Recommendations for Statistical Approach

Heavy Flavor Workshop
Zoom, 17 Sep 2020

https://indico.cern.ch/event/947843/

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

Outline

General comments on statistical approach

- Maximum Likelihood, Least squares
- Nuisance parameters
- Confidence intervals

Treatment of systematics

- Profile Likelihood
- Marginal Likelihood

"Errors on Errors" à la

- GDC, Eur. Phys. J. C (2019) 79:133, arXiv:1809.05778

Least Squares ← Maximum Likelihood

HFLAV averaging based on Least Squares, which follows from method of Maximum Likelihood e.g. if independent measured $y_i \sim$ Gaussian($f(x_i;\boldsymbol{\theta})$, σ_i)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - f(x_i;\boldsymbol{\theta}))^2/2\sigma_i^2}$$

$$-2\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - f(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2} + \text{const.}$$

Warning – tails of Gaussian fall off very fast;
"outliers" have strong influence on parameter estimates.

Systematic errors ↔ nuisance parameters

Suppose we have primary measurements $\boldsymbol{y}$,
want to infer parameters of interest $\boldsymbol{\mu}$ with model $P(\boldsymbol{y}|\boldsymbol{\mu})$.

Model not perfect – need nuisance parameters → $P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta})$.

"Converts" sys. error to part of overall stat. error

Nuisance parameters decrease sensitivity to parameters of interest, so constrain using independent control measurements $\boldsymbol{u}$, which follow $P(\boldsymbol{u}|\boldsymbol{\theta})$. So now full likelihood is:

$$L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\mathbf{y},\mathbf{u}|\boldsymbol{\mu},\boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta})P(\mathbf{u}|\boldsymbol{\theta})$$

Often the control measurements $\boldsymbol{u}$ are estimates of the corresponding parameters $\boldsymbol{\theta}$, e.g., with Gaussian pdf.

Combining measurements

Starting point: reconstruct likelihood for full set of observed data and use to find a new limit. E.g. for independent data:

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = \prod_i L_i(\boldsymbol{\mu}_i, \boldsymbol{\theta}_i)$$

parameters of interest

nuisance parameters

Maybe likelihoods not available in full, need to approximate.

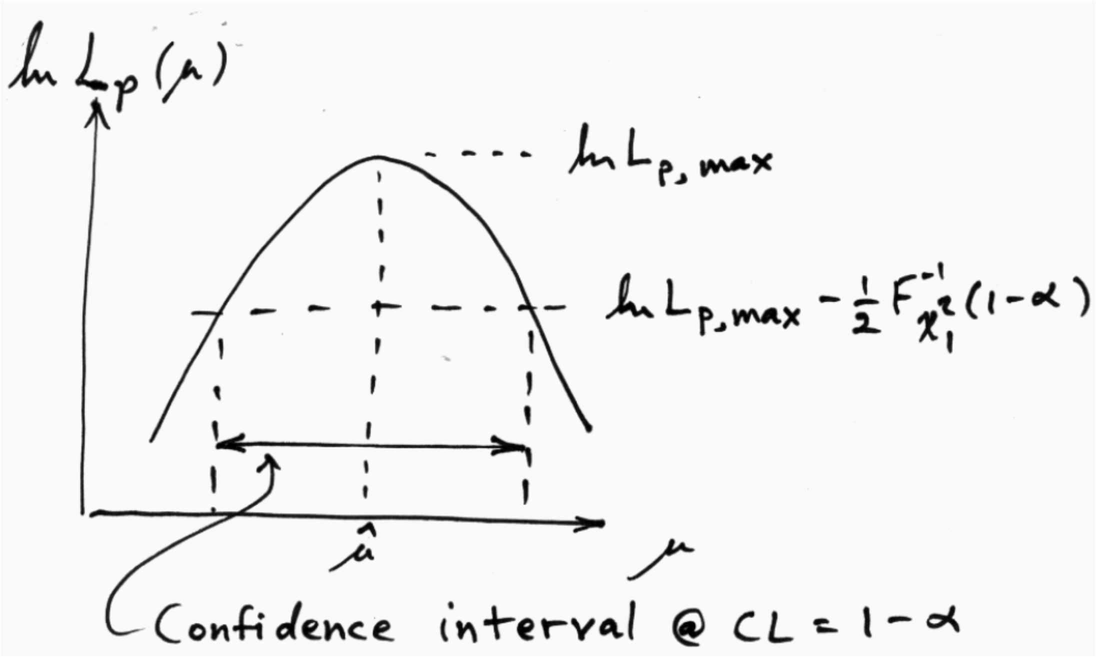
Need some assumptions about parameters common to different terms in the likelihood.

Same basic approach for limits and averages: start with best available approximation for the full likelihood.

Uncertainty of fitted parameters

Suppose parameter of interest μ, nuisance parameter θ, *confidence interval* for μ from "profile likelihood":

$$L_{\mathrm{p}}(\mu) = L(\mu, \hat{\hat{\theta}}) \qquad \hat{\hat{\theta}}(\mu) = \underset{\theta}{\operatorname{argmax}}\, L(\mu, \theta)$$



Width of interval in usual LS fit *independent* of goodness of fit.

Profiling nuisance parameters

Constrained (profiled) nuisance par.: $\hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\mu}, \boldsymbol{\theta})$

Test $\boldsymbol{\mu}$ using profile likelihood ratio $t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}$

p-value of $\boldsymbol{\mu}$: $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}, \boldsymbol{\theta})\, dt_{\boldsymbol{\mu}}$

Wilks' theorem:

pdf $f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu},\boldsymbol{\theta})$ in "large sample limit" + regularity conditions is chi-square with n_{dof} = num. par. of interest, independent of nuisance par. $\boldsymbol{\theta}$.

To the extent that asymptotic pdf is good approx., inference about $\boldsymbol{\mu}$ is independent of the nuisance parameters.

Marginalizing nuisance parameters

Alternatively, test $\boldsymbol{\mu}$ with a statistic such as the likelihood ratio using fixed (or estimated) values of nuisance parameters,

$$q_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}, \hat{\boldsymbol{\theta}})}{L(\boldsymbol{\mu}_0, \hat{\boldsymbol{\theta}})} \quad , \qquad \text{here} \quad L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta})$$

Assign prior (Bayesian) pdf to nuisance parameters $\pi(\boldsymbol{\theta})$ based on the control measurements:

$$\pi(\boldsymbol{\theta}) \propto P(\mathbf{u}|\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}) \quad , \qquad \pi_0(\boldsymbol{\theta}) \text{ e.g. const.}$$

The effect of the systematic uncertainty is built in by averaging over $\boldsymbol{\theta}$ with respect to $\pi(\boldsymbol{\theta})$ ($\rightarrow$ "marginalize" over nuisance par.):

$$f(q_{\boldsymbol{\mu}}|\boldsymbol{\mu}) = \int f(q_{\boldsymbol{\mu}}|\boldsymbol{\mu}, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \quad \longleftarrow$$

use for p-values, CLs, etc.

Profile vs. marginal likelihood approach

Strict frequentist approach: test points in ($\boldsymbol{\mu}$, $\boldsymbol{\theta}$) space and reject $\boldsymbol{\mu}$ if $p_{\boldsymbol{\mu}}(\boldsymbol{\theta}) \leq \alpha$ for all $\boldsymbol{\theta}$.

Obtained automatically if asymptotics à la Wilks holds.
Holds approximately if asymptotics not too badly broken.

Approximate alternative: reject $\boldsymbol{\mu}$ if $p_{\boldsymbol{\mu}}(\hat{\hat{\boldsymbol{\theta}}}) \leq \alpha$,
profile construction (Cranmer); hybrid resampling (B. Sen et al.)

Marginal approach does not test individual points in ($\boldsymbol{\mu}$, $\boldsymbol{\theta}$) space, rather the model for a given $\boldsymbol{\mu}$ averaged over $\boldsymbol{\theta}$.

Marginal likelihood usually requires MC integration to find pdf of test statistic (e.g., q_{μ}) needed to find p-value.

Results from profile and marginal likelihoods equal in simple cases and very similar in "most cases of practical interest".

HFLAV approach for a bad fit

In cases where χ^2/dof > 1, we do not usually scale the resulting uncertainty, in contrast to what is done by the Particle Data Group [5]. Rather, we examine the systematic uncertainties of each measurement to better understand them. Unless we find systematic discrepancies among the measurements, we do not apply any additional correction to the calculated uncertainty.

HFLAV, arXiv:1909.12524

Reasonable but... to be rigorous one would prefer that the statistical procedure not be subject to a posteriori modifications. If one has reason to suspect that potential biases could be present, then this should be built into the original statistical model.

GDC comment

"Errors on Errors"

APRIL 15, 1932 — PHYSICAL REVIEW — VOLUME 40

THE CALCULATION OF ERRORS BY THE METHOD OF LEAST SQUARES

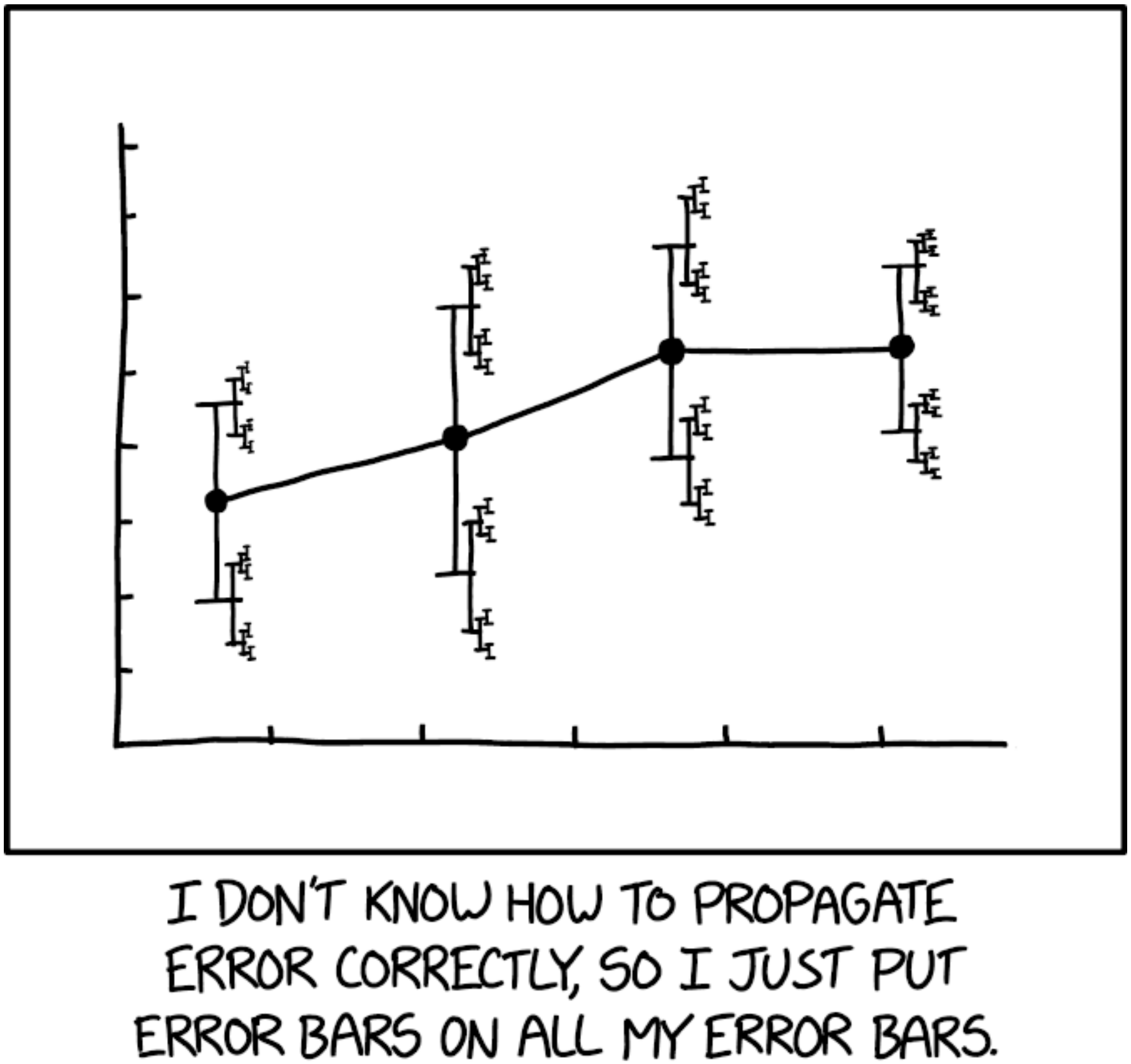
By Raymond T. Birge
University of California, Berkeley

(Received February 18, 1932)

Abstract

Present status of least squares' calculations.—There are three possible stages in any least squares' calculation, involving respectively the evaluation of (1) the most probable values of certain quantities from a set of experimental data, (2) the reliability or probable error of each quantity so calculated, (3) the reliability or probable error of the probable errors so calculated. Stages (2) and (3) are not adequately treated in most texts, and are frequently omitted or misused, in actual work. The present article is concerned mainly with these two stages.

→ PDG "scale factor method" ≈ scale sys. errors with common factor until χ^2_{min} = appropriate no. of degrees of freedom.


I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

Formulation of the problem

Suppose measurements $\boldsymbol{y}$ have probability (density) $P(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\theta})$,

$\boldsymbol{\mu}$ = parameters of interest

$\boldsymbol{\theta}$ = nuisance parameters

To provide info on nuisance parameters, often treat their best estimates $\boldsymbol{u}$ as indep. Gaussian distributed r.v.s., giving likelihood

$$L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\mathbf{y},\mathbf{u}|\boldsymbol{\mu},\boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta})P(\mathbf{u}|\boldsymbol{\theta})$$

$$= P(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{u_i}} e^{-(u_i-\theta_i)^2/2\sigma_{u_i}^2}$$

or log-likelihood (up to additive const.)

$$\ln L(\boldsymbol{\mu},\boldsymbol{\theta}) = \ln P(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\theta}) - \frac{1}{2}\sum_{i=1}^{N} \frac{(u_i-\theta_i)^2}{\sigma_{u_i}^2}$$

Systematic errors and their uncertainty

Often the θ_i could represent a systematic bias and its best estimate u_i in the real measurement is zero.

The $\sigma_{u,i}$ are the corresponding "systematic errors".

Sometimes $\sigma_{u,i}$ is well known, e.g., it is itself a statistical error known from sample size of a control measurement.

Other times the u_i are from an indirect measurement, Gaussian model approximate and/or the $\sigma_{u,i}$ are not exactly known.

Or sometimes $\sigma_{u,i}$ is at best a guess that represents an uncertainty in the underlying model ("theoretical error").

In any case we can allow that the $\sigma_{u,i}$ are not known in general with perfect accuracy.

Gamma model for variance estimates

Suppose we want to treat the systematic errors as uncertain, so let the $\sigma_{u,i}$ be adjustable nuisance parameters.

Suppose we have estimates s_i for $\sigma_{u,i}$ or equivalently $v_i = s_i^2$, is an estimate of $\sigma_{u,i}^2$.

Model the v_i as independent and gamma distributed:

$$f(v;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}$$

$$E[v] = \frac{\alpha}{\beta}$$

$$V[v] = \frac{\alpha}{\beta^2}$$

Set α and β so that they give desired relative uncertainty r in σ_u.

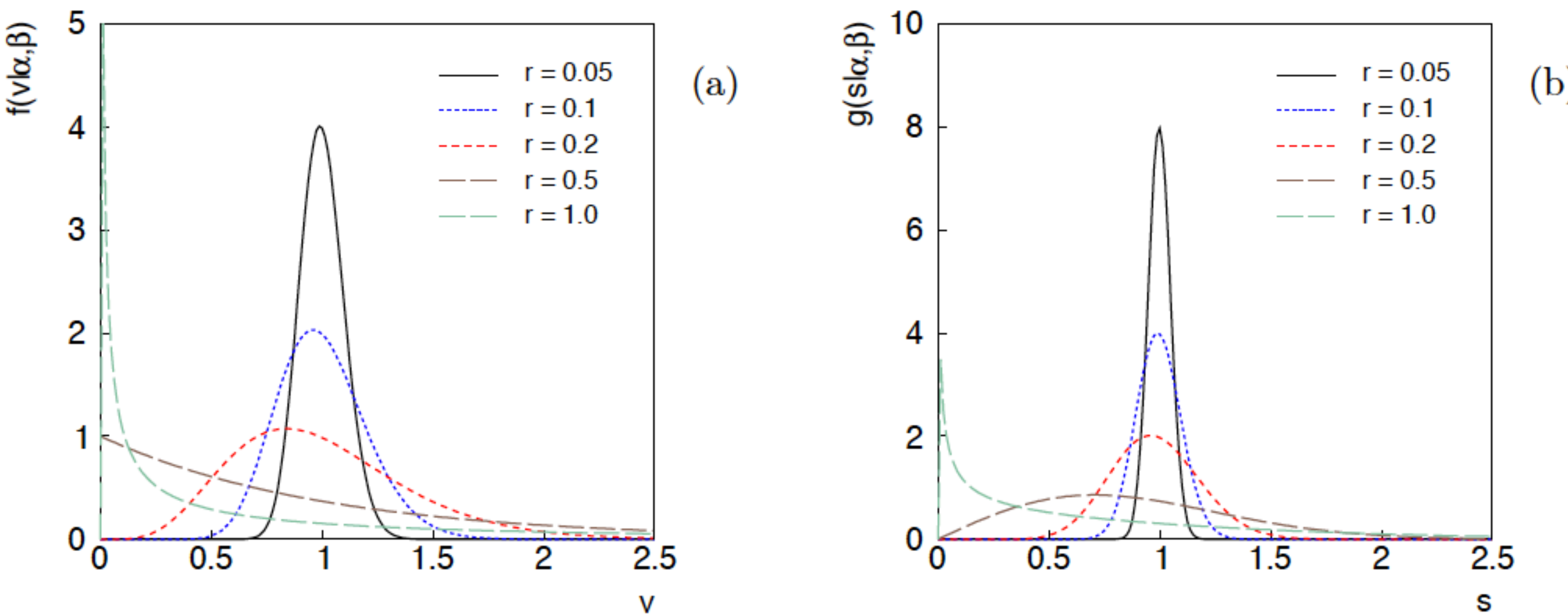
Similar to method 2 in W.J. Browne and D. Draper, Bayesian Analysis, Volume 1, Number 3 (2006), 473-514.

Distributions of v and $s = \sqrt{v}$

For α, β of gamma distribution, $\alpha_i = \frac{1}{4r_i^2}$, $\beta_i = \frac{1}{4r_i^2\sigma_{u_i}^2}$

$$r_i \equiv \frac{1}{2}\frac{\sigma_{v_i}}{E[v_i]} = \frac{1}{2}\frac{\sigma_{v_i}}{\sigma_{u_i}^2} \approx \frac{\sigma_{s_i}}{E[s_i]}$$

relative "error on error"



Motivation for gamma model

If one were to have n independent observations $u_1,\ldots,u_n$, with all $u \sim \text{Gauss}(\theta, \sigma_u^2)$, and we use the sample variance

$$v = \frac{1}{n-1} \sum_{i=1}^{n} (u_i - \overline{u})^2$$

to estimate σ_u^2, then $(n-1)v/\sigma_u^2$ follows a chi-square distribution for $n-1$ degrees of freedom, which is a special case of the gamma distribution ($\alpha = n/2$, $\beta = 1/2$). (In general one doesn't have a sample of u_i values, but if this were to be how v was estimated, the gamma model would follow.)

Furthermore choice of the gamma distribution for v allows one to profile over the nuisance parameters σ_u^2 in closed form and leads to a simple profile likelihood.

Likelihood for gamma error model

$$L(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\sigma}_{\mathbf{u}}^2) = P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_{u_i}^2}} e^{-(u_i - \theta_i)^2 / 2\sigma_{u_i}^2}$$

$$\times \quad \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} v_i^{\alpha_i - 1} e^{-\beta_i v_i} \ .$$

Treated like data:	$y_1,\ldots,y_L$	(the primary measurements)
	$u_1,\ldots,u_N$	(estimates of nuisance par.)
	$v_1,\ldots,v_N$	(estimates of variances of estimates of NP)
Adjustable parameters:	$\mu_1,\ldots,\mu_{\mathrm{M}}$	(parameters of interest)
	$\theta_1,\ldots,\theta_N$	(nuisance parameters)
	$\sigma_{u,1},\ldots,\sigma_{u,N}$	(sys. errors = std. dev. of of NP estimates)
Fixed parameters:	$r_1,\ldots,r_N$	(rel. err. in estimate of $\sigma_{u,i}$)

Profiling over systematic errors

We can profile over the $\sigma_{u,i}$ in closed form

$$\widehat{\widehat{\sigma^2}}_{u_i} = \underset{\sigma^2_{u_i}}{\operatorname{argmax}}\, L(\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\sigma}^2_{\mathbf{u}}) = \frac{v_i + 2r_i^2 (u_i - \theta_i)^2}{1 + 2r_i^2}$$

which gives the profile log-likelihood (up to additive const.)

$$\boxed{\begin{aligned} \ln L'(\mu, \boldsymbol{\theta}) &= \ln L(\mu, \boldsymbol{\theta}, \widehat{\widehat{\boldsymbol{\sigma}^2}}_{\mathbf{u}}) \\ &= \ln P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) - \frac{1}{2} \sum_{i=1}^{N} \left(1 + \frac{1}{2r_i^2}\right) \ln \left[1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right] \end{aligned}}$$

In limit of small r_i and $v_i \to \sigma_{u,i}{}^2$, the log terms revert back to the quadratic form seen with known $\sigma_{u,i}$.

Equivalent likelihood from Student's t

We can arrive at same likelihood by defining $$z_i \equiv \frac{u_i - \theta_i}{\sqrt{v_i}}$$

Since $u_i \sim$ Gauss and $v_i \sim$ Gamma, $z_i \sim$ Student's t

$$f(z_i|\nu_i) = \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\sqrt{\nu_i \pi}\,\Gamma(\nu_i/2)} \left(1 + \frac{z_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}} \quad \text{with} \quad \nu_i = \frac{1}{2r_i^2}$$

Resulting likelihood same as profile $L'(\boldsymbol{\mu},\boldsymbol{\theta})$ from gamma model

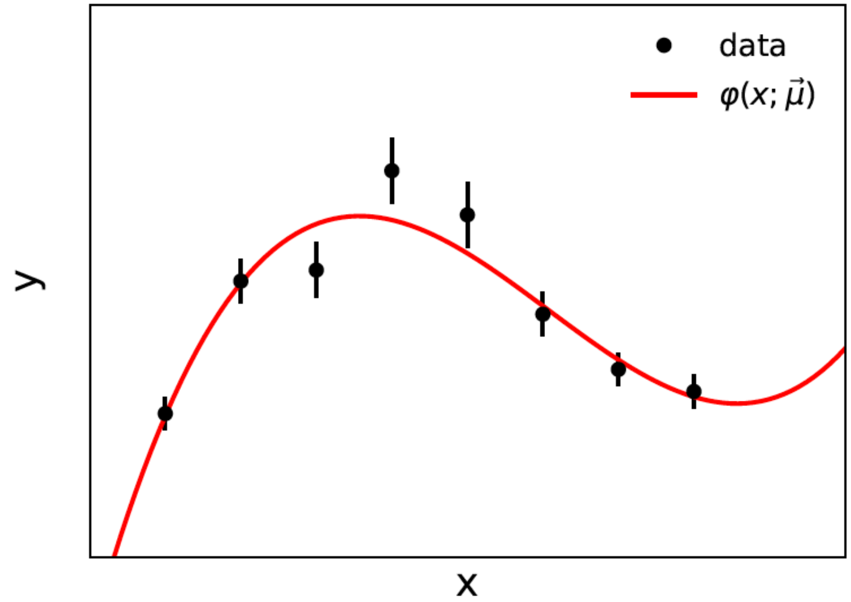
$$L(\boldsymbol{\mu}, \boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\theta}) \prod_{i=1}^{N} \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\sqrt{\nu_i \pi}\,\Gamma(\nu_i/2)} \left(1 + \frac{z_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}$$

Curve fitting, averages

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

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

$$V[y_i] = \sigma_{y_i}^2 \;.$$



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

$\boldsymbol{\theta}$ are bias parameters constrained by control measurements $u_i \sim$ Gauss(θ_i, $\sigma_{u,i}$), so that if $\sigma_{u,i}$ are known we have

$$-2\ln L(\boldsymbol{\mu},\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[\frac{(y_i - \varphi(x_i;\boldsymbol{\mu}) - \theta_i)^2}{\sigma_{y_i}^2} + \frac{(u_i-\theta_i)^2}{\sigma_{u_i}^2}\right]$$

Profiling over θ_i with known $\sigma_{u,i}$

Profiling over the bias parameters θ_i for known $\sigma_{u,i}$ gives usual least-squares (BLUE)

$$-2\ln L'(\boldsymbol{\mu}) = \sum_{i=1}^{N} \frac{(y_i - \varphi(x_i;\boldsymbol{\mu}) - u_i)^2}{\sigma_{y_i}^2 + \sigma_{u_i}^2} \equiv \chi^2(\boldsymbol{\mu})$$

Widely used technique for curve fitting in Particle Physics.

Generally in real measurement, $u_i = 0$.

Generalized to case of correlated y_i and u_i by summing statistical and systematic covariance matrices.

Curve fitting with uncertain $\sigma_{u,i}$

Suppose now $\sigma_{u,i}^2$ are adjustable parameters with gamma distributed estimates v_i.

Retaining the θ_i but profiling over $\sigma_{u,i}^2$ gives

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

Profiled values of θ_i from solution to cubic equations

$$\theta_i^3 \quad + \quad [-2u_i - y_i + \varphi_i]\,\theta_i^2 + \left[\frac{v_i + (1+2r_i^2)\sigma_{y_i}^2}{2r_i^2} + 2u_i(y_i-\varphi_i) + u_i^2\right]\theta_i$$

$$+ \quad \left[(\varphi_i - y_i)\left(\frac{v_i}{2r_i^2} + u_i^2\right) - \frac{(1+2r_i^2)\sigma_{y_i}^2 u_i}{2r_i^2}\right] = 0\,, \quad i = 1,\ldots,N\,,$$

Goodness of fit

Can quantify goodness of fit with statistic

$$q = -2 \ln \frac{L'(\hat{\boldsymbol{\mu}}, \hat{\hat{\boldsymbol{\theta}}})}{L'(\hat{\boldsymbol{\varphi}}, \hat{\boldsymbol{\theta}})}$$

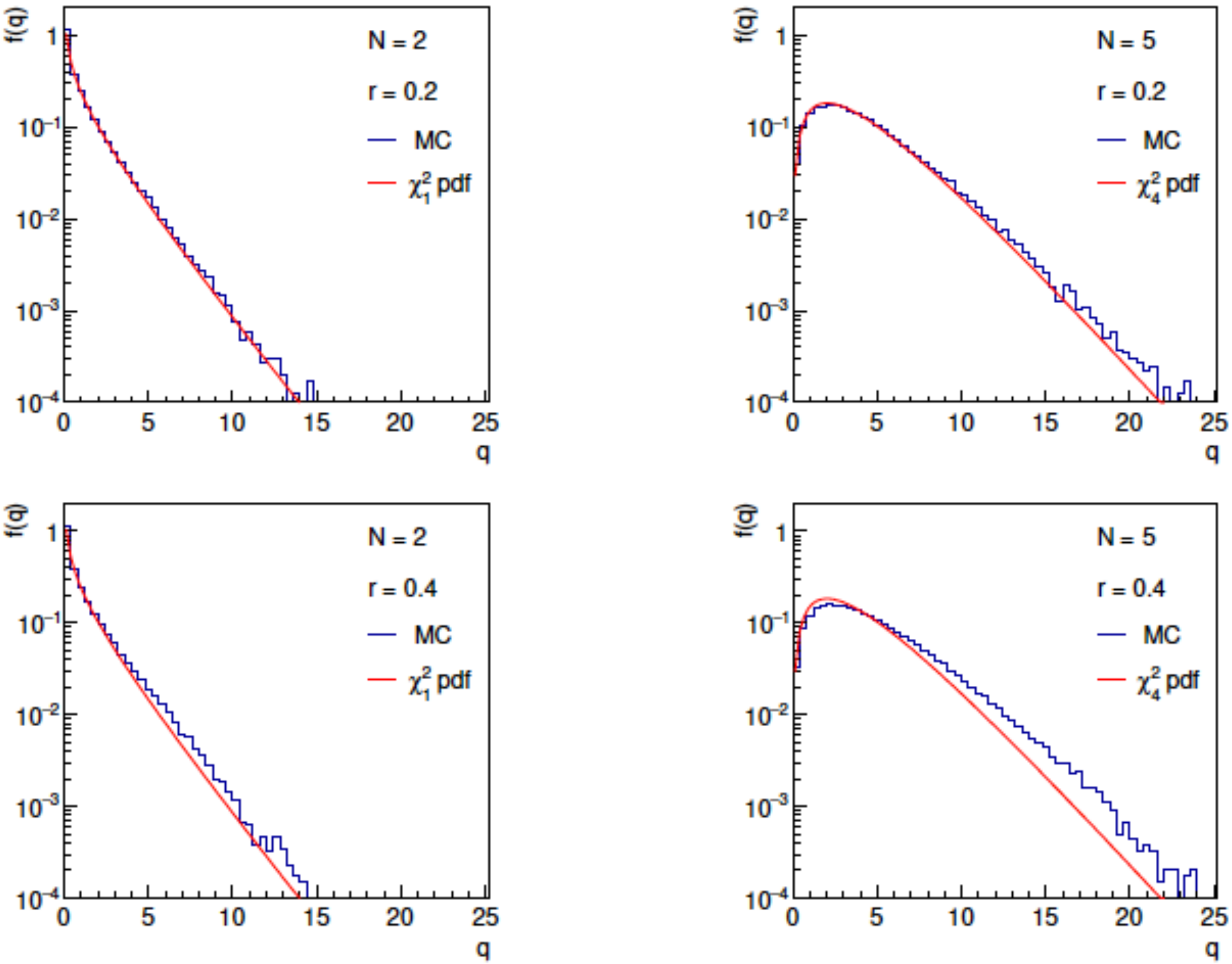
$$= \min_{\boldsymbol{\mu}, \boldsymbol{\theta}} \sum_{i=1}^{N} \left[\frac{(y_i - \varphi(x_i; \boldsymbol{\mu}) - \theta_i)^2}{\sigma_{y_i}^2} + \left(1 + \frac{1}{2r_i^2}\right) \ln\left(1 + 2r_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right)\right]$$

where $L'(\boldsymbol{\varphi}, \boldsymbol{\theta})$ has an adjustable φ_i for each y_i (the saturated model).

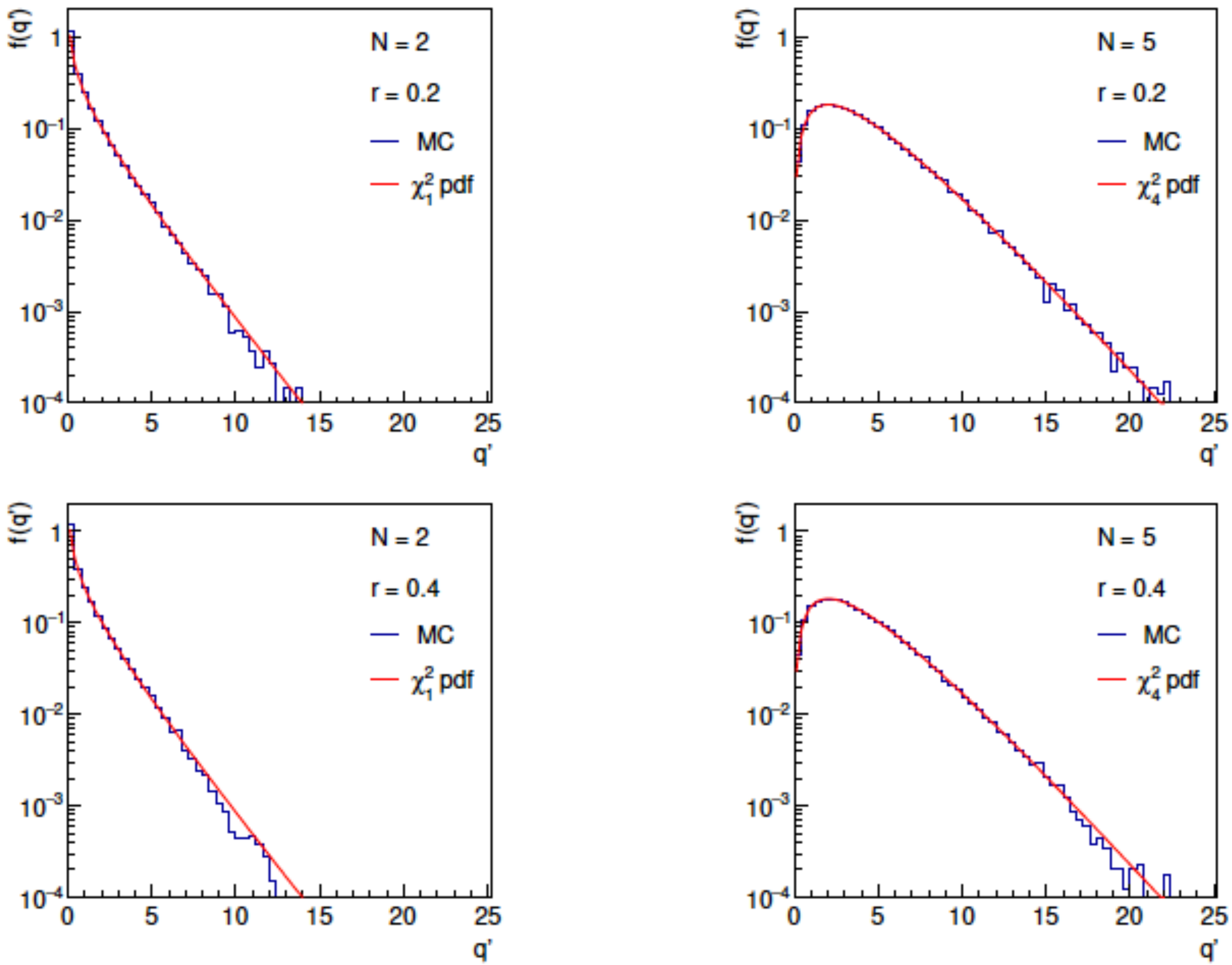
Asymptotically should have $q \sim$ chi-squared($N\text{–}M$).

For increasing r_i, may need Bartlett correction or MC.

Distributions of q



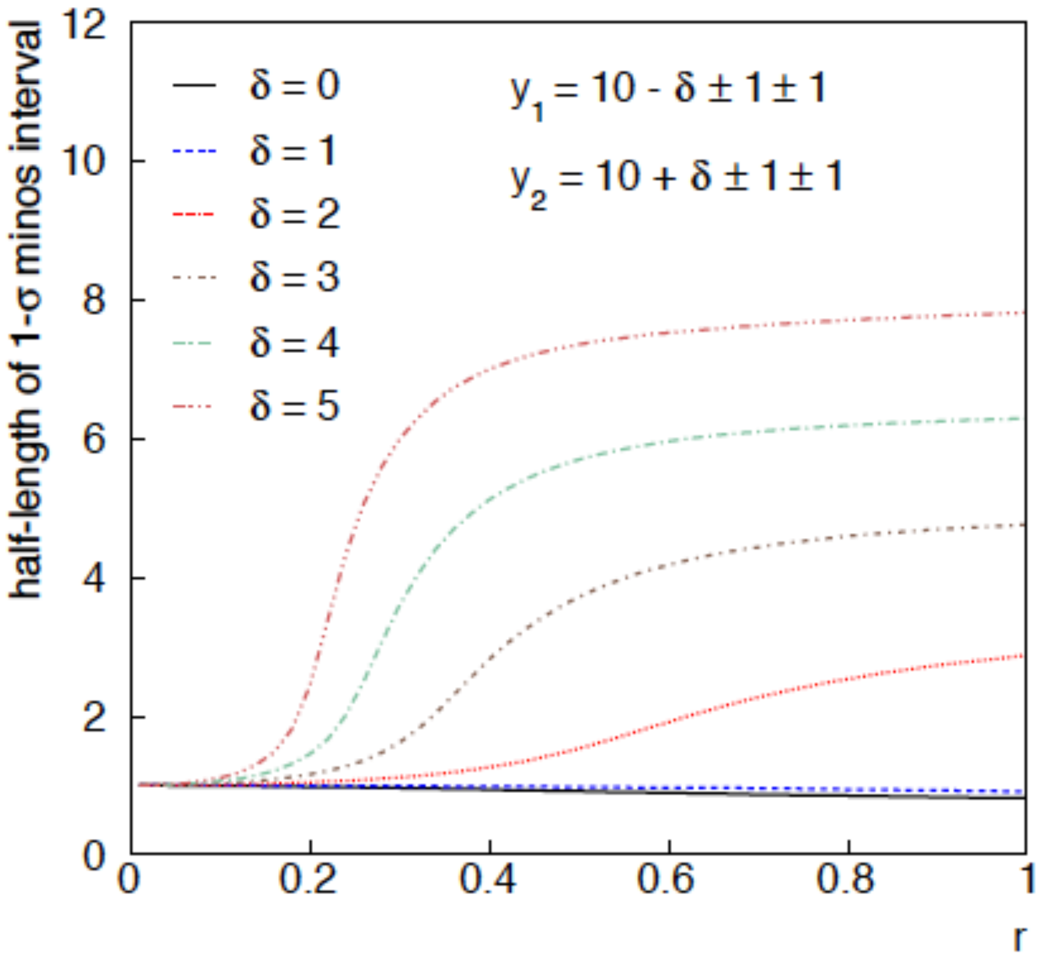
Distributions of Bartlett-corrected q'



Example: average of two measurements

MINOS interval (= approx. confidence interval) based on

$$\ln L'(\mu) = \ln L'(\hat{\mu}) - Q_\alpha/2 \quad \text{with} \quad Q_\alpha = F^{-1}_{\chi^2}(1-\alpha; n)$$

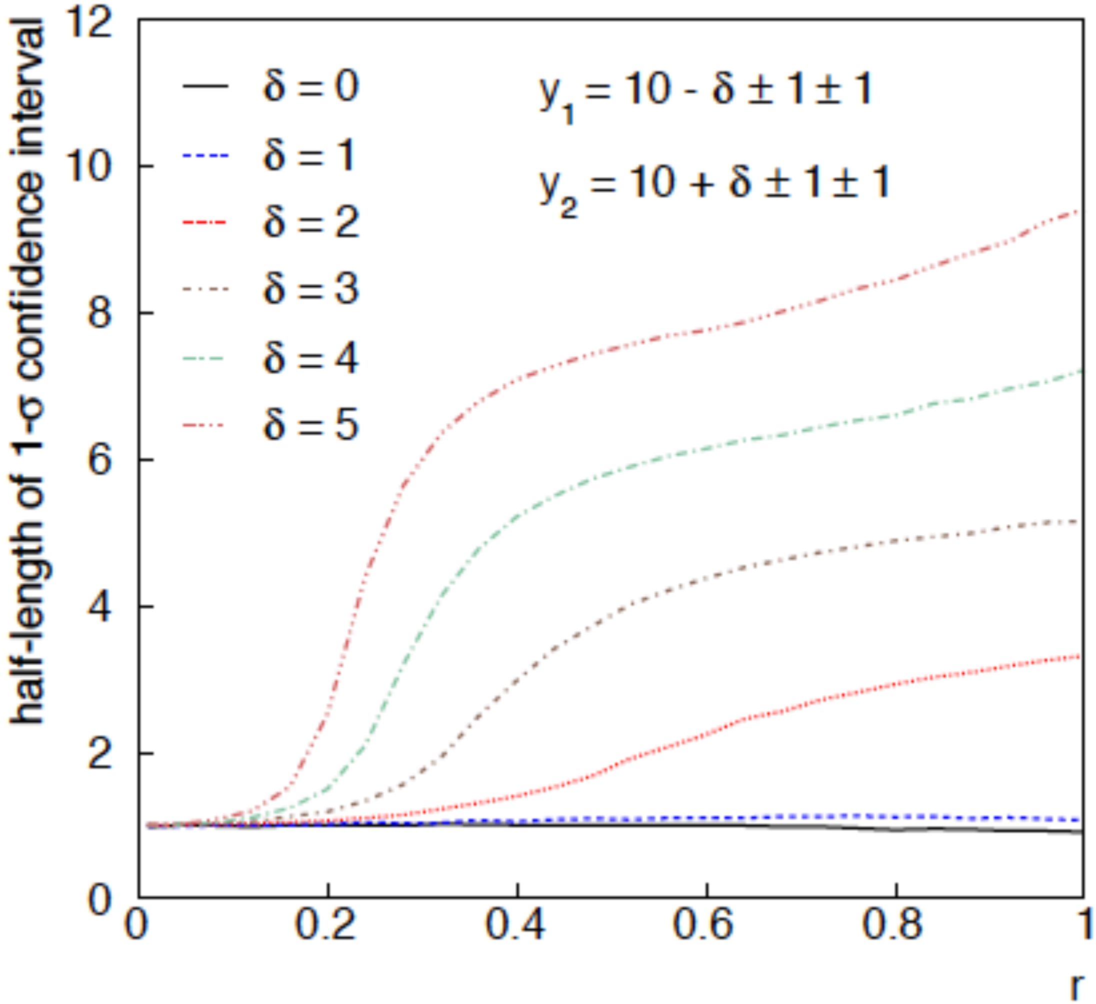


Increased discrepancy between values to be averaged gives larger interval.

Interval length saturates at ~level of absolute discrepancy between input values.

relative error on sys. error

Same with interval from $p_\mu = \alpha$ with nuisance parameters profiled at μ



Coverage of intervals

Consider previous average of two numbers but now generate for $i = 1, 2$ data values

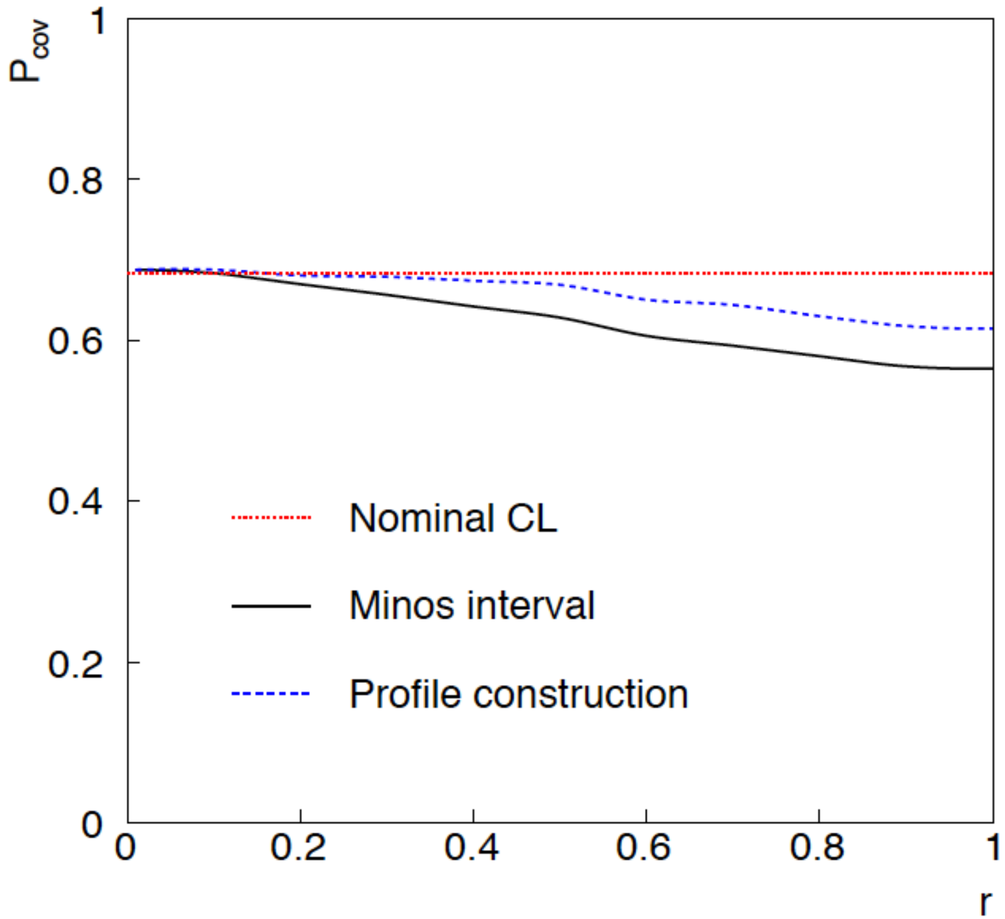
$$y_i \sim \text{Gauss}(\mu, \sigma_{y,i})$$

$$u_i \sim \text{Gauss}(0, \sigma_{u,i})$$

$$v_i \sim \text{Gamma}(\sigma_{u,i}, r_i)$$

$$\sigma_{y,i} = \sigma_{u,i} = 1$$

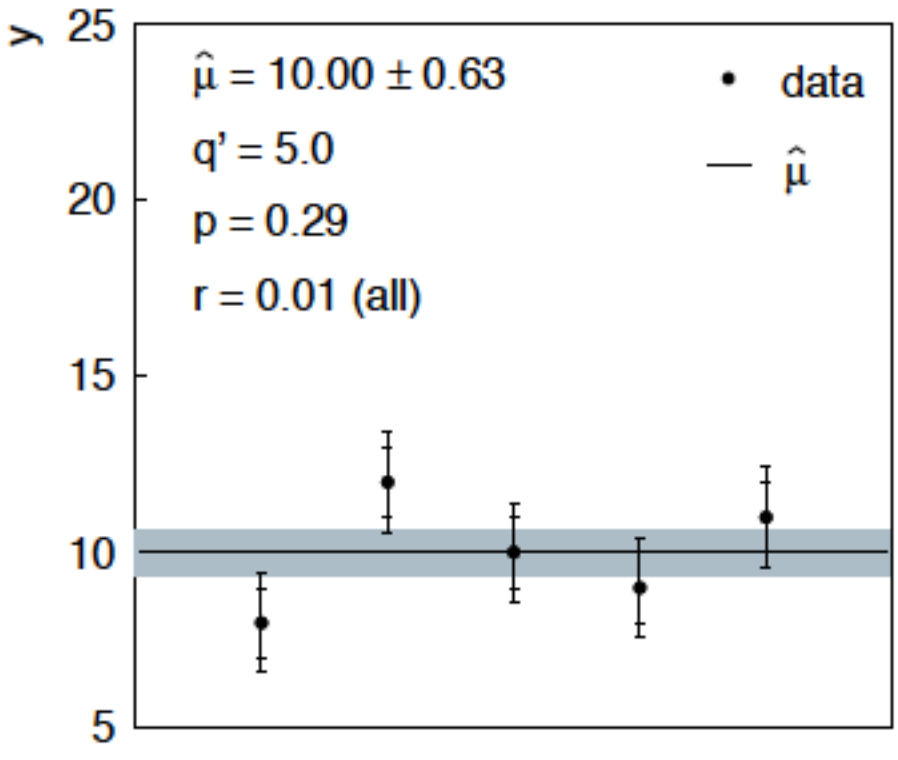
and look at the probability that the interval covers the true value of μ.



Coverage stays reasonable to $r \sim 0.5$, even not bad for Profile Construction out to $r \sim 1$.

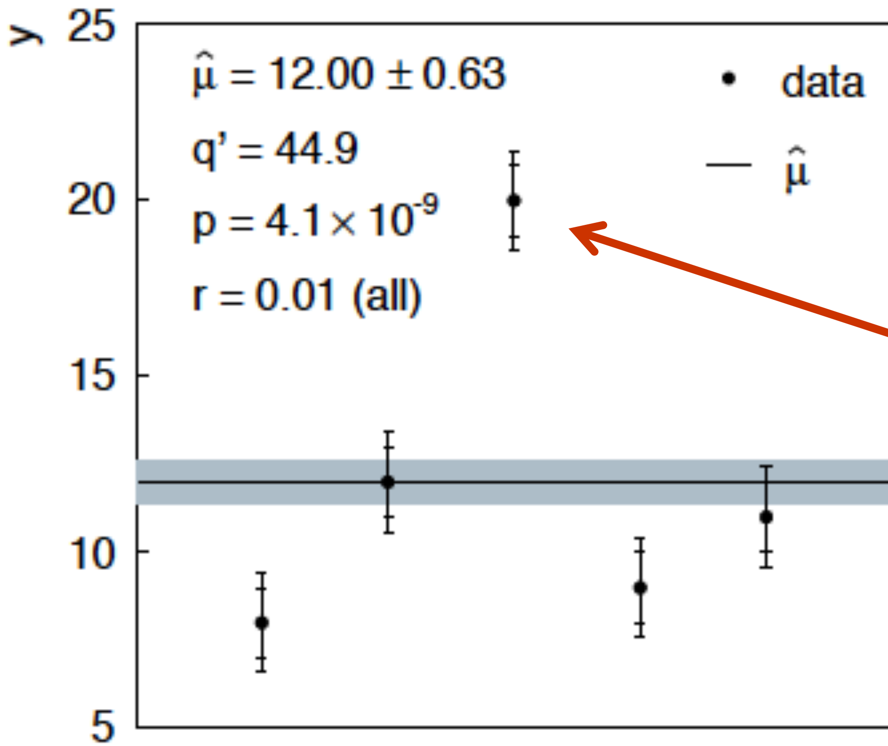
Sensitivity of average to outliers

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).



Sensitivity of average to outliers (2)

Now suppose the measurement at 10 was actually at 20:

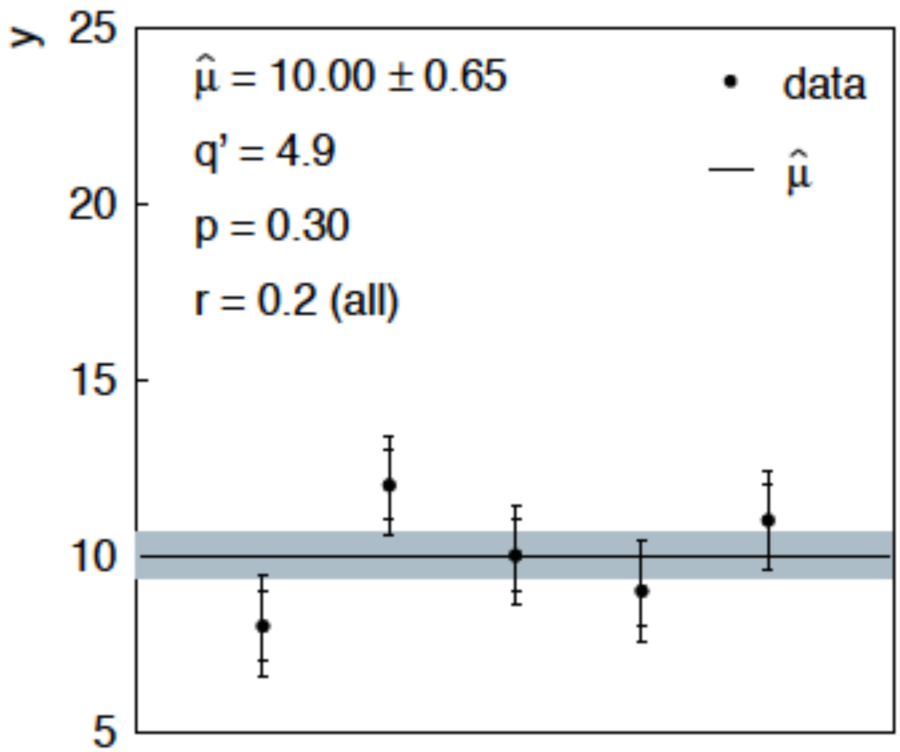


"outlier"

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

Average with all $r = 0.2$

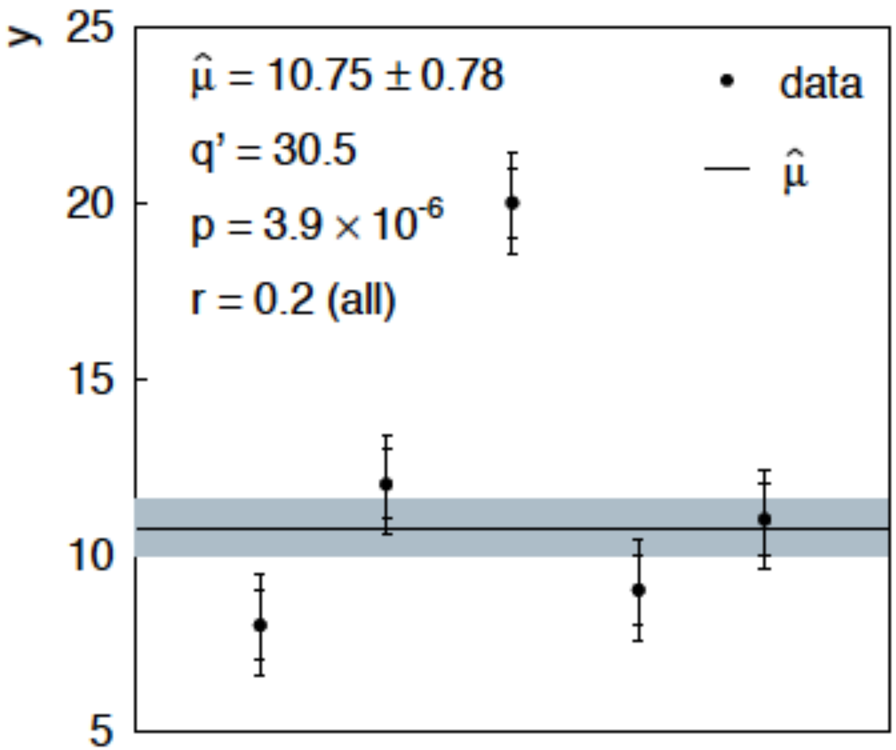
If we assign to each measurement $r = 0.2$,



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 → 20)



Estimate →10.75 (outlier pulls much less).

Half-size of interval → 0.78 (inflated because of bad g.o.f.).

Naive approach to errors on errors

Naively one might think that the error on the error in the previous example could be taken into account conservatively by inflating the systematic errors, i.e.,

$$\sigma_{u_i} \rightarrow \sigma_{u_i}(1 + r_i)$$

But this gives

$\hat{\mu} = 10.00 \pm 0.70$ without outlier (middle meas. 10)

$\hat{\mu} = 12.00 \pm 0.70$ with outlier (middle meas. 20)

So the sensitivity to the outlier is not reduced and the size of the confidence interval is still independent of goodness of fit.

Correlated uncertainties

The phrase "correlated uncertainties" usually means that a single nuisance parameter affects the distribution (e.g., the mean) of more than one measurement.

For example, consider measurements $\boldsymbol{y}$, parameters of interest $\boldsymbol{\mu}$, nuisance parameters $\boldsymbol{\theta}$ with

$$E[y_i] = \varphi_i(\boldsymbol{\mu}, \boldsymbol{\theta}) \approx \varphi_i(\boldsymbol{\mu}) + \sum_{j=1}^{N} R_{ij}\theta_j$$

That is, the θ_i are defined here as contributing to a bias and the (known) factors R_{ij} determine how much θ_j affects y_i.

As before suppose one has independent control measurements $u_i \sim$ Gauss(θ_i, σ_{ui}).

Correlated uncertainties (2)

The total bias of y_i can be defined as $$b_i = \sum_{j=1}^{N} R_{ij} \theta_j$$

which can be estimated with $$\hat{b}_i = \sum_{j=1}^{N} R_{ij} u_j$$

These estimators are correlated having covariance

$$U_{ij} = \mathrm{cov}[\hat{b}_i, \hat{b}_j] = \sum_{k=1}^{N} R_{ik} R_{jk} V[u_k]$$

In this sense the present method treats "correlated uncertainties", i.e., the control measurements u_i are independent, but nuisance parameters affect multiple measurements, and thus bias estimates are correlated.

Discussion / Conclusions (1)

My view: general approach of HFLAV is perfectly reasonable.

Report confidence interval/region (plus covariance?), systematics from profiling nuisance parameters.

Clarify treatment of common nuisance parameters.

Recommend some tweaks of notation and vocabulary:

Greek letters for parameters, Latin letters for data
CL → *p*-value

Recommend avoiding a posteriori changes to model in case of bad fit (→ "errors on errors"?)

Discussion / Conclusions (2)

Gamma model for variance estimates 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:

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

Discussion / Conclusions (3)

Asymptotics can break for increased error-on-error, may need Bartlett correction or MC.

Method assumes that meaningful r_i values can be assigned and is valuable when systematic errors are not well known but enough "expert opinion" is available to do so.

Alternatively one could try to fit a global r to all systematic errors, analogous to PDG scale factor method or meta-analysis à la DerSimonian and Laird. (→ future work).

Could also use e.g. as "stress test" – crank up the r_i values until significance of result degrades and ask if you really trust the assigned systematic errors at that level.

Extra slides

Curve Fitting History: Least Squares

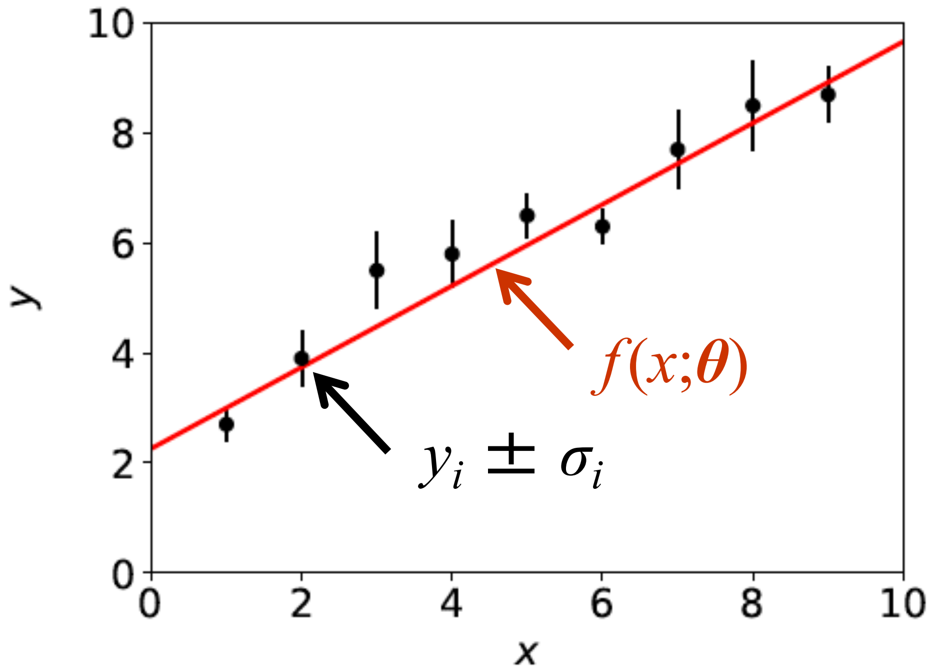
Method of Least Squares by Laplace, Gauss, Legendre, Galton...

C.F. Gauss, Theoria Combinationis Observationum Erroribus Minimis Obnoxiae, Commentationes Societatis Regiae Scientiarium Gottingensis Recectiores Vol. V (MDCCCXXIII).

To fit curve $f(x;\boldsymbol{\theta})$ to data $y_i \pm \sigma_i$, adjust parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$ to minimize

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

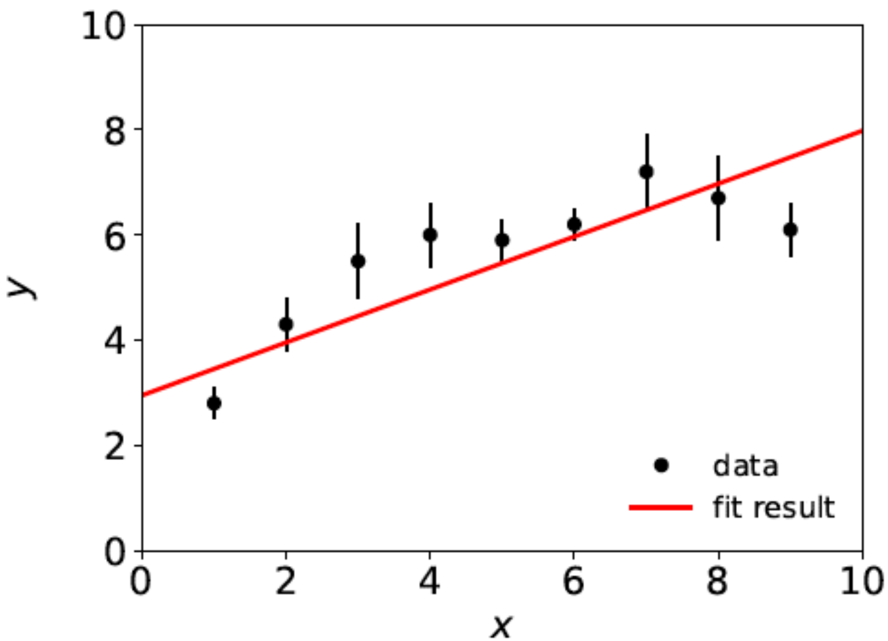
Assumes σ_i known.



Goodness of fit

If the hypothesized model $f(x;\boldsymbol{\theta})$ is correct, $\chi^2_{\min}$ should follow a chi-square distribution for N (# meas.) – M (# fitted par.) degrees of freedom; expectation value = $N - M$.

Suppose initial guess for model is: $f(x;\boldsymbol{\theta}) = \theta_0 + \theta_1 x$



$\chi^2_{\min} = 20.9$,
$N - M = 9 - 2 = 7$,
so goodness of fit is "poor".

This is an indication that the model is inadequate, and thus the values it predicts will have a "systematic error".

Systematic errors ↔ nuisance parameters

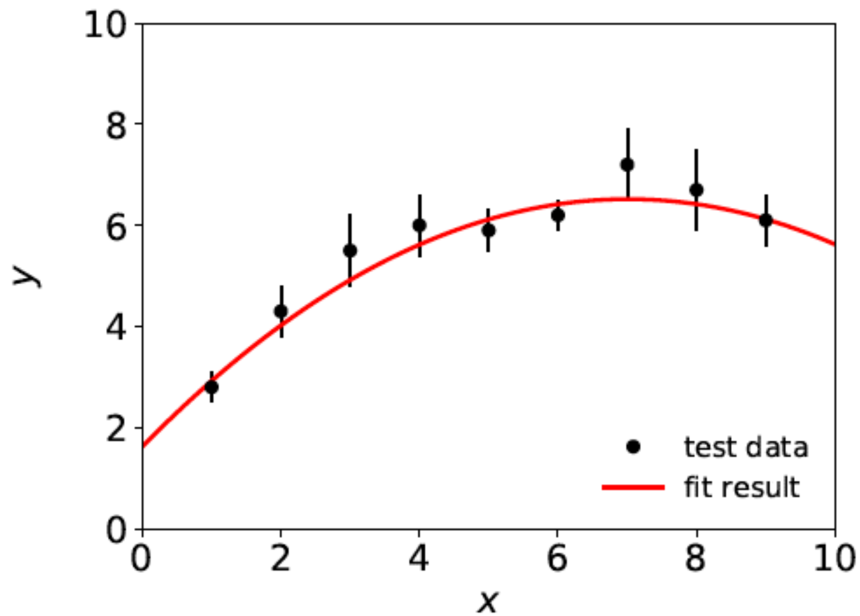
Solution: fix the model, generally by inserting additional adjustable parameters ("nuisance parameters"). Try, e.g.,

$$f(x;\boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2$$

$\chi^2_{\min} = 3.5,\ N - M = 6$

For some point in the enlarged parameter space we hope the model is now ~correct.

Sys. error gone?

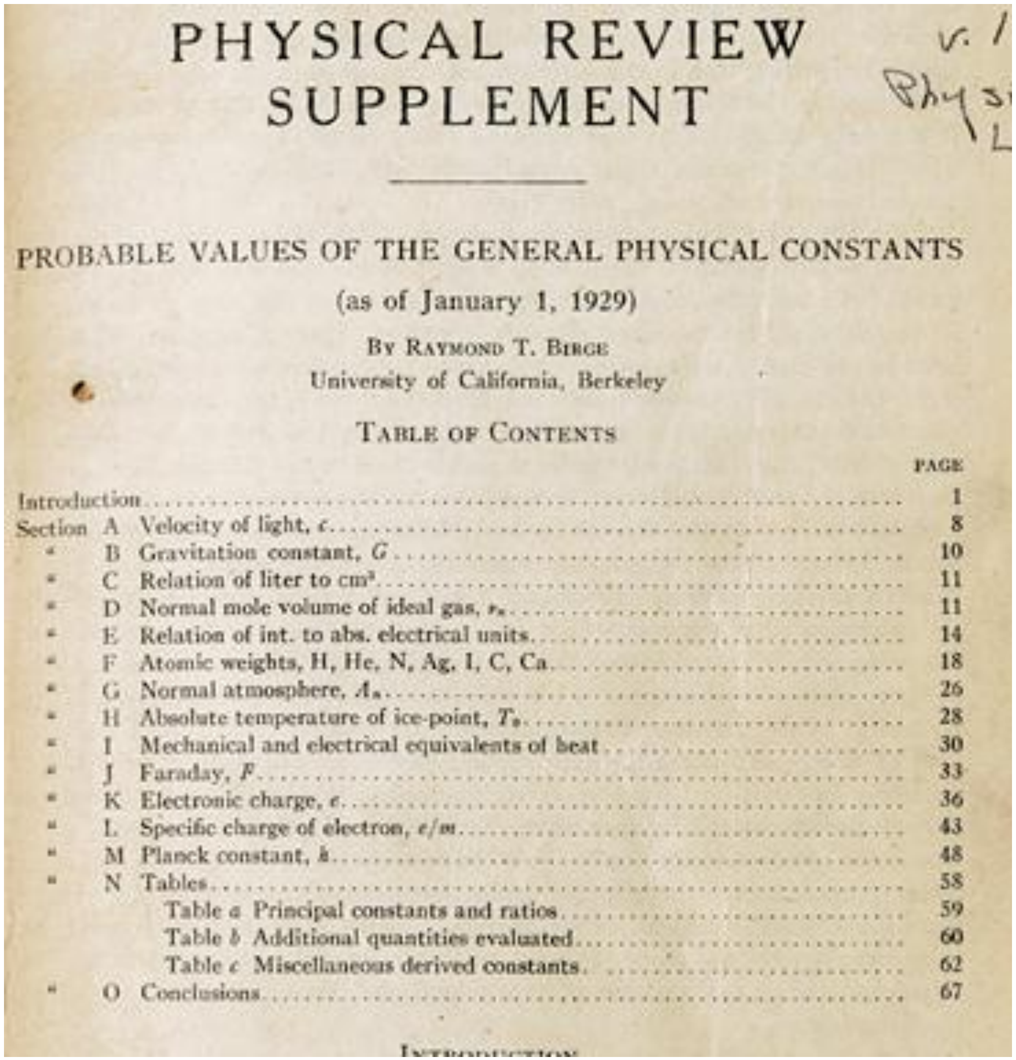


Estimators for all parameters correlated, and as a consequence the presence of the nuisance parameters inflates the statistical errors of the parameter(s) of interest.

http://bancroft.berkeley.edu/Exhibits/physics/learning01.html

Least Squares for Averaging

= fit of horizontal line



PHYSICAL REVIEW
SUPPLEMENT

PROBABLE VALUES OF THE GENERAL PHYSICAL CONSTANTS

(as of January 1, 1929)

BY RAYMOND T. BIRGE
University of California, Berkeley

TABLE OF CONTENTS

Raymond T. Birge,
Probable Values of the General Physical Constants (as of January 1, 1929),
Physical Review Supplement, Vol 1, Number 1, July 1929

Forerunner of the Particle Data Group

Errors on theory errors, e.g., in QCD

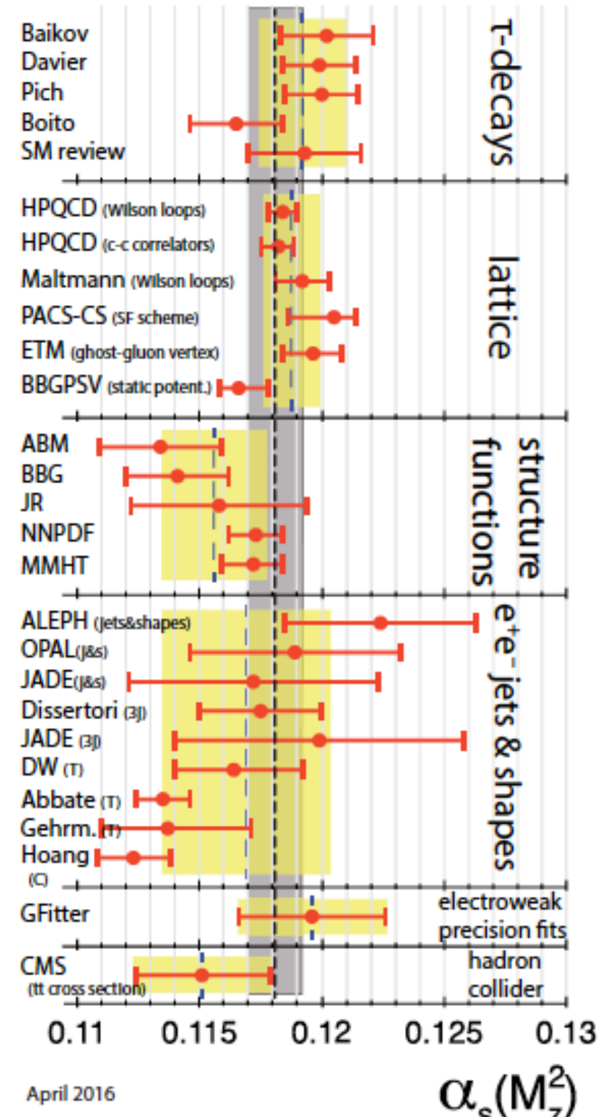
Uncertainties related to theoretical predictions are notoriously difficult to quantify, e.g., in QCD may come from variation of renormalization scale in some "appropriate range".

Problematic e.g. for α_s →

If, e.g., some (theory) errors are underestimated, one may obtain poor goodness of fit, but size of confidence interval from usual recipe will not reflect this.

An outlier with an underestimated error bar can have an inordinately strong influence on the average.

M. Tanabashi et al. (Particle Data Group), Phys. Rev. D 98, 030001 (2018)



Developments of LS for Averaging

Much work in HEP and elsewhere on application/extension of least squares to the problem of averaging or meta-analysis, e.g.,

A. C. Aitken, *On Least Squares and Linear Combinations of Observations*, Proc. Roy. Soc. Edinburgh **55** (1935) 42.

L. Lyons, D. Gibaut and P. Clifford, *How to Combine Correlated Estimates of a Single Physical Quantity*, Nucl. Instr. Meth. **A270** (1988) 110.

A. Valassi, *Combining Correlated Measurements of Several Different Physical Quantities*, Nucl. Instr. Meth. **A500** (2003) 391.

R. Nisius, *On the combination of correlated estimates of a physics observable*, Eur. Phys. J. C **74** (2014) 3004.

R. DerSimonian and N. Laird, *Meta-analysis in clinical trials*, Controlled Clinical Trials **7** (1986) 177-188.

Single-measurement model

As a simplest example consider

$y \sim \text{Gauss}(\mu, \sigma^2)$,

$v \sim \text{Gamma}(\alpha, \beta)$, $\quad \alpha = \frac{1}{4r^2}, \quad \beta = \frac{1}{4r^2\sigma^2}$

$$L(\mu, \sigma^2) = f(y, v|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} \frac{\beta^\alpha}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}$$

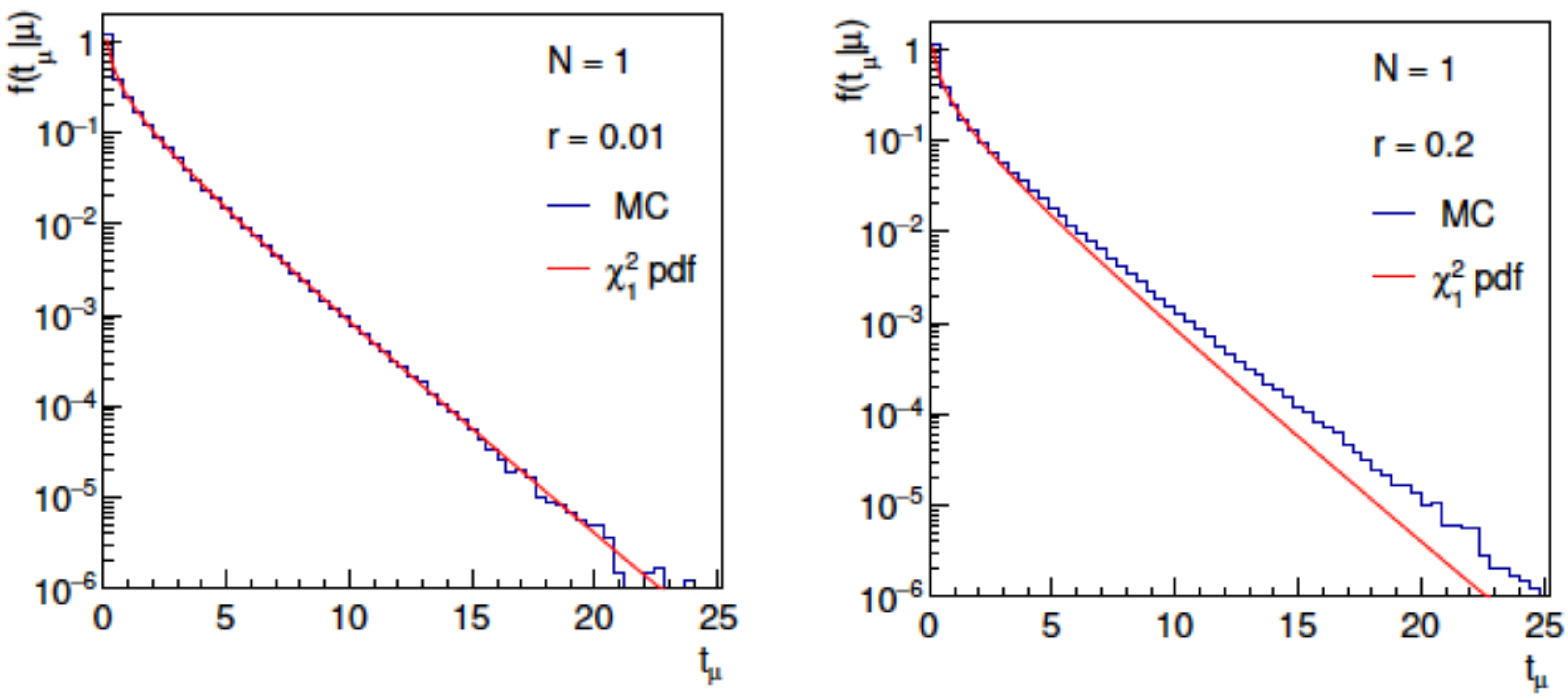
Test values of μ with $t_\mu = -2 \ln \lambda(\mu)$ with $\lambda(\mu) = \frac{L(\mu, \widehat{\widehat{\sigma^2}}(\mu))}{L(\hat{\mu}, \widehat{\sigma^2})}$

$$t_\mu = \left(1 + \frac{1}{2r^2}\right) \ln\left[1 + 2r^2 \frac{(y-\mu)^2}{v}\right]$$

Distribution of t_μ

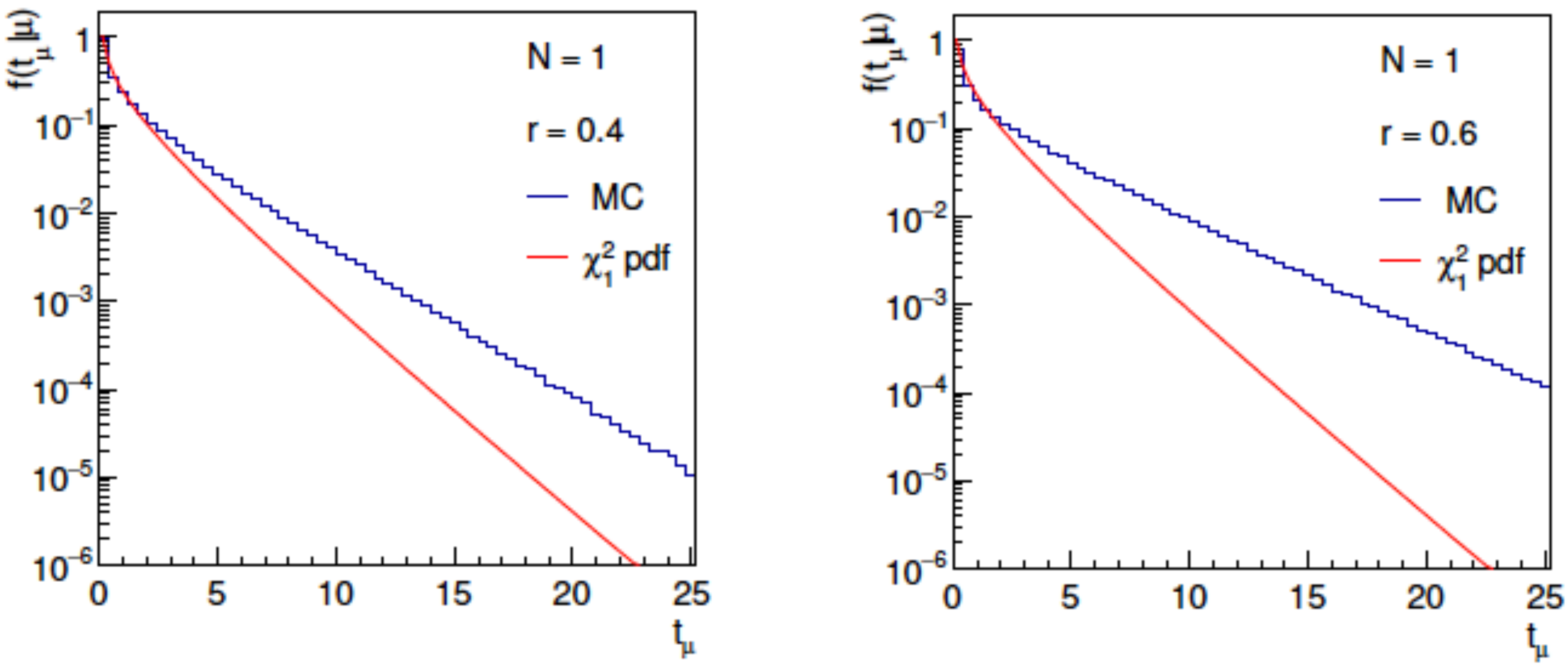
From Wilks' theorem, in the asymptotic limit we should find $t_\mu \sim$ chi-squared(1).

Here "asymptotic limit" means all estimators ~Gauss, which means $r \to 0$. For increasing r, clear deviations visible:



Distribution of t_μ (2)

For larger r, breakdown of asymptotics gets worse:



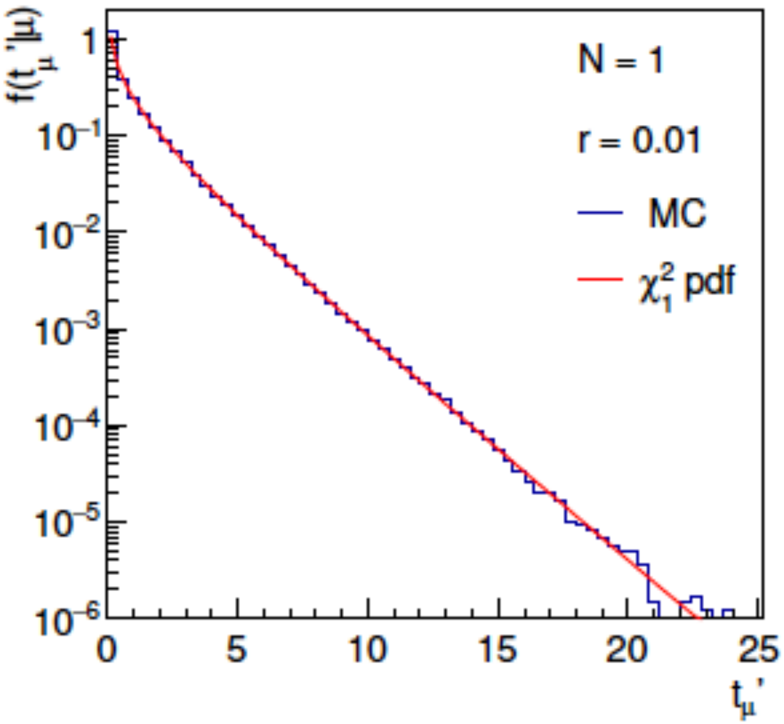
Values of r ~ several tenths are relevant so we cannot in general rely on asymptotics to get confidence intervals, p-values, etc.

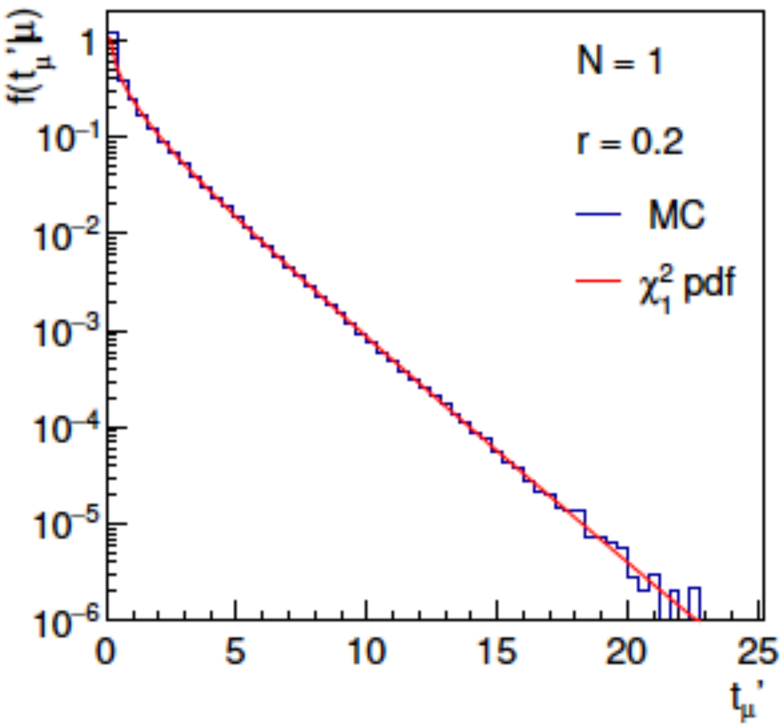
Bartlett corrections

One can modify t_μ defining $$t'_\mu = \frac{n_\mathrm{d}}{E[t_\mu]} t_\mu$$

such that the new statistic's distribution is better approximated by chi-squared for n_d degrees of freedom (Bartlett, 1937).

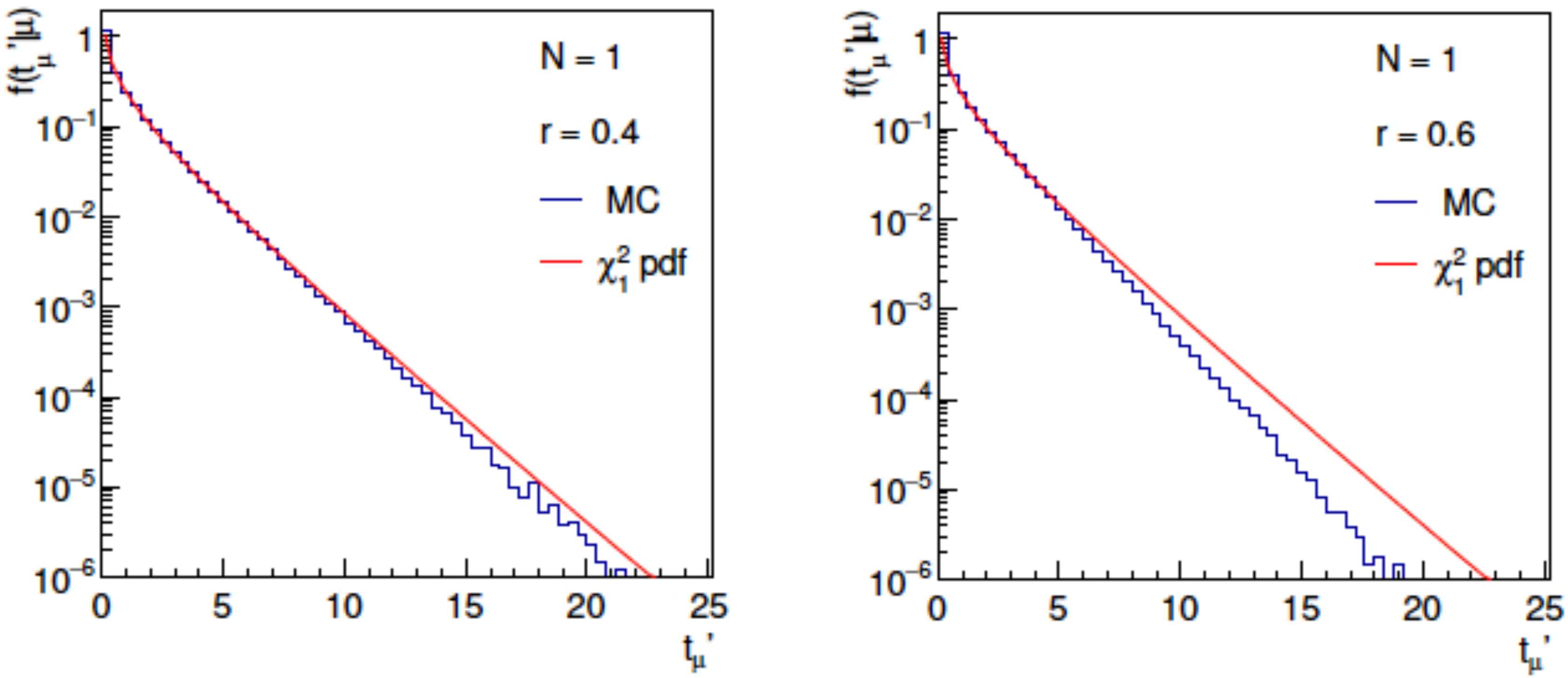
For this example $E[t_\mu] \approx 1 + 3r^2 + 2r^4$ works well:



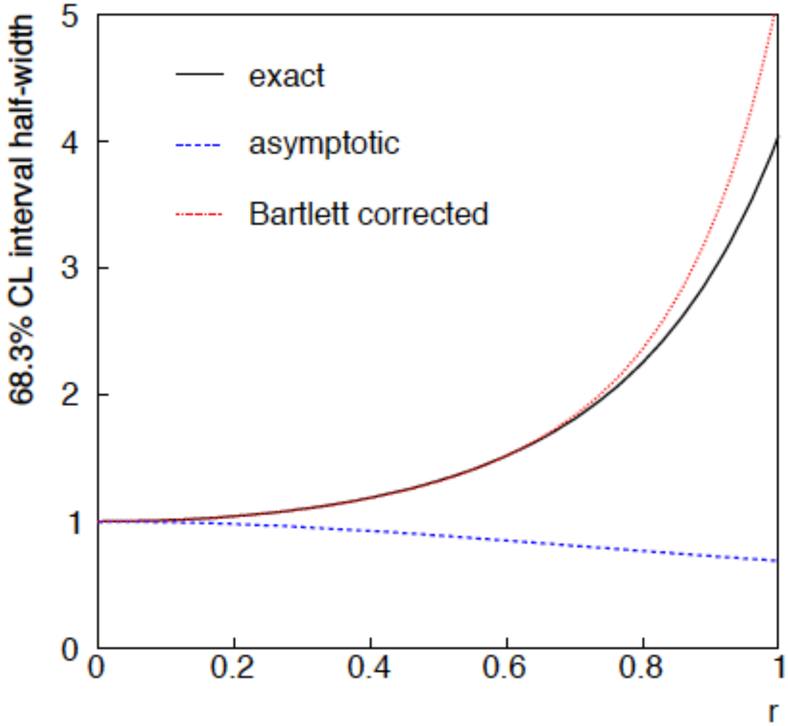


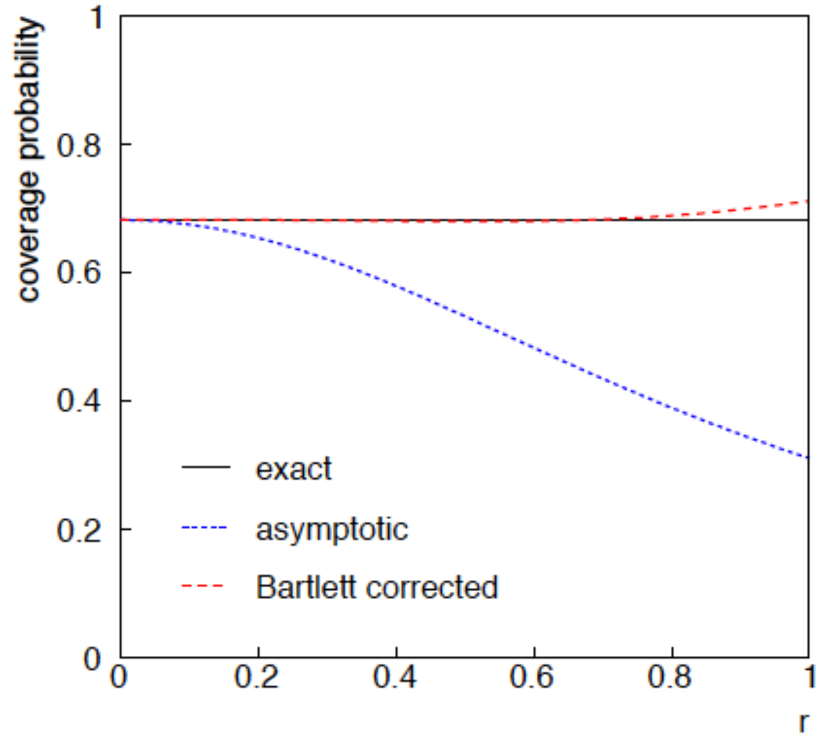
Bartlett corrections (2)

Good agreement for r ~ several tenths out to $\sqrt{t_\mu'}$ ~ several, i.e., good for significances of several sigma:



68.3% CL confidence interval for μ





Gamma model for estimates of variance

Suppose the estimated variance v was obtained as the sample variance from n observations of a Gaussian distributed bias estimate u.

In this case one can show v is gamma distributed with

$$\alpha = \frac{n-1}{2} \qquad \beta = \frac{n-1}{2\sigma_u^2}$$

We can relate α and β to the relative uncertainty r in the systematic uncertainty as reflected by the standard deviation of the sampling distribution of s, σ_s

$$r = \frac{\sigma_s}{E[s]} = \frac{1}{2}\frac{\sigma_v}{E[v]}$$

Exact relation between r parameter and relative error on error

r parameter defined as: $$r \equiv \frac{1}{2}\frac{\sigma_v}{E[v]} \approx \frac{\sigma_s}{E[s]}$$

$v \sim \text{Gamma}(\alpha, \beta)$ so $s = \sqrt{v}$ follows a Nakagami distribution

$$g(s|\alpha,\beta) = \left|\frac{dv}{ds}\right| f(v(s)|\alpha,\beta) = \frac{2\beta^{\alpha}}{\Gamma(\alpha)} s^{2\alpha-1} e^{-\beta s^2}$$

$$E[s] = \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)\sqrt{\beta}}$$

$$V[s] = \frac{\alpha}{\beta} - \frac{1}{\beta}\left(\frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}\right)^2$$

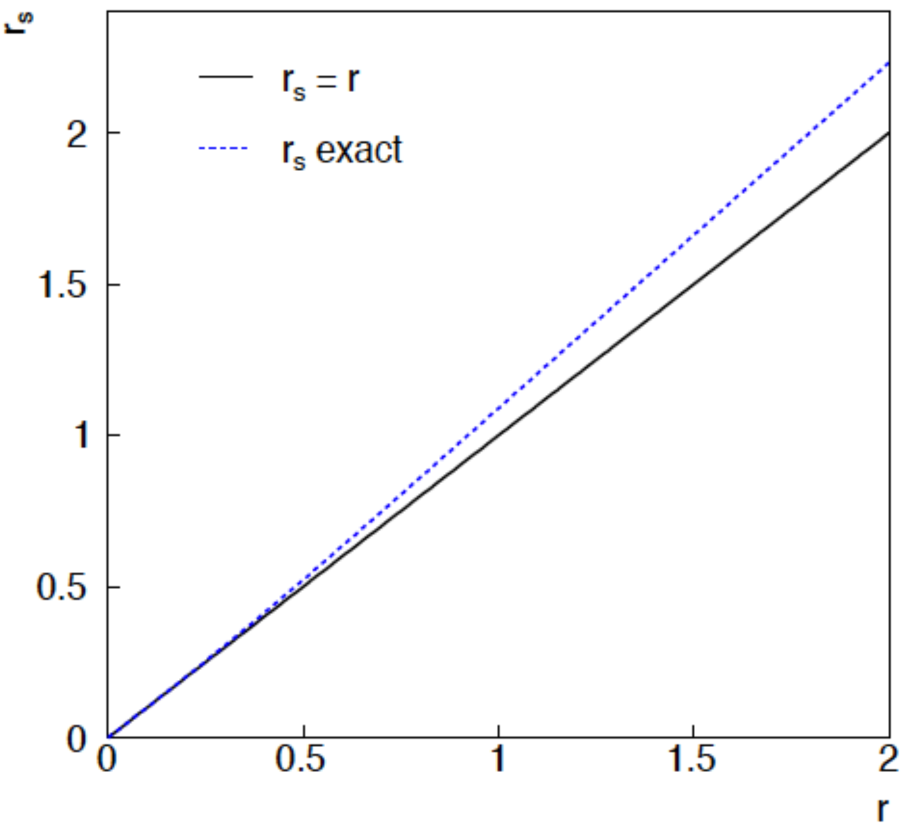
Exact relation between r parameter and relative error on error (2)

The exact relation between the error and the error r_s and the parameter r is therefore

$$r_s \equiv \frac{\sqrt{V[s]}}{E[s]}$$

$$= \sqrt{\alpha \left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + \frac{1}{2})} \right)^2 - 1}$$

$$\alpha = 1/4r^2$$



$\rightarrow$ $r_s \approx r$ good for $r \lessapprox 1$.

PDG scale factor

Suppose we do not want to take the quoted errors as known constants. Scale the variances by a factor ϕ,

$$\sigma_i^2 \to \phi \sigma_i^2$$

The likelihood function becomes

$$L(\mu, \phi) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\phi\sigma_i^2}} \exp\left[-\frac{1}{2} \frac{(y_i - \mu)^2}{\phi\sigma_i^2} \right]$$

The estimator for μ is the same as before; for ϕ ML gives

$$\hat{\phi}_{\mathrm{ML}} = \frac{\chi^2(\hat{\mu})}{N}$$ which has a bias; $$\hat{\phi} = \frac{\chi^2(\hat{\mu})}{N-1}$$ is unbiased.

The variance of $\hat{\mu}$ is inflated by ϕ:

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

Bayesian approach

G. Cowan, *Bayesian Statistical Methods for Parton Analyses*, in *Proceedings of the 14th International Workshop on Deep Inelastic Scattering (DIS2006)*, M. Kuze, K. Nagano, and K. Tokushuku (eds.), Tsukuba, 2006.

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

and (usually) covariances: $V_{ij}^{\text{stat}}, \; V_{ij}^{\text{sys}}$.

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

control variable parameters bias

Frequentist approach: $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

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

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 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!)

Bayesian approach with 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 $\sigma_s(s_i)$ reflects 'error on the error'.

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

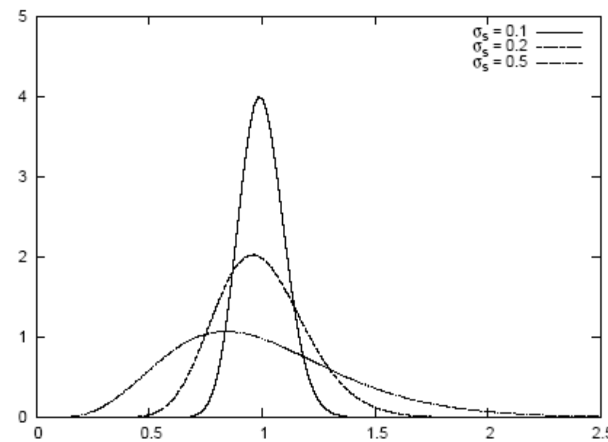
A simple unimodal probability density for $0 < s < 1$ with adjustable mean and variance is the Gamma distribution:

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

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

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

$\pi_s(s)$

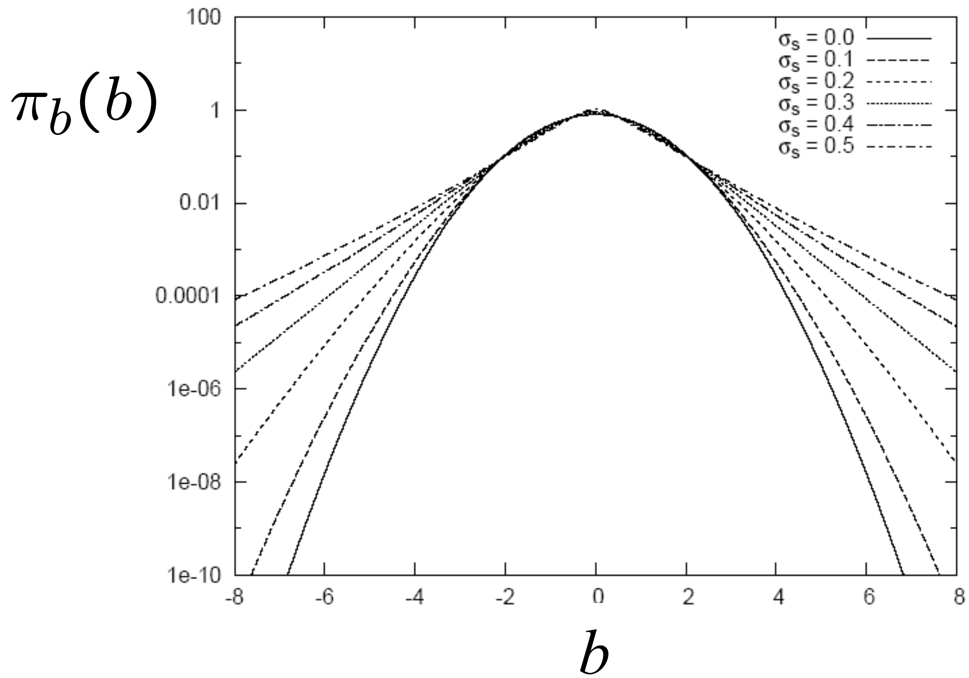


s

In fact if we took $\pi_s(s) \sim$ *inverse Gamma*, we could find $\pi_b(b)$ 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 ($\sigma_s = 0$) $\quad P(|b| > 4\sigma_{\text{sys}}) = 6.3 \times 10^{-5}$

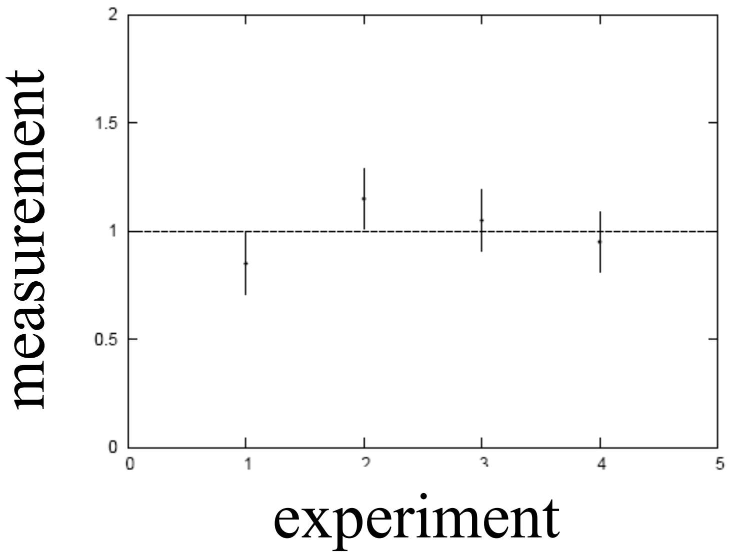
$\sigma_s = 0.5$ $\quad P(|b| > 4\sigma_{\text{sys}}) = 0.65\%$

A simple test

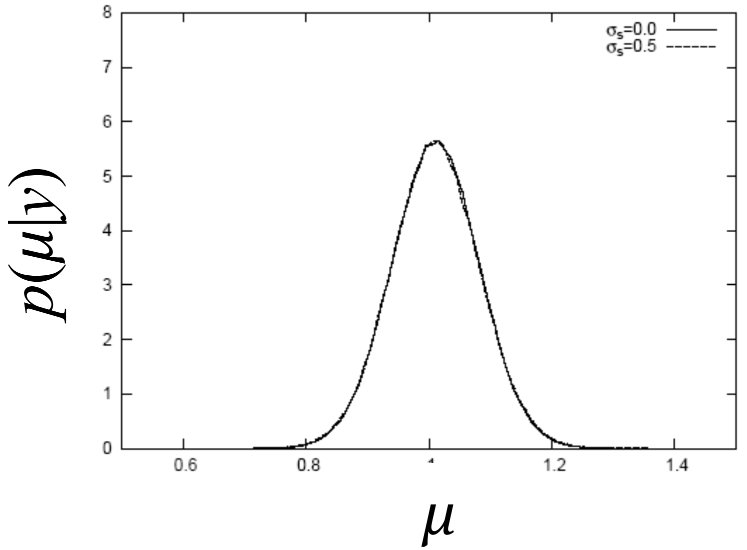
Suppose a fit effectively averages four measurements.

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

Case #1: data appear compatible



Posterior $p(\mu|y)$:



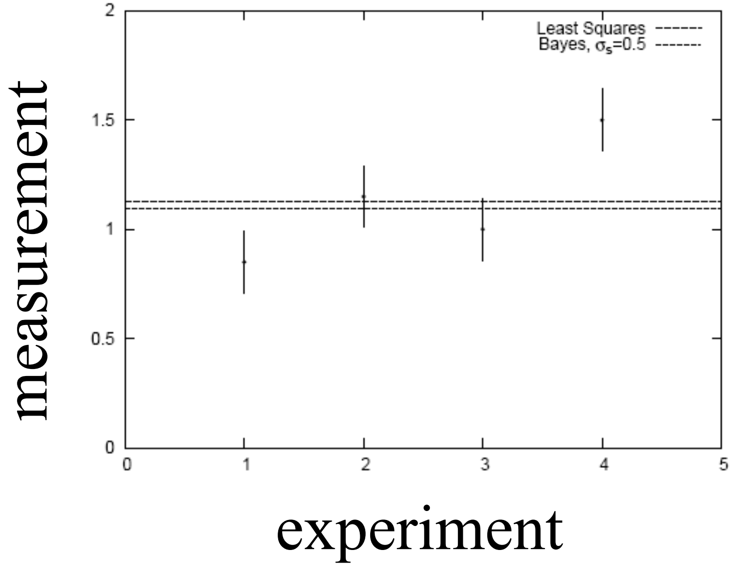
Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

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

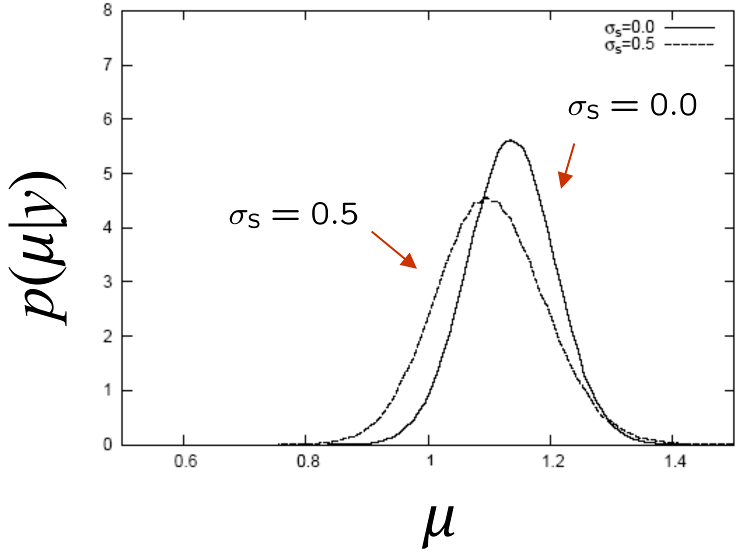
$\sigma_s = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$

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. See also

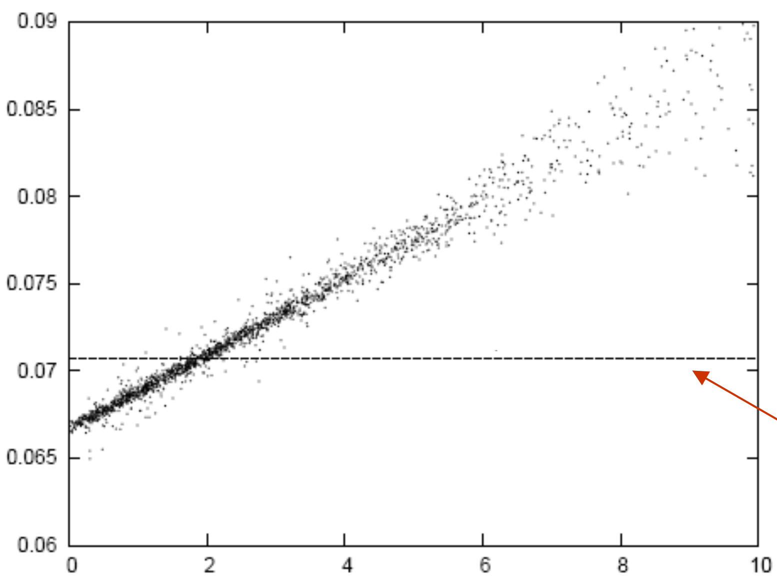
G. D'Agostini, *Sceptical combination of experimental results: General considerations and application to epsilon-prime/epsilon*, arXiv:hep-ex/9910036 (1999).

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).

post-erior σ_μ



χ^2

2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

σ_μ from least squares