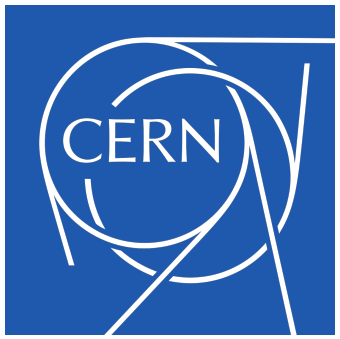
Statistics for Particle Physicists
Lecture 4



Academic Training Lectures
CERN / Zoom
21-24 June 2021

https://indico.cern.ch/event/1040096/

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

Outline

Lecture 1: Introduction, probability, parameter estimation

Lecture 2: Hypothesis tests, limits

Lecture 3: Systematic uncertainties, experimental sensitivity

→ Lecture 4: Bayesian methods, Student's *t* regression

Example: fitting a straight line

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

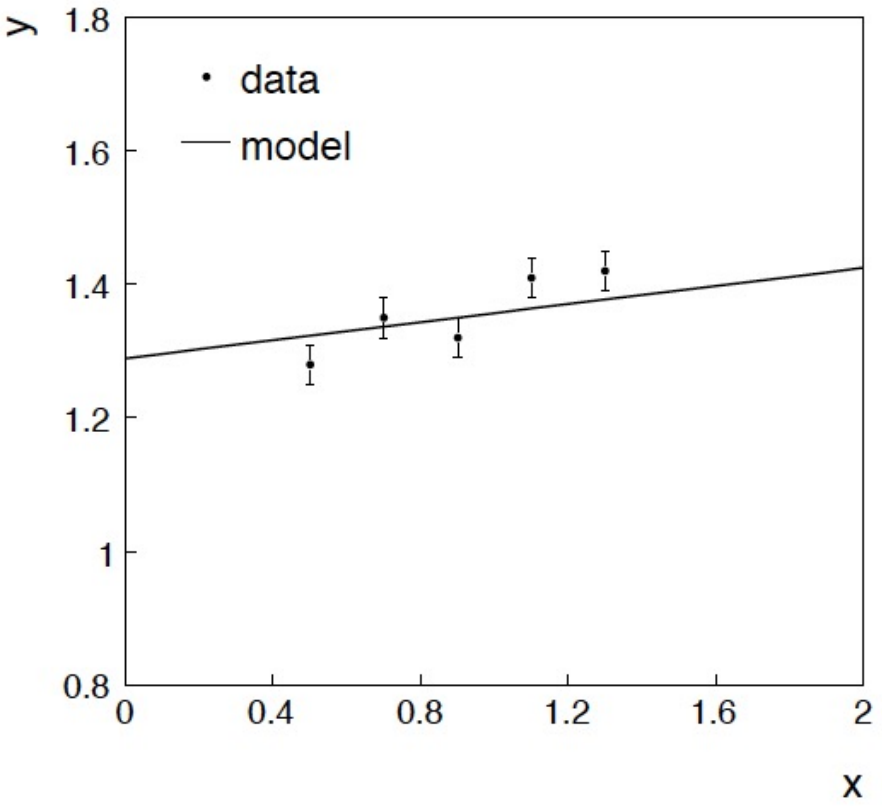
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

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

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

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

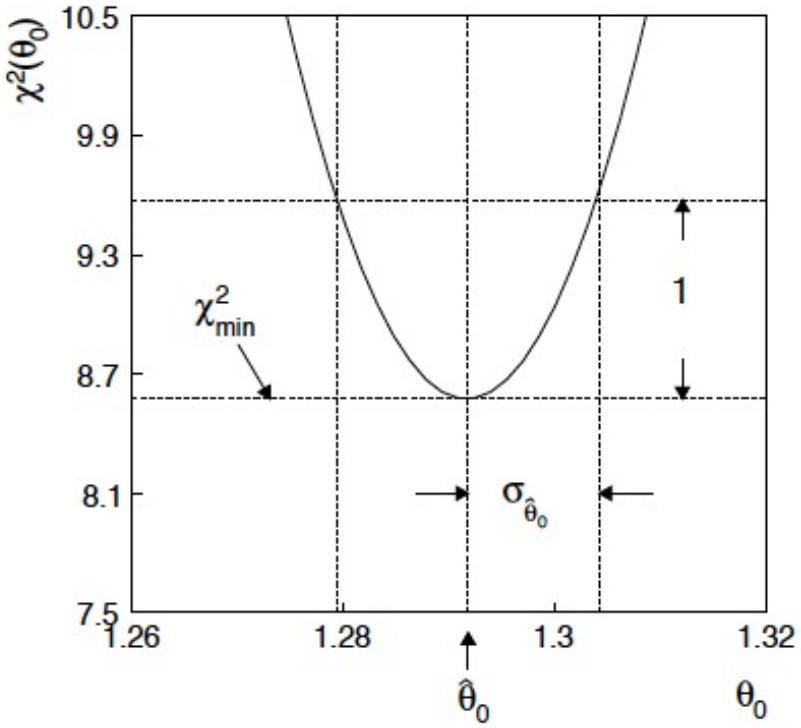
$$\chi^2(\theta_0) = -2\ln L(\theta_0) + \text{const} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2} .$$

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



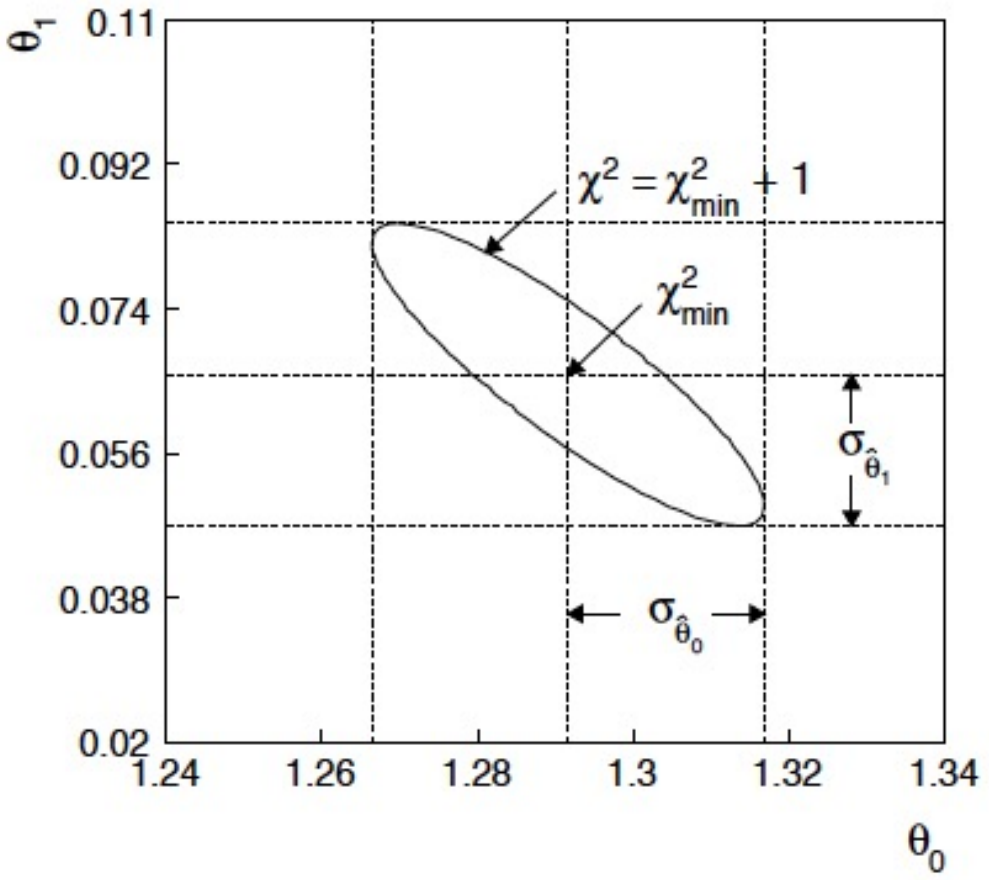
ML (or LS) fit of θ_0 and θ_1

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

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.

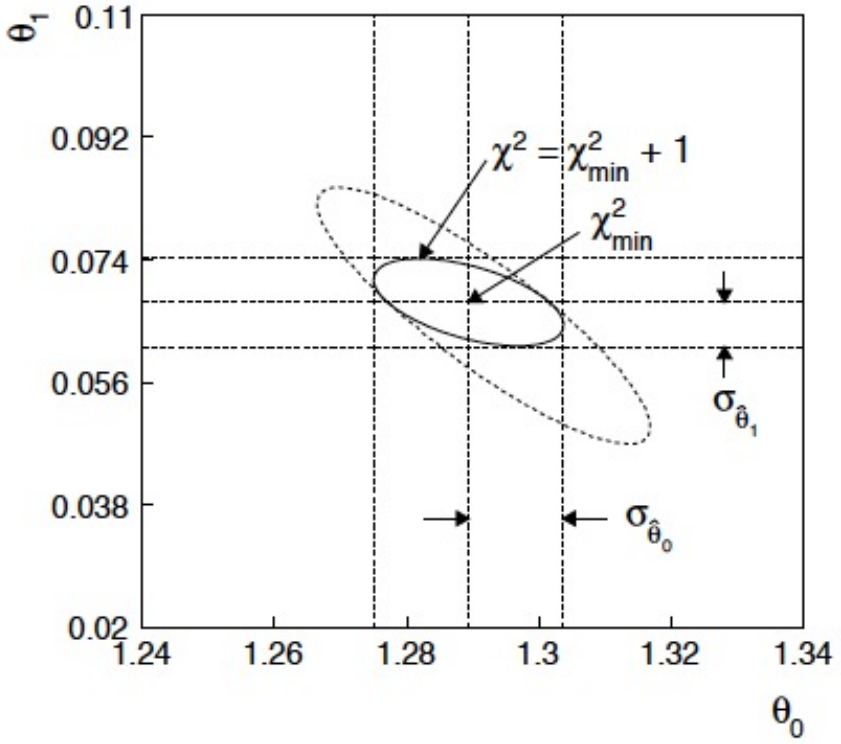


If we have a measurement $t_1 \sim \text{Gauss}\,(\theta_1, \sigma_{t_1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_{t_1}^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i - \mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right]$$

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

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



Profiling

The $\ln L = \ln L_{\max} - ½$ contour in the (θ_0, θ_1) plane is a confidence region at CL = 39.3%.

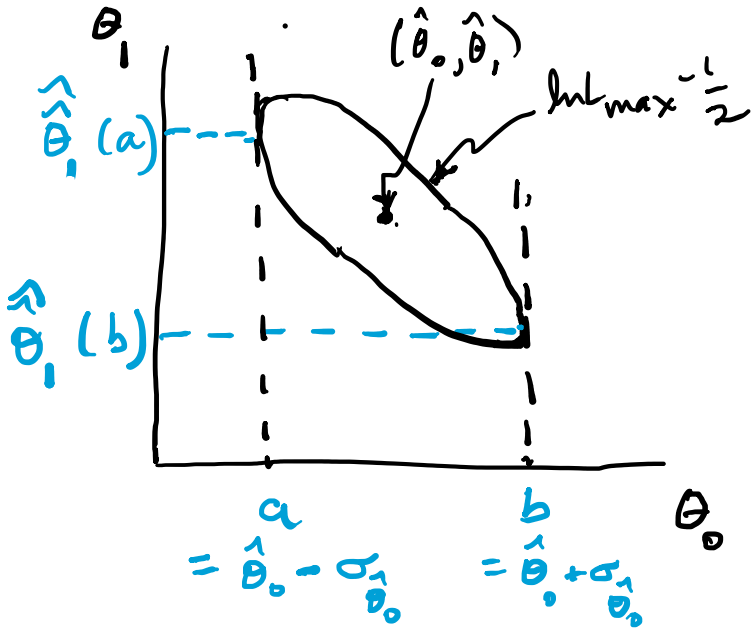
Furthermore if one wants to know only about, say, θ_0, then the interval in θ_0 corresponding to $\ln L = \ln L_{\max} - ½$ is a confidence interval at CL = 68.3% (i.e., ± 1 std. dev.).

I.e., form the interval for θ_0 using

$$\ln L(\theta_0, \hat{\hat{\theta}}_1(\theta_0)) = \ln L_{\max} - 1/2$$

where θ_1 is replaced by its "profiled" value

$$\hat{\hat{\theta}}_1(\theta_0) = \underset{\theta_1}{\operatorname{argmax}}\, L(\theta_0, \theta_1)$$



Reminder of 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, $\rightarrow$ likelihood $L(x|\theta)$.

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.

Bayesian approach: $y_i \sim \text{Gauss}(\mu(x_i;\theta_0,\theta_1), \sigma_i)$

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

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

← suppose knowledge of θ_0 has no influence on knowledge of θ_1

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

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

$$\pi_1(\theta_1) = p(\theta_1|t_1) \propto p(t_1|\theta_1)\pi_{\text{Ur}}(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_t^2} \times \text{const.}$$

prior after t_1, before $\boldsymbol{y}$

Ur = "primordial" prior

Likelihood for control measurement t_1

Bayesian example: $y_i \sim \text{Gauss}(\mu(x_i;\theta_0,\theta_1), \sigma_i)$

Putting the ingredients 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 $\propto$ likelihood $\times$ prior

Note here the likelihood only reflects the measurements $\boldsymbol{y}$.

The information from the control measurement t_1 has been put into the prior for θ_1.

We would get the same result using the likelihood $P(\boldsymbol{y},t|\theta_0,\theta_1)$ and the constant "Ur-prior" for θ_1.

Marginalizing the posterior pdf

We then integrate (marginalize) $p(\theta_0, \theta_1 | \boldsymbol{y})$ to find $p(\theta_0 | \boldsymbol{y})$:

$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

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

$$p(\theta_0|\mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0^2}}$$

$$\hat{\theta}_0 = \text{same as MLE}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \qquad \text{(same as for MLE)}$$

For this example, numbers come out same as in frequentist approach, but interpretation different.

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 $\boldsymbol{\theta}$ but look only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\boldsymbol{\theta})$, generate a sequence of points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots$

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

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

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

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

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).

Still works if $p(\boldsymbol{\theta})$ is known only as a proportionality, which is usually what we have from Bayes' theorem: $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

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

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

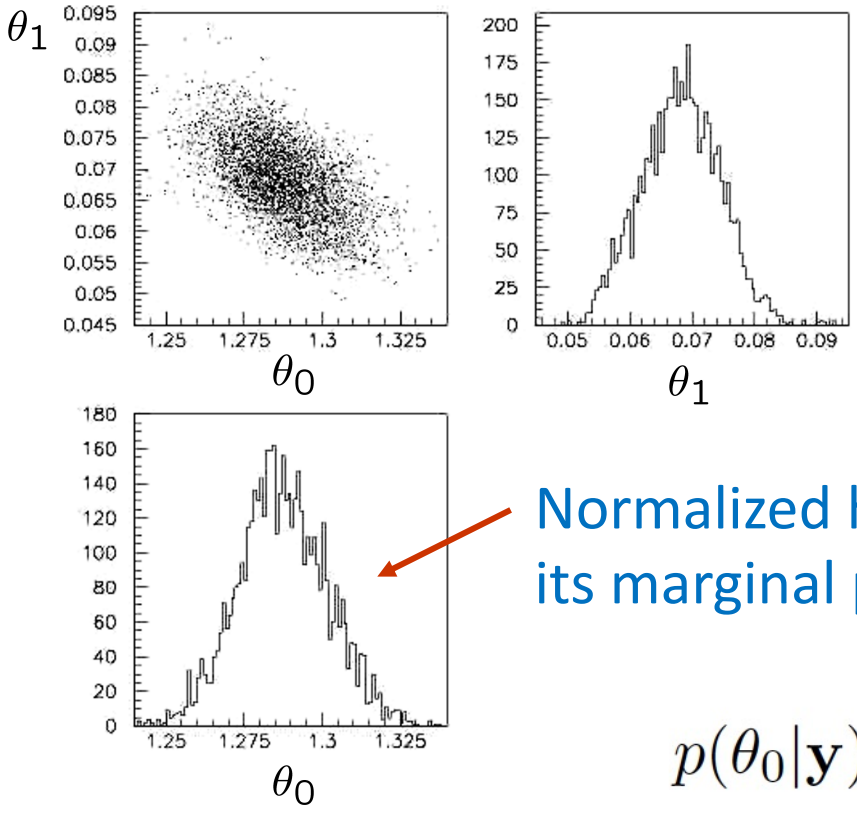
I.e. if the proposed step is to a point of higher $p(\boldsymbol{\theta})$, take it;

if not, only take the step with probability $p(\boldsymbol{\theta})/p(\boldsymbol{\theta}_0)$.

If proposed step rejected, repeat the current point.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Normalized histogram of θ_0 gives its marginal posterior pdf:

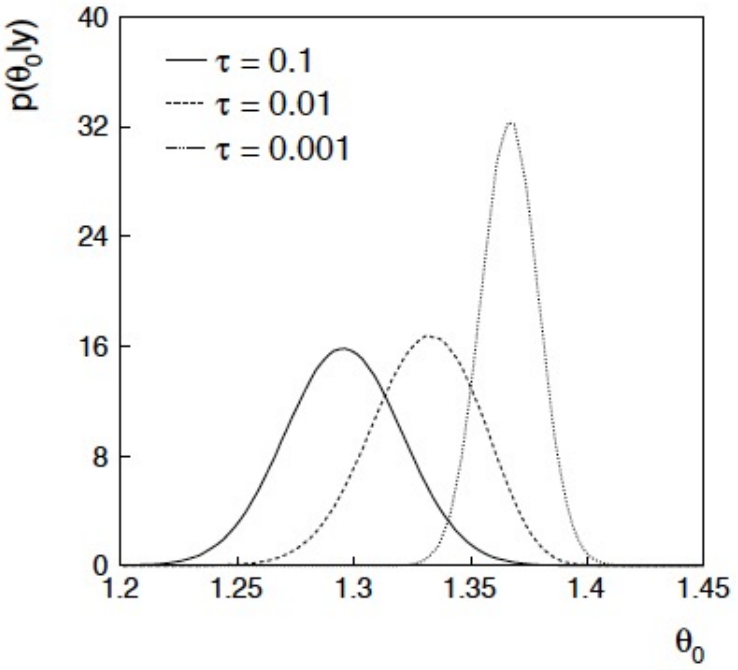
$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, an "expert" 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.

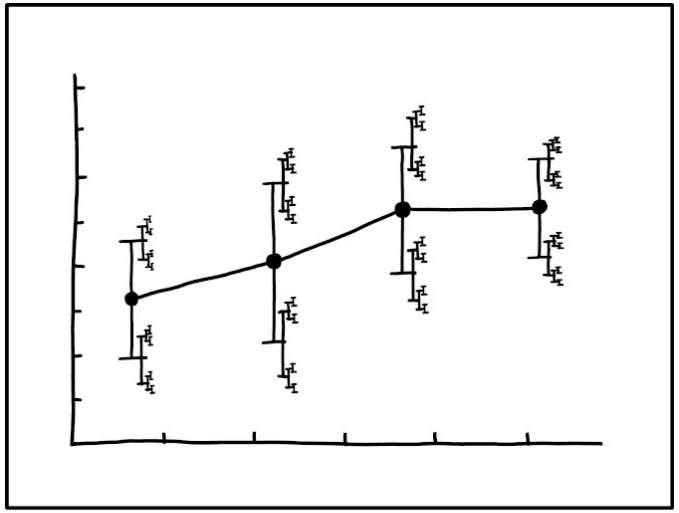
Look also at result from variety of priors.

Motivation, disclosure, etc.

For several years I've been pushing the idea that the uncertainties on estimated systematic errors ("errors on errors") should play a role in HEP analyses, particularly for combinations.

Details in: G. Cowan, *Statistical Models with Uncertain Error Parameters*, Eur. Phys. J. C (2019) 79:133, arXiv:1809.05778

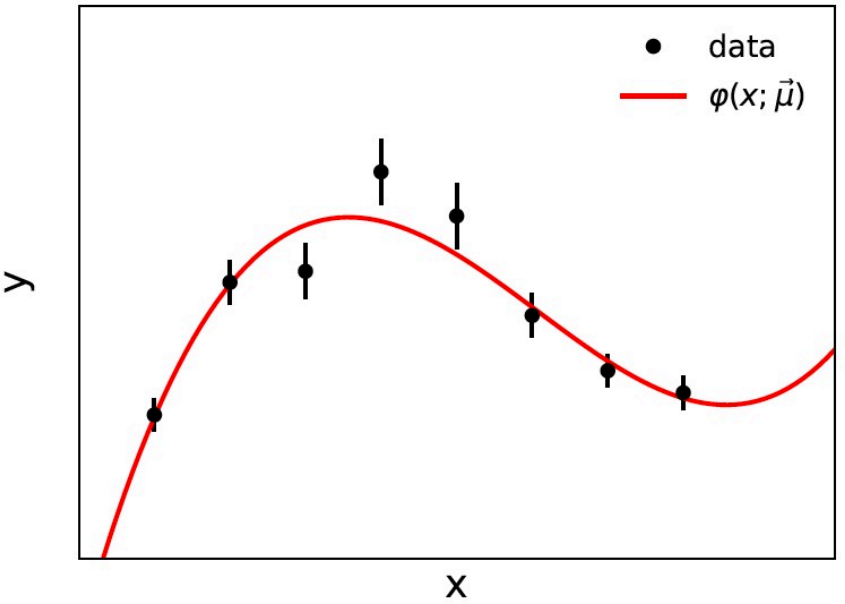
https://xkcd.com/2110/



It turns out that models that use errors on errors have qualitatively new, interesting, desirable features:

- Sensitivity to outliers reduced.
- Confidence intervals sensitive to goodness of fit.

Prototype example: curve fitting, averages



Suppose independent
$y_i \sim$ Gauss, $i = 1,\ldots,N$, with

$$E[y_i] = \varphi(x_i; \boldsymbol{\mu})$$

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

$\boldsymbol{\mu}$ are the parameters in the fit function $\varphi(x;\boldsymbol{\mu})$.

If we take the σ_i as known, we have the usual log-likelihood

$$\ln L(\boldsymbol{\mu}) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - \varphi(x_i; \boldsymbol{\mu}))^2}{\sigma_i^2}$$

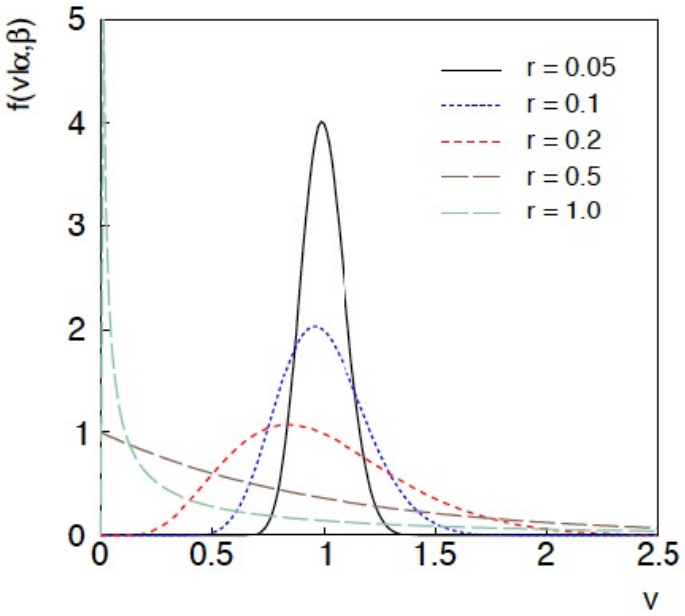
which leads to the Least Squares estimators for $\boldsymbol{\mu}$.

Model with uncertain σ_i^2

If the σ_i^2 are uncertain, we can take them as adjustable parameters.

The estimated variances $v_i = s_i^2$ are modeled as gamma distributed.

The likelihood becomes



$$L(\boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(y_i - \varphi(x_i;\boldsymbol{\mu}))^2/2\sigma_i^2} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} v_i^{\alpha_i - 1} e^{-\beta_i v_i}$$

Want $\quad E[v_i] = \sigma_i^2 \qquad \frac{\sigma_{s_i}}{E[s_i]} \approx r_i \qquad (s_i = \sqrt{v_i})$

$\rightarrow \qquad \alpha_i = \frac{1}{4r_i^2} \qquad \beta_i = \frac{\alpha_i}{\sigma_i^2}$

Profile log-likelihood

One can profile over the σ_i^2 in close form.

The log-profile-likelihood is

$$\ln L'(\boldsymbol{\mu}) = \ln L(\boldsymbol{\mu}, \widehat{\widehat{\boldsymbol{\sigma}^2}}) = -\frac{1}{2}\sum_{i=1}^{N}\left(1 + \frac{1}{2r_i^2}\right)\ln\left[1 + 2r_i^2\frac{(y_i - \varphi(x_i;\boldsymbol{\mu}))^2}{v_i}\right]$$

Quadratic terms replace by sum of logs.

Equivalent to replacing Gauss pdf for y_i by Student's t, $\nu_{\rm dof} = 1/2r_i^2$

Confidence interval for $\boldsymbol{\mu}$ becomes sensitive to goodness-of-fit (increases if data internally inconsistent).

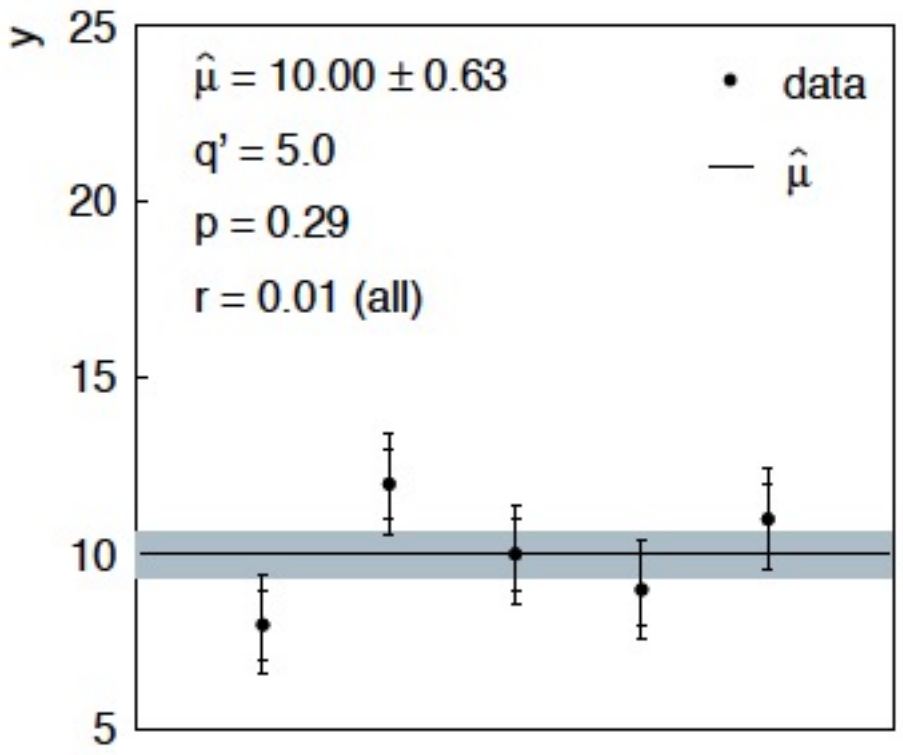
Fitted curve less sensitive to outliers.

Simple program for Student's t average: stave.py

http://www.pp.rhul.ac.uk/~cowan/stat/stave/

Sensitivity of average to outliers

Suppose we average 5 values, $y = 8, 9, 10, 11, 12$, all with stat. and sys. errors of 1.0, and suppose negligible error on error (here take $r = 0.01$ for all).

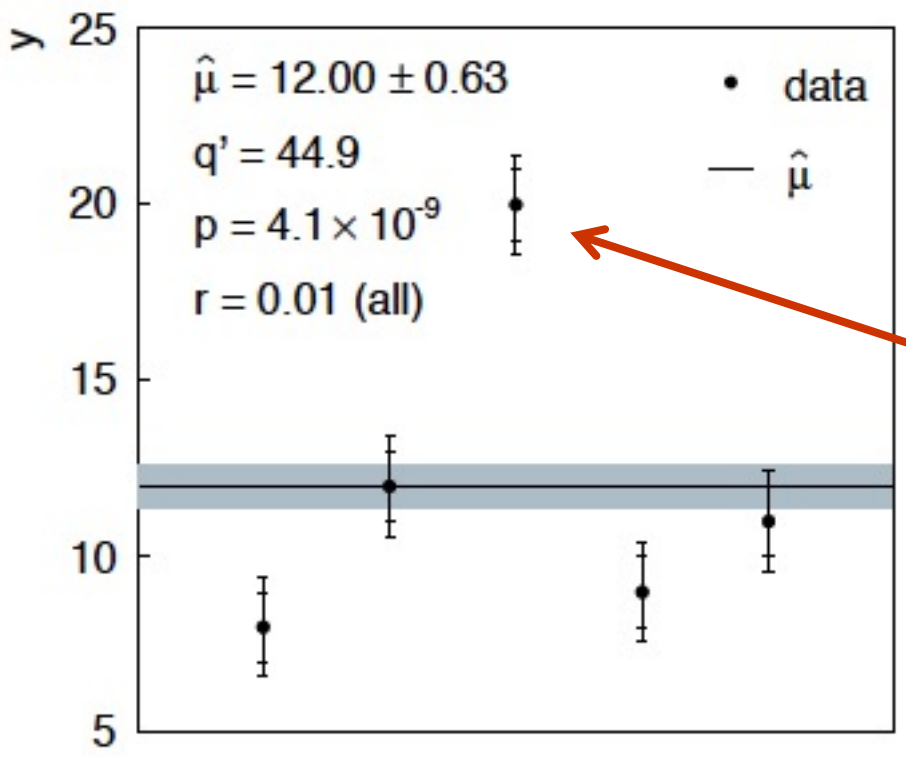


inner error bars
$= \sigma_{y,i}$

outer error bars
$= (\sigma_{y,i}^2 + \sigma_{u,i}^2)^{1/2}$

Sensitivity of average to outliers (2)

Now suppose the measurement at 10 had come out at 20:

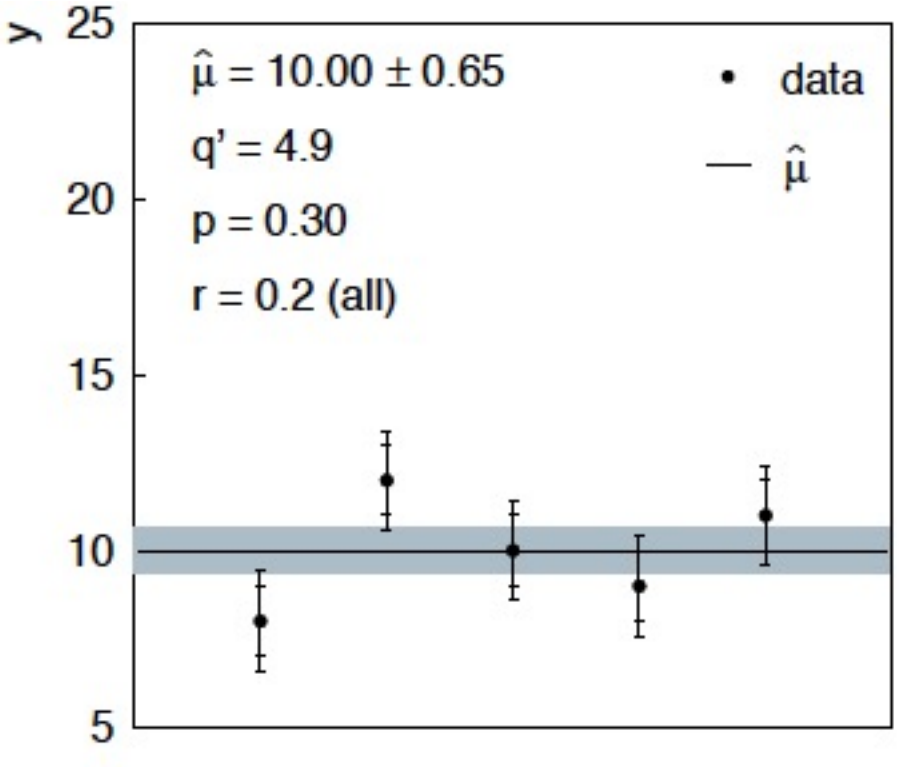


"outlier"

Estimate pulled up to 12.0, size of confidence interval ~unchanged (would be exactly unchanged with $r \to 0$).

Average with all $r = 0.2$

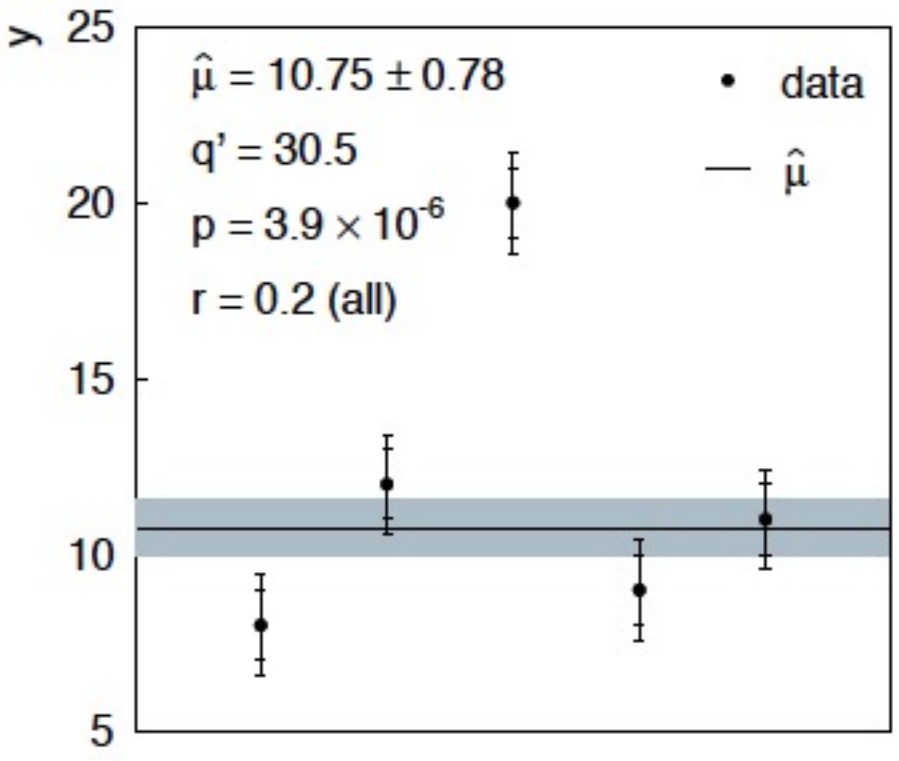
If we assign to each measurement r = 0.2,



Estimate still at 10.00, size of interval moves 0.63 → 0.65

Average with all $r = 0.2$ with outlier

Same now with the outlier (middle measurement $10 \rightarrow 20$)



Estimate $\rightarrow 10.75$ (outlier pulls much less).

Half-size of interval $\rightarrow 0.78$ (inflated because of bad g.o.f.).

Discussion

Gamma variance model gives confidence intervals that increase in size when the data are internally inconsistent, and gives decreased sensitivity to outliers (known property of Student's t based regression).

Equivalence with Student's t model, $\nu = 1/2r^2$ degrees of freedom.

Simple profile likelihood – quadratic terms replaced by logarithmic:

$$\frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2} \quad \rightarrow \quad \left(1 + \frac{1}{2r_i^2}\right) \ln\left[1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right]$$

Discussion (2)

Method assumes that meaningful r_i values can be assigned and there is enough "expert knowledge" is available to do so.

I.e. best if the experts publish some information on the reliability of their reported systematics.

Could the public likelihood standard include at least the possibility to include this information?

Finally

Four lectures only enough for a brief introduction to:

- Parameter estimation, maximum likelihood
- Hypothesis tests, *p*-values
- Limits (confidence intervals/regions)
- Systematics (nuisance parameters)
- Asymptotics (Wilks' theorem)
- Bayesian parameter estimation
- Student's *t* regression, gamma variance model

Final thought: once the basic formalism is fixed, most of the work focuses on writing down the likelihood, e.g., $P(\boldsymbol{x}|\boldsymbol{\theta})$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches).

Extra slides