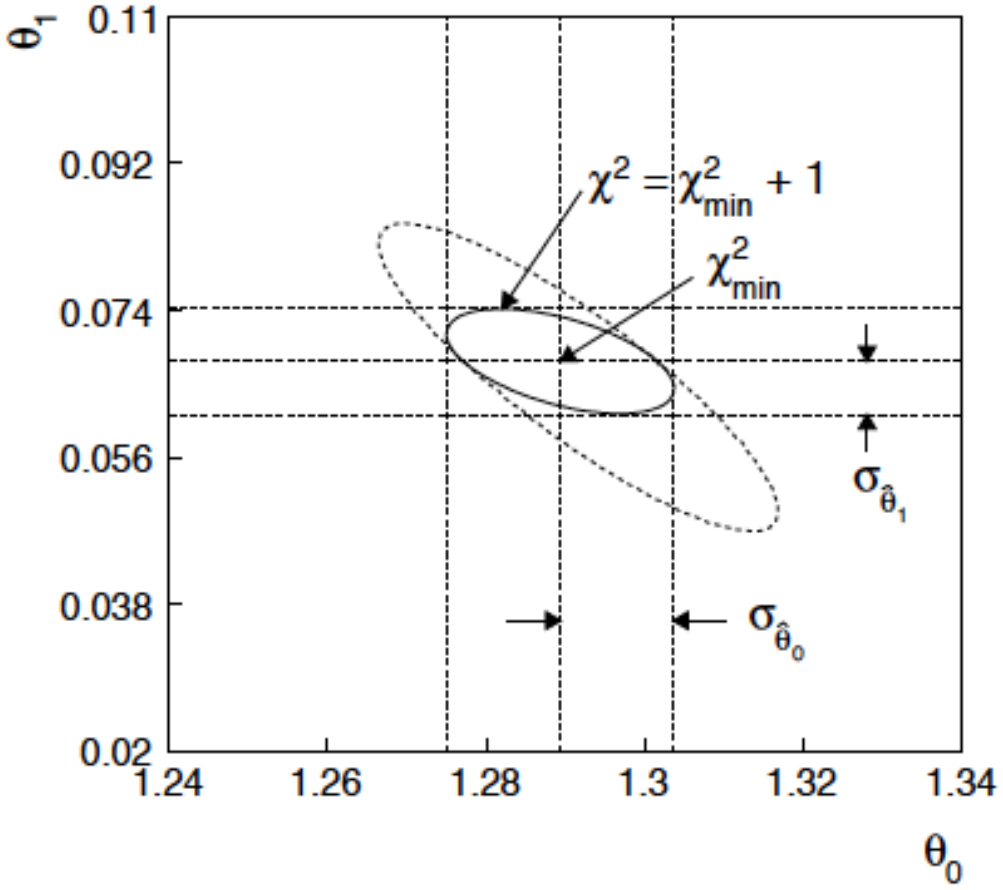
Correlation between $\hat{\theta}_0$, $\hat{\theta}_1$ causes errors to increase.



If we have a measurement $t_1 \sim \text{Gauss}\,(\theta_1, \sigma_{t_1})$

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

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.,

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\,\pi_1(\theta_1)$$

$$\pi_0(\theta_0) = \text{const.}$$

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

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

← 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

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

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

Marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta|x) = \int p(\theta, \nu|x)\, d\nu$$

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.

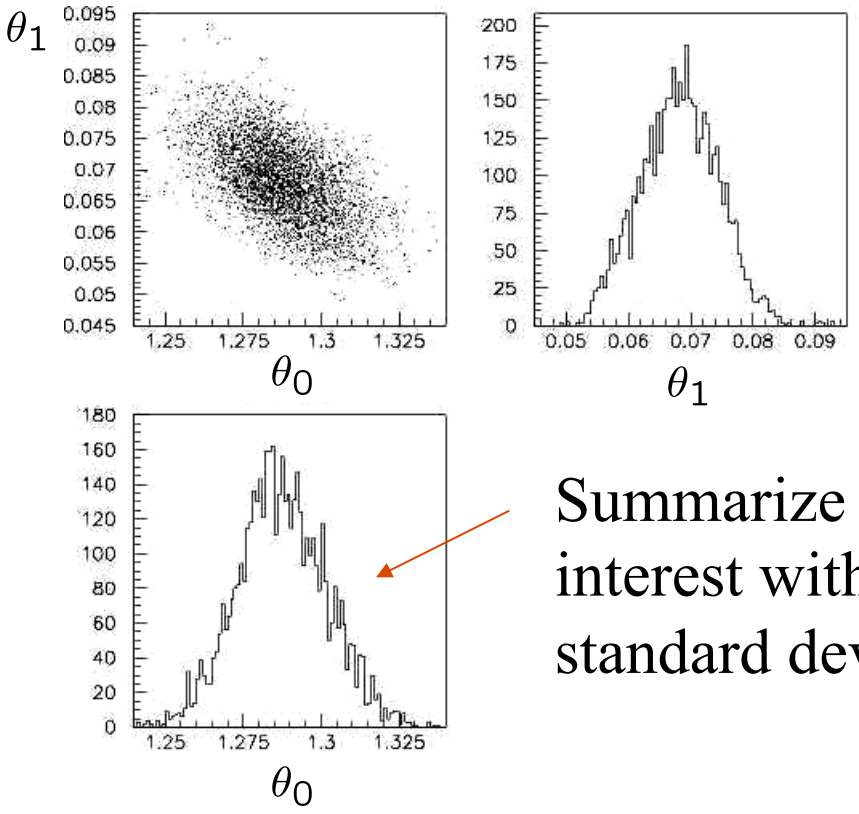
MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than naive $\sqrt{n}$.

Basic idea: sample full multidimensional parameter space;
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.

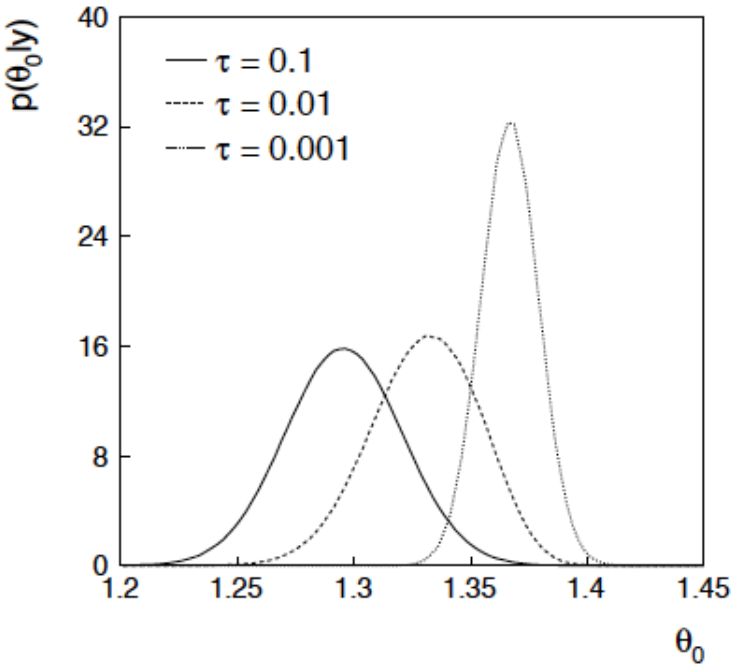
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

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.

The error on the error

Some systematic errors are well determined

Error from finite Monte Carlo sample

Some are less obvious

Do analysis in n 'equally valid' ways and
extract systematic error from 'spread' in results.

Some are educated guesses

Guess possible size of missing terms in perturbation series;

vary renormalization scale $(\mu/2 < Q < 2\mu$?)

Can we incorporate the 'error on the error'?

(cf. G. D'Agostini 1999; Dose & von der Linden 1999)

A more general fit (symbolic)

Given measurements: $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$,

and (usually) covariances: V_{ij}^{stat}, V_{ij}^{sys} .

Predicted value: $\mu(x_i; \theta)$, expectation value $E[y_i] = \mu(x_i; \theta) + b_i$

control variable parameters bias

Often take: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

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

Equivalent to maximizing $L(\theta)$ » $e^{-\chi^2/2}$, i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

Its Bayesian equivalent

Take $L(\vec{y}|\vec{\theta},\vec{b}) \sim \exp\left[-\frac{1}{2}(\vec{y}-\vec{\mu}(\theta)-\vec{b})^T\, V_{\text{stat}}^{-1}\,(\vec{y}-\vec{\mu}(\theta)-\vec{b})\right]$

$\pi_b(\vec{b}) \sim \exp\left[-\frac{1}{2}\,\vec{b}^T\, V_{\text{sys}}^{-1}\,\vec{b}\right]$

$\pi_\theta(\theta) \sim \text{const.}$

Joint probability for all parameters

and use Bayes' theorem: $p(\theta,\vec{b}|\vec{y}) \propto L(\vec{y}|\theta,\vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$

To get desired probability for θ, integrate (marginalize) over $\boldsymbol{b}$:

$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\, d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, σ_θ same as from $\chi^2 = \chi^2_{\text{min}} + 1$. (Back where we started!)

Motivating a non-Gaussian prior $\pi_b(b)$

Suppose now the experiment is characterized by

$$y_i, \quad \sigma_i^{\text{stat}}, \quad \sigma_i^{\text{sys}}, \quad s_i, \quad i = 1, \ldots, n \,,$$

where s_i is an (unreported) factor by which the systematic error is over/under-estimated.

Assume correct error for a Gaussian $\pi_b(b)$ would be $s_i\sigma_i^{\text{sys}}$, so

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi}s_i\sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i\sigma_i^{\text{sys}})^2}\right] \pi_s(s_i)\, ds_i$$

Width of $\pi_s(s_i)$ reflects 'error on the error'.

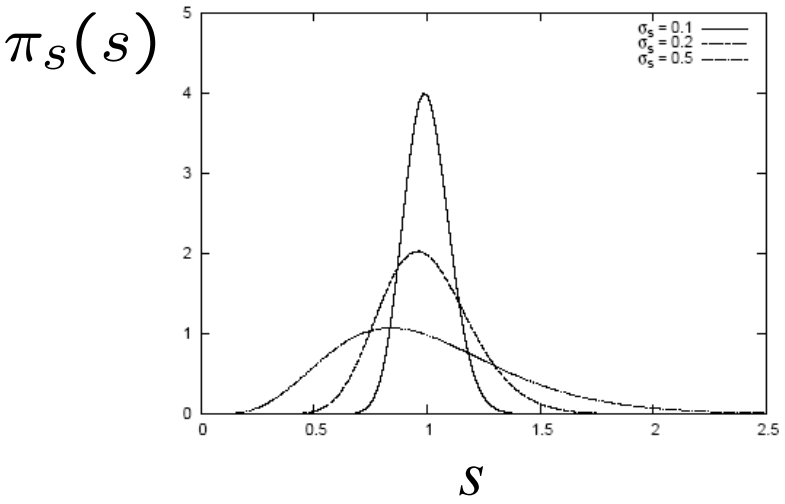
Error-on-error function $\pi_s(s)$

A simple unimodal probability density for $0 < s < 1$ with adjustable mean and variance is the Gamma distribution:

$$\pi_s(s) = \frac{a(as)^{b-1}e^{-as}}{\Gamma(b)}$$

mean = b/a
variance = b/a^2

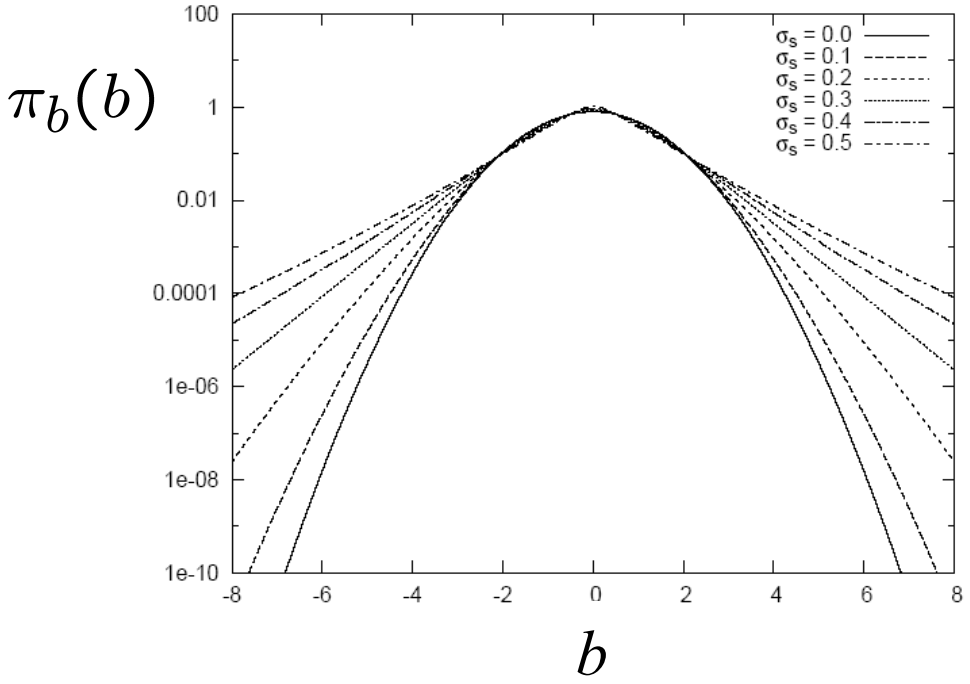
Want e.g. expectation value of 1 and adjustable standard deviation σ_s , i.e., $a = b = 1/\sigma_s^2$



In fact if we took $\pi_s(s) \sim$ *inverse Gamma*, we could integrate $\pi_b(b)$ in closed form (cf. D'Agostini, Dose, von Linden). But Gamma seems more natural & numerical treatment not too painful.

Prior for bias $\pi_b(b)$ now has longer tails

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi} s_i \sigma_i^{\text{sys}}} \exp\left[-\frac{1}{2}\frac{b_i^2}{(s_i\sigma_i^{\text{sys}})^2}\right] \pi_s(s_i)\, ds_i$$



Gaussian ($\sigma_s = 0$) $\quad P(|b| > 4\sigma_{\text{sys}}) = 6.3 \times 10^{-5}$

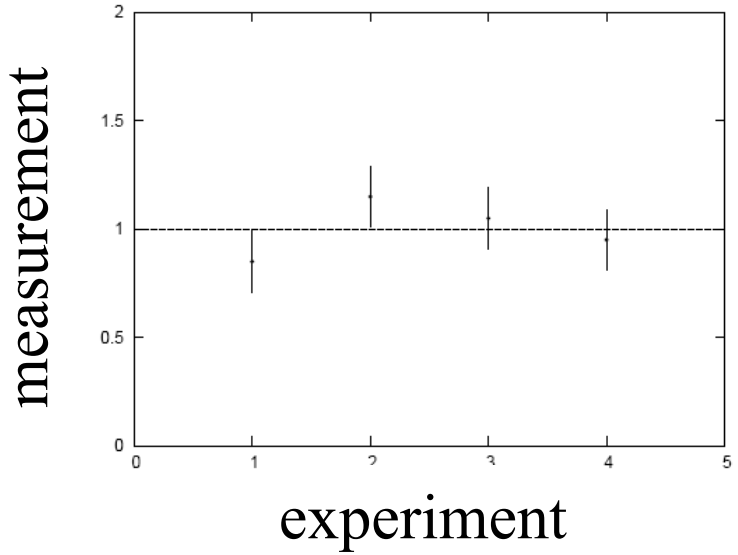
$\sigma_s = 0.5$ $\quad P(|b| > 4\sigma_{\text{sys}}) = 0.65\%$

A simple test

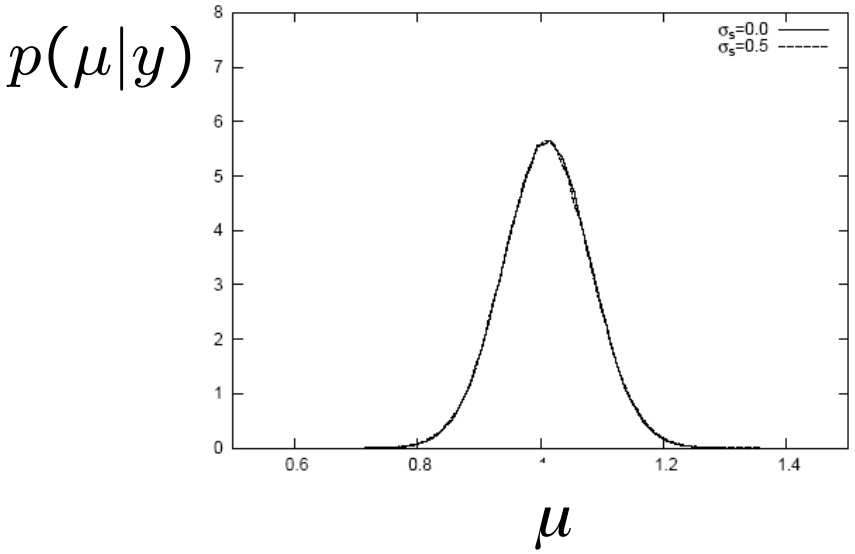
Suppose fit effectively averages four measurements.

Take $\sigma_{\text{sys}} = \sigma_{\text{stat}} = 0.1$, uncorrelated.

Case #1: data appear compatible



Posterior $p(\mu|y)$:



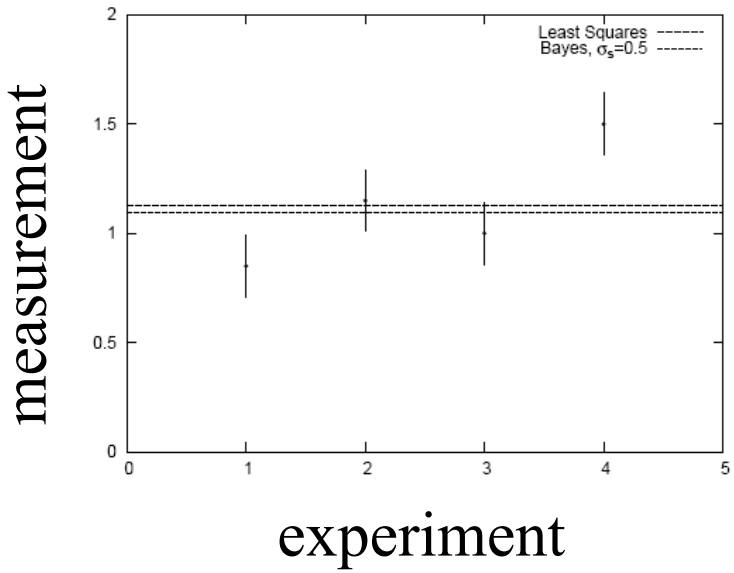
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$\sigma_s = 0.0: \quad \hat{\mu} = 1.000 \pm 0.071$

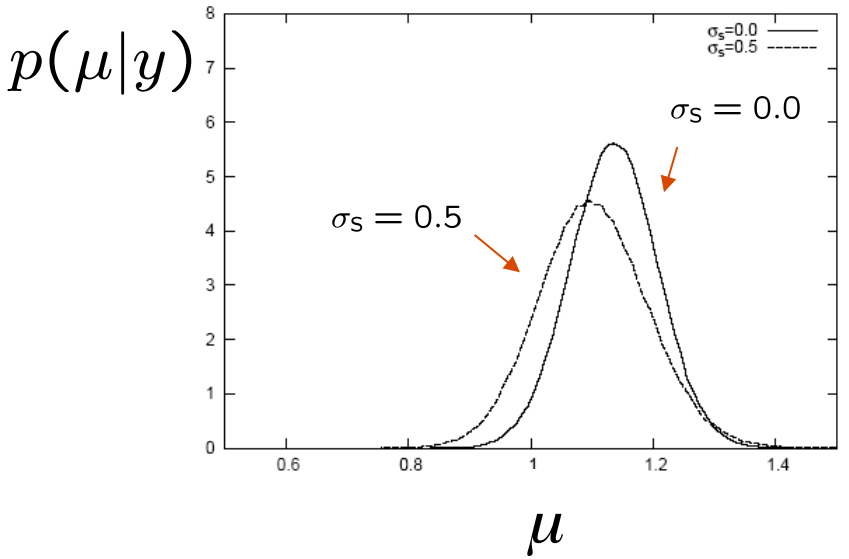
$\sigma_s = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



$$\sigma_s = 0.0: \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_s = 0.5: \quad \hat{\mu} = 1.093 \pm 0.089$$

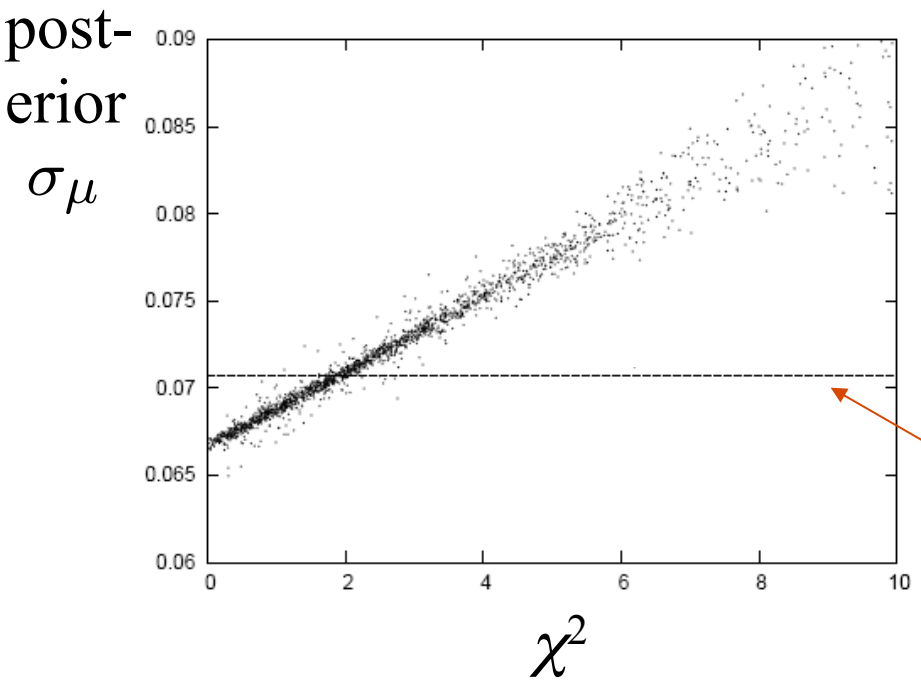
→ Bayesian fit less sensitive to outlier.

→ Error now connected to goodness-of-fit.

Goodness-of-fit vs. size of error

In LS fit, value of minimized χ^2 does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high χ^2 corresponds to a larger error (and vice versa).



Particle Data Group averages

K.A. Olive et al. (Particle Data Group), Chin. Phys. C, 38, 090001 (2014); pdg.lbl.gov

The PDG needs pragmatic solutions for averages where the reported information may be incomplete/inconsistent.

Often this involves taking the quadratic sum of statistical and systematic uncertainties for LS averages.

If asymmetric errors (confidence intervals) are reported, PDG has a recipe to reconstruct a model based on a Gaussian-like function where sigma is a continuous function of the mean.

Exclusion of inconsistent data "sometimes quite subjective".

If min. chi-squared is much larger than the number of degrees of freedom $N_{\rm dof} = N - 1$, scale up the input errors a factor

$$S = \left[\chi^2/(N-1)\right]^{1/2}$$

so that new $\chi^2 = N_{\rm dof}$. Error on the average increased by S.

Summary on combinations

The basic idea of combining measurements is to write down the model that describes all of the available experimental outcomes.

If the original data are not available but only parameter estimates, then one treats the estimates (and their covariances) as "the data". Often a multivariate Gaussian model is adequate for these.

If the reported values are limits, there are few meaningful options.

PDG does not combine limits unless the can be "deconstructed" back into a Gaussian measurement.

ATLAS/CMS 2011 combination of Higgs limits used the histograms of event counts (not the individual limits) to construct a full model (ATLAS-CONF-2011-157, CMS PAS HIG-11-023).

Important point is to publish enough information so that meaningful combinations can be carried out.

Extra slides

A quick review of frequentist statistical tests

Consider a hypothesis H_0 and alternative H_1.

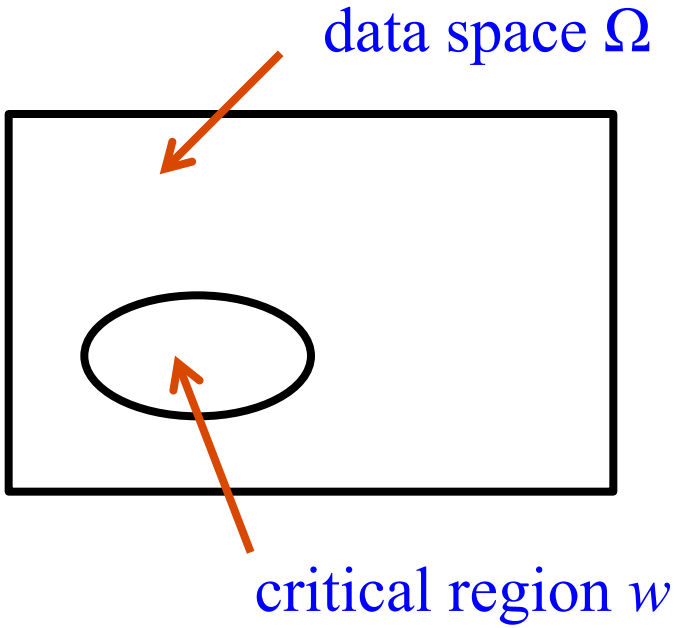
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

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

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.

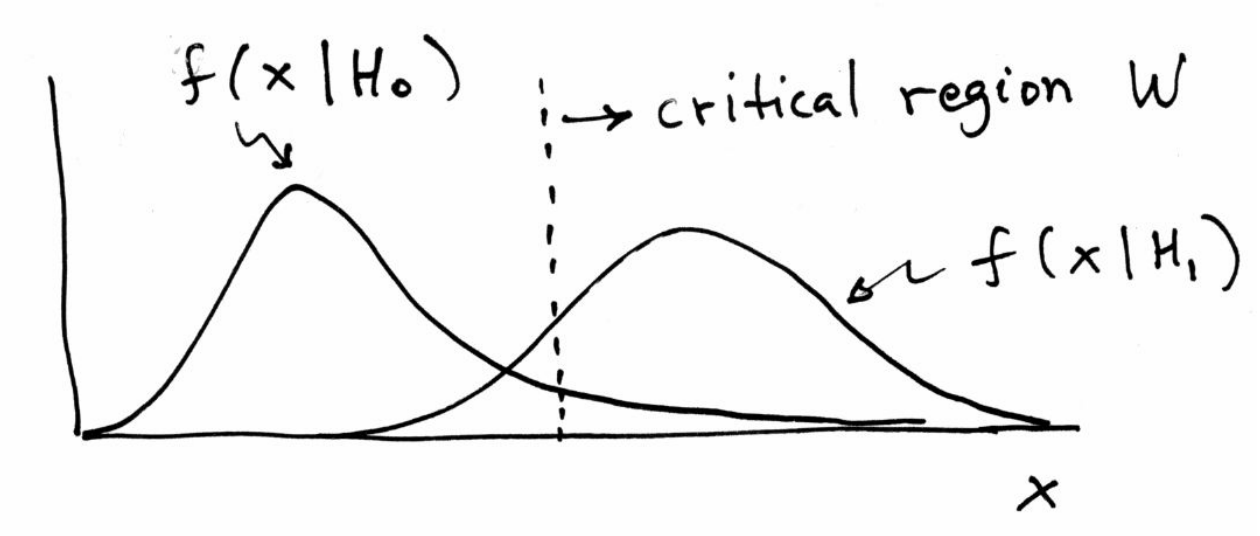


Definition of a test (2)

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

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.

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

$$P(x \in \mathrm{S} - W \mid H_1) = \beta$$

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

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

p-values

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: $\vec{x}_{\text{obs}}$

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

Express level of compatibility by giving the *p*-value for H:

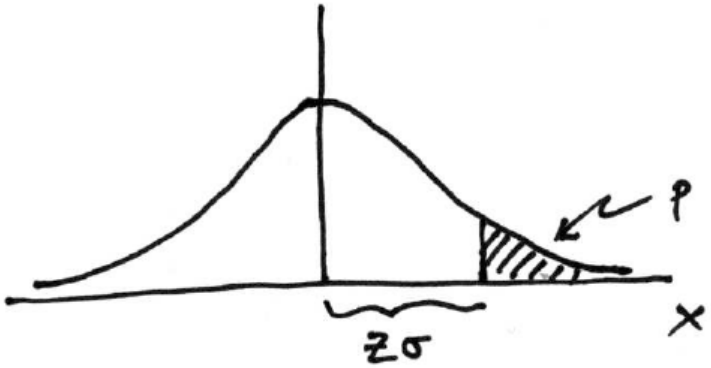
p = probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

This is not the probability that H is true!

Requires one to say what part of data space constitutes lesser compatibility with H than the observed data (implicitly this means that region gives better agreement with some alternative).

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 $p = 2.9 \times 10^{-7}$.

Using a p-value to define test of H_0

One can show the distribution of the p-value of H, under assumption of H, is uniform in [0,1].

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 can 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.

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

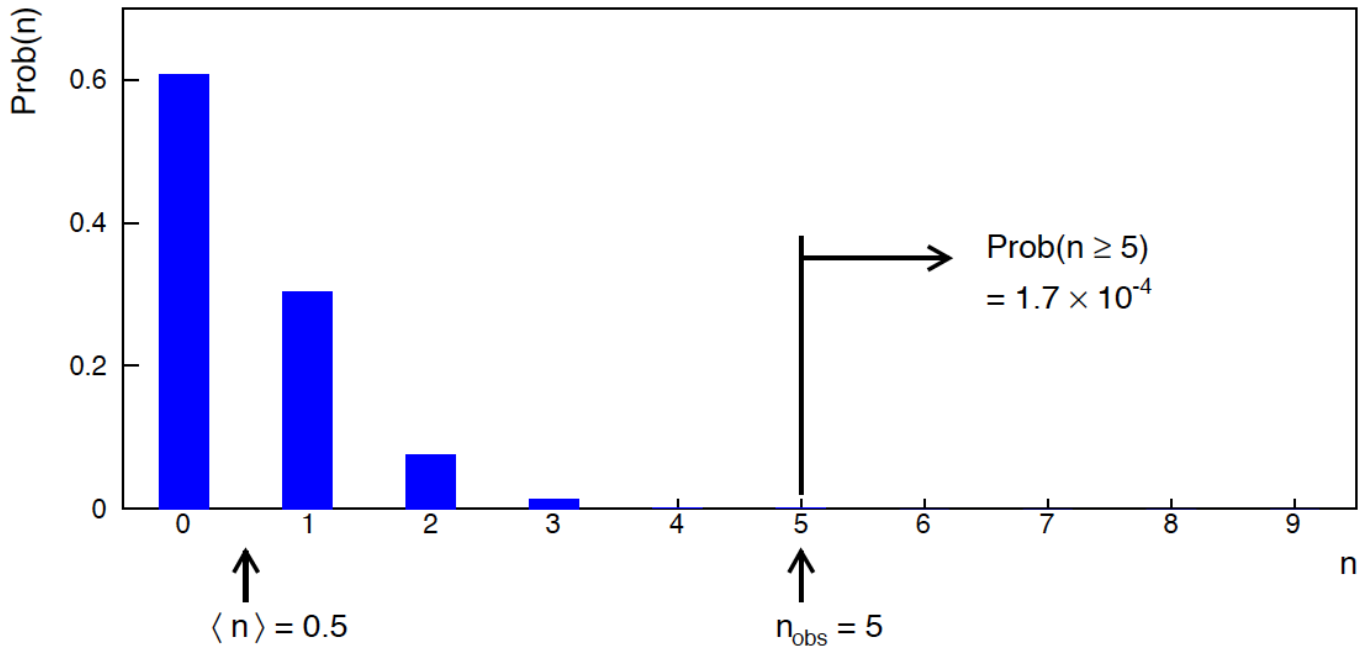
Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

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

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

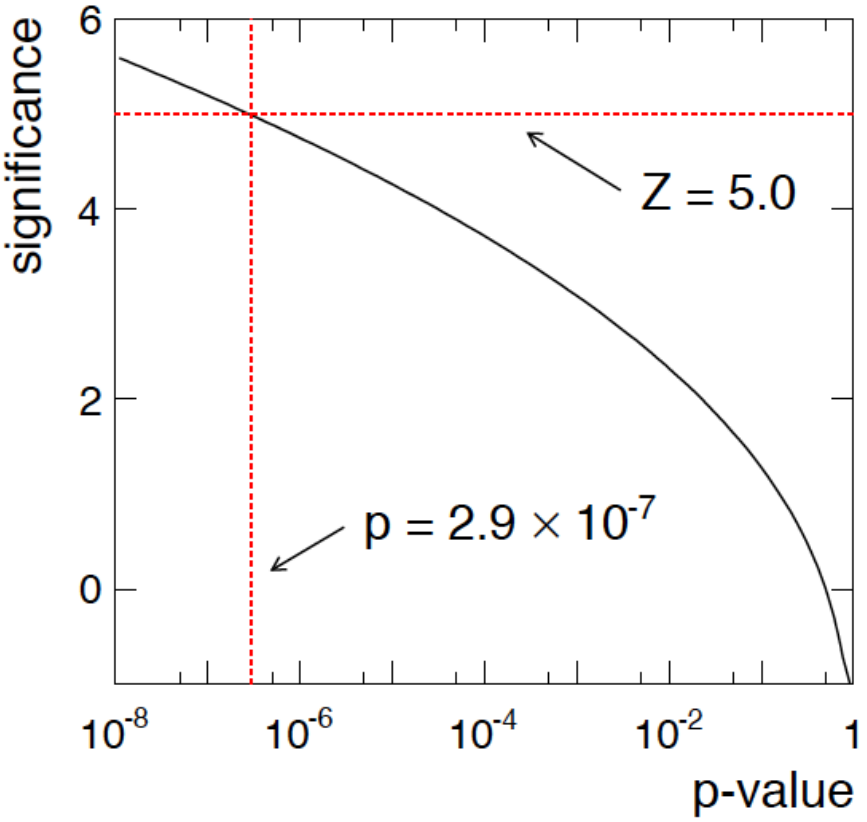
$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$



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.

Confidence intervals by inverting a test

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

> Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.
>
> If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

> set of θ values that would not be rejected in a test of size α (confidence level is $1 - \alpha$).

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

Equivalently, the parameter values in the confidence interval have p-values of at least α.

Frequentist upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p-value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at CL = $1 - \alpha$ found from

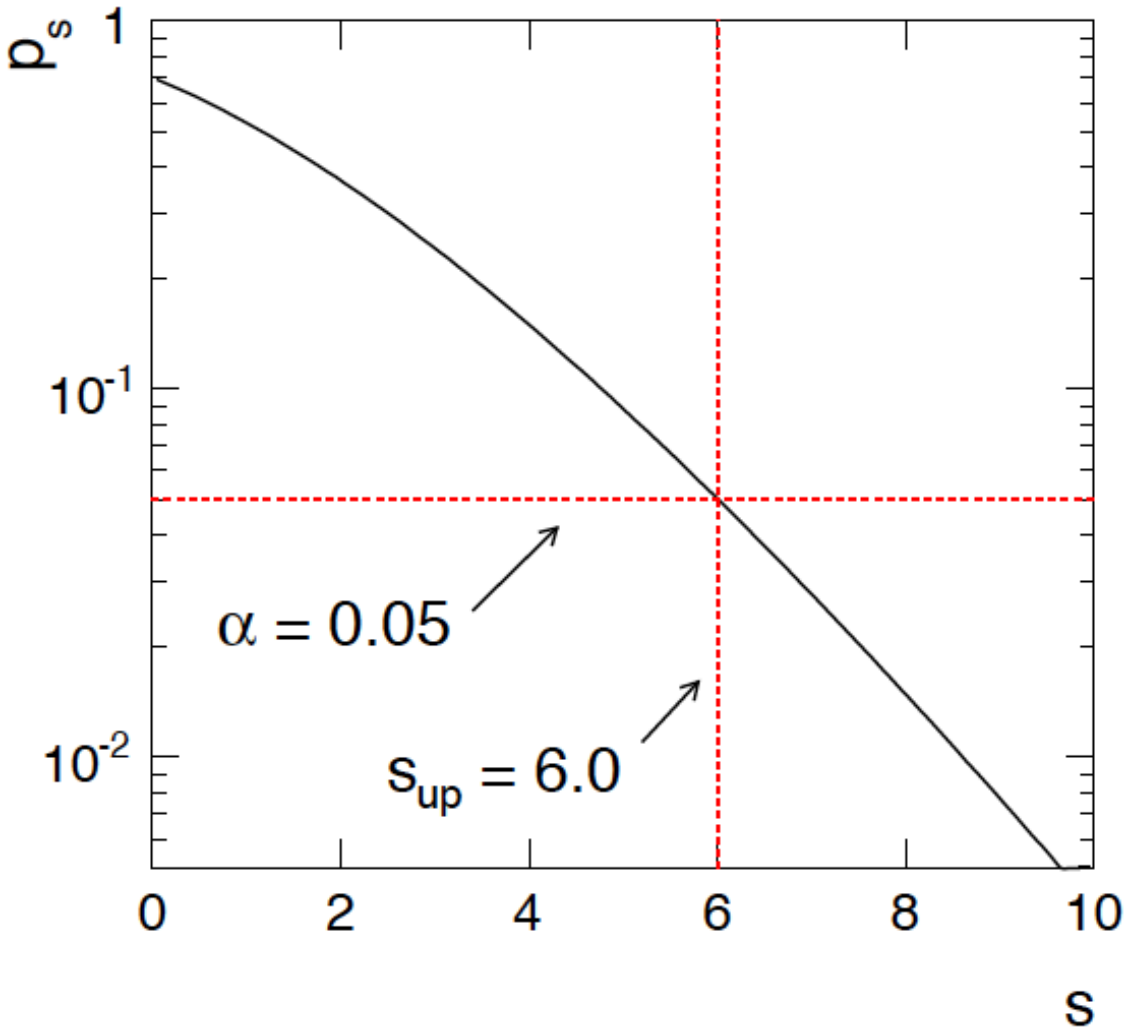
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F^{-1}_{\chi^2}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

Frequentist upper limit on Poisson parameter

Upper limit s_{up} at CL = 1 – α found from $p_s = \alpha$.



$n_{obs} = 5$,

$b = 4.5$

Frequentist treatment of nuisance parameters in a test

Suppose we test a value of θ with the profile likelihood ratio:

$$t_\theta = -2 \ln \frac{L(\theta, \hat{\hat{\nu}}(\theta))}{L(\hat{\theta}, \hat{\nu})}$$

We want a p-value of θ:

$$p_\theta = \int_{t_{\theta,\mathrm{obs}}}^{\infty} f(t_\theta|\theta, \nu)\, dt_\theta$$

Wilks' theorem says in the large sample limit (and under some additional conditions) $f(t_\theta|\theta,\nu)$ is a chi-square distribution with number of degrees of freedom equal to number of parameters of interest (number of components in θ).

Simple recipe for p-value; holds regardless of the values of the nuisance parameters!

Frequentist treatment of nuisance parameters in a test (2)

But for a finite data sample, $f(t_\theta|\theta,\nu)$ still depends on ν.

So what is the rule for saying whether we reject θ?

Exact approach is to reject θ only if $p_\theta < \alpha$ (5%) for all possible ν.

Some values of θ might not be excluded for a value of ν known to be disfavoured.

Less values of θ rejected → larger interval → higher probability to cover true value ("over-coverage").

But why do we say some values of ν are disfavoured? If this is because of other measurements (say, data y) then include y in the model:

$$P(x, y|\theta, \nu) = P_x(x|\theta, \nu)P_y(y|\nu)$$

Now ν better constrained, new interval for θ smaller.

Profile construction ("hybrid resampling")

K. Cranmer, PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics, 2008. oai:cds.cern.ch:1021125, cdsweb.cern.ch/record/1099969.

Approximate procedure is to reject θ if $p_\theta \leq \alpha$ where the p-value is computed assuming the value of the nuisance parameter that best fits the data for the specified θ (the profiled values):

$$\hat{\hat{\nu}}(\theta) = \underset{\nu}{\operatorname{argmax}}\, L(\theta, \nu)$$

The resulting confidence interval will have the correct coverage for the points $(\theta, \hat{\hat{\nu}}(\theta))$

Elsewhere it may under- or over-cover, but this is usually as good as we can do (check with MC if crucial or small sample problem).

Bayesian treatment of nuisance parameters

Conceptually straightforward: all parameters have a prior: $\pi(\theta, \nu)$

Often $\pi(\theta, \nu) = \pi_\theta(\theta)\pi_\nu(\nu)$

Often $\pi_\theta(\theta)$ "non-informative" (broad compared to likelihood).

Usually $\pi_\nu(\nu)$ "informative", reflects best available info. on ν.

Use with likelihood in Bayes' theorem:

$$p(\theta, \nu|x) \propto L(x|\theta, \nu)\pi(\theta, \nu)$$

To find $p(\theta|x)$, marginalize (integrate) over nuisance param.:

$$p(\theta|x) = \int p(\theta, \nu|x)\, d\nu$$

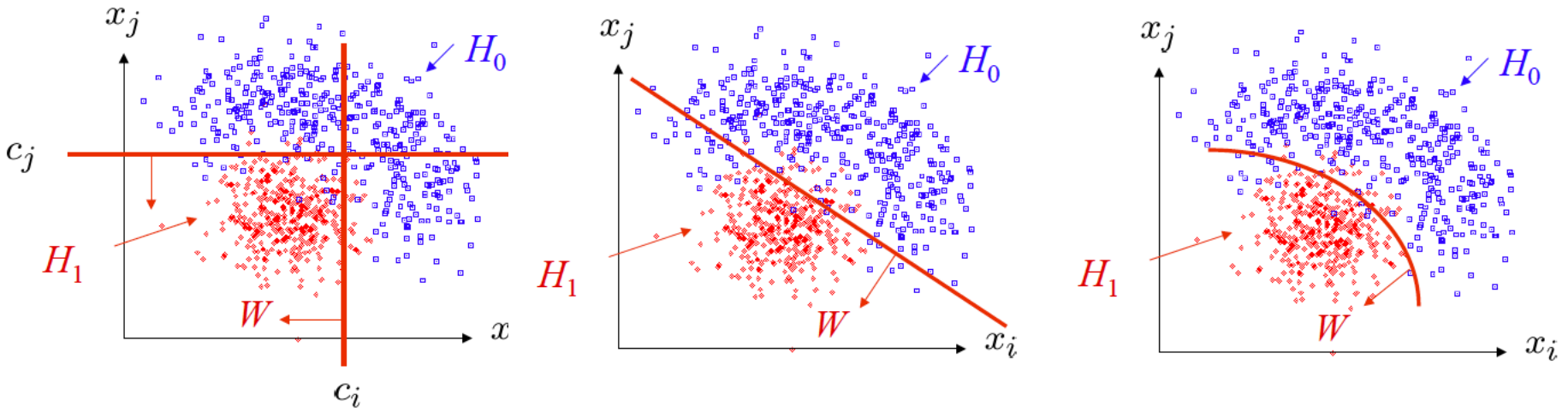
Prototype analysis in HEP

Each event yields a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons, x_2 = p_t of jet, ...

$\vec{x}$ follows some n-dimensional joint pdf, which depends on the type of event produced, i.e., signal or background.

1) What kind of decision boundary best separates the two classes?



2) What is optimal test of hypothesis that event sample contains only background?

Test statistics

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

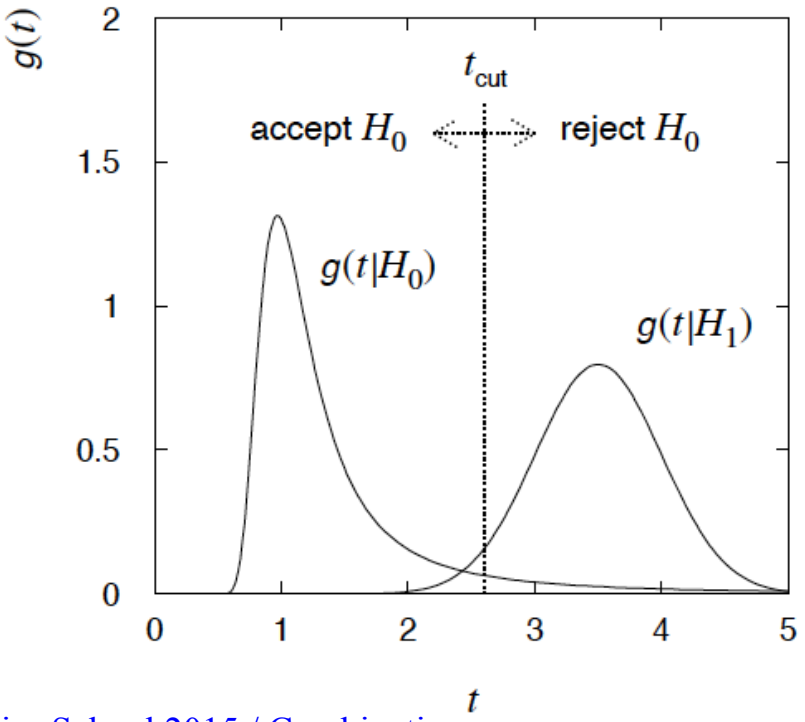
$$t(x_1, \ldots, x_n) = t_{\text{cut}}$$

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.



Test statistic based on likelihood ratio

For multivariate data $\boldsymbol{x}$, not obvious how to construct best test.

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test of H_0, (background) versus H_1, (signal) the critical region should have

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} > c$$

inside the region, and $\leq c$ outside, where c is a constant which depends on the size of the test α.

Equivalently, optimal scalar test statistic is $t(\mathbf{x}) = \dfrac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$

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

Ingredients for a frequentist test

In general to carry out a test we need to know the distribution of the test statistic $t(x)$, and this means we need the full model $P(x|H)$.

Often one can construct a test statistic whose distribution approaches a well-defined form (almost) independent of the distribution of the data, e.g., likelihood ratio to test a value of θ:

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

In the large sample limit t_θ follows a chi-square distribution with number of degrees of freedom = number of components in θ (Wilks' theorem).

So here one doesn't need the full model $P(x|\theta)$, only the observed value of t_θ.

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 $\alpha = \min\left[1, \dfrac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)}\right]$

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

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.