Probability & Statistics
Lecture 3

Data Science in Fundamental Physics
Santiago de Compostela
3-5 June 2024

https://igfae.usc.es/datascience2024/school/

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

Outline

Lecture 1: Probability, Bayes vs. Frequentist

Frequentist parameter estimation

Hypothesis tests

Lecture 2: *p*-values

Confidence limits

Systematic uncertainties

Bayesian parameter estimation

→ Lecture 3: Prototype search analysis

Significance, sensitivity

Errors on errors

Bayes factors

Prototype search analysis

Search for signal in a region of phase space; result is histogram of some variable x giving numbers:

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\rm tot} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta}_s)\, dx \,, \qquad b_i = b_{\rm tot} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta}_b)\, dx \,.$$

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

nuisance parameters ($\boldsymbol{\theta}_\mathrm{s}$, $\boldsymbol{\theta}_\mathrm{b}$,$b_\mathrm{tot}$)

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

maximizes L for specified μ

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

maximize L

Define critical region of test of μ by the region of data space that gives the lowest values of $\lambda(\mu)$.

Important advantage of profile LR is that its distribution becomes independent of nuisance parameters in large sample limit.

Test statistic for discovery

Suppose relevant alternative to background-only ($\mu = 0$) is $\mu \geq 0$.

So take critical region for test of $\mu = 0$ corresponding to high q_0 and $\hat{\mu}$ > 0 (data characteristic for $\mu \geq 0$).

That is, to test background-only hypothesis define statistic

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only large (positive) observed signal strength is evidence against the background-only hypothesis.

Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

Distribution of q_0 in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a “half chi-square” distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

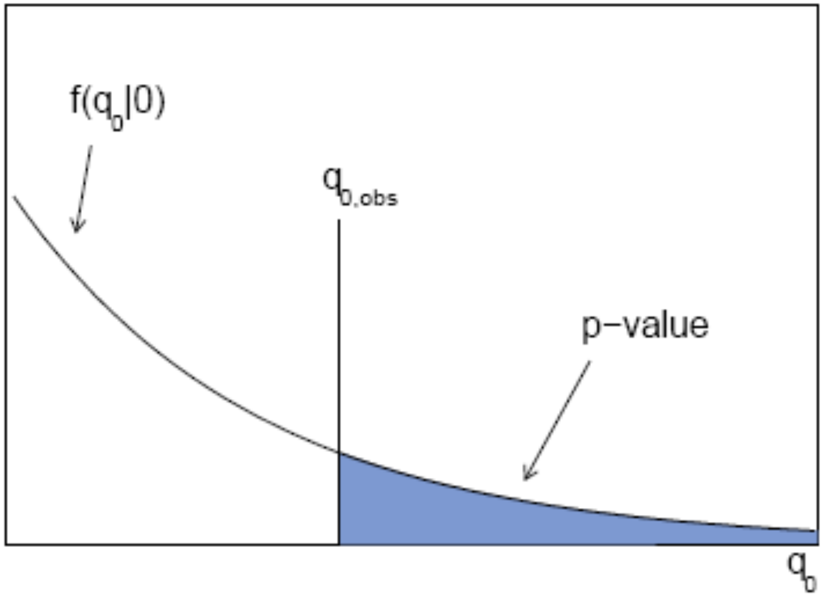
In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

p-value for discovery

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p-value for an observed $q_{0,\text{obs}}$ is

$$p_0 = \int_{q_{0,\text{obs}}}^{\infty} f(q_0|0)\, dq_0$$

use e.g. asymptotic formula



From p-value get equivalent significance,

$$Z = \Phi^{-1}(1 - p)$$

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of q_0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The *p*-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance Z is simply

$$Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}$$

Monte Carlo test of asymptotic formula

$$n \sim \text{Poisson}(\mu s + b)$$

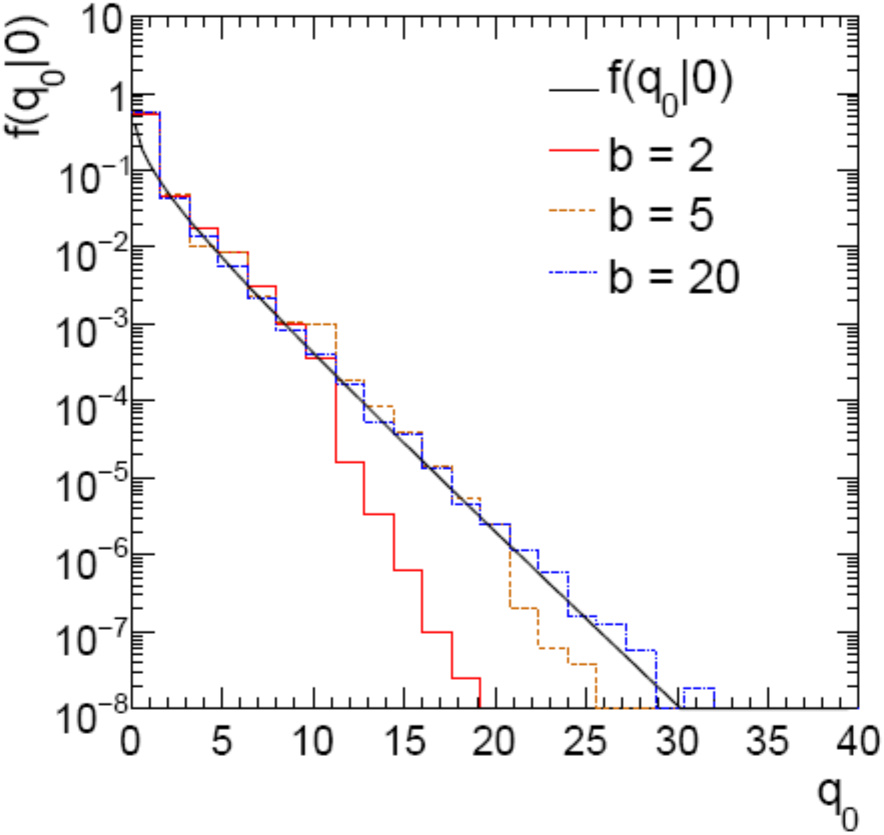
$$m \sim \text{Poisson}(\tau b)$$

$\mu =$ param. of interest

$b =$ nuisance parameter

Here take s known, $\tau = 1$.

Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.

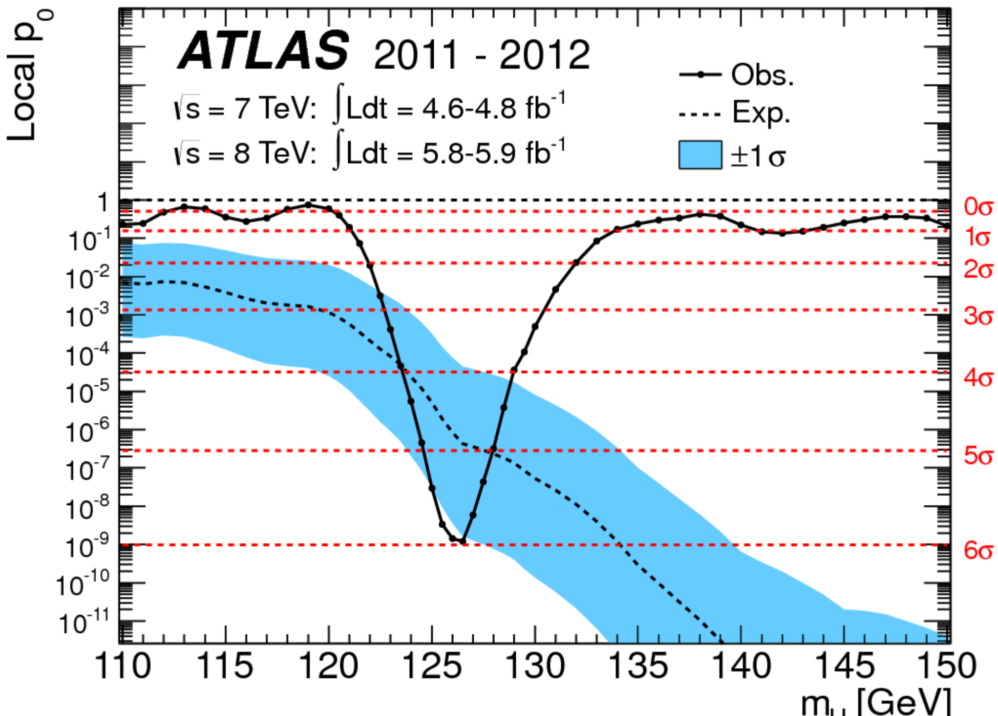


How to read the p_0 plot

The "local" p_0 means the p-value of the background-only hypothesis obtained from the test of $\mu = 0$ at each individual m_{H}, without any correct for the Look-Elsewhere Effect.

The "Expected" (dashed) curve gives the median p_0 under assumption of the SM Higgs (μ = 1) at each m_{H}.

ATLAS, Phys. Lett. B 716 (2012) 1-29



The blue band gives the width of the distribution ($\pm 1\sigma$) of significances under assumption of the SM Higgs.

Test statistic for upper limits

For purposes of setting an upper limit on μ use

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized μ :

From observed q_μ find p-value: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

Large sample approximation: $$p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)$$

To find upper limit at CL = 1–α, set $p_\mu = \alpha$ and solve for μ.

Monte Carlo test of asymptotic formulae

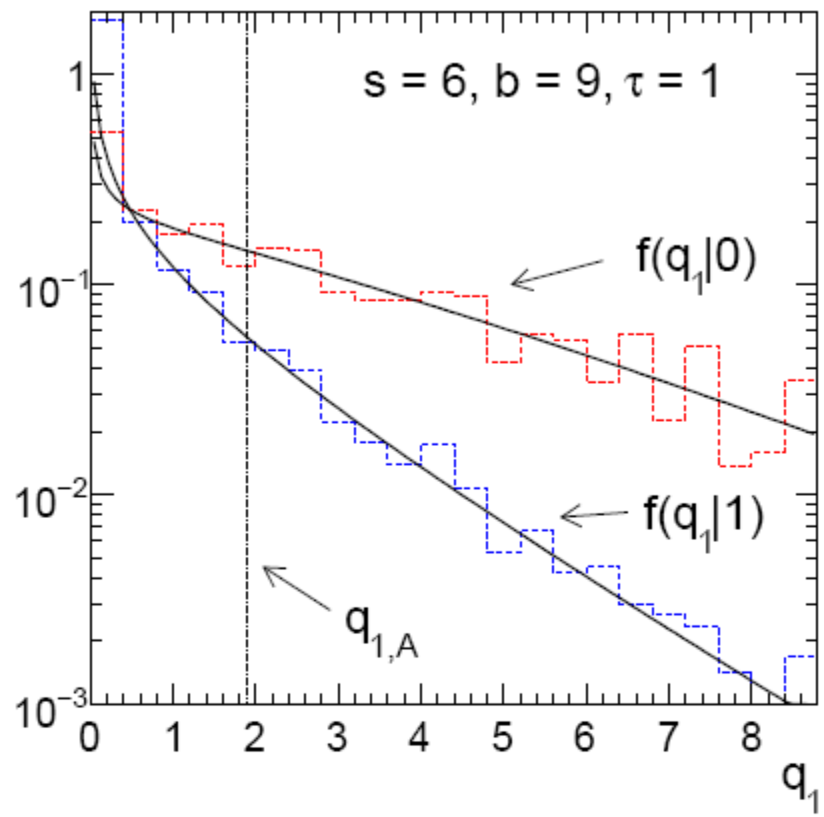
Consider again $n \sim$ Poisson($\mu s + b$), $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of $\mu = 1$.

Typically interested in 95% CL, i.e., p-value threshold = 0.05, i.e., $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median$[q_1|0]$ gives "exclusion sensitivity".

Here asymptotic formulae good for $s = 6$, $b = 9$.



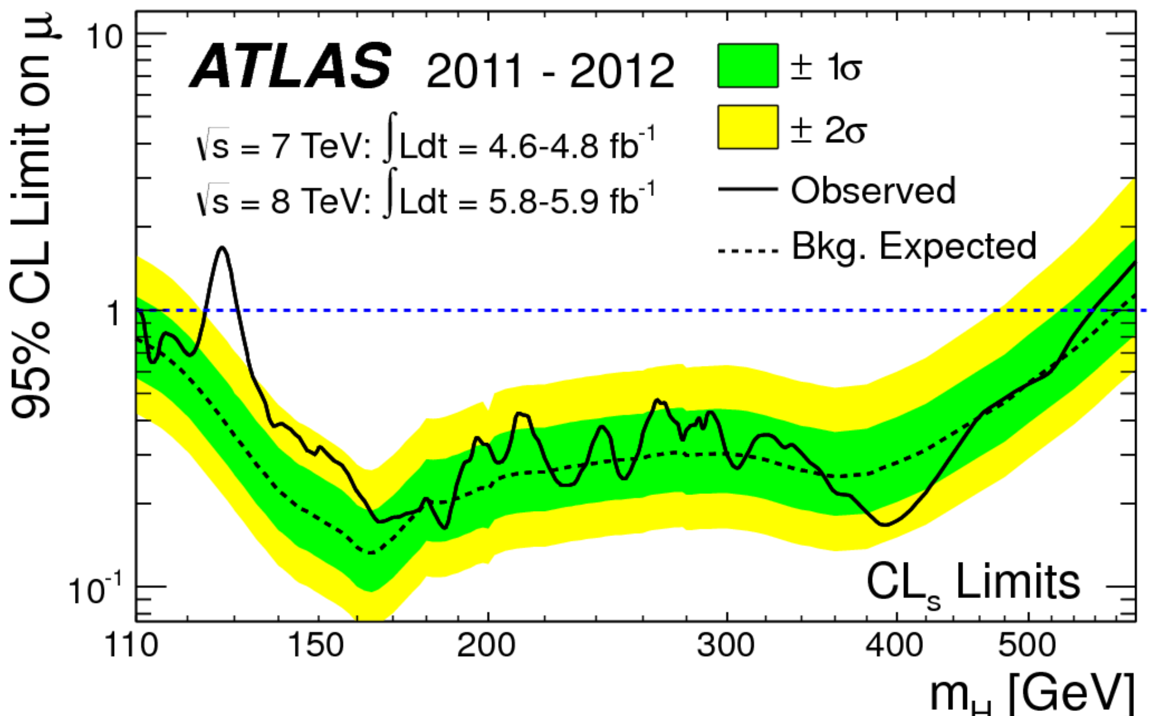
How to read the green and yellow limit plots

For every value of m_{H}, find the upper limit on μ.

Also for each m_{H}, determine the distribution of upper limits μ_{up} one would obtain under the hypothesis of $\mu = 0$.

The dashed curve is the median μ_{up}, and the green (yellow) bands give the $\pm$ 1σ (2σ) regions of this distribution.

ATLAS, Phys. Lett. B 716 (2012) 1-29



Expected discovery significance for counting experiment with background uncertainty

I. Discovery sensitivity for counting experiment with b known:

(a) $$\frac{s}{\sqrt{b}}$$

(b) Profile likelihood ratio test & Asimov: $$\sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}$$

II. Discovery sensitivity with uncertainty in b, σ_b:

(a) $$\frac{s}{\sqrt{b+\sigma_b^2}}$$

(b) Profile likelihood ratio test & Asimov:

$$\left[2\left((s+b)\ln\left[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}\right]-\frac{b^2}{\sigma_b^2}\ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right]\right)\right]^{1/2}$$

Counting experiment with known background

Count a number of events $n \sim$ Poisson(s+b), where

s = expected number of events from signal,

b = expected number of background events.

To test for discovery of signal compute p-value of $s = 0$ hypothesis,

$$p = P(n \geq n_{\rm obs}|b) = \sum_{n=n_{\rm obs}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; 2n_{\rm obs})$$

Usually convert to equivalent significance: $Z = \Phi^{-1}(1-p)$
where Φ is the standard Gaussian cumulative distribution, e.g.,
$Z > 5$ (a 5 sigma effect) means $p < 2.9 \times 10^{-7}$.

To characterize sensitivity to discovery, give expected (mean or median) Z under assumption of a given s.

$s/\sqrt{b}$ for expected discovery significance

For large $s + b$, $n \to x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob($x > x_{obs} \mid s = 0$),:

$$p_0 = 1 - \Phi\left(\frac{x_{\rm obs} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\rm obs} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\text{median}[Z_0|s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for significance

Poisson likelihood for parameter *s* is

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

For now no nuisance params.

To test for discovery use profile likelihood ratio:

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{s} \geq 0 \,, \\ 0 & \hat{s} < 0 \,. \end{cases}$$

$$\lambda(s) = \frac{L(s, \hat{\hat{\theta}}(s))}{L(\hat{s}, \hat{\theta})}$$

So the likelihood ratio statistic for testing $s = 0$ is

$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})} = 2 \left(n \ln \frac{n}{b} + b - n \right) \quad \text{for } n > b, \; 0 \text{ otherwise}$$

Approximate Poisson significance (continued)

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \quad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

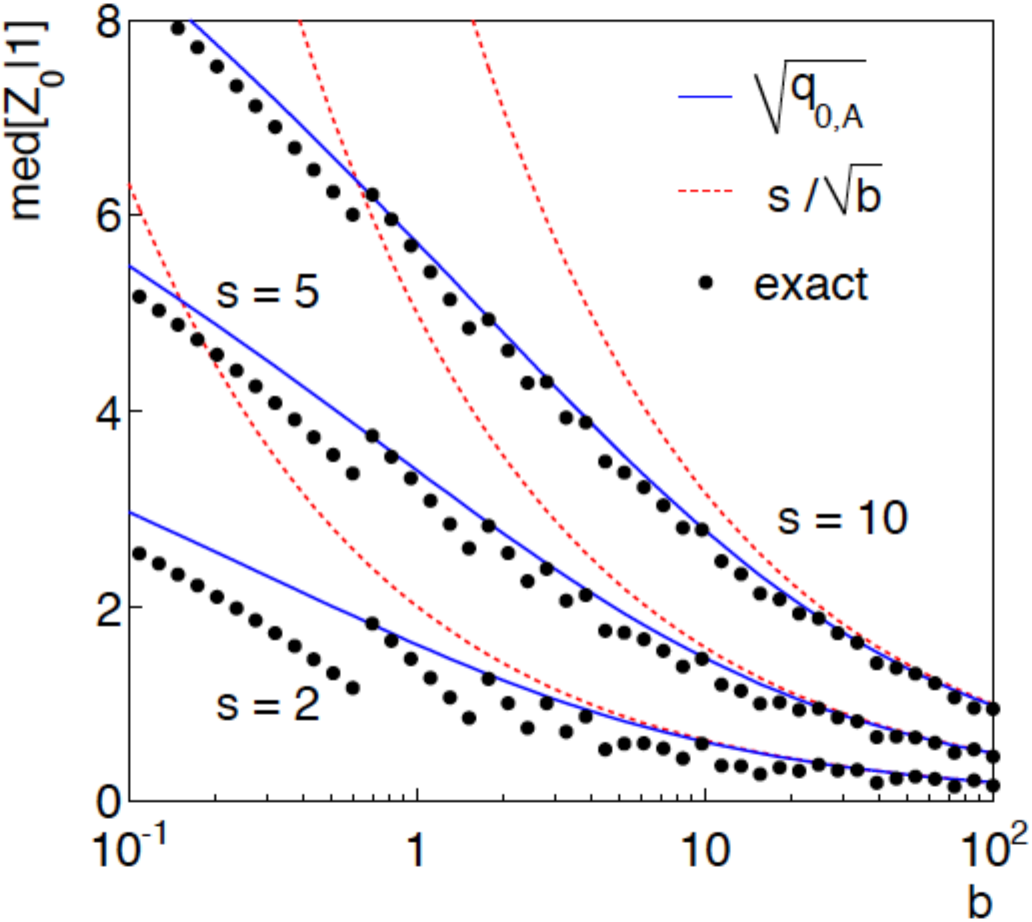
To find median[$Z|s$], let $n \rightarrow s + b$ (i.e., the Asimov data set):

$$Z_{\mathrm{A}} = \sqrt{2\left((s+b)\ln\left(1 + \frac{s}{b}\right) - s\right)}$$

This reduces to $s/\sqrt{b}$ for $s << b$.

$n \sim$ Poisson(s+b), median significance, assuming s, of the hypothesis $s = 0$

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727



"Exact" values from MC, jumps due to discrete data.

Asimov $\sqrt{q_{0,\mathrm{A}}}$ good approx. for broad range of s, b.

$s/\sqrt{b}$ only good for $s \ll b$.

Extending $s/\sqrt{b}$ to case where b uncertain

The intuitive explanation of $s/\sqrt{b}$ is that it compares the signal, s, to the standard deviation of n assuming no signal, $\sqrt{b}$.

Now suppose the value of b is uncertain, characterized by a standard deviation σ_b.

A reasonable guess is to replace $\sqrt{b}$ by the quadratic sum of $\sqrt{b}$ and σ_b, i.e.,

$$\mathrm{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where σ_b cannot be neglected.

Profile likelihood with b uncertain

This is the well studied "on/off" problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...

Measure two Poisson distributed values:

$n \sim$ Poisson(s+b) (primary or "search" measurement)

$m \sim$ Poisson(τb) (control measurement, τ known)

The likelihood function is

$$L(s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{(\tau b)^m}{m!} e^{-\tau b}$$

Use this to construct profile likelihood ratio (b is nuisance parameter):

$$\lambda(0) = \frac{L(0, \hat{\hat{b}}(0))}{L(\hat{s}, \hat{b})}$$

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\hat{s} = n - m/\tau \,,$$

$$\hat{b} = m/\tau \,,$$

$$\hat{\hat{b}}(s) = \frac{n + m - (1+\tau)s + \sqrt{(n + m - (1+\tau)s)^2 + 4(1+\tau)sm}}{2(1+\tau)} \,.$$

and in particular to test for discovery ($s = 0$),

$$\hat{\hat{b}}(0) = \frac{n+m}{1+\tau}$$

Asymptotic significance

Use profile likelihood ratio for q_0, and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$Z = \sqrt{q_0}$$

$$= \left[-2\left(n \ln\left[\frac{n+m}{(1+\tau)n}\right] + m \ln\left[\frac{\tau(n+m)}{(1+\tau)m}\right]\right)\right]^{1/2}$$

for $n > \hat{b}$ and $Z = 0$ otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

Asimov approximation for median significance

To get median discovery significance, replace n, m by their expectation values assuming background-plus-signal model:

$$n \to s + b$$

$$m \to \tau b$$

$$Z_{\mathrm{A}} = \left[-2 \left((s+b) \ln \left[\frac{s + (1+\tau)b}{(1+\tau)(s+b)} \right] + \tau b \ln \left[1 + \frac{s}{(1+\tau)b} \right] \right) \right]^{1/2}$$

Or use the variance of $\hat{b} = m/\tau$, $V[\hat{b}] \equiv \sigma_b^2 = \dfrac{b}{\tau}$, to eliminate τ:

$$Z_{\mathrm{A}} = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}$$

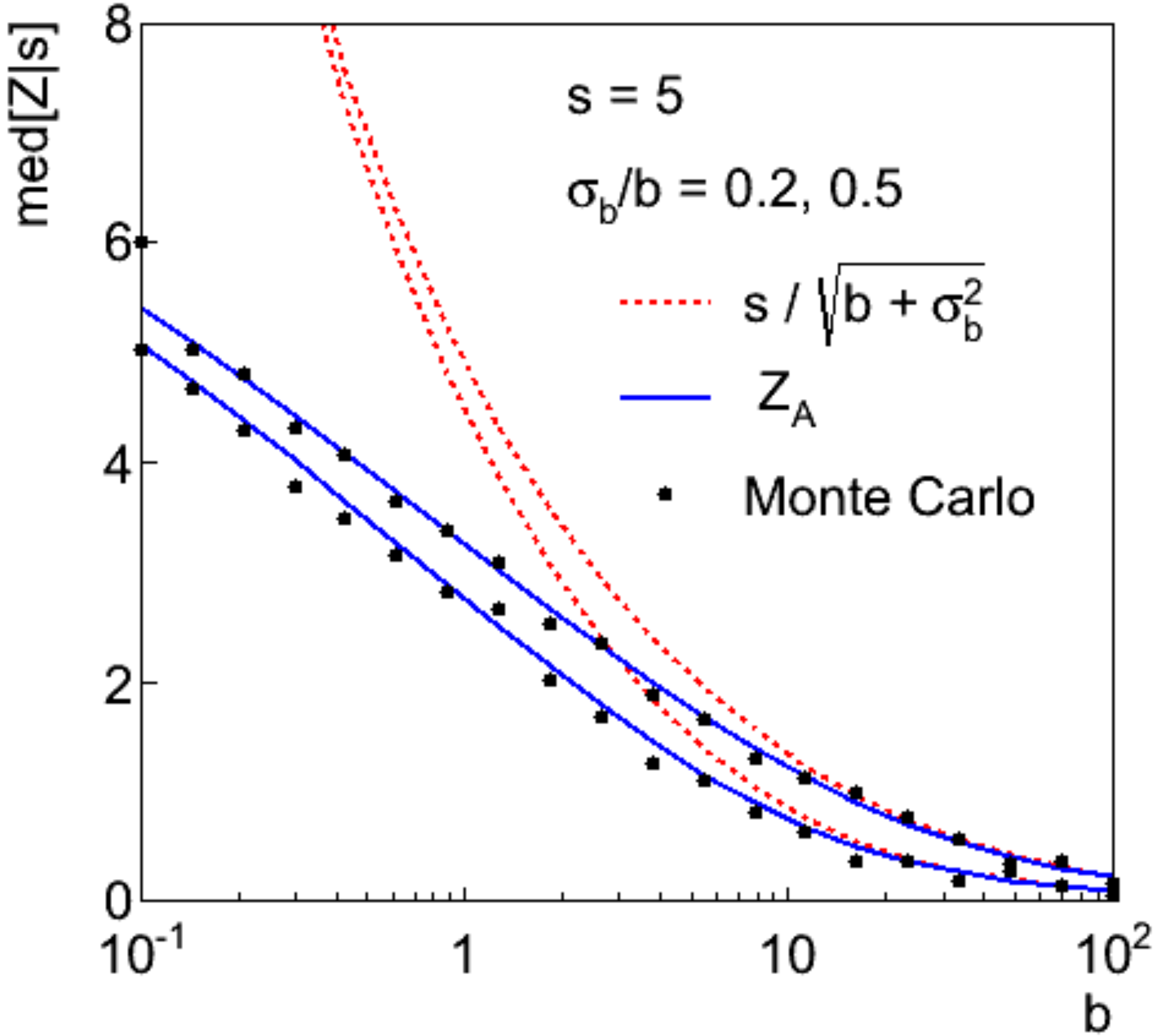
Limiting cases

Expanding the Asimov formula in powers of s/b and σ_b^2/b ($= 1/\tau$) gives

$$Z_{\mathrm{A}} = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b)\right)$$

So the "intuitive" formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set.

Testing the formulae: $s = 5$



Using sensitivity to optimize a cut

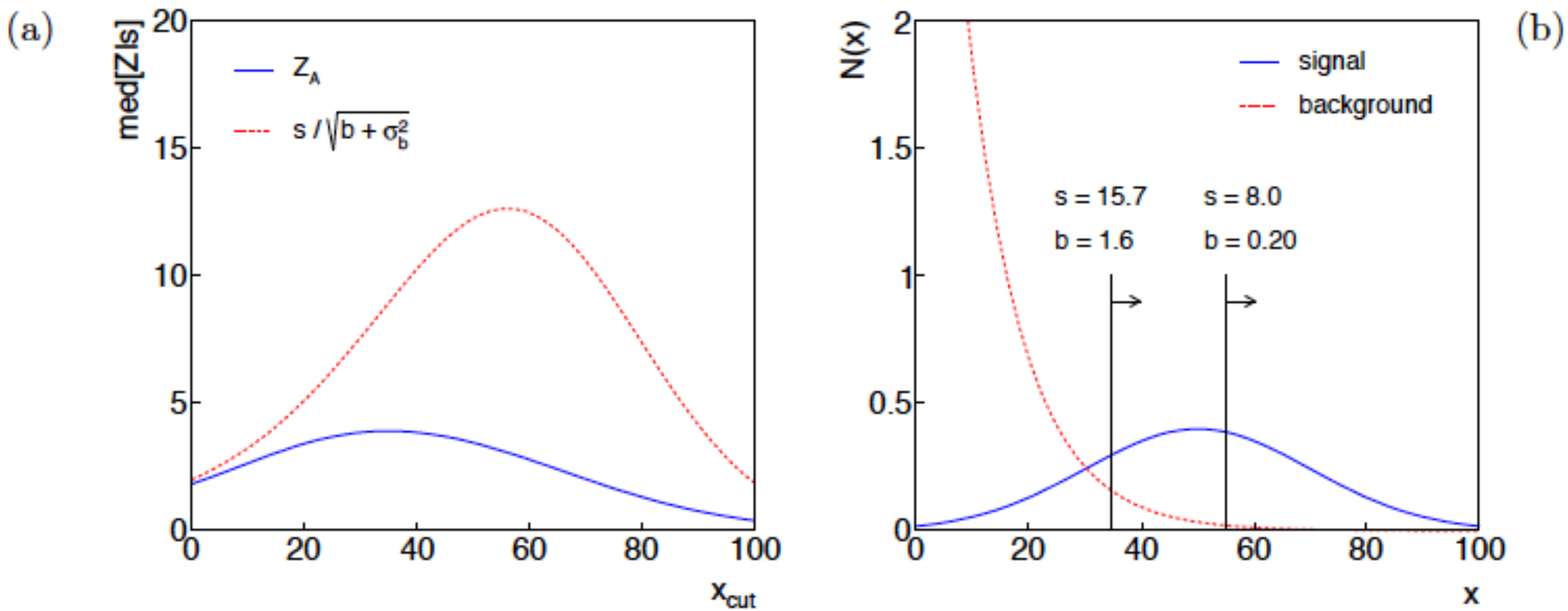


Figure 1: (a) The expected significance as a function of the cut value $x_{\rm cut}$; (b) the distributions of signal and background with the optimal cut value indicated.

"Errors on errors"

The uncertainties on estimated systematic errors ("errors on errors") can in general play an important role in many analyses, see:

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

E. Canonero, A. Brazzale and G. Cowan, *Higher-order asymptotic corrections and their application to the Gamma Variance Model*, Eur. Phys. J. C (2023) 83:1100, arXiv:2304.10574

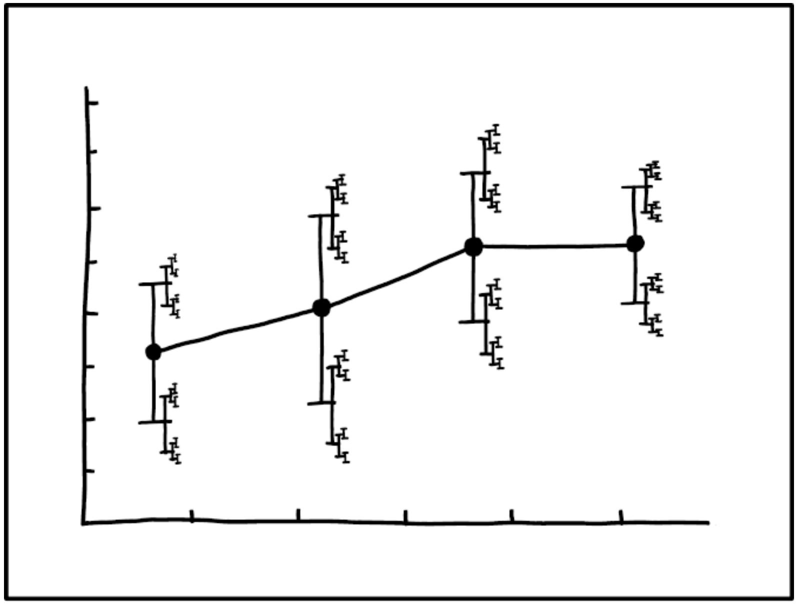
It turns out that models that use errors on errors have qualitatively new, interesting, desirable features:

Sensitivity to outliers reduced.

Confidence intervals sensitive to goodness of fit.

Effect on goodness of fit, p-values, significance.

https://xkcd.com/2110/



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 mean and width for $f(v)$:

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

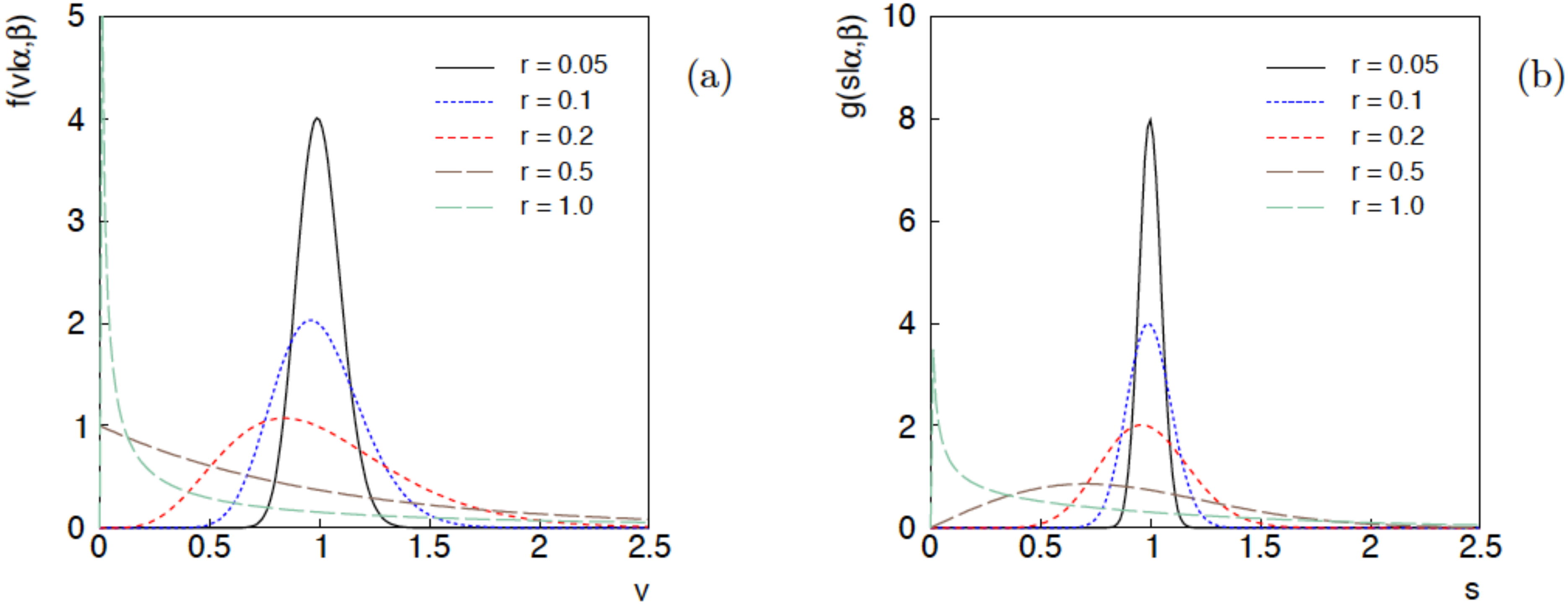
$r = 1/2\sqrt{\alpha} \approx$ relative "error on the error" = $\sigma_s/E[s]$.

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

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

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

← relative “error on error”



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.)

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

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

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

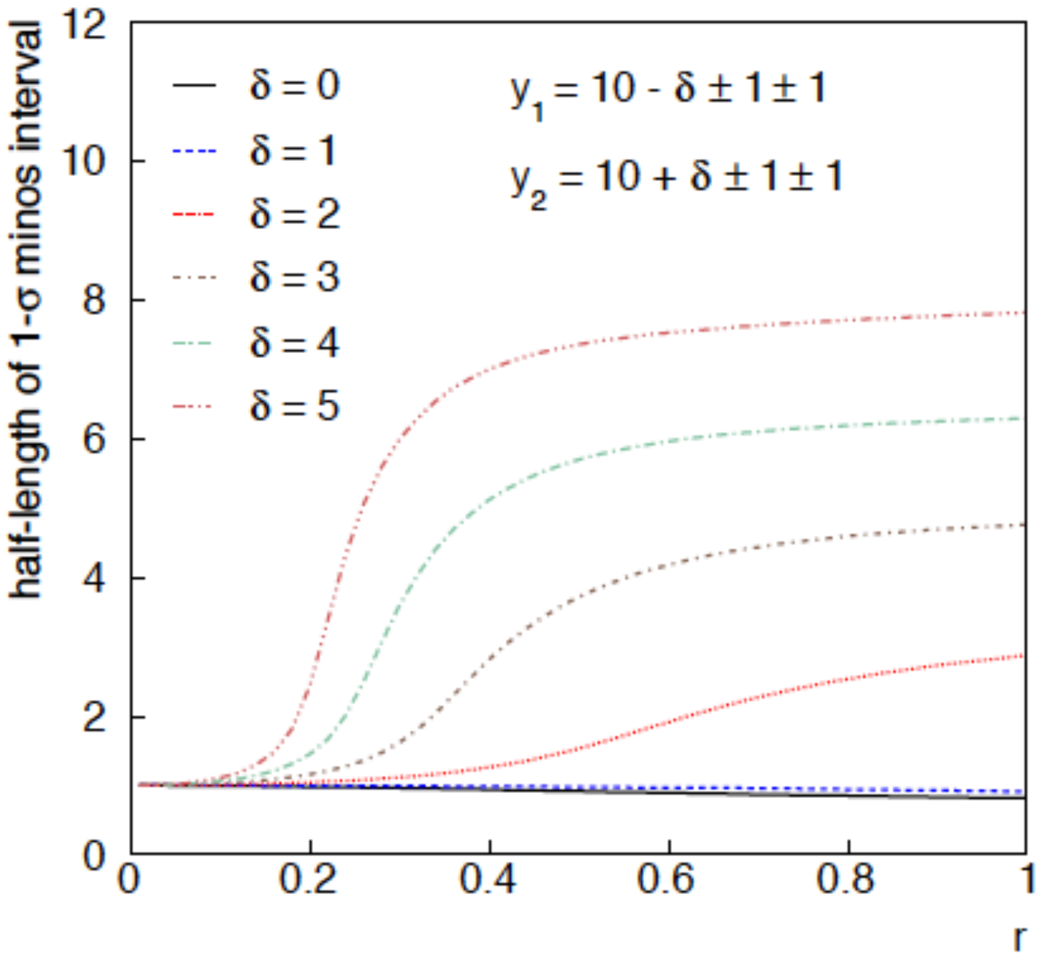
Resulting likelihood same as profile $L'(\boldsymbol{\mu},\boldsymbol{\theta})$ from gamma model

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

Example: average of two measurements

Approximate (”MINOS”) confidence interval based on

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



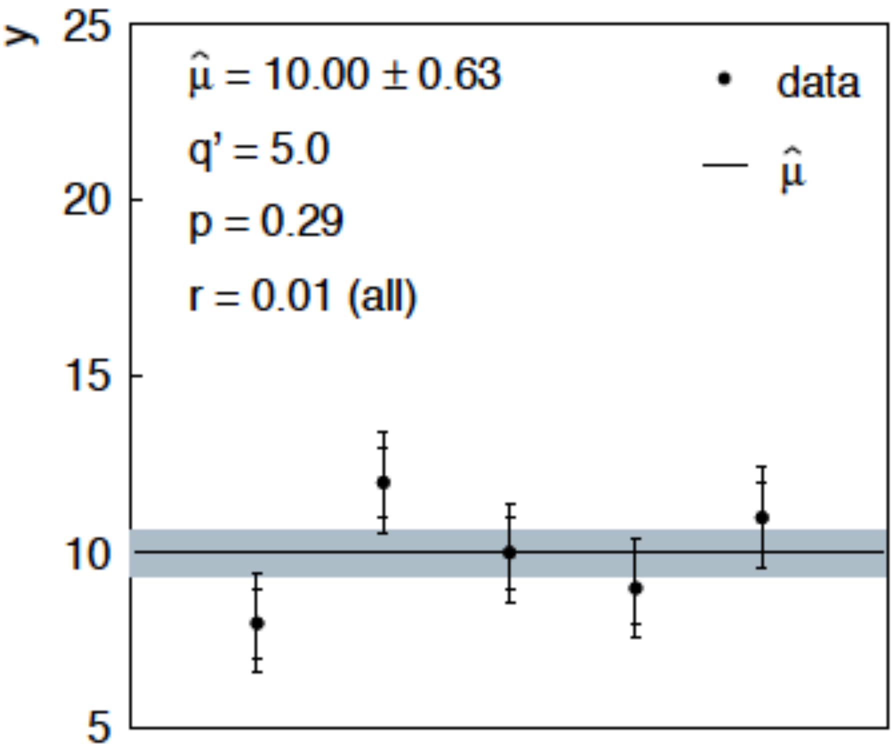
Increased discrepancy between values to be averaged gives larger interval.

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

relative error on sys. error

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).

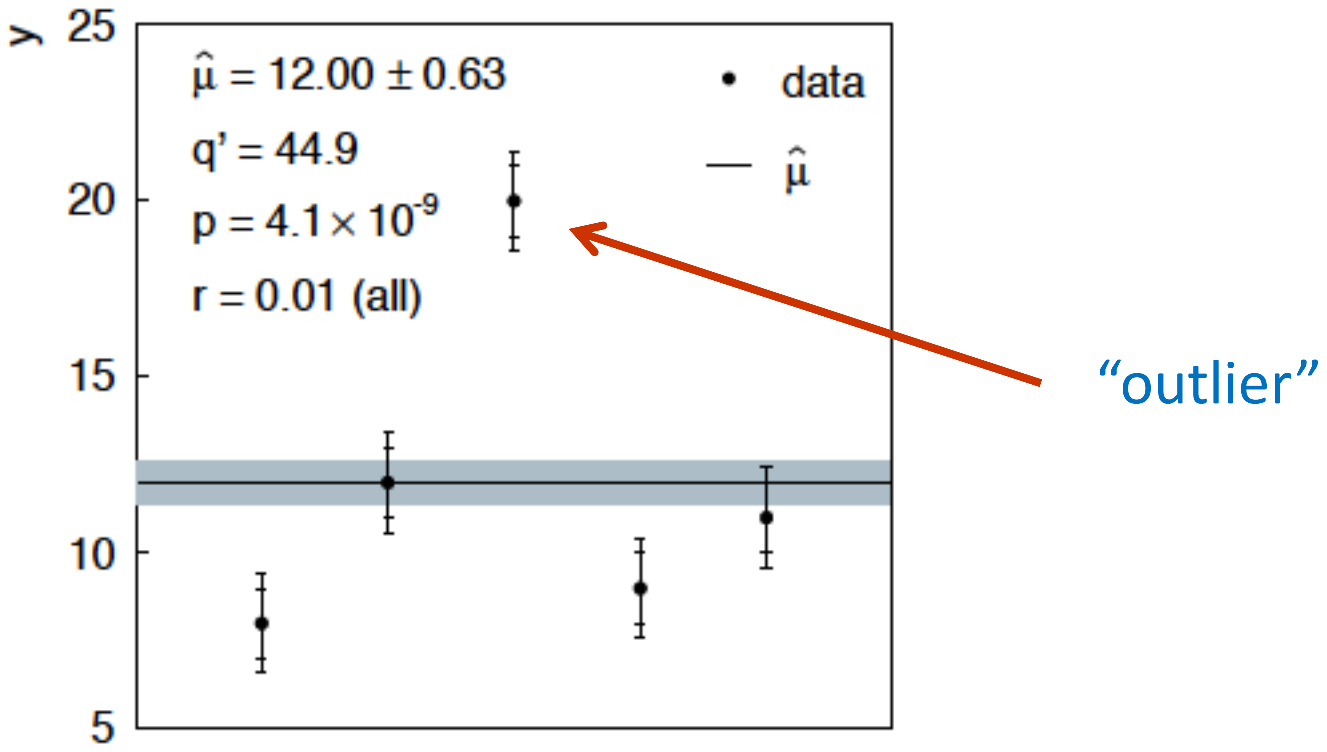


inner error bars $= \sigma_{y,i}$

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

Sensitivity of average to outliers (2)

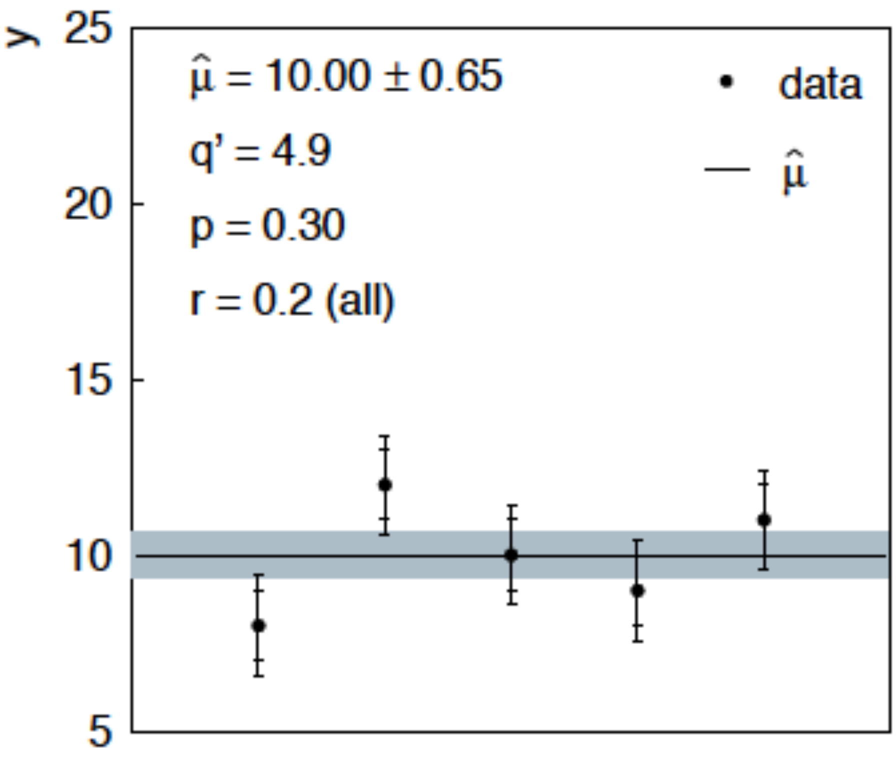
Now suppose the measurement at 10 had come out at 20:



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

Average with all $r = 0.2$

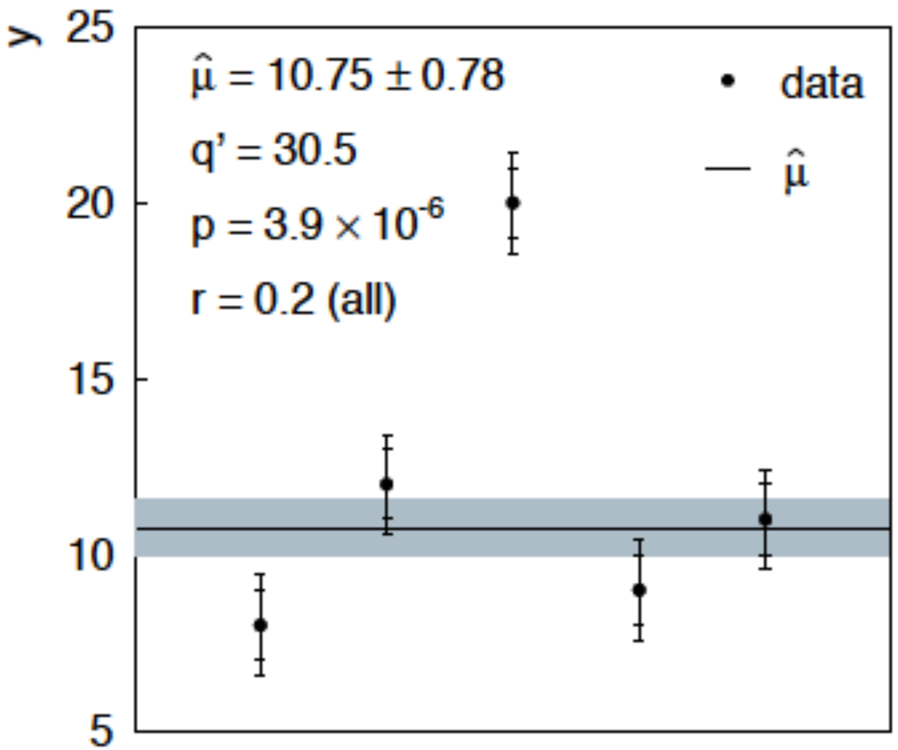
If we assign to each measurement $r = 0.2$,



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 $10 \to 20$)



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.,

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

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.

Discussion on Gamma Variance Model

Other features of Gamma Variance Model (see EPJC (2019) 79:133 and the extra slides)

averages/fits become less sensitive to outliers;

confidence intervals linked to goodness of fit;

straightforward to include multiple correlated error sources.

But... is part of the reason for requiring 5σ for discovery not to account for uncertainties in assigned errors? Is there a trade-off between "errors on errors" and the requirement for discovery?

Best to have most realistic model. If the estimated errors are indeed uncertain, this should be reflected in the model.

Bottom line – it is very difficult to establish convincing evidence for a new physics if relevant uncertainties are estimated in an ad hoc way. We need robust procedures for their assignment.

Finally

Three lectures only enough for a brief discussion of:

Parameter estimation

Hypothesis tests ($\rightarrow$ path to Machine Learning)

Limits (confidence intervals/regions)

Systematics (nuisance parameters)

Bayesian methods, MCMC

A bit beyond... (“errors on errors”)

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) so often best to invest most of your time with it.

Extra Slides

Bayesian model selection

Fundamentally the probability of a hypothesis H_i in the Bayesian approach is given by its posterior probability given the data: $P(H_i|\boldsymbol{x})$.

Finding this requires assignment of prior probabilities to all hypotheses that are considered.

We can give the posterior *odds* (ratio of probabilities) for any pair of hypotheses H_i and H_j (use Bayes' theorem; factors of $P(\boldsymbol{x})$ cancel):

$$\frac{P(H_i|\mathbf{x})}{P(H_j|\mathbf{x})} = \frac{P(\mathbf{x}|H_i)}{P(\mathbf{x}|H_j)} \frac{\pi(H_i)}{\pi(H_j)}$$

posterior odds　　Bayes factor　　prior odds

See: Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

The Bayes factor

The Bayes factor B_{ij} is the likelihood ratio of the two hypotheses:

$$B_{ij} = \frac{P(\mathbf{x}|H_i)}{P(\mathbf{x}|H_j)}$$

= posterior odds if one takes prior odds equal to one.

The Bayes factor is regarded as measuring the weight of evidence of the data in support of H_i over H_j. and can be used much like a p-value (or Z value).

The Jeffreys scale, analogous to the 5σ rule in Particle Physics:

B_{10}	Evidence against H_0
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Marginal likelihood (evidence)

If the model H_i contains internal parameters $\boldsymbol{\theta}_i$, then these must be characterized by a prior pdf $\pi_i(\boldsymbol{\theta}_i|H_i)$ and marginalized:

$$P(\mathbf{x}|H_i) = \int P(\mathbf{x}, \boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i = \int P(\mathbf{x}|H_i, \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i$$

This is called the “marginal likelihood” or “evidence” of H_i.

It is independent of the overall prior probability of H_i

$$\pi(H_i) = \int \pi(H_i, \boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i$$

but it depends on the prior pdf for the model’s internal parameters $\boldsymbol{\theta}_i$:

$$\pi_i(\boldsymbol{\theta}_i|H_i) = \frac{\pi(H_i, \boldsymbol{\theta}_i)}{\pi(H_i)}$$

Bayes factor for models with internal parameters

The Bayes factor is thus the ratio of marginal likelihoods for the two models:

$$B_{ij} = \frac{P(\mathbf{x}|H_i)}{P(\mathbf{x}|H_j)} = \frac{\int P(\mathbf{x}|H_i, \boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i}{\int P(\mathbf{x}|H_j, \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j|H_j)\, d\boldsymbol{\theta}_j}$$

Simplifying the notation, the numerator and denominator are both of the form

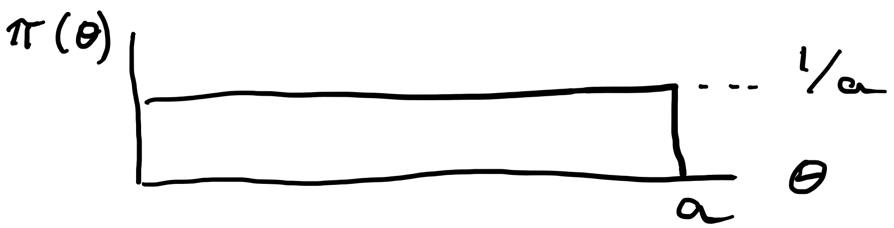
$$m = \int P(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

For high-dimensional $\boldsymbol{\theta}$ these integrals can be very difficult to compute (more on this later).

Priors for Bayes factors

Prior pdfs for the marginal likelihoods used in Bayes factors cannot be improper, i.e., they cannot be defined only up to an arbitrary normalization constant, in which case B_{ij} would not be well defined.

Suppose we try to take a ”non-informative” prior to be constant out to some large cut-off, in the hope that the Bayes factor will decouple from it:



In such cases we find that the Bayes factor remains sensitive to the cut-off even for $a \to \infty$.

So all priors used for Bayes factors must reflect a meaningful degrees of uncertainty about the parameters.

Bayes factor for Poisson counting experiment

Suppose $n \sim \mathrm{Poisson}(s + b)$ with b known. We want to compare

$$H_0 \ : \ s = 0 \ ,$$

$$H_1 \ : \ s > 0 \ .$$

The likelihoods of H_0 and H_1 are

$$L(n|H_0) = \frac{b^n}{n!} e^{-b}$$

$$L(n|s, H_1) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Bayes factor for Poisson counting experiment (2)

Suppose the prior pdf for the parameter s in H_1 is:

$$\pi(s|H_1) = \frac{1}{s_{\rm max}} \qquad (0 \leq s \leq s_{\rm max})$$

The posterior probability for s given n is, assuming H_1,

$$p(s|n, H_1) = \frac{L(n|s, H_1)\pi(s|H_1)}{\int L(n|s, H_1)\pi(s|H_1)\, ds}$$

$$= \frac{(s+b)^n e^{-(s+b)}}{\int_0^{s_{\rm max}} (s+b)^n e^{-(s+b)}\, ds} \qquad (0 \leq s \leq s_{\rm max})$$

$$= \frac{(s+b)^n e^{-(s+b)}}{\gamma(n+1, s_{\rm max}+b) - \gamma(n+1, b)}$$

γ = lower incomplete gamma function

Bayes factor for Poisson counting experiment (3)

In the limit $s_{\max} \to \infty$ this goes to

$$p(s|n, H_1) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(n+1) - \gamma(n+1, b)}$$

where $$\gamma(a, x) = \int_0^x t^{a-1} e^{-t} \, dt$$

is the lower incomplete gamma function.

Thus the posterior pdf for s given n under assumption of H_1 decouples from $s_{\max}$ in the limit $s_{\max} \to \infty$, and hence we can use this limiting case e.g. for finding an upper limit (credibility interval) for s.

Bayes factor for Poisson counting experiment (4)

The hypothesis H_0 has no internal parameters so its marginal likelihood is simply $m_0 = L(n|\, H_0)$.

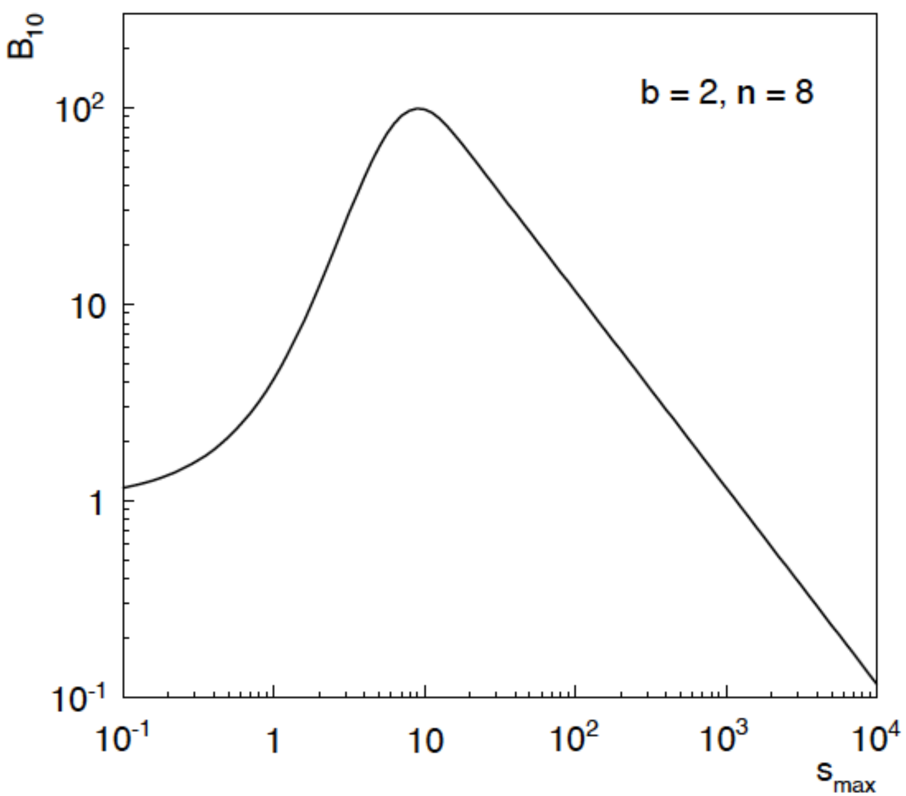
The marginal likelihood of H_1 is

$$
\begin{aligned}
m_1 &= \int L(n|s, H_1)\pi(s|H_1)\, ds \\
&= \frac{1}{n! s_{\max}} \int_0^{s_{\max}} (s+b)^n e^{-(s+b)}\, ds \\
&= \frac{1}{n! s_{\max}} \left(\gamma(n+1, s_{\max}+b) - \gamma(n+1, b) \right)
\end{aligned}
$$

Bayes factor for Poisson counting experiment (5)

So the Bayes factor is

$$B_{10} = \frac{m_1}{m_0} = \frac{1}{s_{\rm max}} \frac{\gamma(n+1, s_{\rm max}+b) - \gamma(n+1,b)}{b^n e^{-b}}$$



Example: $b = 2, n = 8$

As $s_{\rm max}$ increases the data start to favour H_1.

As $s_{\rm max}$ increases further, the larger volume of H_1's parameter space penalizes it (Ockham's razor).

Numerical determination of Bayes factors

Both numerator and denominator of B_{ij} are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta})\,d\vec{\theta}$$ ← 'marginal likelihood'

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

Harmonic Mean (and improvements)
Importance sampling
Parallel tempering (~thermodynamic integration)
Nested Sampling (MultiNest), ...

Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

Cong Han and Bradley Carlin, *Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review*, J. Am. Stat. Assoc. 96 (2001) 1122-1132.

Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge University Press, 2005.

Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\boldsymbol{\theta})$ is normalized to unity so integrate both sides,

posterior expectation

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p[1/L]$$

Therefore sample $\boldsymbol{\theta}$ from the posterior via MCMC and estimate m with one over the average of $1/L$ (the harmonic mean of L).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

Called the "worst Monte Carlo method ever"

https://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/

Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). A variant (cf. Gelfand and Dey):

Rearrange Bayes thm; multiply both sides by arbitrary pdf $f(\boldsymbol{\theta})$:

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over $\boldsymbol{\theta}$:

$$m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} = E_p\left[\frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}\right]$$

Improved convergence if tails of $f(\boldsymbol{\theta})$ fall off faster than $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Note harmonic mean estimator is special case $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$.

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

Adaptive Harmonic Mean Integration

A. Caldwell et al., International Journal of Modern Physics A Vol. 35, No. 24 (2020) 2050142

Want to compute $$I \equiv \int_\Omega f(\lambda) d\lambda$$ (Ω = support of f)

E.g. $f(\lambda) = L(\lambda)\,\pi(\lambda)$ = unnormalized target density; we can sample from this with MCMC.

Define integral over subvolume Δ of Ω with volume V_Δ

$$I_\Delta \equiv \int_\Delta f(\lambda) d\lambda \qquad r \equiv \frac{I_\Delta}{I}$$

Adaptive Harmonic Mean Integration (2)

If $f(\lambda)$ not small in Δ, then we can find I_Δ from harmonic mean:

$$E\left[\frac{1}{f(\lambda)}\right]_{\lambda\in\Delta} = \int_\Delta \frac{1}{f(\lambda)} \frac{f(\lambda)}{I_\Delta}\, d\lambda = \frac{1}{I_\Delta}\int_\Delta d\lambda = \frac{V_\Delta}{I_\Delta} \approx \frac{1}{N_\Delta}\sum_{\lambda_i\in\Delta}\frac{1}{f(\lambda_i)}$$

Sample λ from $f(\lambda)$ using MCMC, estimate $r = I_\Delta/I$ with fraction of points found in Δ:

$$\hat{r} = \frac{N_\Delta}{N_\Omega}$$

Use these to estimate I:

$$\hat{I} = \frac{\hat{I}_\Delta}{\hat{r}} = \frac{N_\Omega V_\Delta}{\sum_{\lambda_i\in\Delta}\frac{1}{f(\lambda_i)}}$$

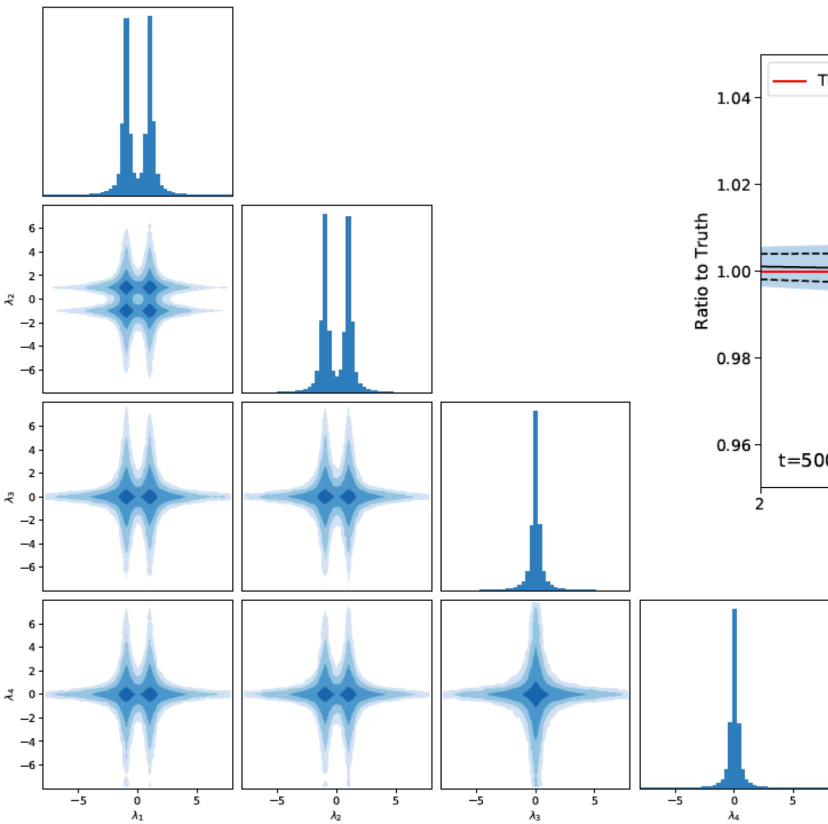
"The task of estimating our integral, therefore reduces to choosing one or several subspaces Δ — typically small regions around local modes of *f*(*λ*). The full space Ω over which the integration ought to be performed can be large or even infinite, while this does not affect the outcome of our integral estimate."

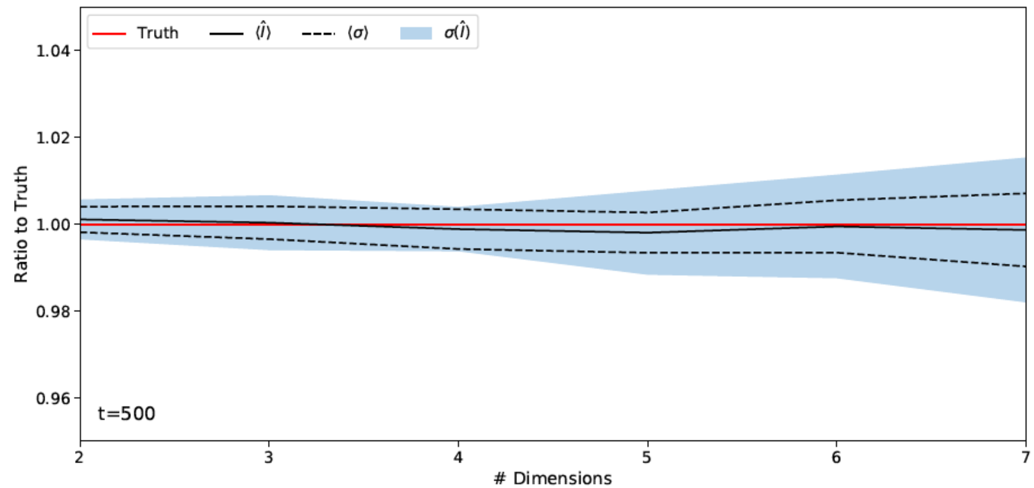
A. Caldwell et al., IJMP A Vol. 35, No. 24 (2020) 2050142

A. Caldwell et al., IJMP A Vol. 35, No. 24 (2020) 2050142

Adaptive Harmonic Mean Integration (3)

Testing AHMI with multimodal multidimensional Cauchy pdf





Challenging pdf because of long tails.

Good results for up to 7 dimensions for MCMC sample size of 10^6.

Software: Bayesian Analysis Toolkit

https://github.com/bat/BAT.jl

Importance sampling

Need pdf $f(\boldsymbol{\theta})$ which we can evaluate at arbitrary $\boldsymbol{\theta}$ and also sample with MC.

The marginal likelihood can be written

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E_f \left[\frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} \right]$$

Sample $\boldsymbol{\theta} \sim f(\boldsymbol{\theta})$, compute average of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/f(\boldsymbol{\theta})$.

Best convergence when $f(\boldsymbol{\theta})$ approximates shape of $L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Use for $f(\boldsymbol{\theta})$ e.g. multivariate Gaussian with mean and covariance estimated from posterior.

Nested sampling

J. Skilling, Bayesian Analysis, No. 4, pp. 833-860 (2006)

We want to compute $Z = \text{evidence} = \int L \, dX$ $\quad L = L(\theta)$

$$dX = \pi(\theta) d\theta$$

Can add up portions of X (equivalently, $\boldsymbol{\theta}$) space in any order. Use

$$X(\lambda) = \int_{L(\theta) > \lambda} \pi(\theta) d\theta$$

X near 1 means low λ, all of $\boldsymbol{\theta}$ space included.

Write inverse function as $L(X(\lambda)) \equiv \lambda$ so that the desired result is

$$Z = \int_0^1 L(X) \, dX$$

Elements of $\boldsymbol{\theta}$ space are sorted by decreasing likelihood.

J. Skilling, Bayesian Analysis, No. 4, pp. 833-860 (2006)

Nested sampling (2)

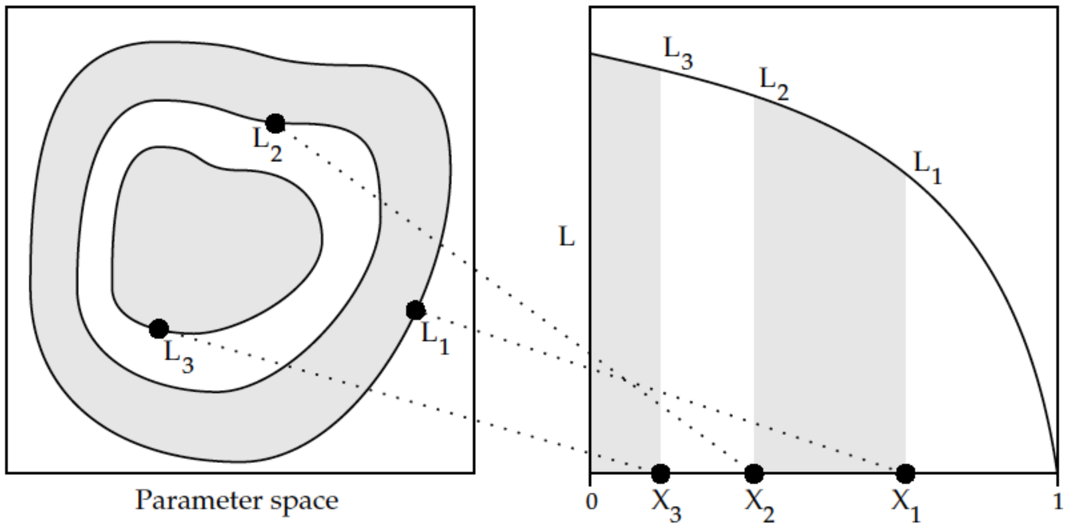


Figure 3: Nested likelihood contours are sorted to enclosed prior mass X.

The evidence Z is the area under the curve of $L(X)$.

Computational challenge is to sample $\boldsymbol{\theta}$ space from prior subject to constraint $L(\boldsymbol{\theta}) > \lambda$. Software: MultiNest

Farhan Feroz, Mike Hobson, Mon. Not. Roy. Astron. Soc., 384, 2, 449-463 (2008); arXiv:0704.3704,

F. Feroz, M.P. Hobson, M. Bridges, Mon. Not. Roy. Astron. Soc. 398: 1601-1614,2009; arXiv:0809.3437

F. Feroz, M.P. Hobson, E. Cameron, A.N. Pettitt, arXiv:1306.2144

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).

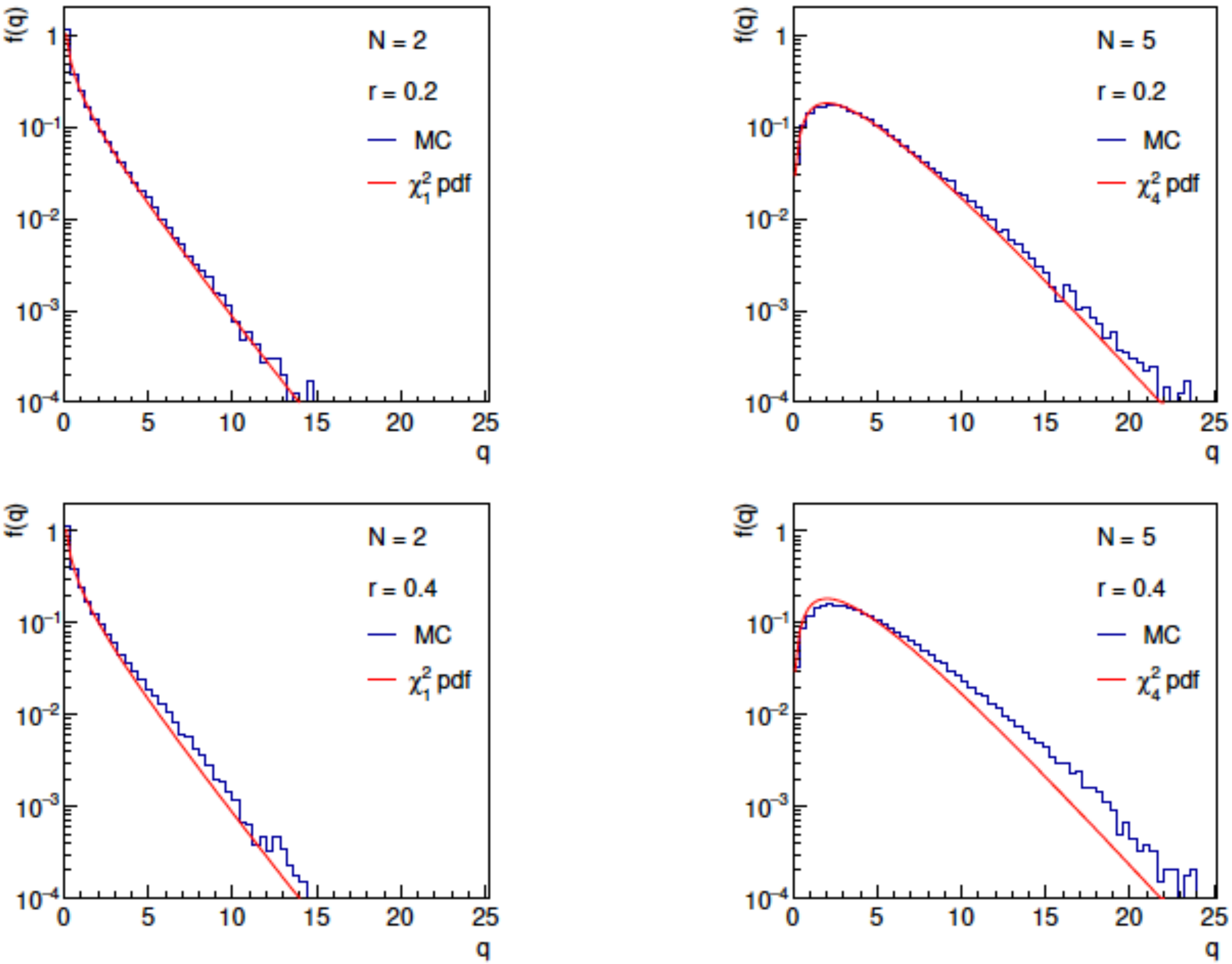
Asymptotically should have $q \sim$ chi-squared(N–M).

For increasing r_i, asymptotic distribution no longer valid.

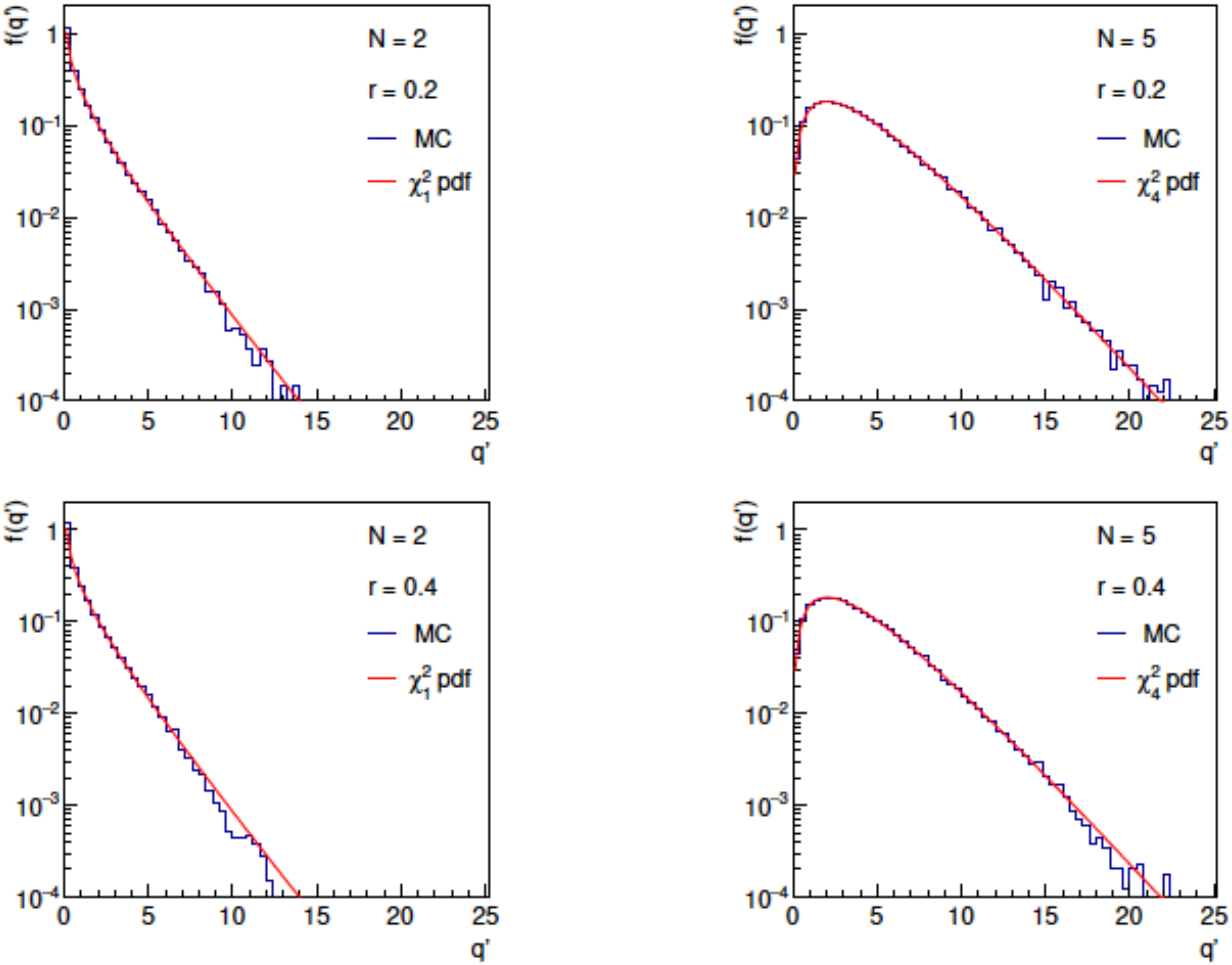
Bartlett (1937) defines modified statistic: $q' = \frac{n_{\rm d}}{E[q]} q$

By construction q' has mean $n_{\rm d} = N$-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'

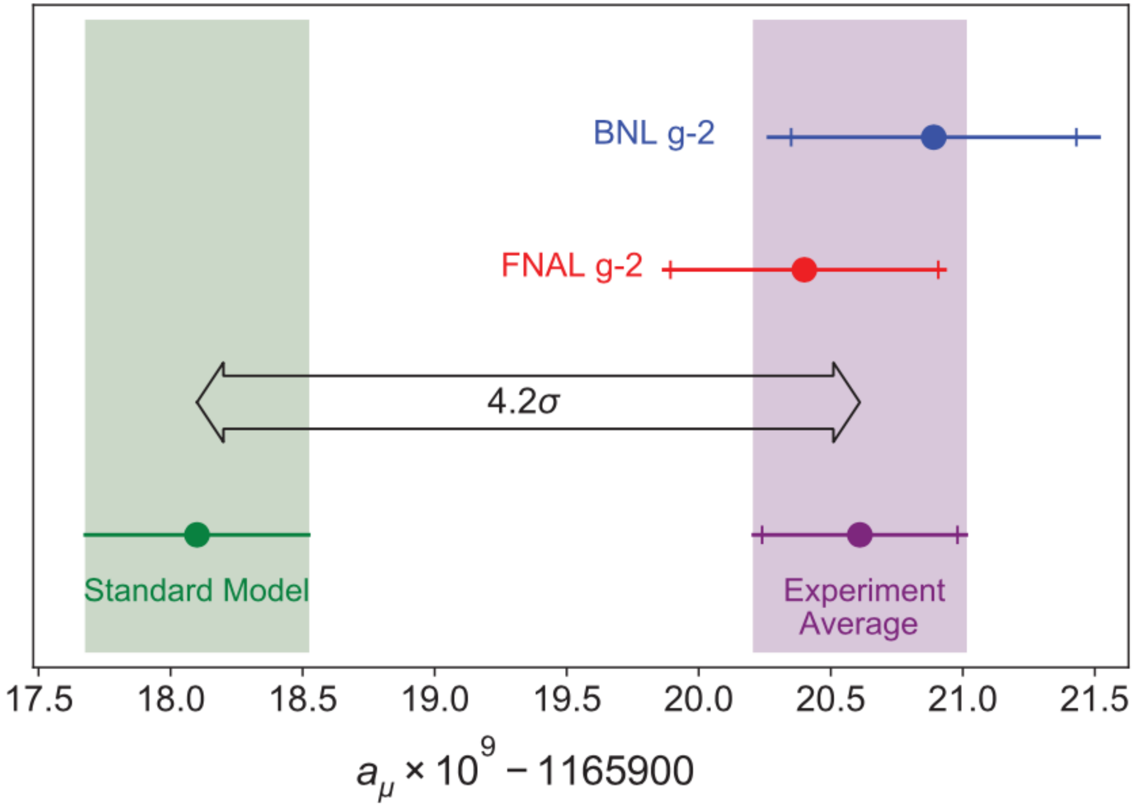


Application to the muon $g-2$ anomaly

G. Cowan, Effect of Systematic Uncertainty Estimation on the Muon g – 2 Anomaly, EPJ Web of Conferences 258, 09002 (2022), arXiv:2107.02652

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.

Muon $g-2$ ingredients

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

the ingredients of the 4.2σ effect are:

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

(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).

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

(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).

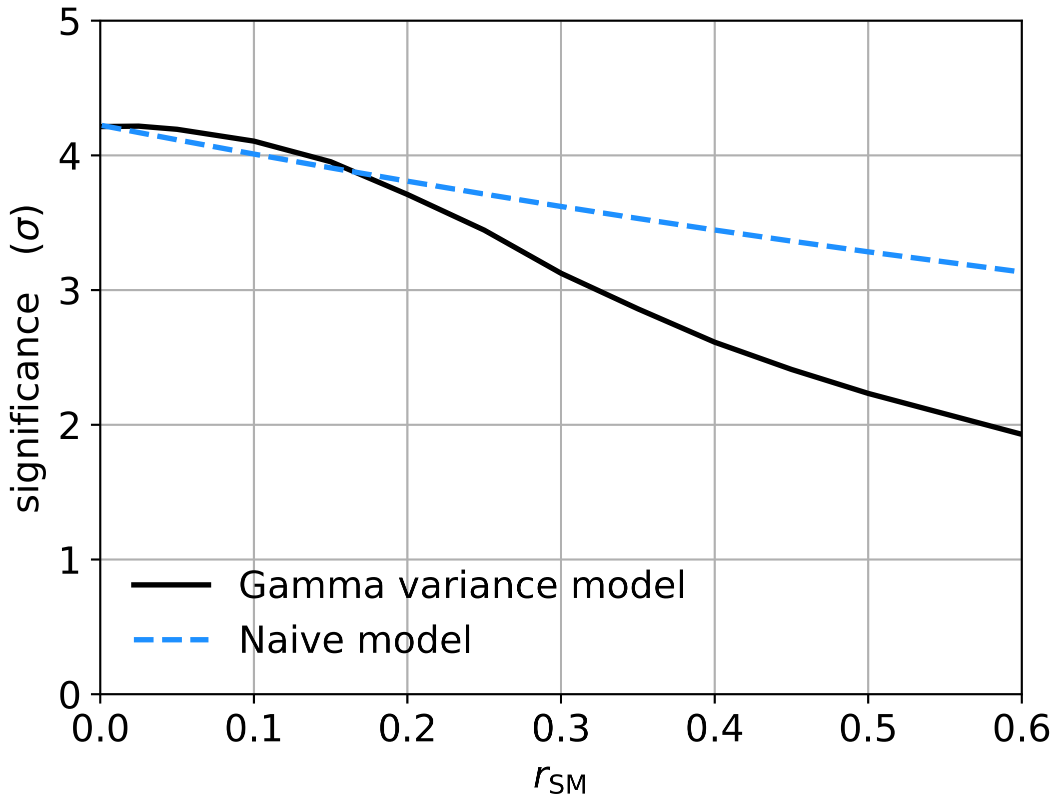
For $r_{\rm SM} > 0$ distribution of q depends on $\sigma^2_{\rm SM}$. For MC use Maximum-Likelihood estimate ("profile construction"):

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

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

# of sigmas

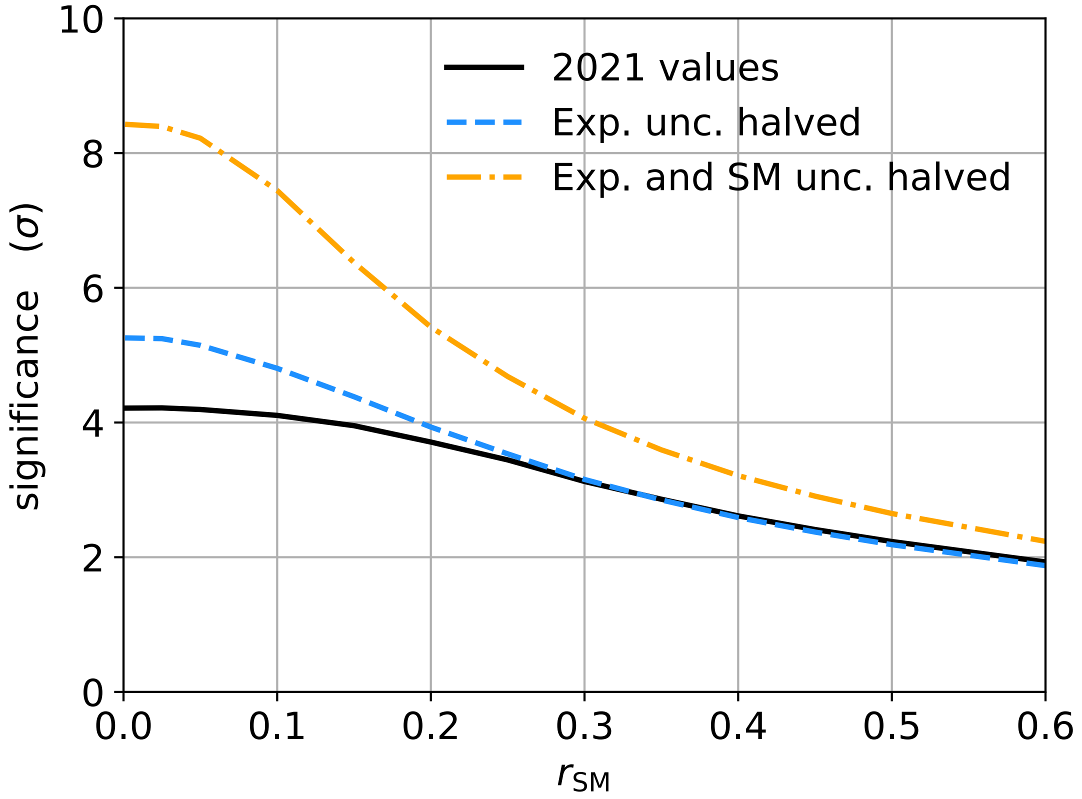
Significance of discrepancy versus $r_{\rm SM}$



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

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

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.