Statistical Data Analysis: Lecture 13

1 Probability, Bayes' theorem

2 Random variables and probability densities

3 Expectation values, error propagation

4 Catalogue of pdfs

5 The Monte Carlo method

6 Statistical tests: general concepts

7 Test statistics, multivariate methods

8 Goodness-of-fit tests

9 Parameter estimation, maximum likelihood

10 More maximum likelihood

11 Method of least squares

12 Interval estimation, setting limits

→ 13 Nuisance parameters, systematic uncertainties

14 Examples of Bayesian approach

Statistical vs. systematic errors

Statistical errors:

How much would the result fluctuate upon repetition of the measurement?

Implies some set of assumptions to define probability of outcome of the measurement.

Systematic errors:

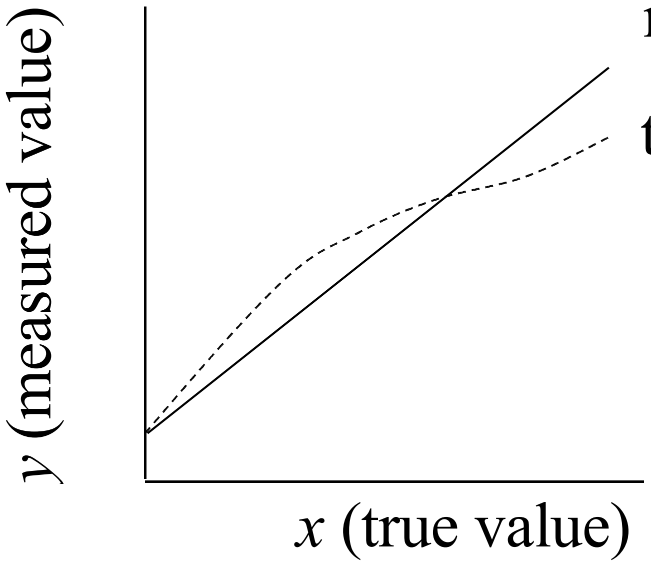
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modelling of measurement apparatus.

The sources of error do not vary upon repetition of the measurement. Often result from uncertain value of, e.g., calibration constants, efficiencies, etc.

Systematic errors and nuisance parameters

Response of measurement apparatus is never modelled perfectly:



model: $y = \alpha + \beta x$

truth: $y = \alpha + \beta x + \gamma x^2 + \varepsilon x^3 + \ldots$

Model can be made to approximate better the truth by including more free parameters.

systematic uncertainty ↔ nuisance parameters

Nuisance parameters

Suppose the outcome of the experiment is some set of data values x (here shorthand for e.g. $x_1, ..., x_n$).

We want to determine a parameter θ, (could be a vector of parameters $\theta_1, ..., \theta_n$).

The probability law for the data x depends on θ :

$L(x| \theta)$ (the likelihood function)

E.g. maximize L to find estimator $\hat{\theta}$.

Now suppose, however, that the vector of parameters:

contains some that are of interest, $\psi_1, \ldots, \psi_n$

and others that are not of interest: $\lambda_1, \ldots, \lambda_m$.

Symbolically: $\theta = (\psi, \lambda)$

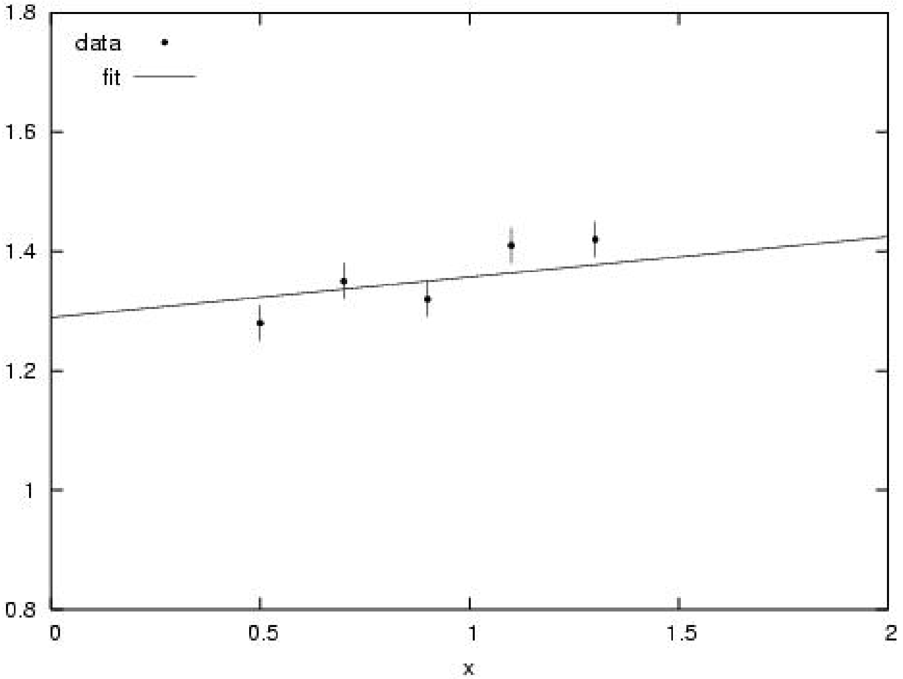
The $\lambda_1, \ldots, \lambda_m$ are called nuisance parameters.

Example #1: fitting a straight line

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

Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

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



assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).

Case #1: θ_1 known a priori

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

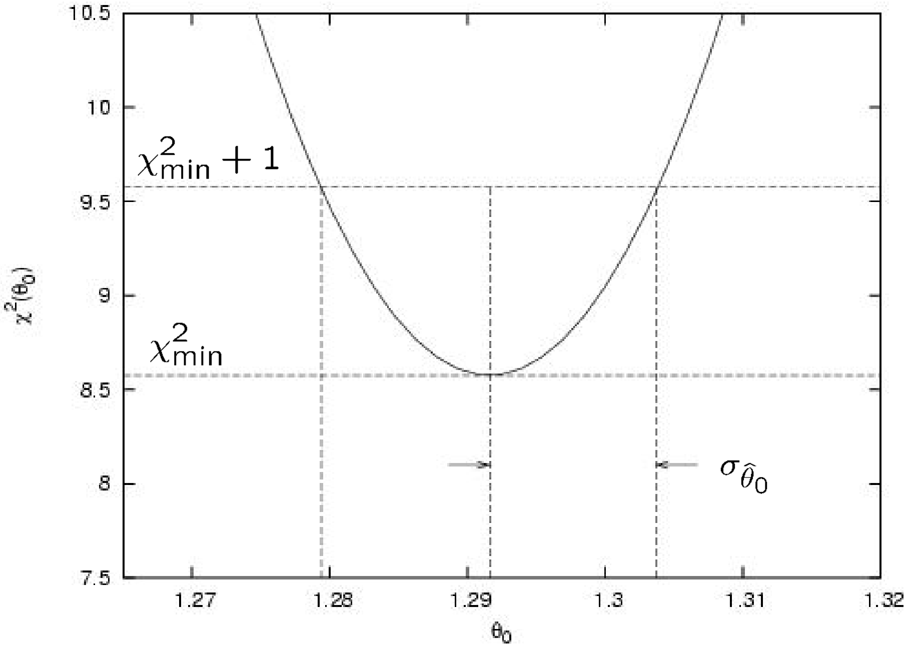
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i, ML same as LS

Minimize χ^2 → estimator $\hat{\theta}_0$.

Come up one unit from $\chi^2_{\min}$

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



Case #2: both θ_0 and θ_1 unknown

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

Standard deviations from

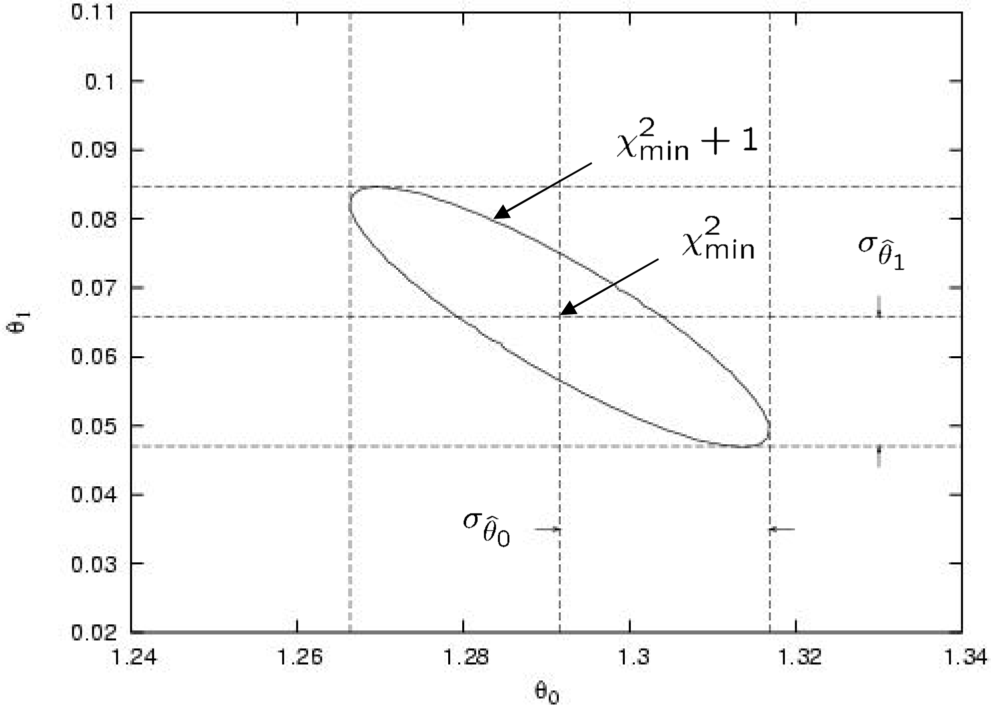
tangent lines to contour

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

Correlation between

$\hat{\theta}_0, \; \hat{\theta}_1$ causes errors

to increase.

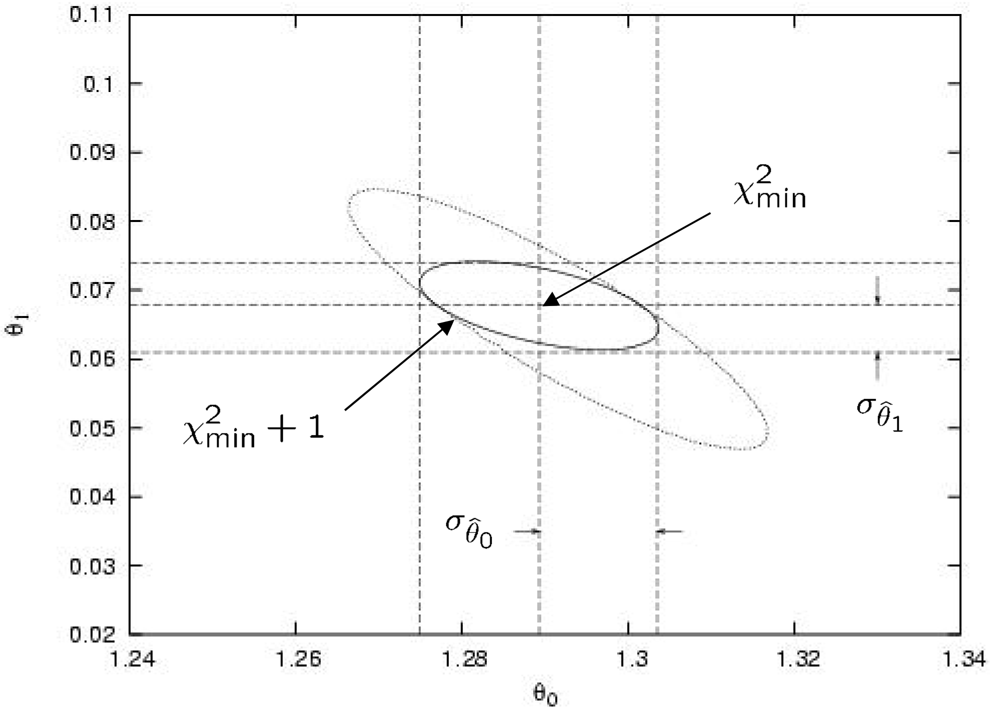


Case #3: we have a measurement t_1 of θ_1

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

The information on θ_1

improves accuracy of $\hat{\theta}_0$.



The profile likelihood

The ‘tangent plane’ method is a special case of using the

profile likelihood: $L'(\theta_0) = L(\theta_0, \hat{\hat{\theta}}_1)$.

$\hat{\hat{\theta}}_1$ is found by maximizing $L\,(\theta_0,\,\theta_1)$ for each θ_0.

Equivalently use $\chi^{2\prime}(\theta_0) = \chi^2(\theta_0, \hat{\hat{\theta}}_1)$.

The interval obtained from $\chi^{2\prime}(\theta_0) = \chi^{2\prime}_{\mathrm{min}} + 1$ is the same as what is obtained from the tangents to $\chi^2(\theta_0, \theta_1) = \chi^2_{\mathrm{min}} + 1$.

Well known in HEP as the ‘MINOS’ method in MINUIT.

Profile likelihood is one of several ‘pseudo-likelihoods’ used in problems with nuisance parameters. See e.g. talk by Rolke at PHYSTAT05.

The Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value θ.

Interpret probability of θ as 'degree of belief' (subjective).

Need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Our experiment has data x, → likelihood function $L(x|\theta)$.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

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

Posterior pdf $p(\theta\,|\,x)$ contains all our knowledge about θ.

Case #4: Bayesian method

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

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

reflects 'prior ignorance', 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) $p(\theta_0, \theta_1 \mid x)$ to find $p(\theta_0 \mid x)$:

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

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

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

Ability to marginalize over nuisance parameters is an important feature of Bayesian statistics.

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.

Google for ‘MCMC’, ‘Metropolis’, ‘Bayesian computation’, ...

MCMC generates correlated sequence of random numbers:

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

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

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

5) If $u \le \alpha$, $\vec{\theta}_1 = \vec{\theta}$, ← 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.

Metropolis-Hastings caveats

Actually one can only prove that the sequence of points follows the desired pdf in the limit where it runs forever.

There may be a "burn-in" period where the sequence does not initially follow $p(\vec{\theta})$.

Unfortunately there are few useful theorems to tell us when the sequence has converged.

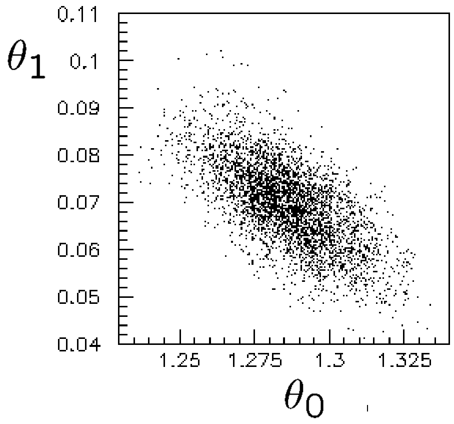
Look at trace plots, autocorrelation.

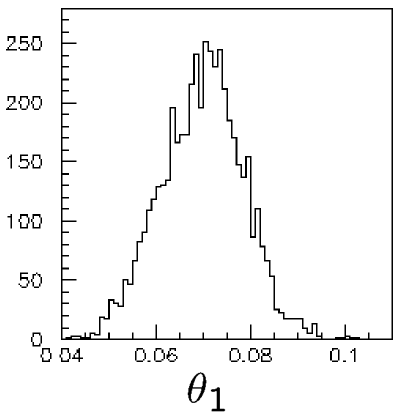
Check result with different proposal density.

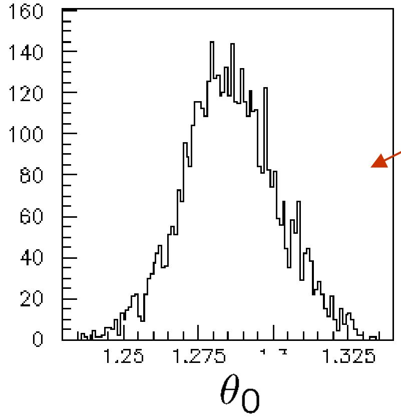
If you think it's converged, try it again starting from 10 different initial points and see if you find same result.

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?)

Case #5: Bayesian method with vague prior

Suppose we don't have a previous measurement of θ_1 but rather some vague information, e.g., a theorist tells us:

$\theta_1 \geq 0$ (essentially certain);

θ_1 should have order of magnitude less than 0.1 'or so'.

Under pressure, the theorist sketches the following prior:

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

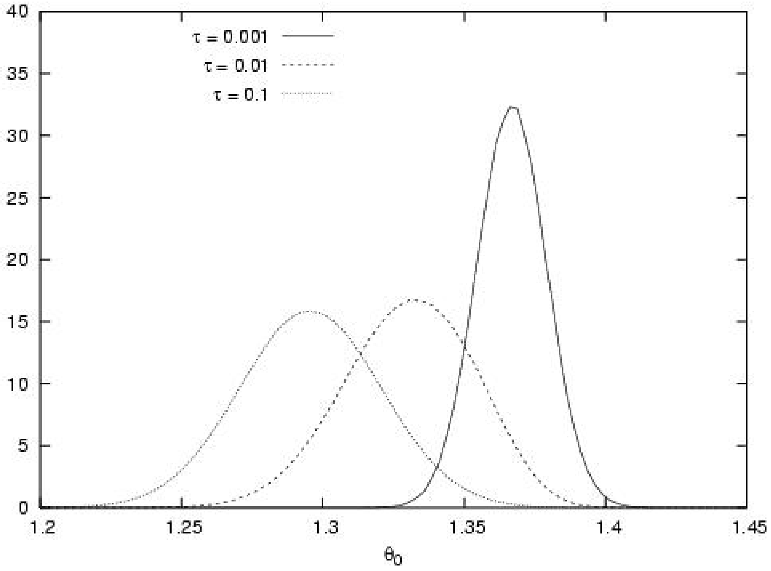
From this we will obtain posterior probabilities for θ_0 (next slide).

We do not need to get the theorist to 'commit' to this prior; final result has 'if-then' character.

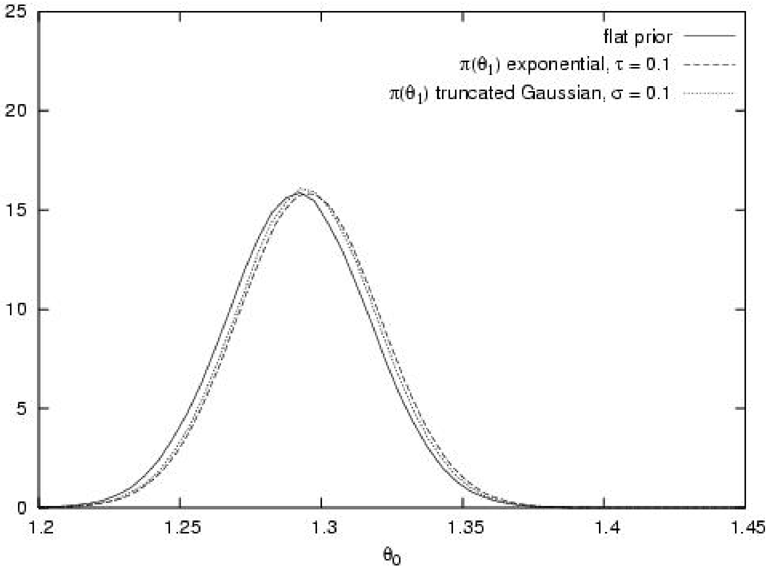
Sensitivity to prior

Vary $\pi(\theta)$ to explore how extreme your prior beliefs would have to be to justify various conclusions (sensitivity analysis).

Try exponential with different mean values...



Try different functional forms...



Example #2: Poisson data with background

Count n events, e.g., in fixed time or integrated luminosity.

s = expected number of signal events

b = expected number of background events

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

Sometimes b known, other times it is in some way uncertain.

Goal: measure or place limits on s, taking into consideration the uncertainty in b.

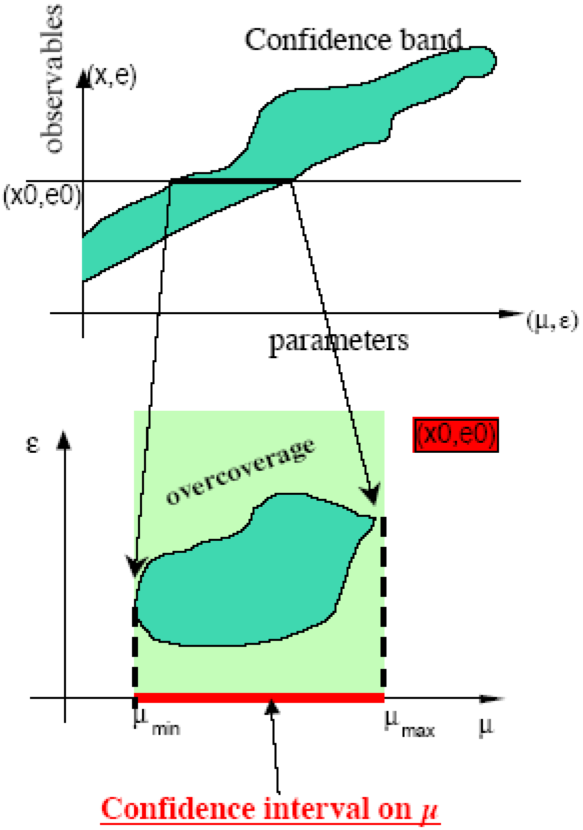
Classical procedure with measured background

Suppose we have a measurement of b, e.g., $b_{\text{meas}} \sim N(b, \sigma_b)$

So the data are really: n events and the value $b_{\text{meas.}}$

In principle the confidence interval recipe can be generalized to two measurements and two parameters.

Difficult and not usually attempted, but see e.g. talks by K. Cranmer at PHYSTAT03, G. Punzi at PHYSTAT05.



G. Punzi, PHYSTAT05

Bayesian limits with uncertainty on *b*

Uncertainty on *b* goes into the prior, e.g.,

$$\begin{aligned}\pi(s,b) &= \pi_s(s)\pi_b(b) \quad \text{(or include correlations as appropriate)} \\ \pi_s(s) &= \text{const}, \quad \sim 1/s, \ldots \quad ? \\ \pi_b(b) &= \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad \text{(or whatever)}\end{aligned}$$

Put this into Bayes' theorem,

$$p(s,b|n) \propto L(n|s,b)\pi(s,b)$$

Marginalize over *b*, then use $p(s|n)$ to find intervals for *s* with any desired probability content.

Controversial part here is prior for signal $\pi_s(s)$ (treatment of nuisance parameters is easy).

Cousins-Highland method

Regard b as ‘random’, characterized by pdf $\pi(b)$.

Makes sense in Bayesian approach, but in frequentist model b is constant (although unknown).

A measurement b_{meas} is random but this is not the mean number of background events, rather, b is.

Compute anyway $P(n;s) = \int P(n;s,b)\pi_b(b)\,db$

This would be the probability for n if Nature were to generate a new value of b upon repetition of the experiment with $\pi_b(b)$.

Now e.g. use this $P(n;s)$ in the classical recipe for upper limit at CL = 1 – β: $\beta = P(n \leq n_{\mathsf{obs}}; s_{\mathsf{up}})$

Result has hybrid Bayesian/frequentist character.

‘Integrated likelihoods’

Consider again signal s and background b, suppose we have uncertainty in b characterized by a prior pdf $\pi_b(b)$.

Define integrated likelihood as $L'(s) = \int L(s,b)\pi_b(b)\,db$, also called modified profile likelihood, in any case not a real likelihood.

Now use this to construct likelihood ratio test and invert to obtain confidence intervals.

Feldman-Cousins & Cousins-Highland (FHC^2), see e.g. J. Conrad et al., Phys. Rev. D67 (2003) 012002 and Conrad/Tegenfeldt PHYSTAT05 talk.

Calculators available (Conrad, Tegenfeldt, Barlow).

Interval from inverting profile LR test

Suppose we have a measurement b_{meas} of b.

Build the likelihood ratio test with profile likelihood:

$$l(s) = \frac{L(n, b_{\mathsf{meas}}|s, \hat{\hat{b}})}{L(n, b_{\mathsf{meas}}|\hat{s}, \hat{b})}$$

and use this to construct confidence intervals.

See PHYSTAT05 talks by Cranmer, Feldman, Cousins, Reid.

Wrapping up lecture 13

We've seen some main ideas about systematic errors,

uncertainties in result arising from model assumptions;

can be quantified by assigning corresponding uncertainties to additional (nuisance) parameters.

Different ways to quantify systematics

Bayesian approach in many ways most natural;

marginalize over nuisance parameters;

important tool: MCMC

Frequentist methods rely on a hypothetical sample space for often non-repeatable phenomena