Errors on Errors in Particle Physics

University of Manchester
Bohr Seminar, 13 Feb 2026

https://indico.global/event/16667/

MANCHESTER
1824
The University of Manchester

ROYAL HOLLOWAY UNIVERSITY OF LONDON

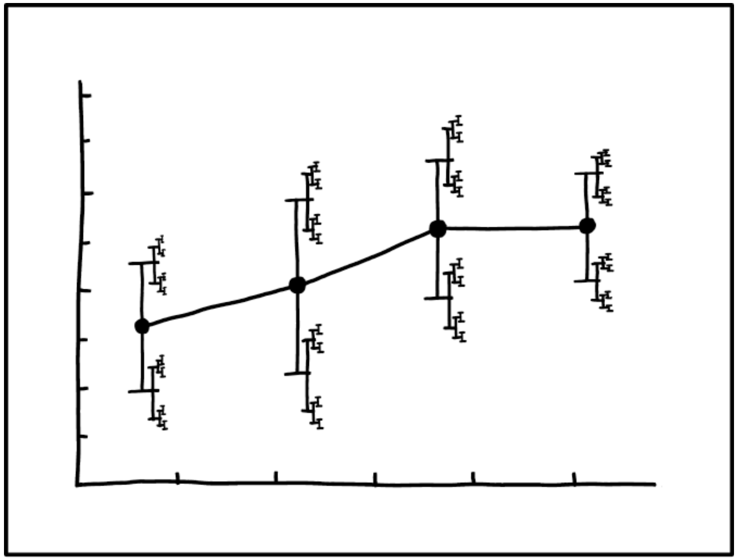
Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Preface

Why error bars on error bars?

- Not just an extra layer of complexity:
- Reduces sensitivity to outliers in fit/combination
- Gives connection between goodness of fit and size of confidence intervals

Randall Munroe, xkcd.com https://xkcd.com/2110/



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

Based on work esp. with Enzo Canonero:

G. Cowan, Eur. Phys. J. C (2019) 79:133; arXiv:1809.05778

G. Cowan, , EPJ Web of Conferences 258, 09002 (2022); arXiv:2107.02652

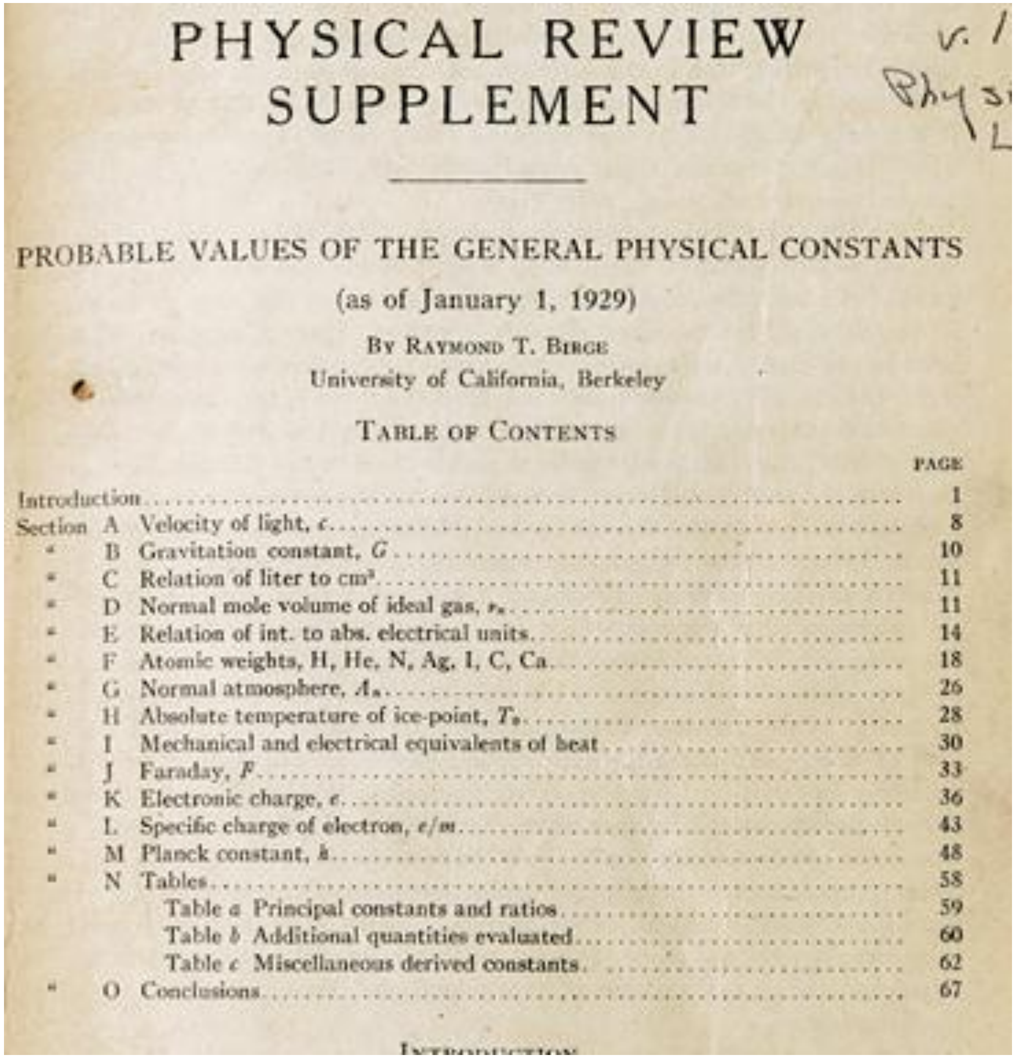
E. Canonero, A. Brazzale and G. Cowan, Eur. Phys. J. C (2023) 83:1100; arXiv:2304.10574

E. Canonero, G. Cowan, Eur. Phys. J. C (2025) 85: 156; arXiv:2407.05322

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

Least Squares for Averaging

= fit of horizontal line



PHYSICAL REVIEW SUPPLEMENT

v. 1

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

Least squares: some issues

The method of least squares requires the standard deviations of the measured quantities, but often these are poorly known.

The uncertainty (e.g. confidence interval) of an LS average does not reflect goodness of fit:

LS average of 9 ± 1 and 11 ± 1 is 10 ± 0.71

LS average of 5 ± 1 and 15 ± 1 is 10 ± 0.71

LS estimators are equivalent to maximum-likelihood assuming Gaussian distributed measurements; but the tails of a Gaussian fall off very fast, not always an appropriate model.

$\rightarrow$ Outliers in LS average have very large influence.

Solution: incorporate the uncertainty in the standard deviations of the measurements into the analysis.

"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 $\chi^2_{\min}$ = appropriate no. of degrees of freedom.

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 distribution 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 mean and width for $f(v)$:

$$E[v] = \sigma_u^2 = \alpha/\beta,$$

$\varepsilon = 1/2\sqrt{\alpha} \approx$ relative "error on the error" = $\sigma_s/E[s]$.

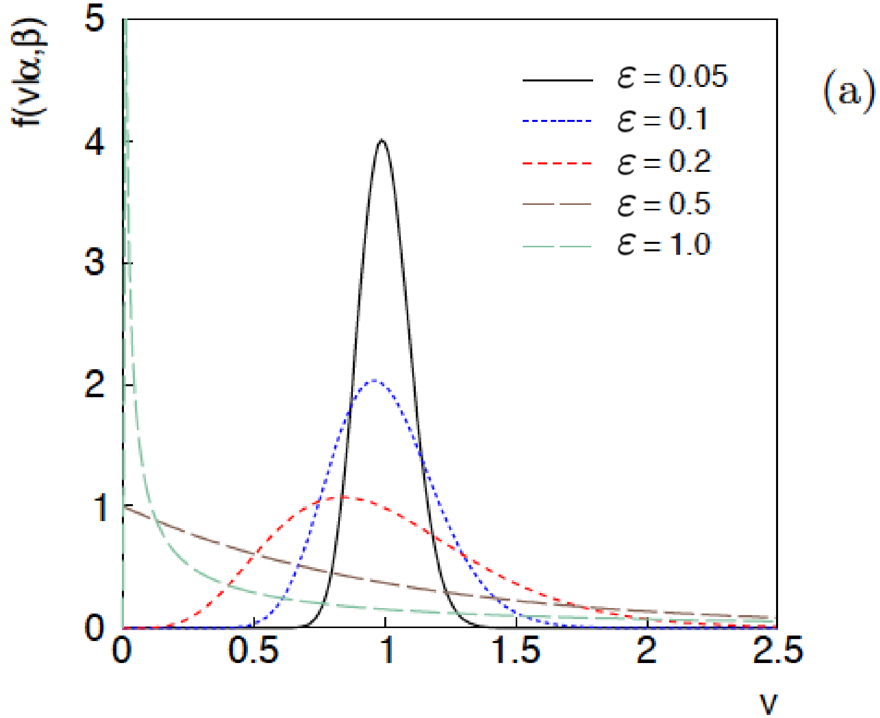
Distributions of v and $s = \sqrt{v}$

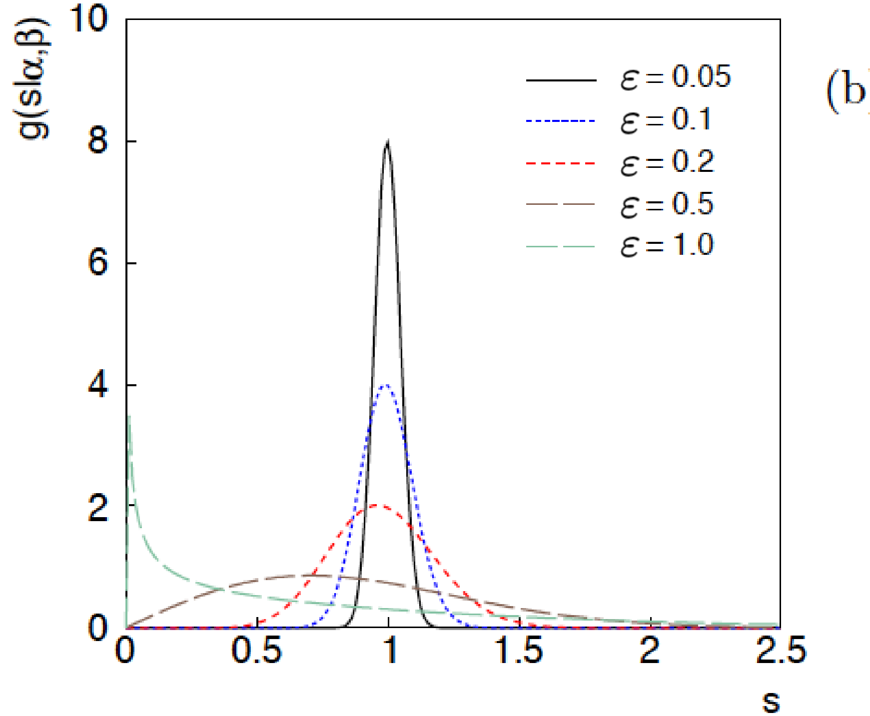
For α, β of gamma distribution,

$$\alpha_i = \frac{1}{4\varepsilon_i^2} \, , \; \beta_i = \frac{1}{4\varepsilon_i^2 \sigma_{u_i}^2}$$

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

relative “error on error”





Motivation for Gamma Variance Model (GVM)

If one were to have n independent observations $u_1,..,u_n$, with all $u \sim$ Gauss(θ, σ_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 variance model

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

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

$\alpha_i = 1/4\varepsilon_i^2$

$\beta_i = \alpha_i/\sigma_{ui}^2$

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: $\mu_1,\ldots,\mu_M$ (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: $\varepsilon_1,\ldots,\varepsilon_N$ (rel. err. in estimate of $\sigma_{u,i}$)

Profiling over systematic errors

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

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

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

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

In limit of small ε_i and $v_i \longrightarrow \sigma_{u,i}{}^2$, 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

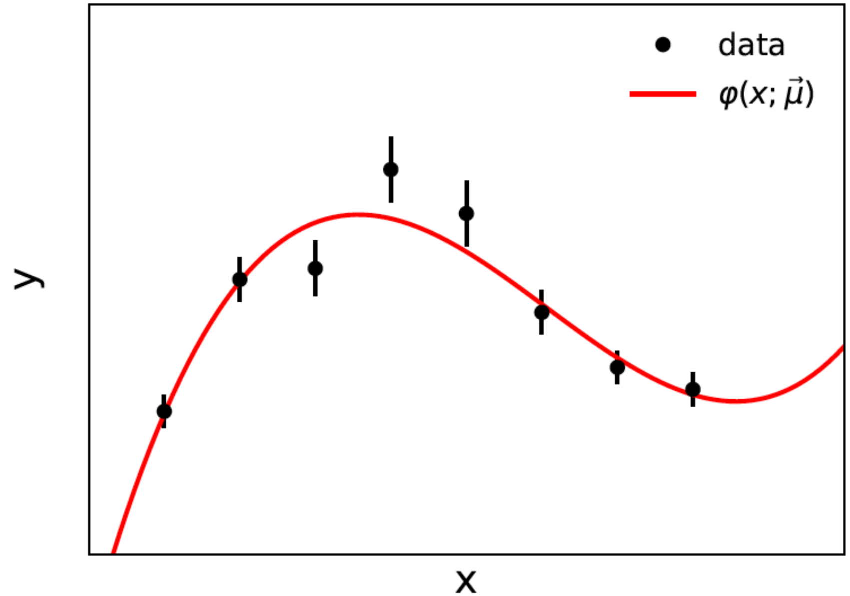
$$f(z_i|\nu_i) = \frac{\Gamma\left(\frac{\nu_i+1}{2}\right)}{\sqrt{\nu_i\pi}\Gamma(\nu_i/2)}\left(1+\frac{z_i^2}{\nu_i}\right)^{-\frac{\nu_i+1}{2}}$$

with $\nu_i = \frac{1}{2\varepsilon_i^2}$

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

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

$$V[y_i] = \sigma_{y_i}^2 \qquad \text{(known).}$$

$\boldsymbol{\mu}$ 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$(\theta_i, \sigma_{u,i})$, 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

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

Profiled values of θ_i from solution to cubic equations:

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

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

Goodness of fit

Can quantify goodness of fit with statistic

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

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

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

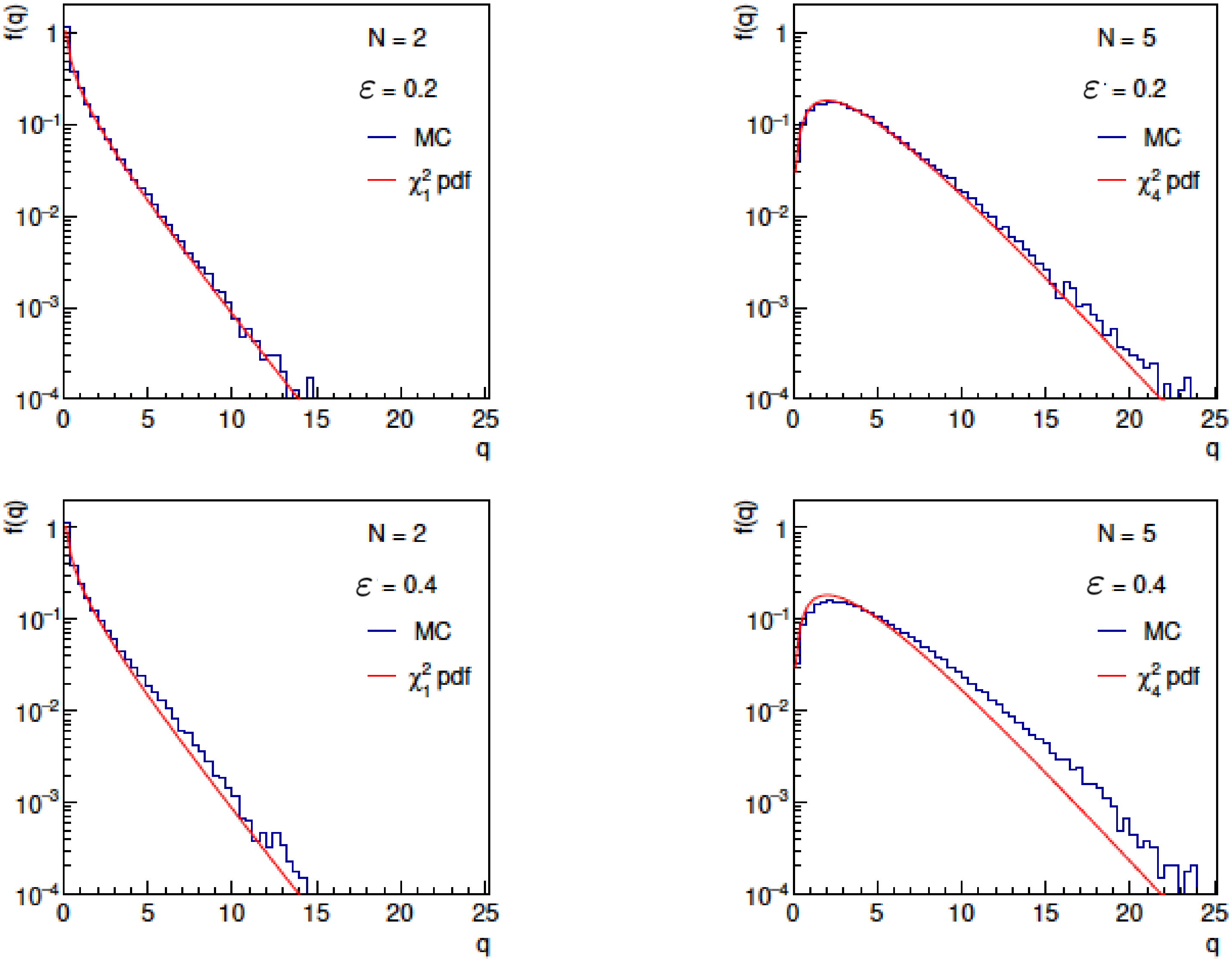
Asymptotically should have $q \sim$ chi-squared($N\!-\!M$).

For increasing ε_i, asymptotic distribution no longer valid.

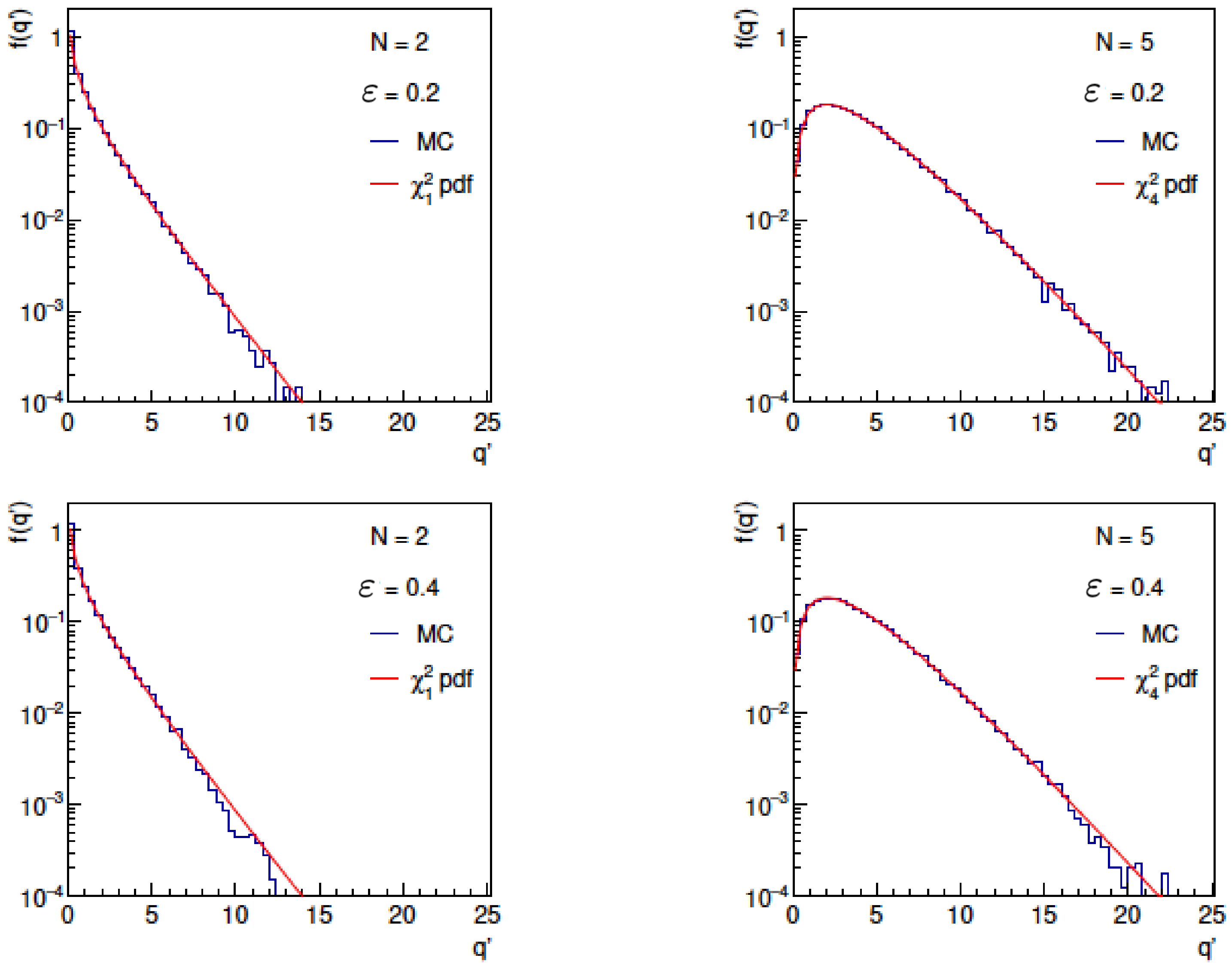
Bartlett (1937) defines modified statistic: $q' = \frac{n_\mathrm{d}}{E[q]} q$

By construction q' has mean $n_\mathrm{d} = N\!-\!M$ and turns out to have a distribution significantly closer to the asymptotic chi-square. (See Canonero et al., Eur. Phys. J. C (2023) 83:1100.)

Distributions of q

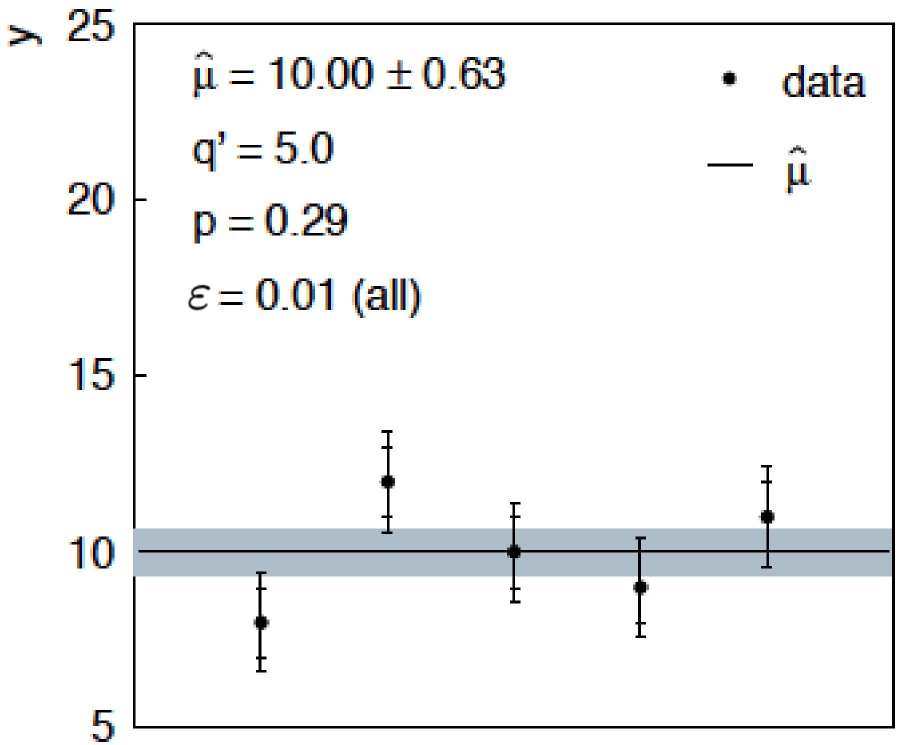


Distributions of Bartlett-corrected q'



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 $\varepsilon = 0.01$ for all).

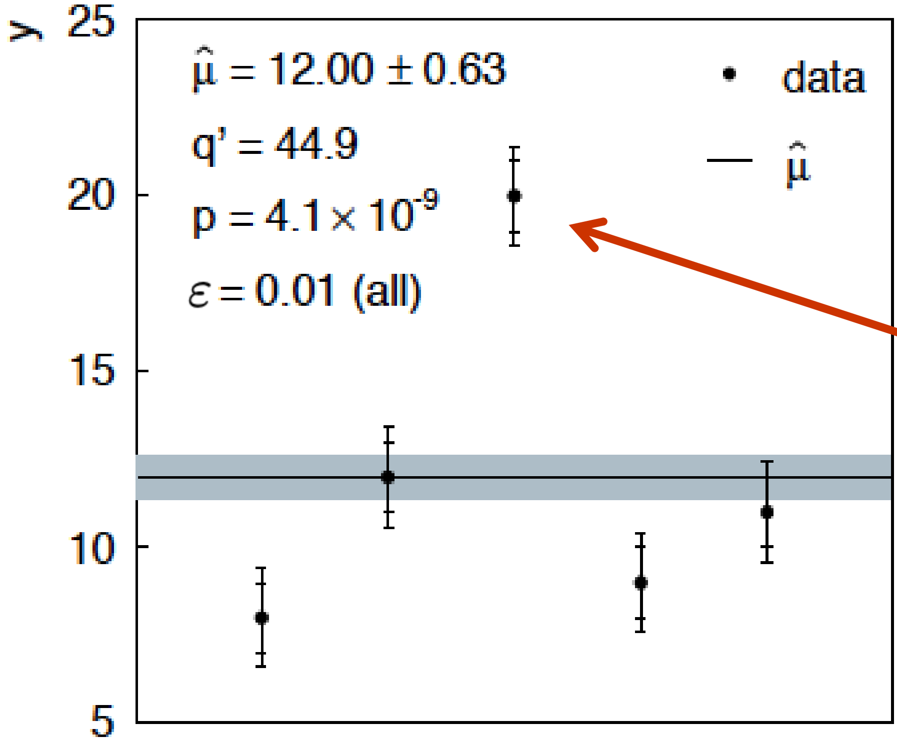


inner error bars $= \sigma_{y,i}$

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:

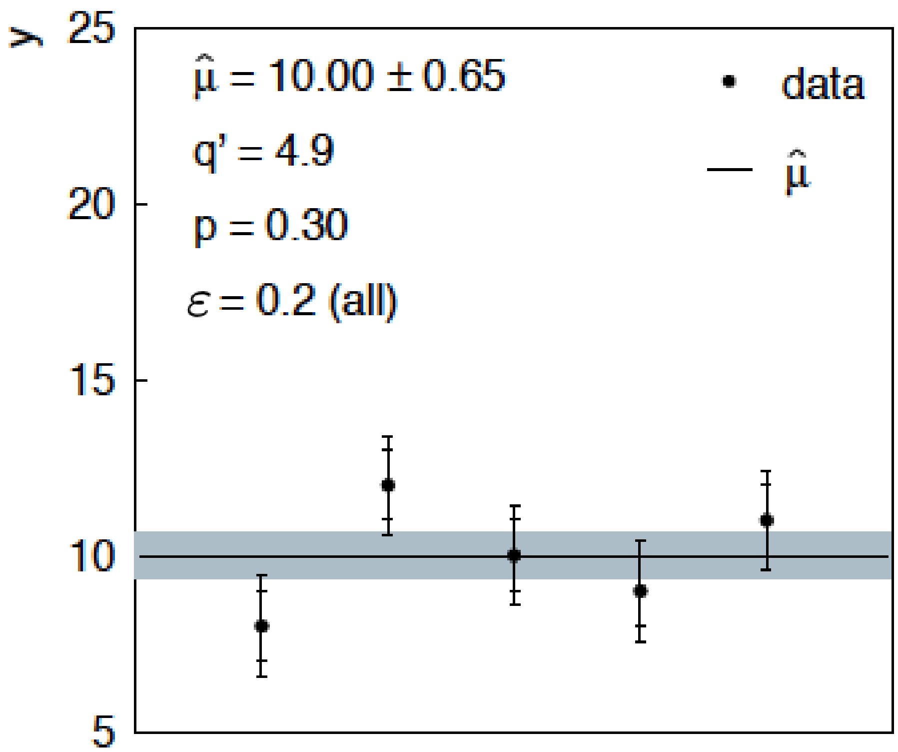


“outlier”

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

Average with all $\varepsilon = 0.2$

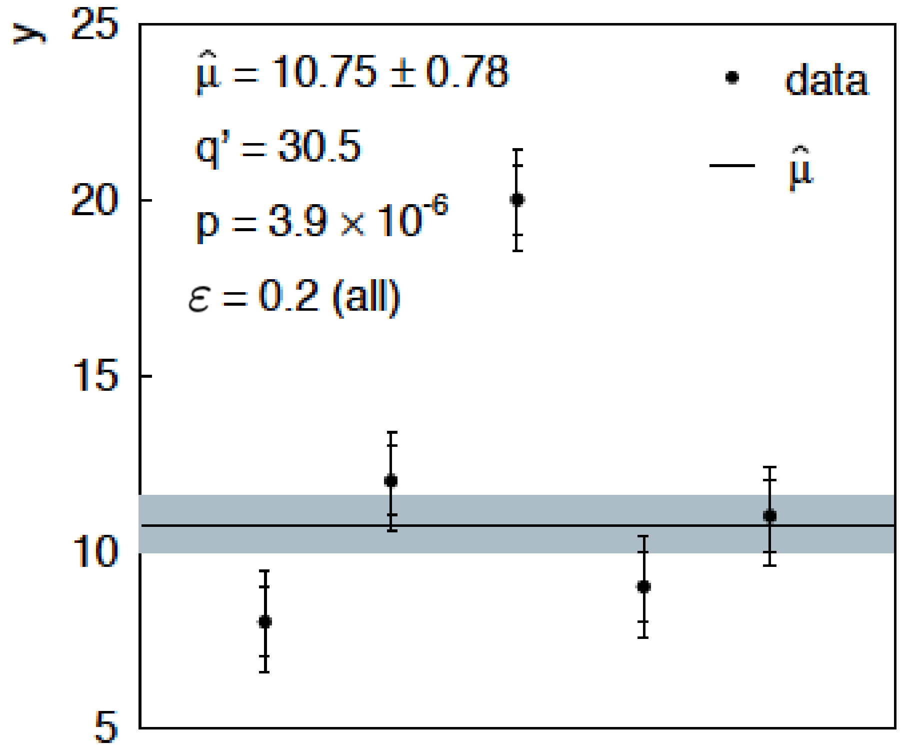
If we assign to each measurement $\varepsilon = 0.2$,



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

Average with all $\varepsilon = 0.2$ with outlier

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



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

Half-size of interval $\rightarrow 0.78$ (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.,

$$\sigma_{u_i} \to \sigma_{u_i}(1+\varepsilon_i)$$

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.

Application to top-quark mass

E. Canonero, G. Cowan, Eur. Phys. J. C (2025) 85: 156; arXiv:2407.05322 and
E. Canonero (thesis) inspirehep.net/literature/2971307

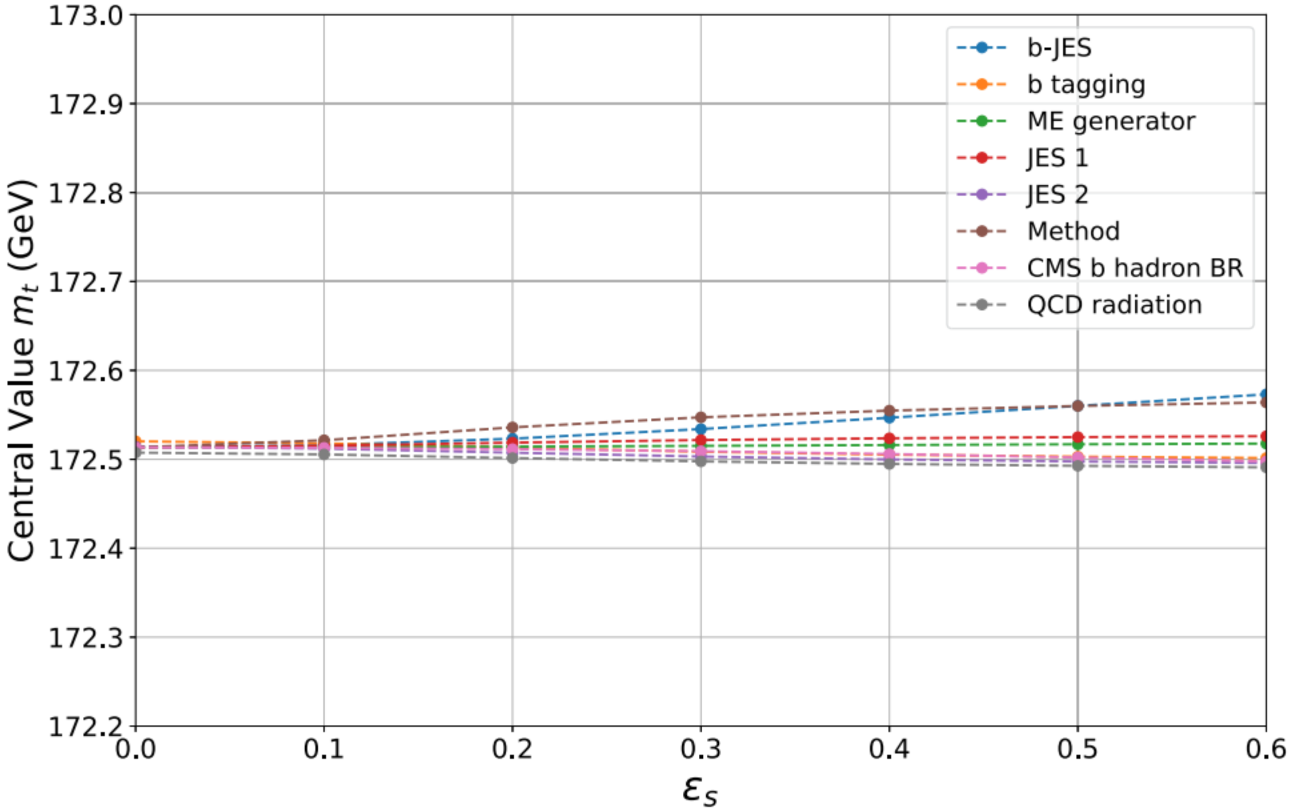
Based on: 2024 7–8 TeV ATLAS–CMS top quark mass combination, https://doi.org/10.1103/physrevlett.132.261902. arXiv:2402.08713

$t\bar{t}$ final state	CM Energy (TeV)	$m_{top} \pm$ (stat) $\pm$ (syst) (GeV)	Total uncertainty (GeV)
All-hadronic [34]	7	$173.49 \pm 0.69 \pm 1.23$	± 1.41
Dileptonic [35]	7	$172.50 \pm 0.43 \pm 1.52$	± 1.58
Lepton+jets [36]	7	$173.49 \pm 0.43 \pm 0.97$	± 1.06
All-hadronic [37]	8	$172.32 \pm 0.25 \pm 0.57$	± 0.62
Dileptonic [37]	8	$172.22 \pm 0.18 \pm 0.94$	± 0.95
Lepton+jets [37]	8	$172.35 \pm 0.16 \pm 0.45$	± 0.48
Single top [38]	8	$172.95 \pm 0.77 \pm 0.93$	± 1.20
J/ψ [39]	8	$173.50 \pm 3.00 \pm 0.94$	± 3.14
Secondary vertex [40]	8	$173.68 \pm 0.20 \pm 1.11$	± 1.12

average: $$m_t = 172.52 \pm 0.14\ (\mathrm{stat})\ \pm 0.30\ (\mathrm{syst})\ \mathrm{GeV}$$

Application to top-quark mass (2)

For the 8 largest systematics, apply errors-on-errors in turn:

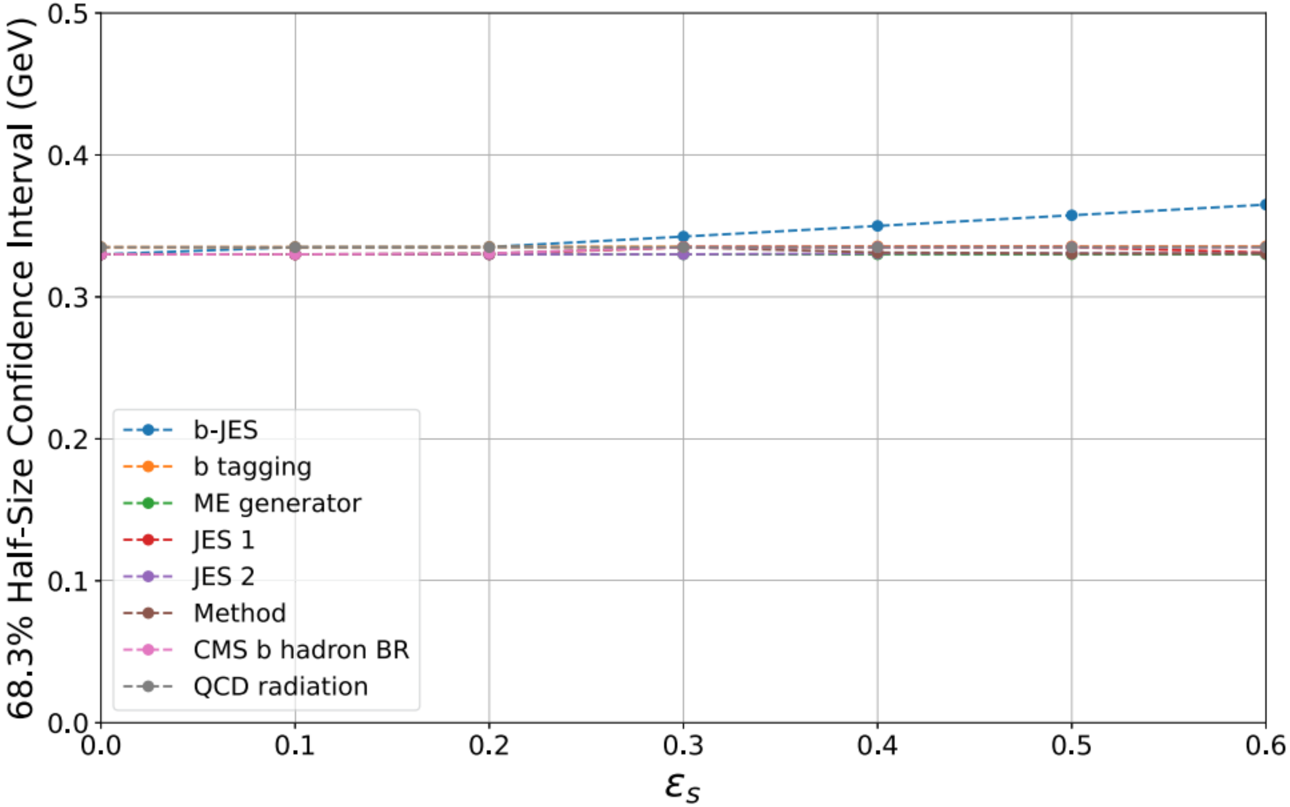


Estimated value of m_t quite stable.

$\varepsilon \longrightarrow 0$ reproduces published average.

Application to top-quark mass (3)

Size of error bar on average also fairly stable (expected since input values in relatively good agreement).

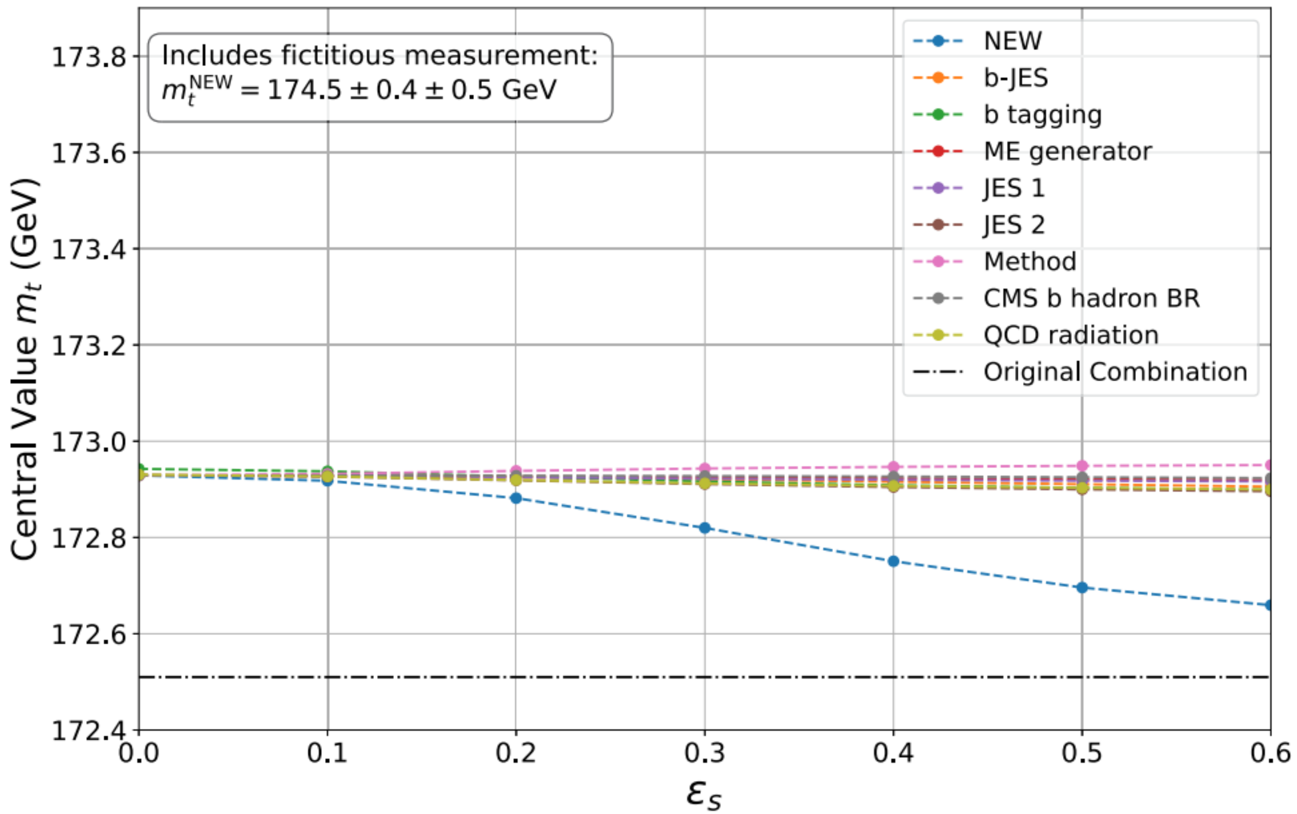


Conclusion: this m_t combination is robust wrt errors-on-errors.

Application to top-quark mass (4)

If there were to be an outlier, things would change:

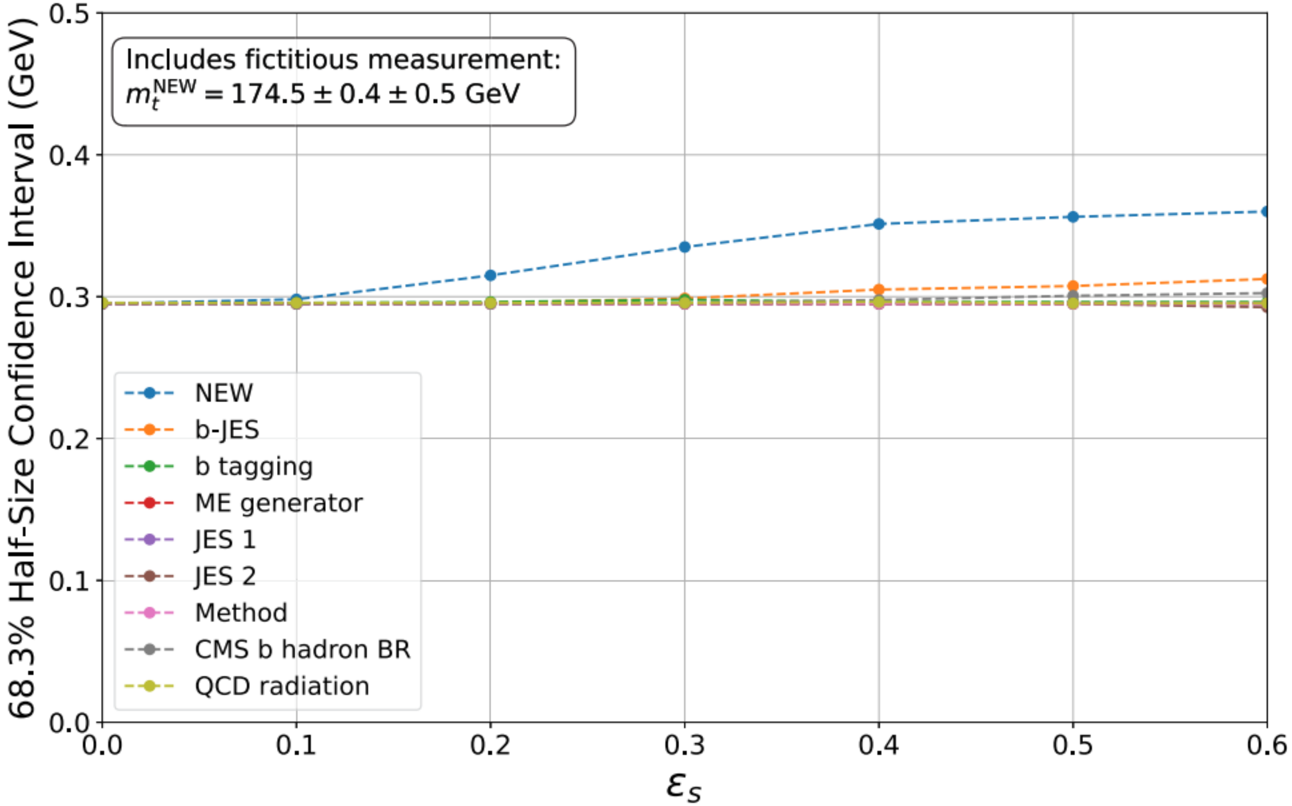
→ include <u>fictitious</u> measurement m_t = 174.5 ± 0.4 ± 0.5 GeV:



If assigned small ε, fictitious outlier pulls average significantly.

Application to top-quark mass (5)

And with the fictitious outlier, error on average is inflated provided one assumes the outliner has large enough ε.

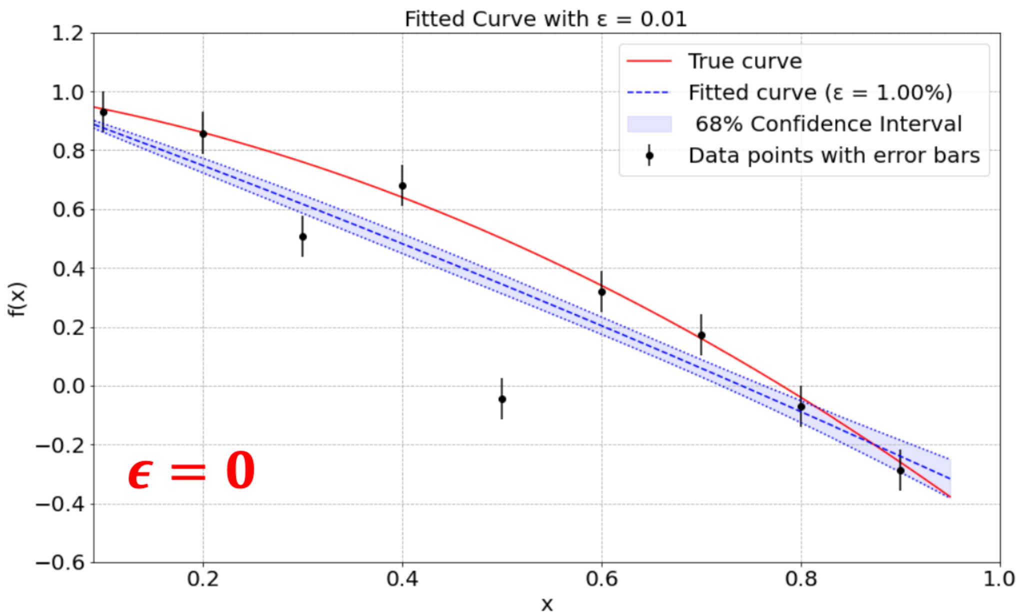


Fitting with outliers (e.g., parton fits)

Fitting of a curve: compatible measurements

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- Fit of a quadratic function with ***two outliers***



$$y_i \sim f(x_i) + \theta_i$$

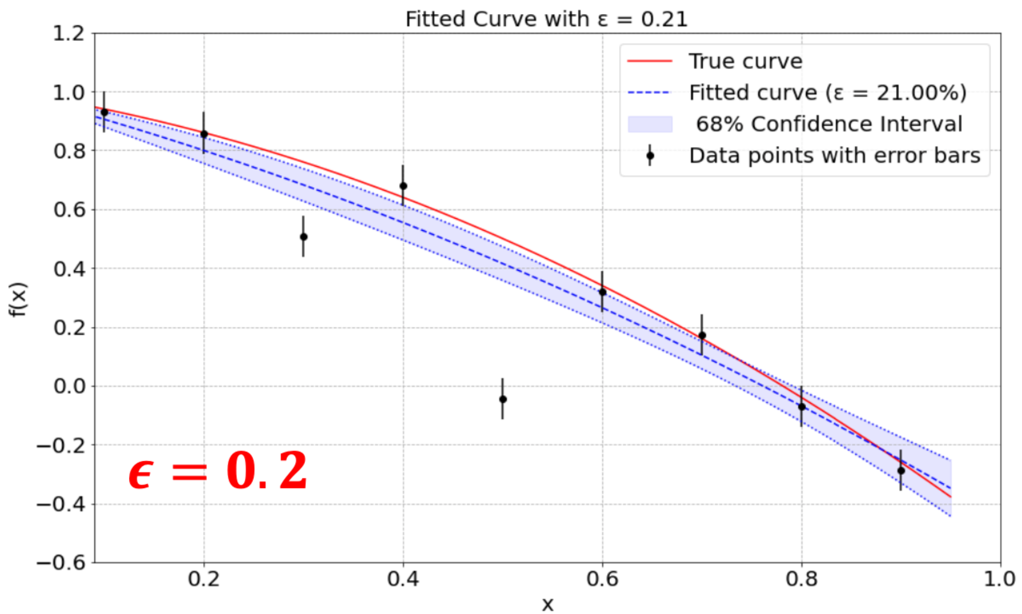
Params of interest

$$f(x_i) = \boldsymbol{a}x_i^2 + \boldsymbol{b}x + \boldsymbol{c}$$

Fitting of a curve: compatible measurements

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- Fit of a quadratic function with ***two outliers***



$$y_i \sim f(x_i) + \theta_i$$

Params of interest

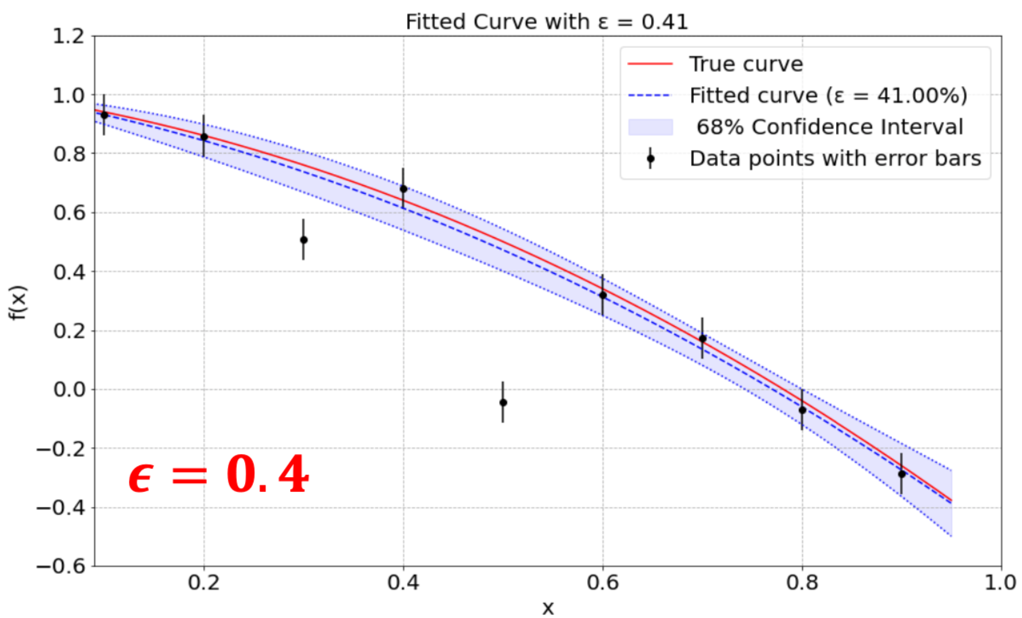
$$f(x_i) = \boldsymbol{a} x_i^2 + \boldsymbol{b} x + \boldsymbol{c}$$

26

Fitting of a curve: compatible measurements

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- Fit of a quadratic function with ***two outliers***



$$y_i \sim f(x_i) + \theta_i$$

Params of interest

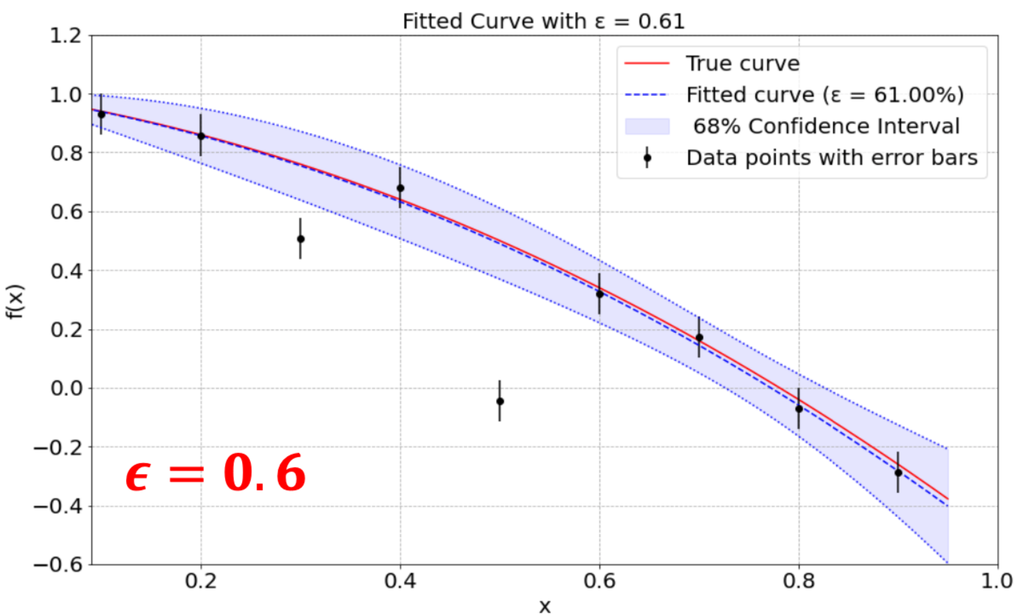
$$f(x_i) = ax_i^2 + bx + c$$

27

Fitting of a curve: compatible measurements

ROYAL HOLLOWAY UNIVERSITY OF LONDON

- Fit of a quadratic function with ***two outliers***



$$y_i \sim f(x_i) + \theta_i$$

Params of interest

$$f(x_i) = \ \boldsymbol{a} x_i^2 + \boldsymbol{b} x + \boldsymbol{c}$$

Conclusion: The model is sensitive to internal compatibility of the data

28

Recent and Ongoing Developments

With Enzo Canonero, esp. Eur. Phys. J. C (2025) 85:156

- Fully extended to problems with correlated systematic uncertainties.
- Fast approximations of expectation values for Bartlett-corrected goodness of fit and confidence intervals.
- Closed-form solutions for estimators ($\longrightarrow$ fast profile-likelihood fitting).
- github.com/EnzoCanonero/GVM-Combinations Combination Toolkit
- Applications to estimates of top, W masses: EC, GDC EPJC (2025) 85:156 and EC thesis: inspirehep.net/literature/2971307
- Investigation of higher-order asymptotics: E. Canonero, A. Brazzale and G. Cowan, Eur. Phys. J. C (2023) 83:1100; arXiv:2304.10574
- Application to muon g-2 anomaly: G. Cowan, , EPJ Web of Conferences 258, 09002 (2022); arXiv:2107.02652

Parton fits (ongoing w/ Enzo Canonero, Georgia Brown)

- Preliminary (EC): errors-on-errors branch in gitlab.cern.ch/fitters/xfitter

Discussion / Conclusions

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, $\nu = 1/2\varepsilon^2$ degrees of freedom.

Simple profile likelihood – quadratic terms replaced by logarithmic:

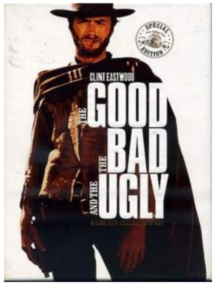
$$\frac{(u_i - \theta_i)^2}{\sigma_{u_i}^2} \quad \rightarrow \quad \left(1 + \frac{1}{2\varepsilon_i^2}\right) \ln\left[1 + 2\varepsilon_i^2 \frac{(u_i - \theta_i)^2}{v_i}\right]$$

Discussion / Conclusions (2)

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

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

Realistic approach: classify systematic uncertainties as:



good $\rightarrow \varepsilon = 0$ (error related to sample size)

bad $\rightarrow \varepsilon = 0.3$ (~justified but still uncertain)

ugly $\rightarrow \varepsilon = 0.6$ (2-point sys., theory uncertainties)

Could also use e.g. as "stress test" – crank up the ε_i values until significance of result degrades and ask if you really trust the assigned systematic errors at that level ($\rightarrow$ muon g-2, m_t, M_W,).

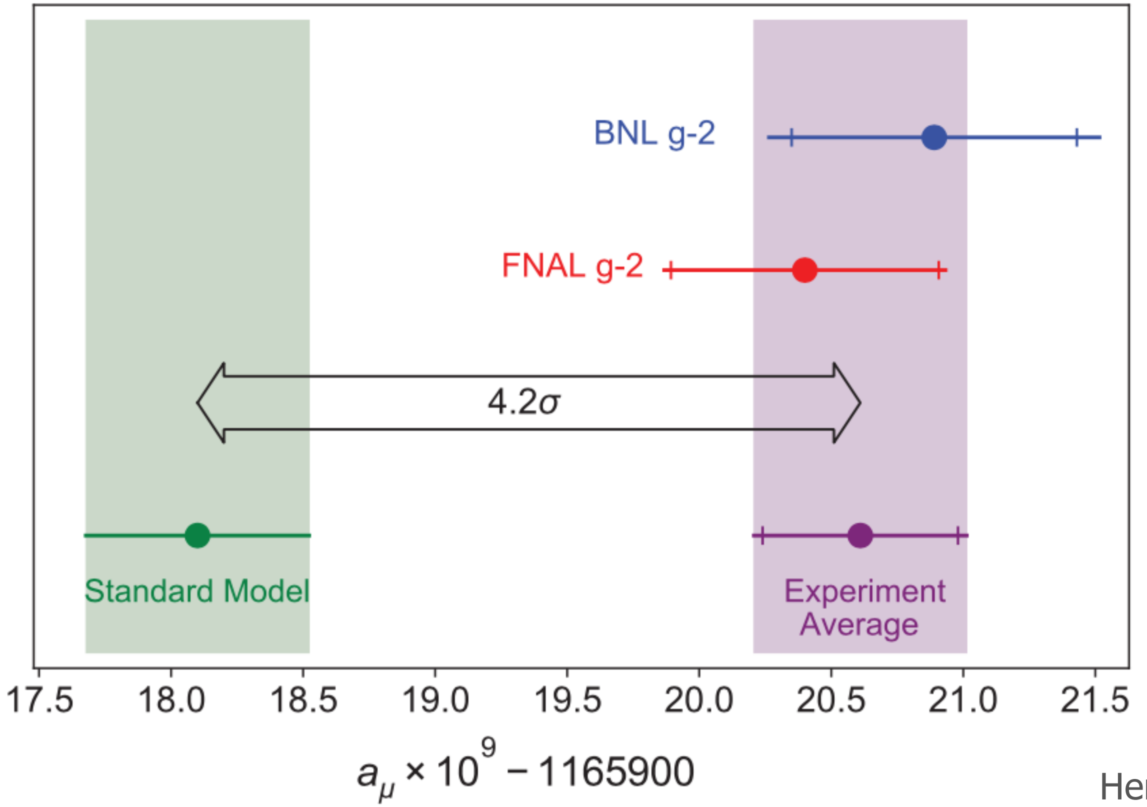
Ongoing: application to parton density fits, MC tuning,...

Extra Slides

Application to the muon $g-2$ anomaly

The recently measured muon $g-2$ (ave. of 2006, 2021) disagrees with the Standard Model prediction with a significance of 4.2σ.

Muon g-2 Collab., PRL 126, 141801 (2021)



Discrepancy significantly reduced by 2021 lattice-based prediction of Borsanyi et al. (BMW).

Current goal is to investigate sensitivity of significance to error assumptions, so for now focus on the 4.2σ problem.

Here using 2021 meausrement; see also D. P. Aguillard *et al.* (The Muon g−2 Collaboration) Phys. Rev. Lett. **131**, 161802 (2023)

Muon $g-2$ ingredients

Using $a_\mu = (g-2)/2$ $y = a_\mu \times 10^9 - 1165900$

the ingredients of the 4.2σ effect are:

$$\boxed{y_{\rm exp} = 20.61 \pm 0.41}$$

(ave. of BNL 2006 and FNAL 2021)

0.37 (stat.) ± 0.17 (sys.)

B. Abi et al. (Muon $g-2$ Collaboration), *Measurement of the Positive Muon Anomalous Magnetic Moment to 0.46 ppm*, Phys. Rev. Lett. 126, 141801 (2021).

G. W. Bennett et al. (Muon $g-2$ Collaboration), *Final report of the E821 muon anomalous magnetic moment measurement at BNL*, Phys. Rev. D 73, 072003 (2006).

$$\boxed{y_{\rm SM} = 18.10 \pm 0.43}$$

(SM pred. by Muon $g-2$ theory initiative)

0.40 (Had. Vac. Pol.) ± 0.18 (Had. Light-by-Light)

T. Aoyama, N. Asmussen, M. Benayoun, J. Bijnens, and T. Blum et al., *The anomalous magnetic moment of the muon in the standard model*, Phys. Rep. 887, 1 (2020).

Suppose σ_{SM} uncertain

Suppose measurement errors well known, but that the SM theory error σ_{SM} (estimated 0.43) could be uncertain.

This is the largest systematic and probably hardest to estimate.

Treat estimate $v_{\mathrm{SM}} = (0.43)^2$ of variance σ^2_{SM} as gamma distributed, width from relative uncertainty parameter r_{SM}.

Maximum-likelihood for mean from minimum of

$$Q(\mu) = -2 \ln \frac{L(\mu)}{L_{\mathrm{sat}}}$$

$$= \frac{(y_{\mathrm{exp}} - \mu)^2}{\sigma^2_{\mathrm{exp}}} + \left(1 + \frac{1}{2r^2_{\mathrm{SM}}}\right) \ln \left[1 + 2r^2_{\mathrm{SM}} \frac{(y_{\mathrm{SM}} - \mu)^2}{v_{\mathrm{SM}}}\right]$$

p-value/significance of common-mean hypothesis

Significance (goodness of fit) from $q = Q(\hat{\mu})$.

Because of non-quadratic term in $Q(\mu)$, distribution of q departs from chi-square(1) for increasing $r_{\rm SM}$.

Best to get distribution of q from Monte Carlo (and speed up with Bartlett correction – see EPJC (2019) 79:133).

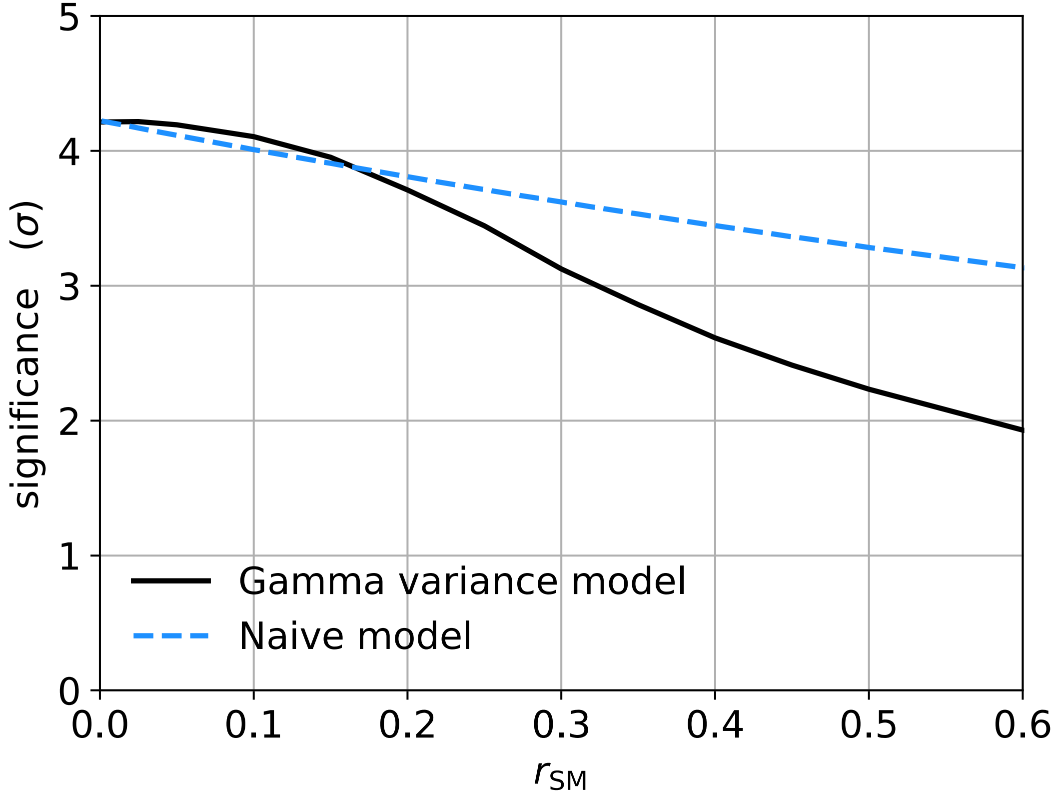
For $r_{\rm SM} > 0$ distribution of q depends on $\sigma^2{}_{\rm SM}$. For MC use Maximum-Likelihood estimate ("profile construction"):

$$\widehat{\sigma^2}_{\rm SM} = \frac{v_{\rm SM} + 2r_{\rm SM}^2 (y_{\rm SM} - \hat{\mu})^2}{1 + 2r_{\rm SM}^2}$$

$$\text{MC} \;\rightarrow\; f(q) \;\rightarrow\; p = \int_{q,\text{obs}}^{\infty} f(q)\, dq \;\rightarrow\; \text{significance } Z = \Phi^{-1}(1 - p/2)$$

# of sigmas

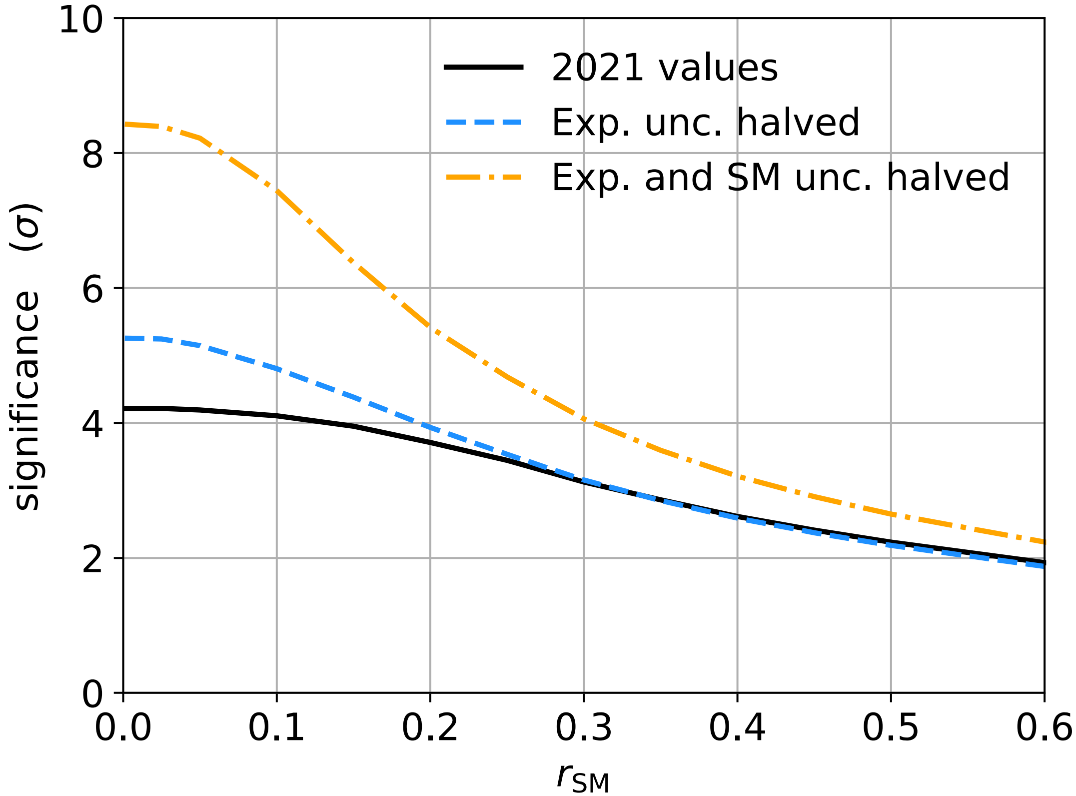
Significance of discrepancy versus r_{SM}



Naive model: use least squares but let $\sigma_{\mathrm{SM}} \rightarrow (1 + r_{\mathrm{SM}})\sigma_{\mathrm{SM}}$

Gamma variance model gives greater decrease in significance for $r_{\mathrm{SM}} \gtrsim 0.2$, e.g., 3.1σ for $r_{\mathrm{SM}} = 0.3$, 2.0σ for $r_{\mathrm{SM}} = 0.6$.

Significance of discrepancy versus r_{SM}



Establishing 4σ effect requires $r_{\mathrm{SM}} \lesssim 0.3$ even if nominal exp. and SM uncertainties become half of present values.

Discussion on muon $g\!-\!2$

Including uncertainties on estimates of uncertainties can have large effect on hypothesis test, esp. for high significance.

To establish e.g. a 5σ effect it is crucial to have both:

small uncertainties

accurate estimates of those uncertainties ($\sim$ 20% level)

This is ultimately because the tails of the Gaussian fall off so quickly.

Gamma Variance Model $\sim$ Student's t likelihood with $\nu = 1/2r^2$ degrees of freedom $\rightarrow$ longer tails than Gaussian.

Ongoing discussion with Bogdan Malaescu of Muon g-2 Theory Initiative on the HVP uncertainty, see, e.g.,

B. Malaescu et al., https://indico.him.uni-mainz.de/event/11/contributions/80/attachments/50/51/amuWorkshop_Correlations_Malaescu.pdf

M. Davier et al., Eur. Phys. J. C 80 (2020) 241 , arXiv:1908.00921

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

$y \sim$ Gauss(μ, σ^2),

$v \sim$ Gamma(α, β), $\qquad \alpha = \frac{1}{4\varepsilon^2}\,, \quad \beta = \frac{1}{4\varepsilon^2\sigma^2}$

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

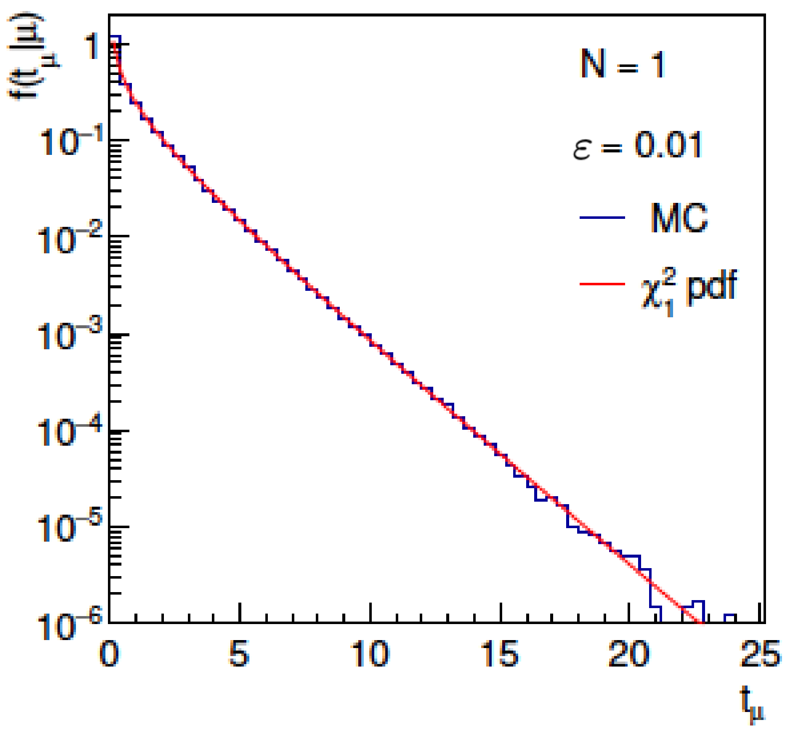
Test values of μ with $t_\mu = -2\ln\lambda(\mu)$ with $\lambda(\mu) = \frac{L(\mu, \widehat{\widehat{\sigma^2}}(\mu))}{L(\hat{\mu}, \widehat{\sigma^2})}$

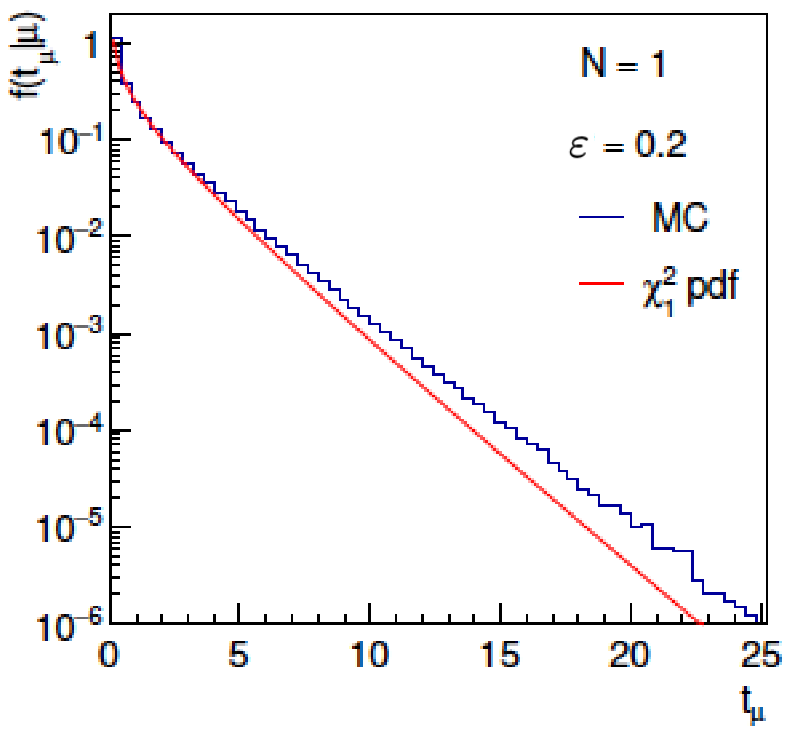
$$t_\mu = \left(1 + \frac{1}{2\varepsilon^2}\right) \ln\left[1 + 2\varepsilon^2 \frac{(y-\mu)^2}{v}\right]$$

Distribution of t_μ

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

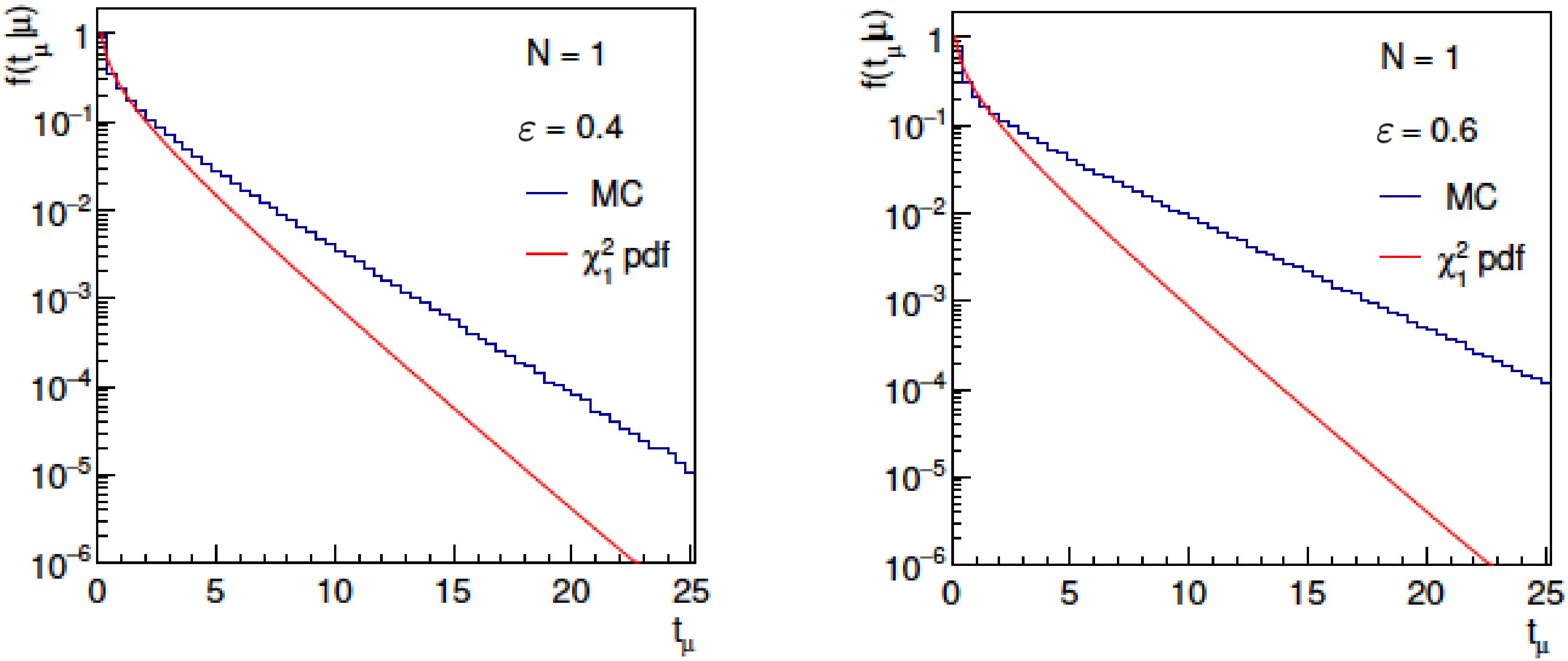
Here "asymptotic limit" means all estimators ~Gauss, which means $\varepsilon \to 0$. For increasing ε, clear deviations visible:





Distribution of t_μ (2)

For larger ε, breakdown of asymptotics gets worse:



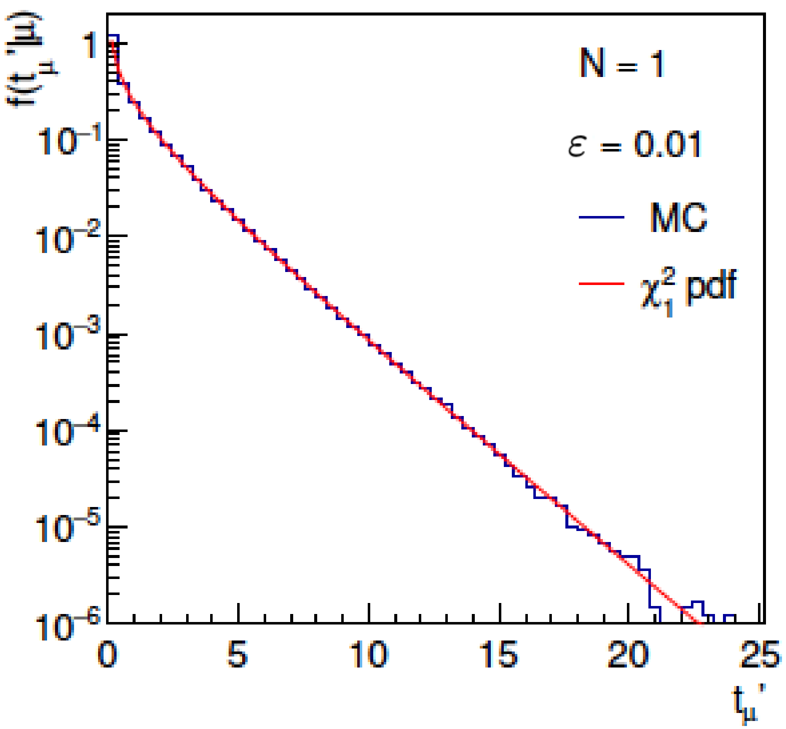
Values of $\varepsilon \sim$ several tenths are relevant so we cannot in general rely on asymptotics to get confidence intervals, p-values, etc.

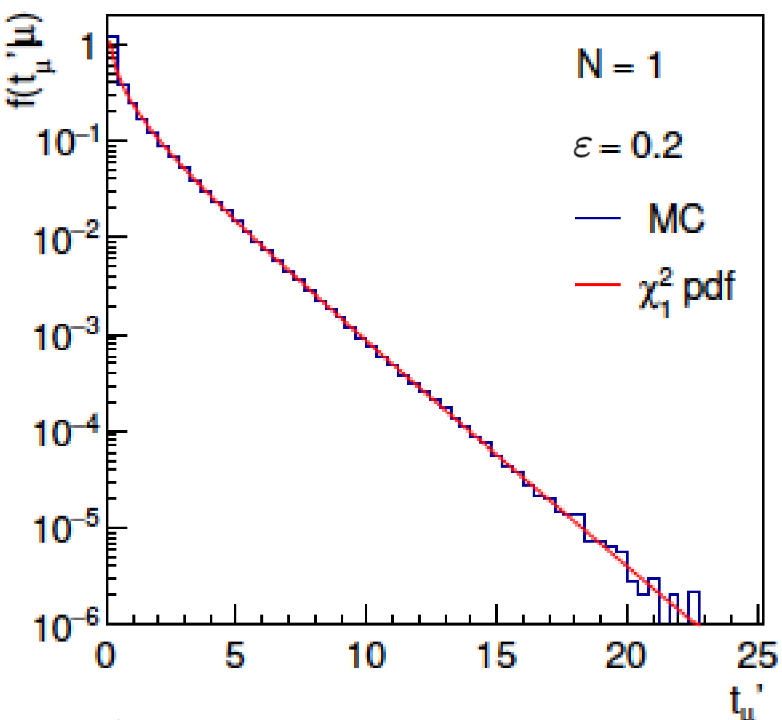
Bartlett corrections

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).

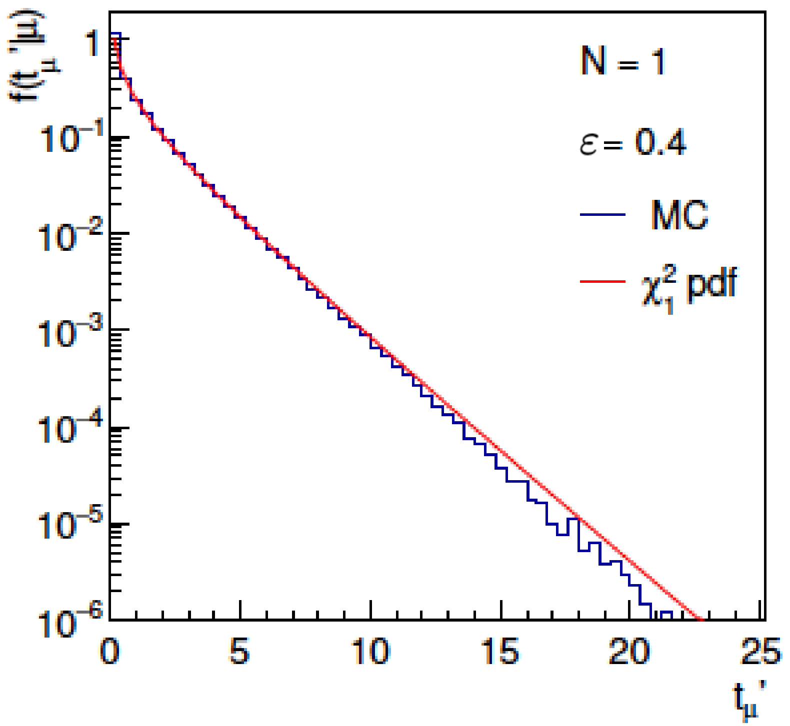
For this example $E[t_\mu] \approx 1 + 3\varepsilon^2 + 2\varepsilon^4$ works well:

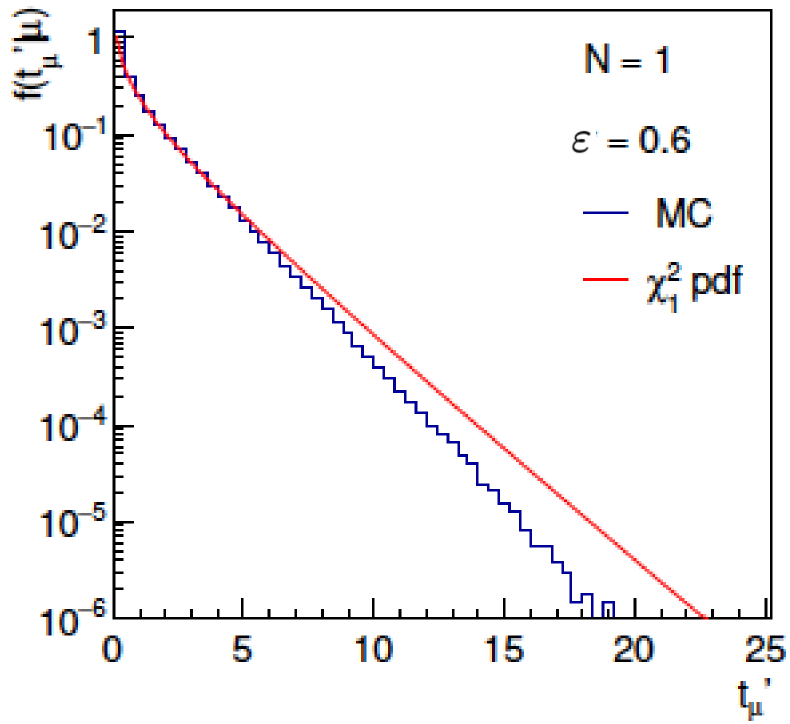




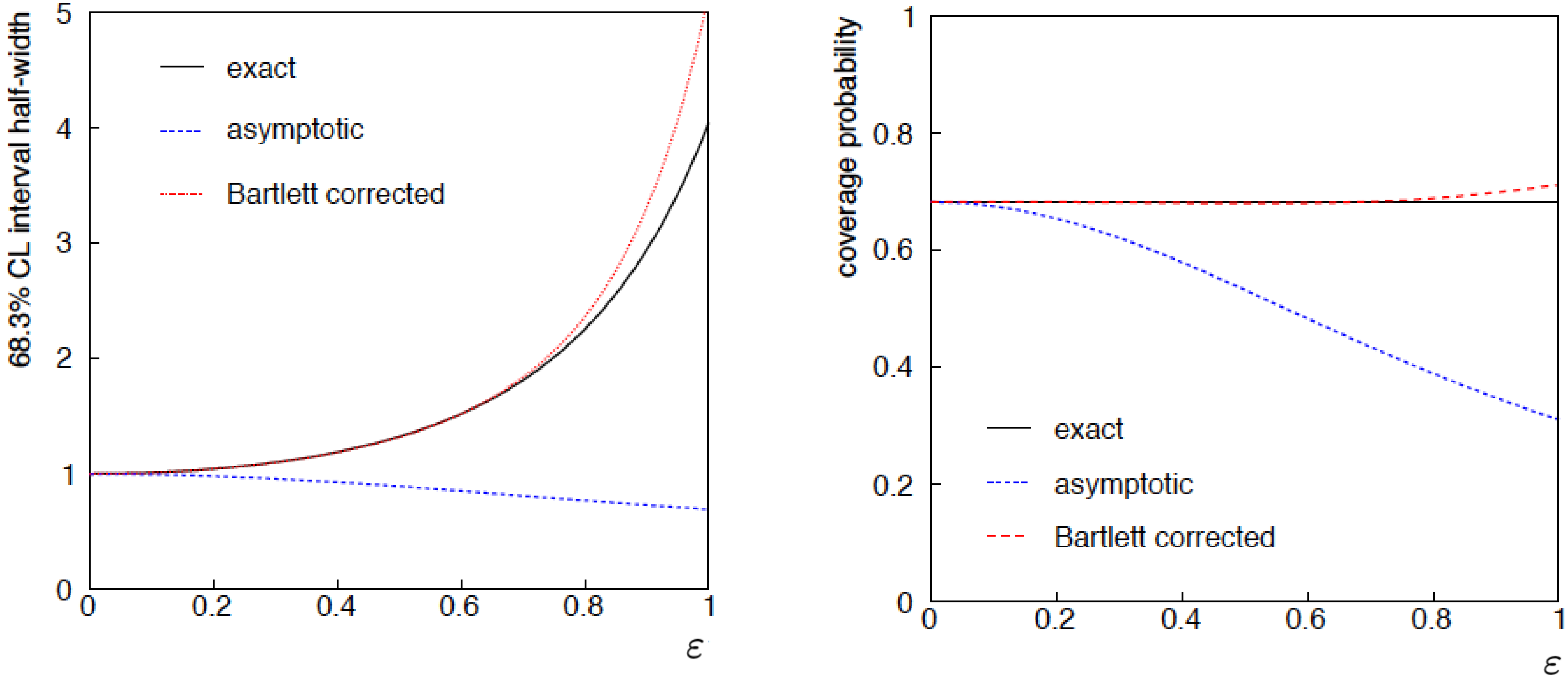
Bartlett corrections (2)

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





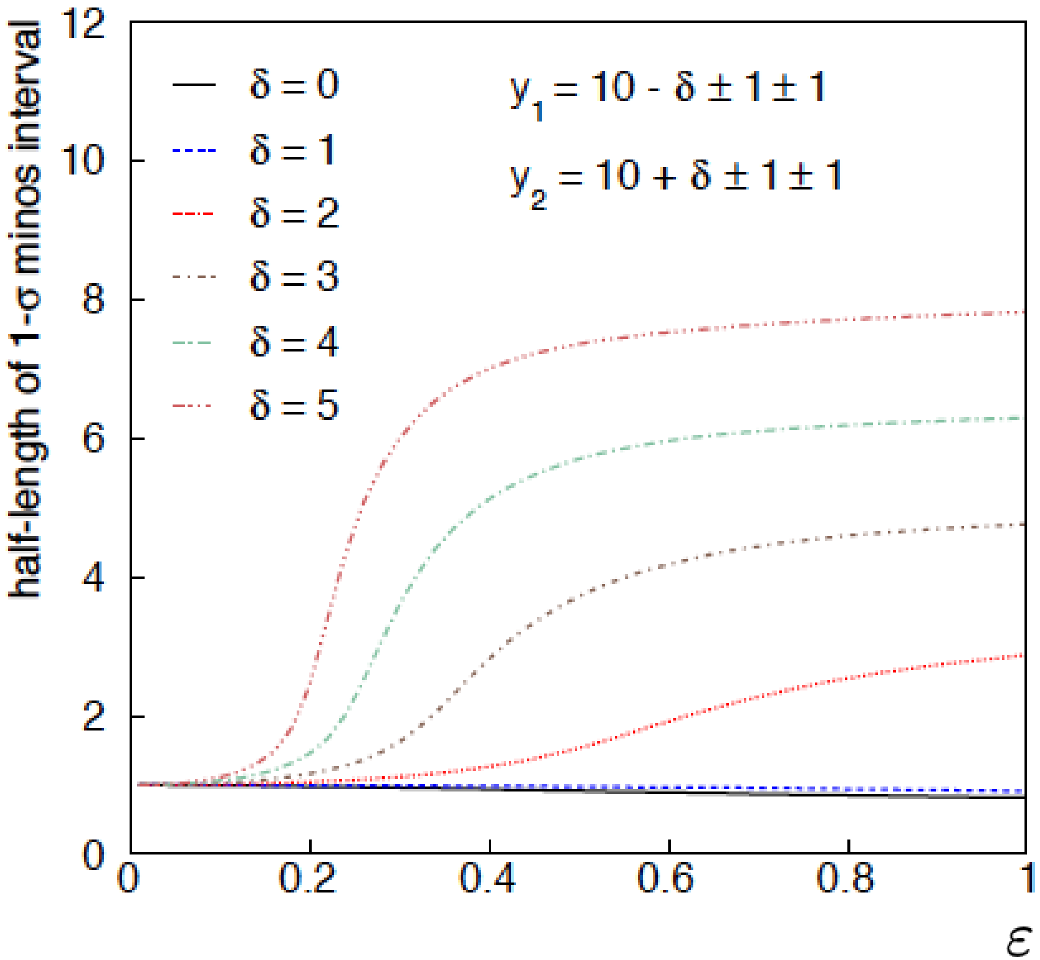
68.3% CL confidence interval for μ



Example: average of two measurements

Approximate ("MINOS") confidence interval based on

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

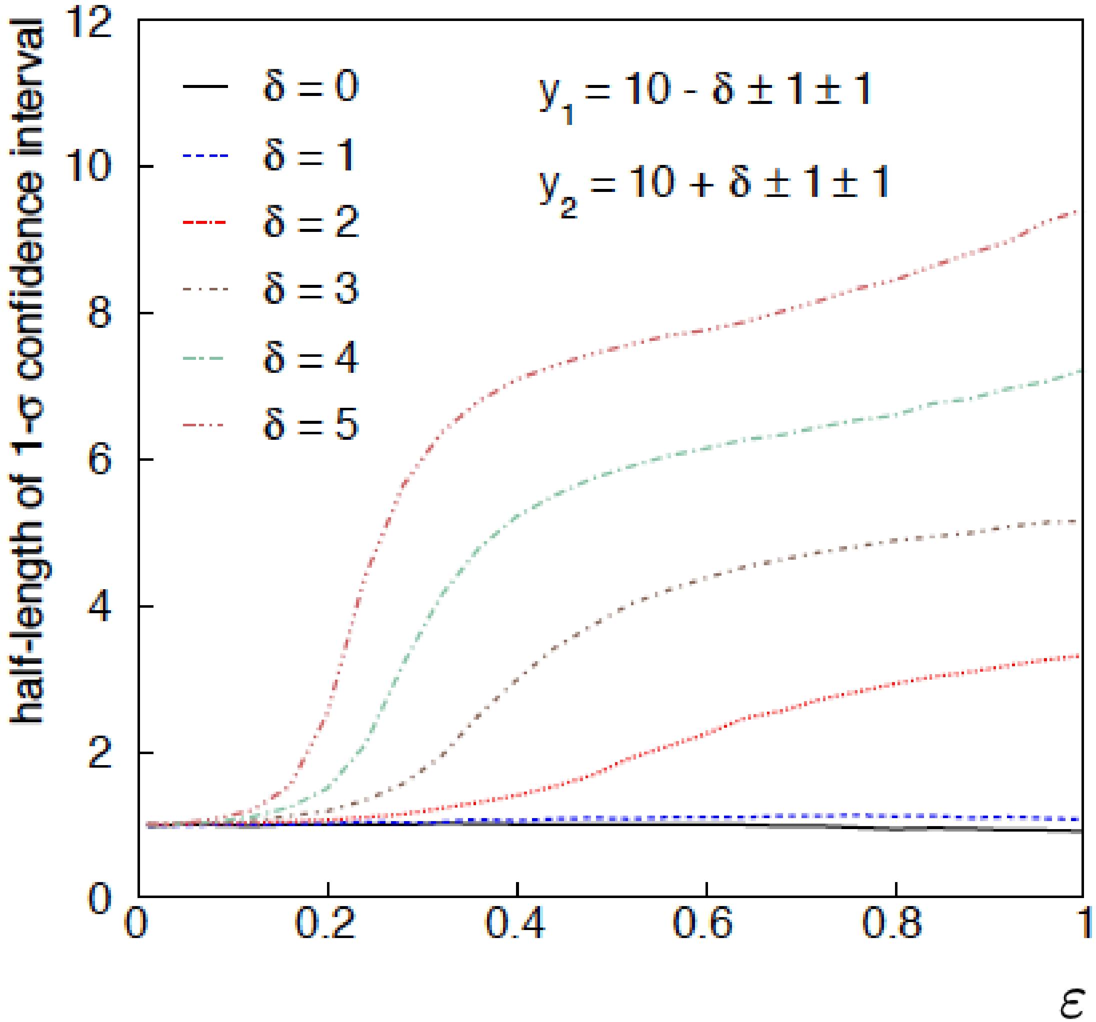


relative error on sys. error

Increased discrepancy between values to be averaged gives larger interval.

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

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

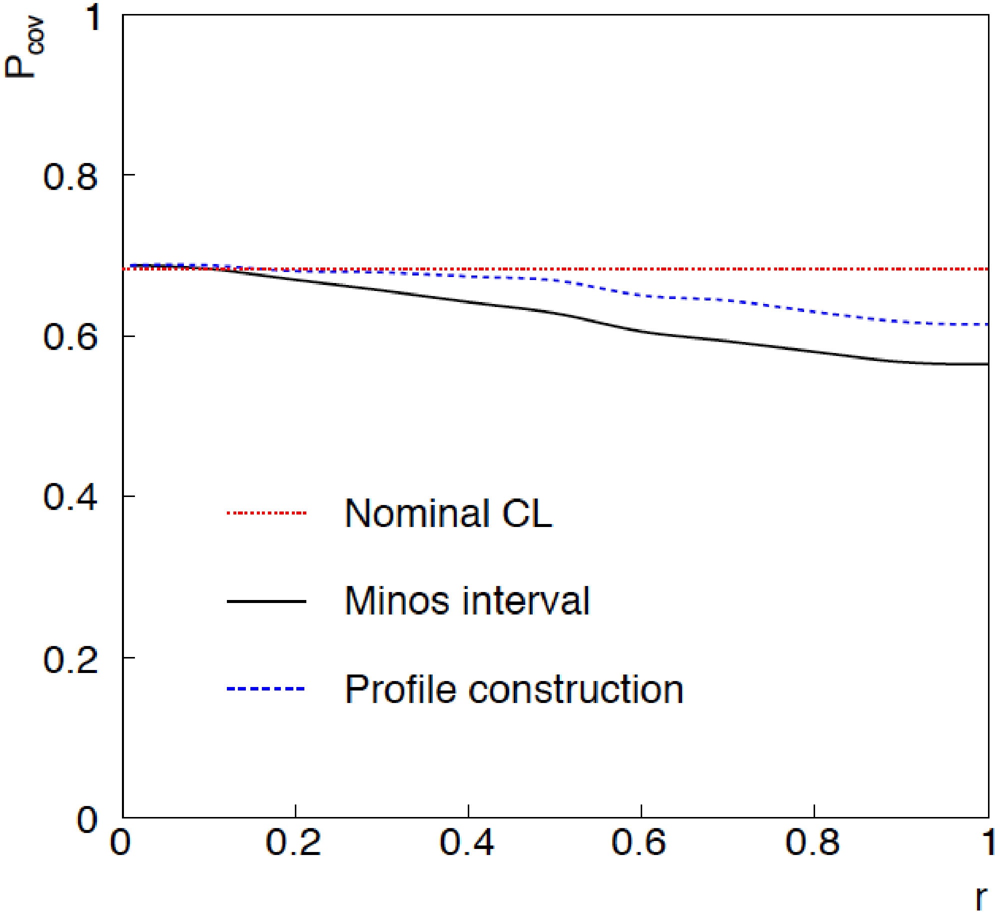
$$y_i \sim \text{Gauss}(\mu, \sigma_{y,i})$$

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

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

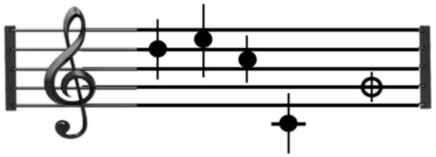
$$\sigma_{y,i} = \sigma_{u,i} = 1$$

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



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

Student's *t* average software



Software: stave.py
https://www.pp.rhul.ac.uk/~cowan/stat/exercises/stave/

The program stave.py implements the Gamma Variance Model (GVM) for averaging N measurements.

For details see G. Cowan, EPJC (2019) 79:133.

In this version the model does not distinguish between statistical and systematic errors.

Confidence interval for the mean μ becomes sensitive to goodness-of-fit (increases if data internally inconsistent).

Estimated mean less sensitive to outliers.

Least Squares vs Gamma Variance Model

Quadratic terms from Least Squares replaced by logarithmic ones:

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

where

y_i = measured value

$v_i = s_i^2$ = estimated variance

ε_i = relative uncertainty on estimate of variance

Equivalent to replacing Gauss pdf for measurements by Student's t, number of degrees of freedom $= 1/2\varepsilon_i^2$

A quick look at stave.py

Set measured values, estimates of std. dev., errors on errors:

```
y = np.array([17., 19., 15., 3.])      # measured values
s = np.array([1.5, 1.5, 1.5, 1.5])     # estimates of std. dev
v = s**2                               # estimates of variances
r = np.array([0.2, 0.2, 0.2, 0.2])     # relative errors on errors
```

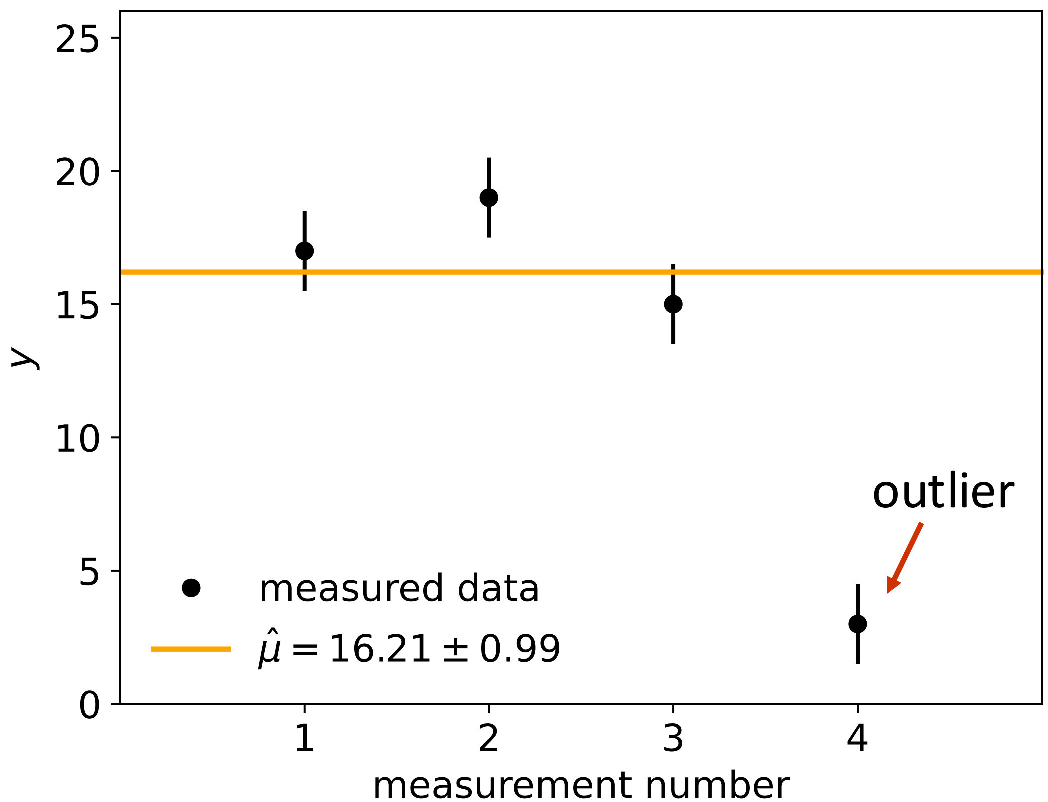
log-likelihood:

```
class NegLogL:

    def __init__(self, y, s, r):
        self.setData(y, s, r)

    def setData(self, y, s, r):
        self.data = y, s, r

    def __call__(self, mu):
        y, s, r = self.data
        v = s ** 2
        lnf = -0.5*(1. + 1./(2.*r**2))*np.log(1. + 2.*(r*(y-mu))**2/v)
        return -np.sum(lnf)
```

Example average with GVM

Suppose four measurements of the parameter μ.

Each reports an estimated standard dev. of $s = 1.5$ and a “relative error on the error” $\varepsilon = 0.2$.



Suggested exercise:

Experiment with different numbers of measurements, different levels of internal consistency, different values for the std. dev. and error on error.