Reflections on 20+ years of F-C: Hypothesis testing of a point null vs a continuous alternative

Bob Cousins

Univ. of California, Los Angeles

PhyStat-Flavour

October, 2020

Based on (a small part of) my writeup,

"Lectures on Statistics in Theory: Prelude to Statistics in Practice"

https://arxiv.org/abs/1807.05996 and references therein.

Section numbers in today's slides refer to this arxiv post.

Notation

x denotes observable(s), can be multi-D

More generally, x is any convenient or useful function of the observable(s), and is called a “statistic” or “test statistic”

μ denotes parameter(s)   (also use θ if there are muons around)

$p(x|\mu)$ is probability/pdf characterizing everything that determines the probabilities (densities) of the observations, from laws of physics to experiment setup and protocol

$p(x|\mu)$ is called the “statistical model” or simply “the model” by statisticians.

Basic notions of confidence intervals (Sec. 6.2)

In two sentences:

Given the model $p(x|\mu)$ and the observed value x_{obs}, for what values of μ is x_{obs} an “extreme” value of x?

Include in the confidence interval $[\mu_1,\mu_2]$ those values of μ for which x_{obs} is *not* “extreme”.

Basic notions of confidence intervals (Sec. 6.2)

In two sentences:

Given the model $p(x|\mu)$ and the observed value x_{obs}, for what values of μ is x_{obs} an "extreme" value of x?
Include in the confidence interval $[\mu_1,\mu_2]$ those values of μ for which x_{obs} is *not* "extreme".

To be well-defined, the first point needs to be supplemented:

1) In order to define "extreme", one needs to choose an *ordering principle* for x applicable to each μ: *high rank means not extreme*.

2) One also needs to specify what *fraction* of values of x are *not* considered extreme. Called the *confidence level* C.L.; α = 1 – C.L.

Basic notions of confidence intervals (cont.)

Three common ordering choices in 1D
(when $p(x|\mu)$ is such that higher μ implies higher average x):

1. **Order x from largest to smallest.**
   Leads to confidence intervals known as *upper limits* on μ.
2. **Order x from smallest to largest. Leads to *lower limits* on μ.**
3. **Order x using *central* quantiles of $p(x|\mu)$.**
   Gives *central* confidence intervals for μ.

These orderings apply only when x is 1D

Basic notions of confidence intervals (cont.)

So, one-sentence definition of confidence interval:

The *confidence interval* $[\mu_1,\mu_2]$ contains those values of μ for which x_{obs} is *not* “extreme” at the chosen C.L., *given the ordering*.

See Section 6.8 (and F-C paper) for graphical equivalent that we call “Neyman’s construction”, and “confidence belts”

[SKIP]

Confidence Intervals and Coverage (Sec. 6.11)

Let μ_t be the unknown true value of μ. In repeated experiments, confidence intervals will have different endpoints $[\mu_1, \mu_2]$, since the endpoints are functions of the randomly sampled x.

A little thought will convince you that a fraction C.L. = 1 – α of confidence intervals so obtained will contain ("cover") the fixed but unknown μ_t . I.e.,

$P(\mu_t \in [\mu_1, \mu_2])$ = C.L. = 1 – α. (Definition of coverage)

[SKIP]

Confidence Intervals and Coverage (Sec. 6.11)

Let μ_t be the unknown true value of μ. In repeated experiments, confidence intervals will have different endpoints $[\mu_1, \mu_2]$, since the endpoints are functions of the randomly sampled x.

A little thought will convince you that a fraction C.L. = $1 - \alpha$ of confidence intervals so obtained will contain ("cover") the fixed but unknown μ_t . I.e.,

$P(\mu_t \in [\mu_1, \mu_2])$ = C.L. = $1 - \alpha$. (Definition of coverage)

In this (frequentist) equation, μ_t is *fixed* and unknown.
The endpoints μ_1,μ_2 are the random variables (!).

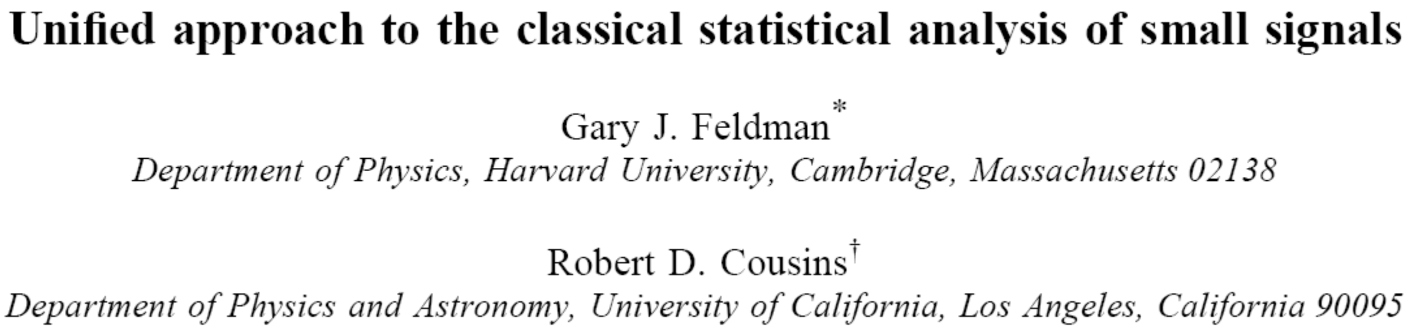
Coverage is a property of the *set* of confidence intervals, not of any one interval.

See backup re Neyman's point that expts need not be the same.

Discrete observations and/or nuisance parameters typically make exact coverage unobtainable – see writeup.

Fourth ordering: Likelihood ratios (Sec. 6.7)

4. Order x using *likelihood ratio* $\mathcal{L}(x|\mu)$ / $\mathcal{L}(x|\mu_{\text{best fit}})$, advocated by F-C.



Unified approach to the classical statistical analysis of small signals

Gary J. Feldman*
Department of Physics, Harvard University, Cambridge, Massachusetts 02138

Robert D. Cousins†
Department of Physics and Astronomy, University of California, Los Angeles, California 90095

Phys. Rev. D57 3873 (1998):

Ordering applies (in principle) for arbitrary dimensions of x, μ.

We looked "everywhere" in literature on confidence intervals, did not see this ordering used for intervals. *Was it really new?*

Instructive twist as our paper was in proof!

For that we must first turn to...

Hypothesis testing

Many special cases, including:

a) Model Selection: A given functional form ("model") vs another functional form.

b) Goodness of Fit: A given functional form against all other (unspecified) functional forms (aka "model checking")

c) Within the *same* functional form, a single value of a parameter (say 0 or 1) vs all other values.
The model with the single value is *nested* within the model with all other values

(Section 2.3)

Nested Hypothesis Testing is *very* common in HEP

There is an undetermined parameter θ in H_1 .
H_0 corresponds to a particular parameter value θ_0
E.g., zero, SM prediction, or ∞.

H_0: $\theta = \theta_0$ (the "point null", or "sharp hypothesis")
H_1: $\theta \neq \theta_0$ (the "continuous alternative").

Examples:

1) **Signal strength θ of previously unobserved physics (SM,BSM):**
   null θ_0 = 0, alternative θ > 0.

2) **$B_s^0 \rightarrow \mu^+\mu^-$ *before* observation, signal strength θ :**
   null θ_0 = 0, alternative θ > 0.

3) **$B_s^0 \rightarrow \mu^+\mu^-$ *after* observation, signal strength θ :**
   null θ_0 = SM prediction, alternative is any other $\theta \neq \theta_0$

(Section 7.3)

Classical Frequentist Hypothesis Testing

For null hypothesis H_0, order possible observations x from least extreme to most extreme, using an ordering principle (which can depend on H_1 as well). Choose a cutoff α (smallish number).

Then "reject" H_0 if observed x_{obs} is in the *most* extreme fraction α of observations x (generated under H_0). By construction,

α = probability (with x generated according to H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error)

[See elsewhere for Type II error prob β]

Classical Frequentist Hypothesis Testing

For null hypothesis H_0, order possible observations x from least extreme to most extreme, using an ordering principle (which can depend on H_1 as well). Choose a cutoff α (smallish number).

Then "reject" H_0 if observed x_{obs} is in the *most* extreme fraction α of observations x (generated under H_0). By construction,

α = probability (with x generated according to H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error)

[See elsewhere for Type II error prob β]

α is a *pre-data* assessment of risk. *After* data are obtained, the *p-value* is the smallest value of α for which H_0 would be rejected, *had it been specified in advance*.

This is numerically (if not philosophically) the same as definition used e.g. by Fisher and often taught: "*p-value* is probability under H_0 of obtaining x as extreme *or more extreme* than observed x_0." [See backup slides and google regarding p-values.]

Nested Hypothesis Testing: Duality with Intervals

In classical/frequentist formalism (but *not* Bayesian formalism), the theory of these hypo tests maps to that of confidence intervals:

Having observed data x_{obs}, suppose the 95% C.L. confidence interval for μ is $[\mu_1,\mu_2]$.

This contains all values of μ for which observed x_{obs} is ranked in the *least* extreme 95% of possible outcomes x according to $p(x|\mu)$ and the ordering principle in use.

Nested Hypothesis Testing: Duality with Intervals

In classical/frequentist formalism (but *not* Bayesian formalism), the theory of these hypo tests maps to that of confidence intervals:

Having observed data x_{obs}, suppose the 95% C.L. confidence interval for μ is $[\mu_1,\mu_2]$.
This contains all values of μ for which observed x_{obs} is ranked in the *least* extreme 95% of possible outcomes x according to $p(x|\mu)$ and the ordering principle in use.

Now suppose we test $H_0{:}\mu=\mu_0$ vs $H_1{:}\mu\neq\mu_0$ at Type I error prob $\alpha=5\%$. We reject H_0 if x_{obs} is ranked in the *most* extreme 5% of x according to $p(x|\mu)$ and the ordering principle in use.

Nested Hypothesis Testing: Duality with Intervals

In classical/frequentist formalism (but *not* Bayesian formalism), the theory of these hypo tests maps to that of confidence intervals:

Having observed data x_{obs}, suppose the 95% C.L. confidence interval for μ is $[\mu_1,\mu_2]$.
This contains all values of μ for which observed x_{obs} is ranked in the *least* extreme 95% of possible outcomes x according to $p(x|\mu)$ and the ordering principle in use.

Now suppose we test $H_0:\mu=\mu_0$ vs $H_1:\mu\neq\mu_0$ at Type I error prob α=5%. We reject H_0 if x_{obs} is ranked in the *most* extreme 5% of x according to $p(x|\mu)$ and the ordering principle in use.

Comparing the two procedures, we see:

Reject H_0 at α=5% iff μ_0 is *not* in 95% C.L. conf. interval $[\mu_1,\mu_2]$.

Use of the duality is referred to as "inverting a test" to obtain confidence intervals, and vice versa. (Section 7.4)

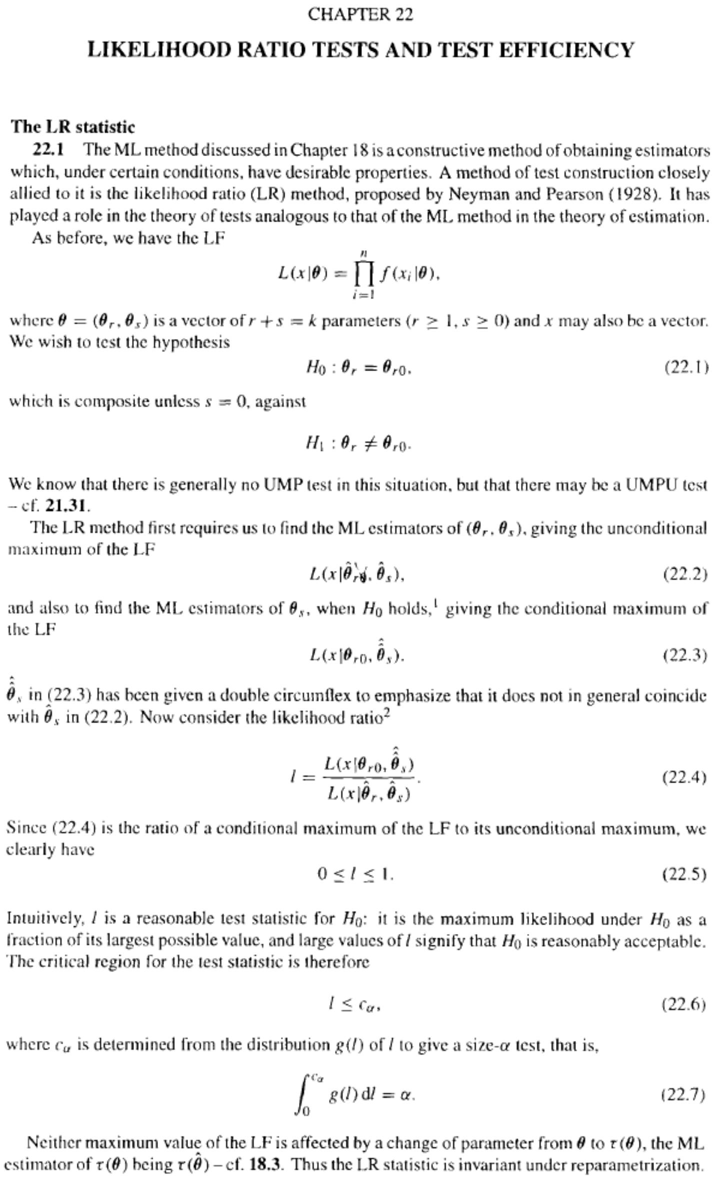
Duality in Nested Hypothesis Testing

While F-C was “in proof”, Gary realized that “our” intervals were simply those obtained by “inverting” the classic “exact” LR hypothesis test (which specifies LR ordering) in Kendall and Stuart.

It was all on 1¼ pages, plus profiling nuisance parameters!

See Gary’s Fermilab talk, “Journeys of an Accidental Statistician”,
http://users.physics.harvard.edu/~feldman/Journeys.pdf

This was of course *good* !
It led to rapid inclusion in PDG RPP.



CHAPTER 22

LIKELIHOOD RATIO TESTS AND TEST EFFICIENCY

The LR statistic

22.1 The ML method discussed in Chapter 18 is a constructive method of obtaining estimators which, under certain conditions, have desirable properties. A method of test construction closely allied to it is the likelihood ratio (LR) method, proposed by Neyman and Pearson (1928). It has played a role in the theory of tests analogous to that of the ML method in the theory of estimation.

As before, we have the LF

$$L(x|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$ is a vector of $r + s = k$ parameters ($r \geq 1, s \geq 0$) and x may also be a vector. We wish to test the hypothesis

$$H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}, \qquad (22.1)$$

which is composite unless $s = 0$, against

$$H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r0}.$$

We know that there is generally no UMP test in this situation, but that there may be a UMPU test – cf. **21.31**.

The LR method first requires us to find the ML estimators of $(\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$, giving the unconditional maximum of the LF

$$L(x|\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s), \qquad (22.2)$$

and also to find the ML estimators of $\boldsymbol{\theta}_s$, when H_0 holds,[1] giving the conditional maximum of the LF

$$L(x|\boldsymbol{\theta}_{r0}, \hat{\hat{\boldsymbol{\theta}}}_s). \qquad (22.3)$$

$\hat{\hat{\boldsymbol{\theta}}}_s$ in (22.3) has been given a double circumflex to emphasize that it does not in general coincide with $\hat{\boldsymbol{\theta}}_s$ in (22.2). Now consider the likelihood ratio[2]

$$l = \frac{L(x|\boldsymbol{\theta}_{r0}, \hat{\hat{\boldsymbol{\theta}}}_s)}{L(x|\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)}. \qquad (22.4)$$

Since (22.4) is the ratio of a conditional maximum of the LF to its unconditional maximum, we clearly have

$$0 \leq l \leq 1. \qquad (22.5)$$

Intuitively, l is a reasonable test statistic for H_0: it is the maximum likelihood under H_0 as a fraction of its largest possible value, and large values of l signify that H_0 is reasonably acceptable. The critical region for the test statistic is therefore

$$l \leq c_\alpha, \qquad (22.6)$$

where c_α is determined from the distribution $g(l)$ of l to give a size-α test, that is,

$$\int_0^{c_\alpha} g(l)\,\mathrm{d}l = \alpha. \qquad (22.7)$$

Neither maximum value of the LF is affected by a change of parameter from $\boldsymbol{\theta}$ to $\tau(\boldsymbol{\theta})$, the ML estimator of $\tau(\boldsymbol{\theta})$ being $\tau(\hat{\boldsymbol{\theta}})$ – cf. **18.3**. Thus the LR statistic is invariant under reparametrization.

Famous confusion re Gaussian $p(x|\mu)$ where $\mu \geq 0$

It is *crucial* to distinguish between the data x, which *can* be negative (no problem), and a parameter μ such as mass or signal strength, for which negative values *do not exist in the model*.

I.e., for mass $\mu < 0$, $p(x|\mu)$ does not exist: You would not know how to simulate the physics of detector response for *mass* < 0. Constraint $\mu \geq 0$ has *nothing* to do with a Bayesian prior for μ !!! It's in the *model* (and hence in $\mathcal{L}(\mu)$).

Famous confusion re Gaussian $p(x|\mu)$ where $\mu \geq 0$

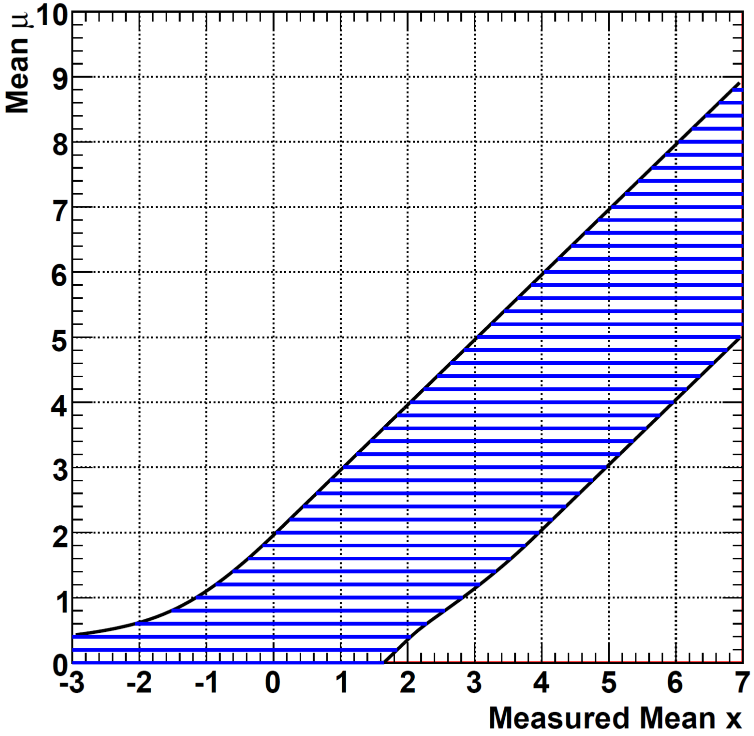
It is *crucial* to distinguish between the data x, which *can* be negative (no problem), and a parameter μ such as mass or signal strength, for which negative values *do not exist in the model*.

I.e., for mass $\mu < 0$, $p(x|\mu)$ does not exist: You would not know how to simulate the physics of detector response for *mass* < 0.

Constraint $\mu \geq 0$ has *nothing* to do with a Bayesian prior for μ !!! It's in the *model* (and hence in $\mathcal{L}(\mu)$).

The confusion is encouraged since we often refer to x as the "measured value of μ", and say that x<0 is "unphysical" – bad habits!

A proper confidence belt has x of both signs, only non-negative $\mu \geq 0$. Example: Construction on right is LR ordering advocated by F-C (Sections 6.9, 14)



>20 years of experience with F-C

Lots of experience in HEP, many find it useful, especially when:

- ★ A model parameter is bounded (mass, cross section, sine/cosine of an angle, etc.); and/or
- ★ Log-likelihood is non-Gaussian (so Wilks's Theorem is inaccurate); multiply connected confidence regions; and/or
- ★ The interesting parameter space or sample space is >1D, where LR ordering a la F-C and K&S is particularly useful, and other orderings are poorly defined (metric dependent)

Flavour experiments have one or more, so various usage.

BTW, for data with a "5-sigma discovery", the F-C "unified approach" reproduces same answer as usual one-tailed test.

>20 years of experience with F-C (cont.)

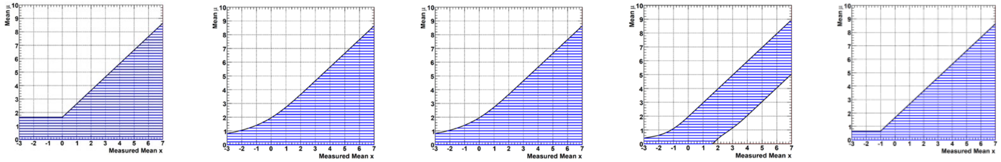
**Main foundational (philosophical) issue, already discussed in the F-C paper, is illustrated by Poisson case with non-zero expected background, zero events observed.
See Section 9.1 of arxiv post (violation of Likelihood Principle, very common in frequentist statistics).**

Main practical issues:

1) **Computational time, especially in presence of nuisance parameters.**
2) **In common with other frequentist methods, there is no automatic way to "eliminate" nuisance parameters that is always satisfactory. (Section 12)**

Comparison to other "contenders" in a prototype problem:
http://www.physics.ucla.edu/~cousins/stats/cousins_bounded_gaussian_virtual_talk_12sep2011.pdf

Bayes, Fisher, Neyman, Neutrino Masses, and the LHC



Bob Cousins

Univ. of California, Los Angeles

Virtual Talk

12 September 2011

http://www.physics.ucla.edu/~cousins/stats/cousins_bounded_gaussian_virtual_talk_12sep2011.pdf

Early uses of F-C in flavour physics

1999 CDF low-statistics measurement of CP parameter sin2β.
Sampled value was near boundary, so natural to use F-C.
Use duality to test sin2β=0 by finding that CL for which 0 is an endpoint: 93%. Equivalent to p-value of 0.07 for F-C ordering.

Measurement of $\sin 2\beta$ from $B \rightarrow J/\psi K_S^0$ with the CDF detector

PHYSICAL REVIEW D, VOLUME 61, 072005

$$\sin 2\beta = 0.79 \pm 0.39(\text{stat}) \pm 0.16(\text{syst}).$$

A scan through the likelihood function as $\sin 2\beta$ is varied is shown in Fig. 7 and demonstrates that the uncertainties follow Gaussian statistics. Using the Feldman-Cousins frequentist approach [30], we calculate a confidence interval of $0.0 < \sin 2\beta < 1$ at 93%. An alternative approach is the Bayesian method, where a flat prior distribution in $\sin 2\beta$ is assumed and a probability that $\sin 2\beta > 0.0$ of 95% is calculated. Finally, if the true value of $\sin 2\beta$ is zero, and the measurement uncertainty is 0.44 (Gaussian uncertainty), the probability of obtaining $\sin 2\beta > 0.79$ is 3.6%.

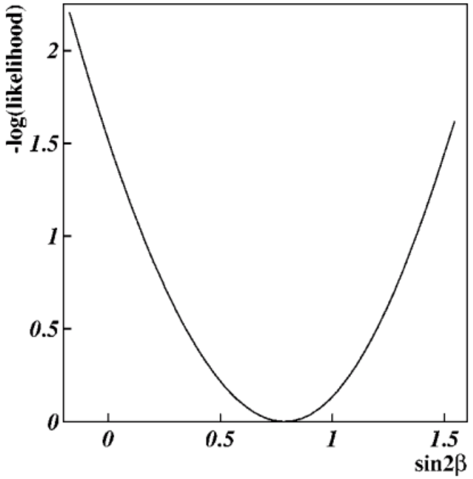
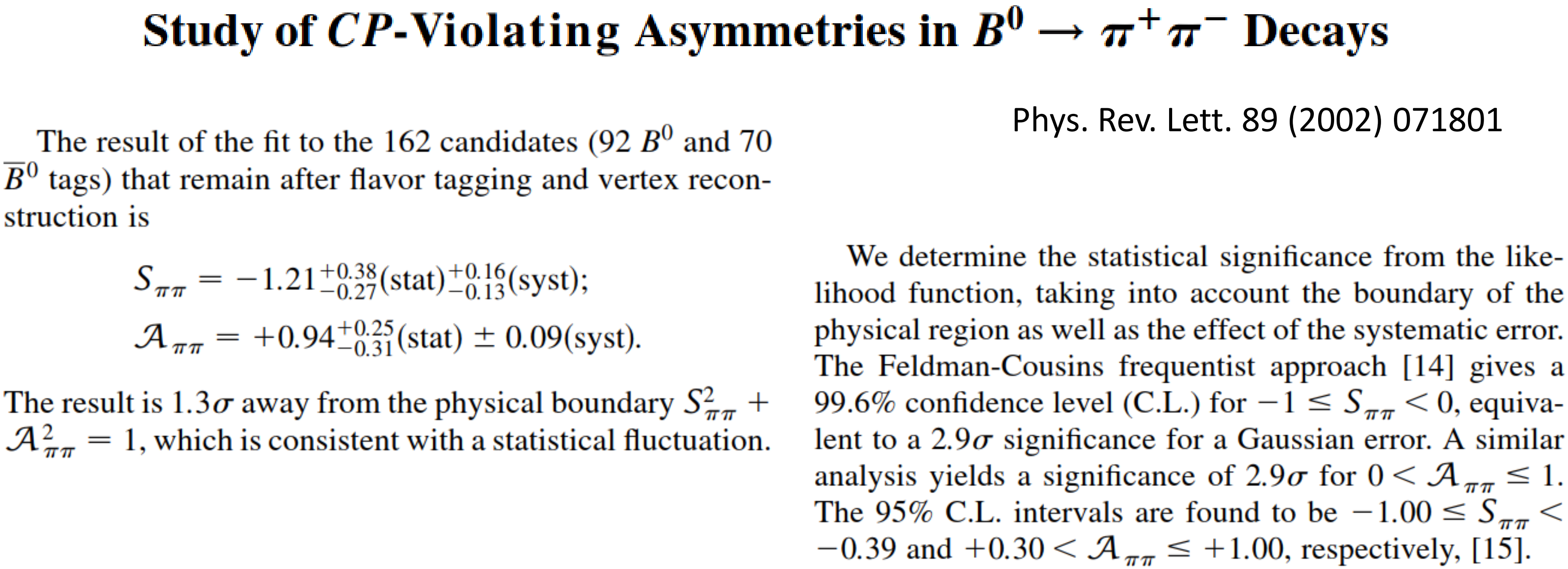


FIG. 7. A scan of the log-likelihood function. The value of $\sin 2\beta$ is scanned, and at each step, the function is minimized.

Early uses of F-C in flavour physics (cont.)

2002 Belle analysis, amplitudes S and A (both 0 if CP conserved).
Bruce Yabsley talk at Durham (post-named) PhyStat 2002.
Physically $A^2 + S^2 \leq 1$. Belle's sampled values were outside.
F-C provided frequentist way to deal with boundary.
Like CDF, Belle used test-interval duality to obtain p-value for testing S=0 and A=0 for FC ordering.



Study of CP-Violating Asymmetries in $B^0 \rightarrow \pi^+\pi^-$ Decays

Phys. Rev. Lett. 89 (2002) 071801

The result of the fit to the 162 candidates (92 B^0 and 70 $\overline{B}^0$ tags) that remain after flavor tagging and vertex reconstruction is

$$S_{\pi\pi} = -1.21^{+0.38}_{-0.27}(\text{stat})^{+0.16}_{-0.13}(\text{syst});$$

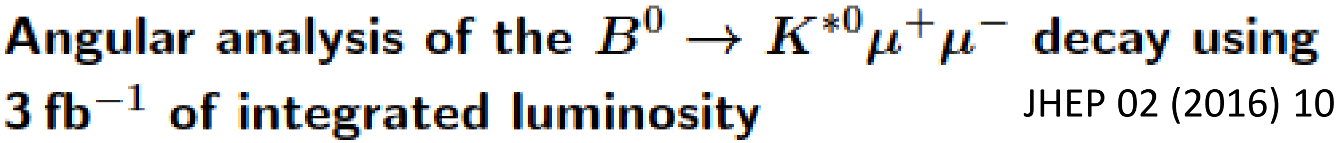
$$\mathcal{A}_{\pi\pi} = +0.94^{+0.25}_{-0.31}(\text{stat}) \pm 0.09(\text{syst}).$$

The result is 1.3σ away from the physical boundary $S^2_{\pi\pi} + \mathcal{A}^2_{\pi\pi} = 1$, which is consistent with a statistical fluctuation.

We determine the statistical significance from the likelihood function, taking into account the boundary of the physical region as well as the effect of the systematic error. The Feldman-Cousins frequentist approach [14] gives a 99.6% confidence level (C.L.) for $-1 \leq S_{\pi\pi} < 0$, equivalent to a 2.9σ significance for a Gaussian error. A similar analysis yields a significance of 2.9σ for $0 < \mathcal{A}_{\pi\pi} \leq 1$. The 95% C.L. intervals are found to be $-1.00 \leq S_{\pi\pi} < -0.39$ and $+0.30 < \mathcal{A}_{\pi\pi} \leq +1.00$, respectively, [15].

More recent uses of F-C in flavour physics

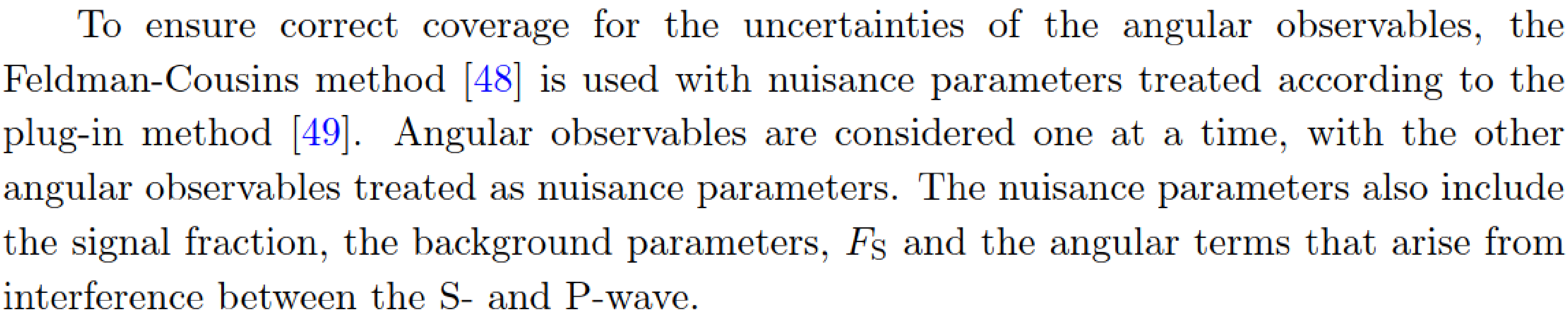
LHCb



Angular analysis of the $B^0 \to K^{*0}\mu^+\mu^-$ decay using $3\,\mathrm{fb}^{-1}$ of integrated luminosity

JHEP 02 (2016) 104

Very complicated analysis…I am familiar with CMS version --- difficult. Multiple analysis approaches, Again, physical constraints mean that F-C can provide coverage (approximate since nuisance params).



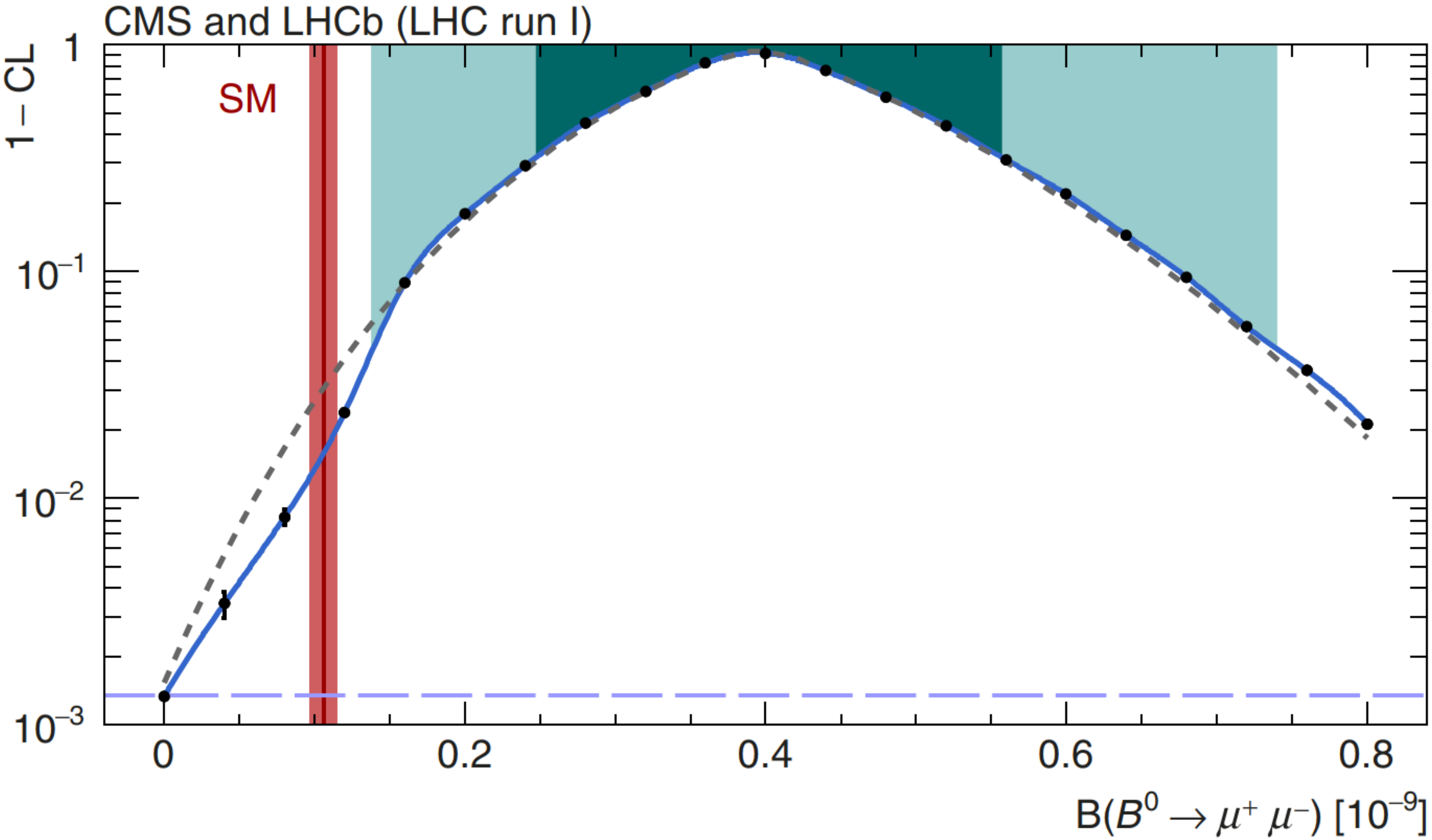
To ensure correct coverage for the uncertainties of the angular observables, the Feldman-Cousins method [48] is used with nuisance parameters treated according to the plug-in method [49]. Angular observables are considered one at a time, with the other angular observables treated as nuisance parameters. The nuisance parameters also include the signal fraction, the background parameters, F_S and the angular terms that arise from interference between the S- and P-wave.

For testing SM, they abandon F-C duality and used $\Delta\chi^2$-based test using EOS package. (I did not try to follow.)

Observation of the rare $B_s^0 \to \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data

Nature 522 (2015) 68

Interesting comparison of F-C and asymptotic Wilks



Extended Data Figure 5 | Confidence level as a function of the $\mathcal{B}(B^0 \to \mu^+\mu^-)$ hypothesis. The value of 1 − CL, where CL is the confidence level obtained with the Feldman–Cousins procedure, as a function of $\mathcal{B}(B^0 \to \mu^+\mu^-)$ is shown in logarithmic scale. The points mark the computed 1 − CL values and the curve is their spline interpolation. The dark and light (cyan) areas define the two-sided ±1σ and ±2σ confidence intervals for the branching fraction, while the dashed horizontal line defines the confidence level for the 3σ one-sided interval. The dashed (grey) curve shows the 1 − CL values computed from the one-dimensional $-2\Delta\ln L$ test statistic using Wilks' theorem. Deviations between these confidence level values and those from the Feldman–Cousins procedure[30] illustrate the degree of approximation implied by the asymptotic assumptions inherent to Wilks' theorem[29].

Back to main theme of this talk:

Hypothesis testing of a point null vs a continuous alternative

Bayesian Hypothesis Testing (Model Selection)

Forget the duality with intervals. Estimation ≠ testing!

Typically follows Chapter 5 of book by Harold Jeffreys: Bayes's Theorem is applied to the models themselves after integrating out *all* parameters, including parameter of interest!

Presented too often as "logical" and therefore simple to use, with great benefits such as automatic "Ockham's razor", etc.

In fact, it is *full of subtleties*. E.g., Jeffreys and followers use *different priors* for integrating out parameter in model selection than for *same* parameter in parameter estimation.

(Sections 5, 10, Appendix A)

Bayesian Hypothesis Testing (Model Selection)

Here I will mainly just say: Beware! There are posted/published applications HEP that are silly (*by Bayesian standards*). A pentaquark example in PRL provoked me to write a Comment: https://arxiv.org/abs/0807.1330 .

For testing point null vs continuous alternative, in asymptotic limit of large sample size, your answer (e.g. probability H_0 is true, or an odds ratio called the Bayes Factor) *remains proportional to the prior pdf of parameter of interest*.

This is *totally different* behavior compared to interval estimation, where the effect of prior p(μ) typically becomes negligible as sample size increases without bound.

Jeffreys-Lindley paradox

For fixed frequentist significance Z (number of “sigma”), the Bayesian posterior probability of H_0, as well as the Bayes Factor, depend directly on the ratio
(width of prior)/(std dev of measurement).
This factor, which provides the famous “Ockham’s razor”, leads to the *Jeffreys-Lindley paradox.*

It implies that, for experiments obtaining the *same* Z, the Bayesian answers depend on sample size (std dev typically goes as 1/sqrt(sample size). Very different behavior!

For a review and comparison to p-values in discovery of Higgs boson, see my paper, “The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics”

(Published in Synthese – long story)
https://arxiv.org/abs/1310.3791 .

[SKIP?]

Priors in Bayesian Hypothesis Testing

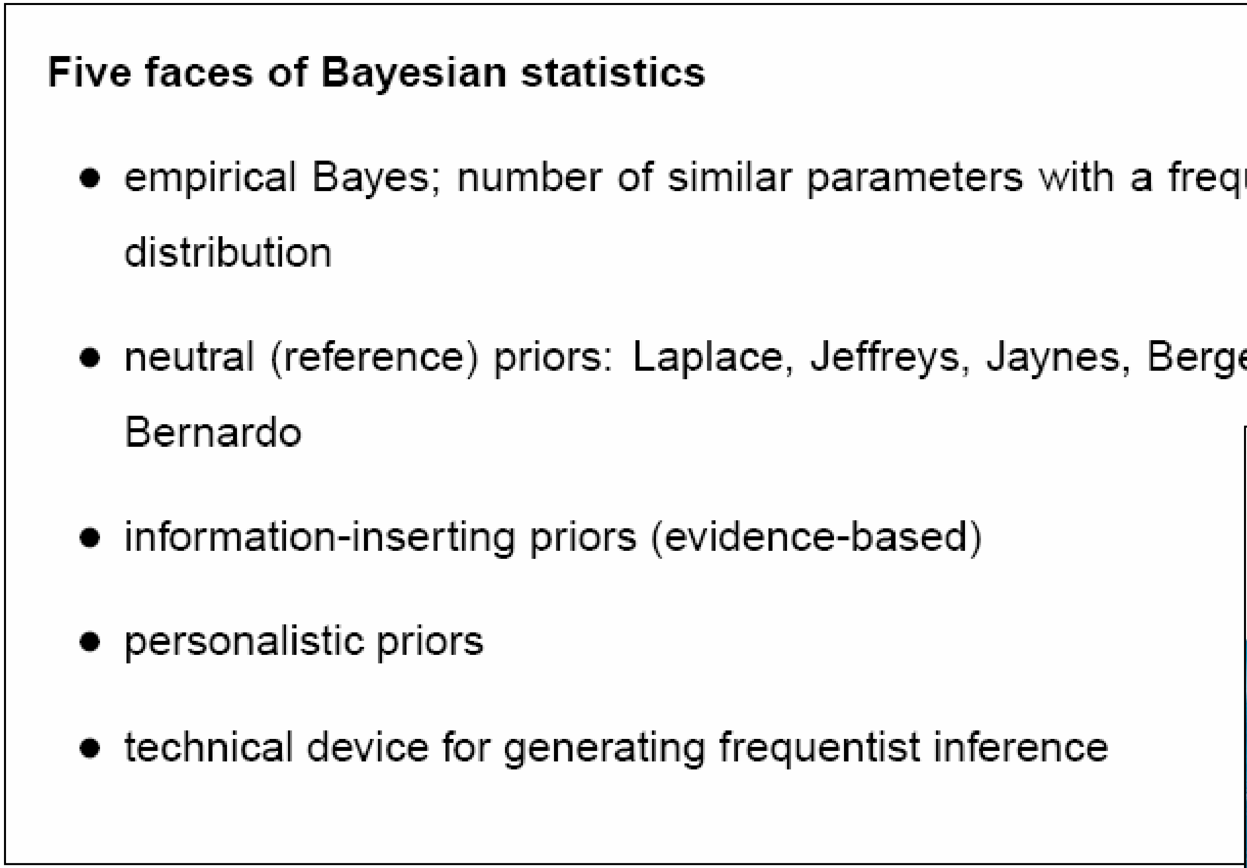
For testing H_0: $\theta = \theta_0$ vs H_1: $\theta \neq \theta_0$, improper priors $g(\theta)$ for θ that work fine for estimation (such as Jeffreys priors) become a disaster.

The ratio, (width of prior)/(std dev of measurement), diverges so H_0 always preferred.

Adding cut-off to make prior $g(\theta)$ proper just gives direct dependence on (arbitrary?) width of prior .
(Contrast with Bayesian point/interval estimation!)

Silly things like prior flat in log of mass as a way to represent "ignorance" are *strongly* informative!
(See any serious Bayesian literature.)

Sir David Cox at PhyStat-LHC 2007



Five faces of Bayesian statistics

- empirical Bayes; number of similar parameters with a frequency distribution
- neutral (reference) priors: Laplace, Jeffreys, Jaynes, Berger and Bernardo
- information-inserting priors (evidence-based)
- personalistic priors
- technical device for generating frequentist inference

PHYSTAT-LHC Workshop

on

Statistical Issues for LHC Physics

CERN Geneva June 27-29, 2007

This Workshop will address statistical topics relevant for LHC Physics analyses.Issues related to discovery, and the associated problems arising from systematic uncertainties, will feature prominently.

Contacts	
Louis Lyons	l.lyons@physics.ox.ac.uk
Albert De Roeck	Albert.de.Roeck@cern.ch

Conference secretaries	
Dorothée Denise	Dorothee.Denise@cern.ch
Kate Ross	Kate.Ross@cern.ch

Further information and registration at http://cern.ch/phystat-lhc

- technical device for generating frequentist inference

This is done, especially for upper limits, in HEP (flat prior on Poisson mean). I consider the following paper from May 2020 to be an example from flavour physics, though it was not completely clear from the published paper. My thanks to the corresponding author for clarifications.

Measurement of the $\Lambda_b^0 \to J/\psi\Lambda$ angular distribution and the Λ_b^0 polarisation in pp collisions JHEP 06 (2020)110

The LHCb collaboration

"The Bayesian analysis procedure has been validated for both small and large values of the polarisation using pseudoexperiments."

"Validated" means good frequentist coverage (!)

“The remaining amplitudes are measured relative to b_+. A uniform prior is assumed on their magnitudes and phases and on P_b. The priors use the ranges [−1, +1] for P_b, [$-\pi$, $+\pi$] for the phases, and the range [0, 20] for the magnitudes of the amplitudes.”

No Bayesian or probability-matching theory was used to choose these priors; it seems that the results are mostly insensitive to the choice, and in any case the coverage was good.

“The 95% credibility intervals are provided in table 6 of the appendix. … The resulting Λ_b^0 polarisation at each centre-of-mass energy is found to be consistent with zero.”

Note: The frequentist definition of “consistent” (inside the interval) was used to test the “point null” of zero.

[SKIP to bottom]

“Table 5. Estimates for the magnitude and phase of the decay amplitudes and the transverse production polarisation of the Λ_b^0 baryons, extracted using the Bayesian analysis. The most probable value (MPV) and the shortest 68% interval containing the most probable value are given.”

Both MPV and “shortest” are metric-dependent. The choices here seem OK, if not optimal, from frequentist perspective.

The 68% credibility interval around the most probable value is [−0.048, 0.005]. This measurement is consistent with, but more precise than, previous measurements of α_b by the ATLAS, CMS and LHCb collaborations [26–28].

Since [26–28] were not Bayesian analyses, it would be interesting to compare recipes. Since this paper had coverage checked, that presumably means answers would be similar.

Conclusion: “Bayesian recipe” with good frequentist coverage can be win-win, since likelihood principle built in.

Workshop Homepage

Workshop on Confidence Limits

27-28 March, 2000
Fermilab 1-West Conference Room

Jim Berger:

M. Kendall, giving the 'old' frequentist viewpoint of Bayesian analysis:

"If they [Bayesians] would only do as he [Bayes] did and publish posthumously, we should all be saved a lot of trouble."

What should be the view today:

Objective Bayesian analysis is the best frequentist tool around.

My advocacy for >10 years (Section 16):

Have in place tools to allow computation of results using a variety of recipes, for problems up to intermediate complexity:

- **Bayesian with analysis of sensitivity to prior**
- **Profile likelihood ratio (Minuit MINOS)**
- **Frequentist construction with approximate treatment of nuisance parameters**
- **Other “favorites” such as LEP’s $\mathrm{CL_S}$ (an HEP invention)**

My advocacy for >10 years (Section 16):

Have in place tools to allow computation of results using a variety of recipes, for problems up to intermediate complexity:

- Bayesian with analysis of sensitivity to prior
- Profile likelihood ratio (Minuit MINOS)
- Frequentist construction with approximate treatment of nuisance parameters
- Other “favorites” such as LEP’s $\mathrm{CL_S}$ (an HEP invention)

The community can (and should) then demand that a result shown with one’s preferred method also be shown with the other methods, *and with sampling properties studied*.

When the methods all agree, we are in asymptopic nirvana.

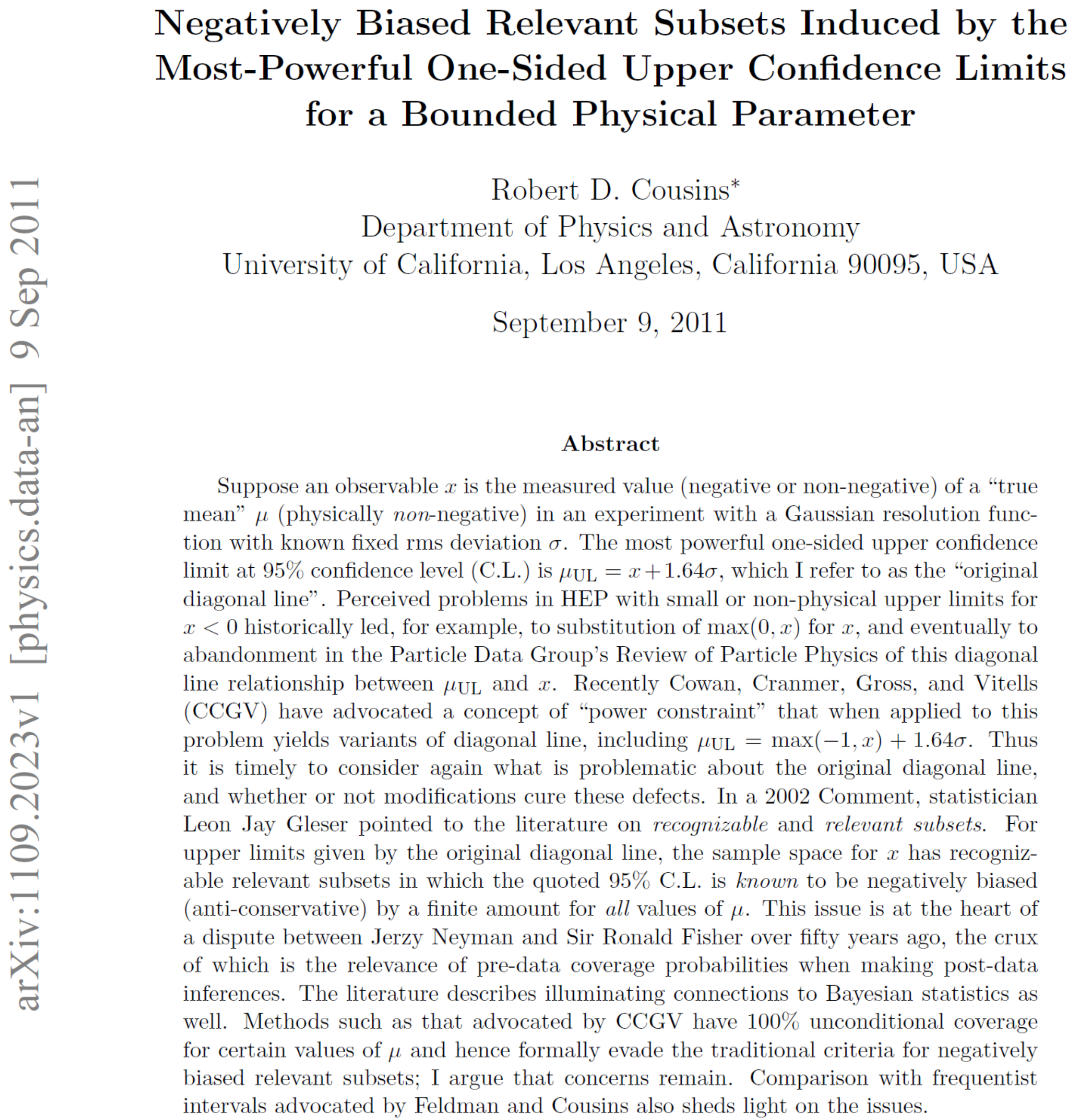
When the methods disagree, we are reminded that the results are answers to different questions, and we learn something! E.g.:

- Bayesian methods can have poor frequentist properties
- Frequentist methods can badly violate likelihood principle

Thanks to all (see note), including my "sponsor", U.S. DOE Office of Science

P.S. On another topic, I wrote a note for students, "What is the likelihood function, and how is it used in particle physics? https://arxiv.org/abs/2010.00356

BACKUP



Negatively Biased Relevant Subsets Induced by the Most-Powerful One-Sided Upper Confidence Limits for a Bounded Physical Parameter

Robert D. Cousins*
Department of Physics and Astronomy
University of California, Los Angeles, California 90095, USA

September 9, 2011

Abstract

Suppose an observable x is the measured value (negative or non-negative) of a "true mean" μ (physically *non*-negative) in an experiment with a Gaussian resolution function with known fixed rms deviation σ. The most powerful one-sided upper confidence limit at 95% confidence level (C.L.) is $\mu_{\mathrm{UL}} = x + 1.64\sigma$, which I refer to as the "original diagonal line". Perceived problems in HEP with small or non-physical upper limits for $x < 0$ historically led, for example, to substitution of $\max(0, x)$ for x, and eventually to abandonment in the Particle Data Group's Review of Particle Physics of this diagonal line relationship between μ_{UL} and x. Recently Cowan, Cranmer, Gross, and Vitells (CCGV) have advocated a concept of "power constraint" that when applied to this problem yields variants of diagonal line, including $\mu_{\mathrm{UL}} = \max(-1, x) + 1.64\sigma$. Thus it is timely to consider again what is problematic about the original diagonal line, and whether or not modifications cure these defects. In a 2002 Comment, statistician Leon Jay Gleser pointed to the literature on *recognizable* and *relevant subsets*. For upper limits given by the original diagonal line, the sample space for x has recognizable relevant subsets in which the quoted 95% C.L. is *known* to be negatively biased (anti-conservative) by a finite amount for *all* values of μ. This issue is at the heart of a dispute between Jerzy Neyman and Sir Ronald Fisher over fifty years ago, the crux of which is the relevance of pre-data coverage probabilities when making post-data inferences. The literature describes illuminating connections to Bayesian statistics as well. Methods such as that advocated by CCGV have 100% unconditional coverage for certain values of μ and hence formally evade the traditional criteria for negatively biased relevant subsets; I argue that concerns remain. Comparison with frequentist intervals advocated by Feldman and Cousins also sheds light on the issues.

Bayesian hypothesis testing for nested case H_0: $\theta=\theta_0$ vs H_1: $\theta\neq\theta_0$

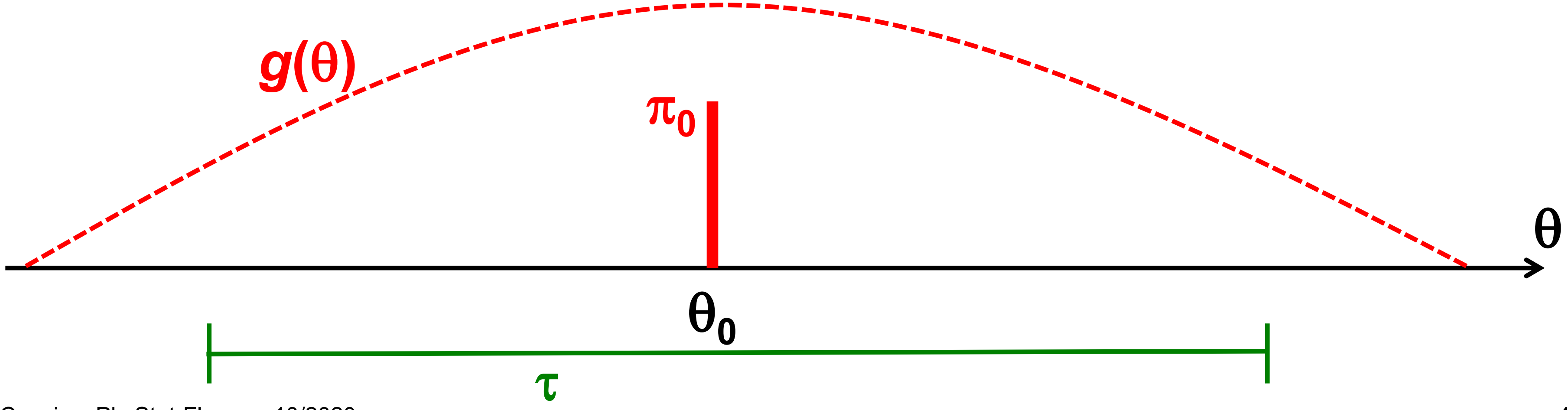
Let π_0 be prior prob for H_0. Then $\pi_1 = 1-\pi_0$ is prior prob for H_1.

Conditional on H_1 true: prior pdf for θ, $g(\theta)$.

π_0 is like bit of Dirac δ-ftn ("probability mass") at $\theta=\theta_0$.
In practice can have a little width:

ε_0 = scale of width of null value(s) of θ

scale τ: extent of prior plausible values in $g(\theta)$



Gaussian model $p(x|\theta)$ with rms σ_{tot} , sampled value x_{obs} .

ML Estimate for θ is $\hat{\theta} = x_{obs}$.

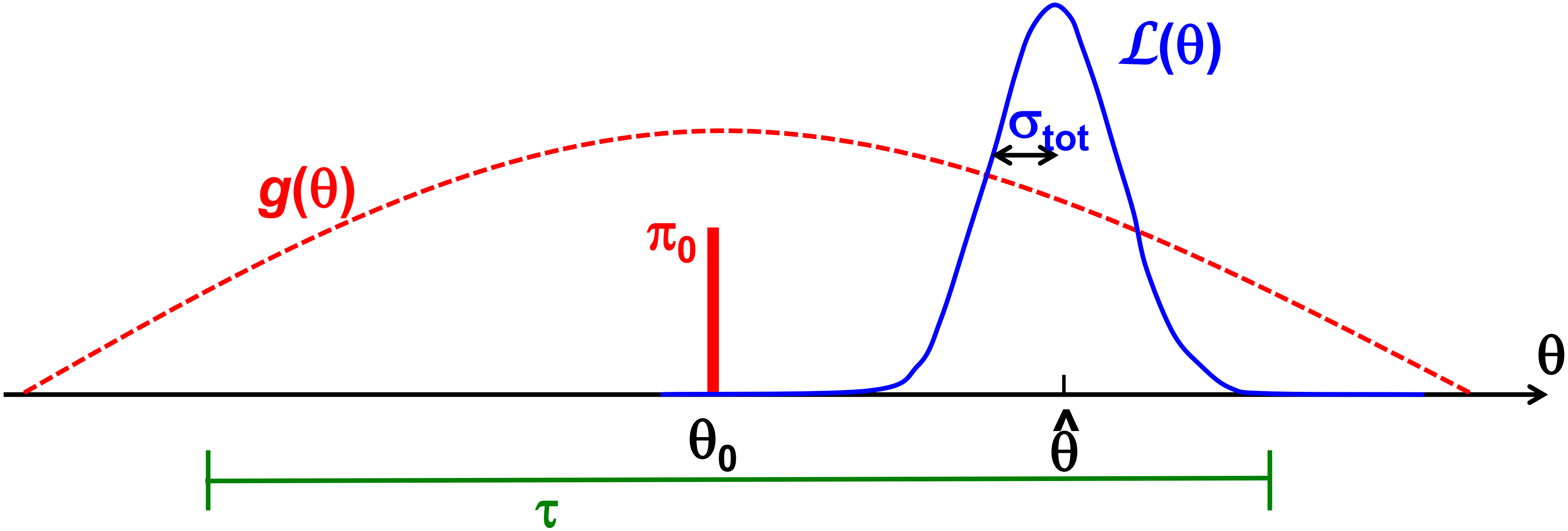
Departure from null in sigma: $Z = (\hat{\theta} - \theta_0)/\sigma_{tot}$

Sketch has $Z \approx 5$.

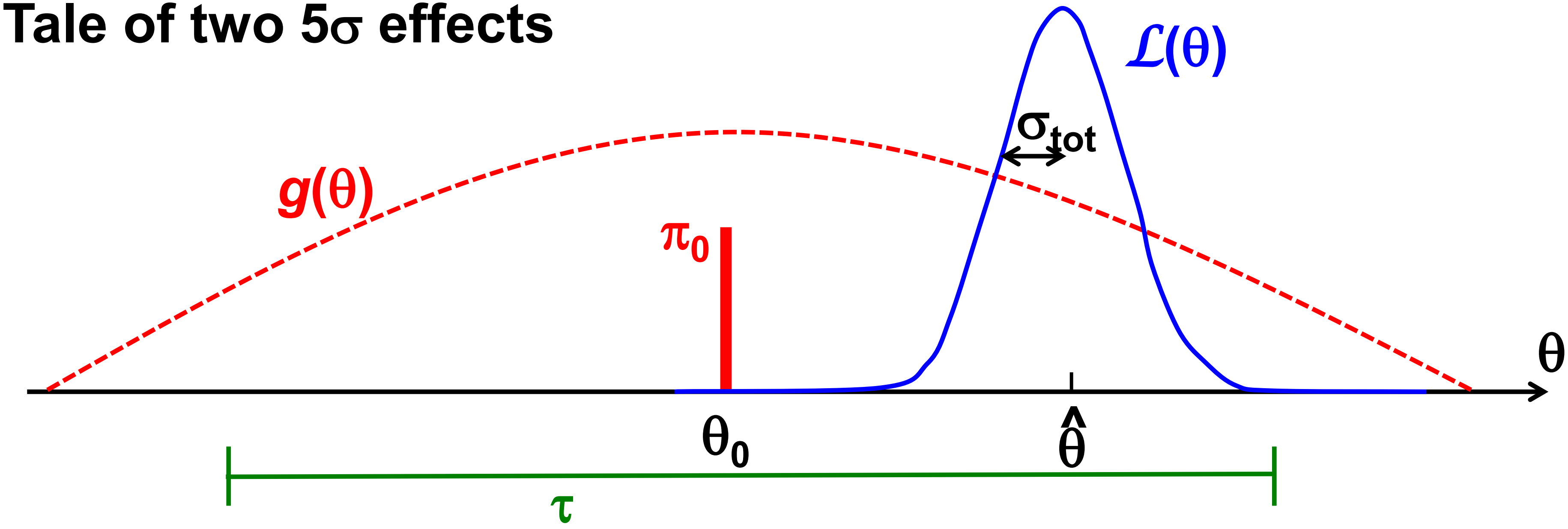
Three independent scales: gets interesting when, as shown,

$\varepsilon_0 << \sigma_{tot} << \tau$.

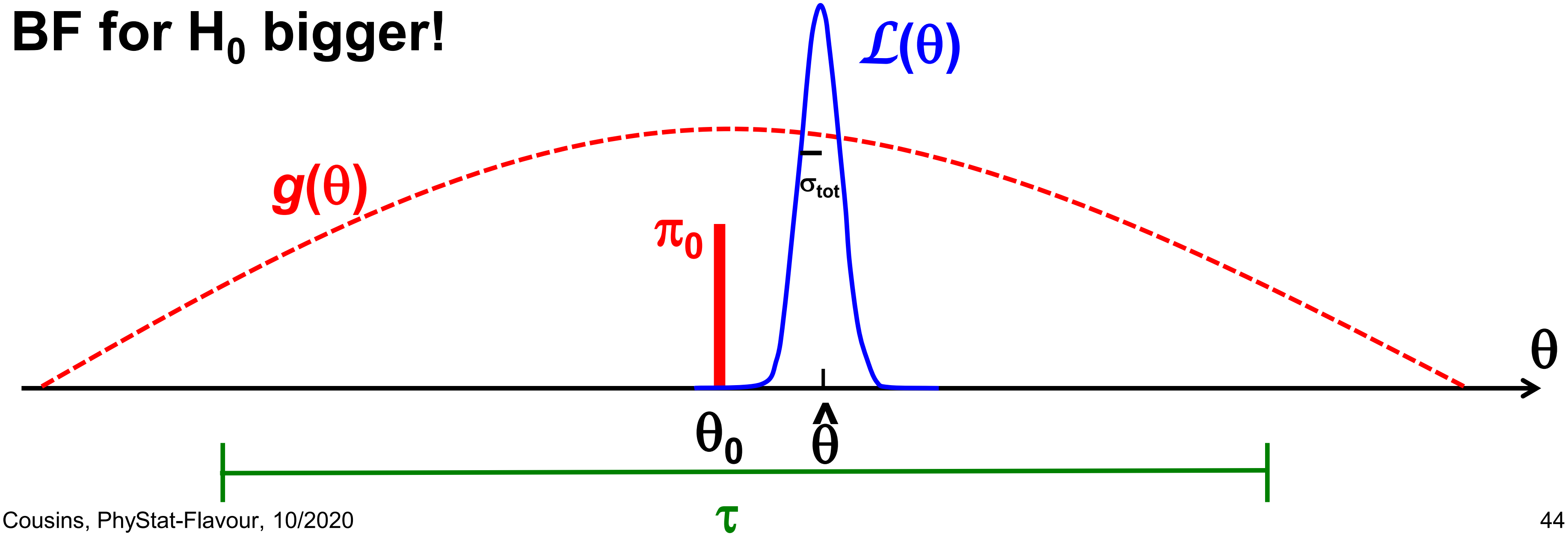
Bayesian posterior prob for H_0, and Bayes Factor are prop to τ/σ_{tot} (1/Ockham factor), *independent of Z!*



Tale of two 5σ effects



σ_{tot} smaller, τ / σ_{tot} larger:
BF for H_0 bigger!



A side note on priors for “Scales”

As my writeup mentions (following Eadie et al. nearly 50 years earlier) various words including “estimation” have different meanings to statisticians than to physicists. Beware!

Since then, I have realized that a disastrous mistake seems to be made by some physicists regarding the word “scale”.

Recall (or learn about): So-called “objective priors” or “default priors” (often called by the misnomer “noninformative priors”) in Bayesian estimation are based on the *measurement model*, i.e., the measuring apparatus and the protocol (stopping rule, etc.). (Jeffreys’s Rule, Bernardo-Berger Reference Priors, etc.)

E.g., if the measuring apparatus has Gaussian resolution for some parameter (say mass-squared), then the default prior for that parameter (for estimation) is uniform, with all that implies.

A side note on priors for “Scales” (cont.)

To a statistician, whether or not a parameter is a *scale parameter* again depends on the *measurement model* (!).

Parameter θ is a scale parameter if the model $p(x|\theta)$ has the form: $p(x|\theta) = (1/\theta)\ f(x/\theta)$.

From this one can partly derive and partly argue* that invariance of prior form under change of scale parameter implies the non-subjective prior $p(\theta) = 1/\theta$, i.e., a prior uniform in $\log(\theta)$.

***See pp. 85-87 of Jim Berger’s book on decision theory for subtleties of derivation. See also Jeffreys pp. 120-123, which he abandons later in the book.**

A side note on priors for “Scales” (cont.)

To a statistician, whether or not a parameter is a *scale parameter* again depends on the *measurement model* (!).

Parameter θ is a scale parameter if the model $p(x|\theta)$ has the form:

$p(x|\theta) = (1/\theta)\ f(x/\theta)$.

From this one can partly derive and partly argue* that invariance of prior form under change of scale parameter implies the non-subjective prior $p(\theta) = 1/\theta$, i.e., a prior uniform in $\log(\theta)$.

To a physicist, a “scale” is a quantity that sets the size of physical quantities like mass, length.

E.g., “What is the DM mass scale?”

So it seems that some physicists make the mistake of saying, “Since mass is a *scale*, I use the prior uniform in log(mass).”

OOPS! This “scale” is not a statistician’s “scale parameter”!

See Comment https://arxiv.org/abs/1703.04585

There is a large literature on frequentist properties of Bayesian (inspired) procedures

Google on:
probability matching priors
Welch and Peers 1963
calibrated Bayes
Bayes non-Bayes compromise
prior predictive p-value
posterior predictive p-value
etc.

A nice introductory review is by M.J. Bayarri and J.O. Berger, "The Interplay of Bayesian and Frequentist Analysis", Statist. Sci. 19 58-80 (2004), doi:10.1214/088342304000000116

We should be doing more of this in HEP, in my opinion.

Coverage of Bayesian estimation procedures

Pre-data, Bayesians have the model $p(x|\mu)$.
Thus, quite apart from imagined repeated experiments (to which they may object) or frequentist definition of probability (to which they may object), a Bayesian can calculate:

As a function of μ, what is the coverage probability of the credible interval $[\mu_1, \mu_2]$ that they will report: what is the probability, given the model $p(x|\mu)$ (with whatever definition of p they use), that their procedure will lead to an interval in which $\mu \in [\mu_1, \mu_2]$.

This is a crucial diagnostic to report to the consumer, especially if default priors are used! (Jim B. says reference priors will work.)

(Of course, one can also average this coverage over μ, weighted by either the prior or the posterior.)

Evaluation of properties of Bayesian hypothesis testing procedures

Similarly, quite apart from imagined repeated experiments or frequentist definition of p, a Bayesian can calculate:

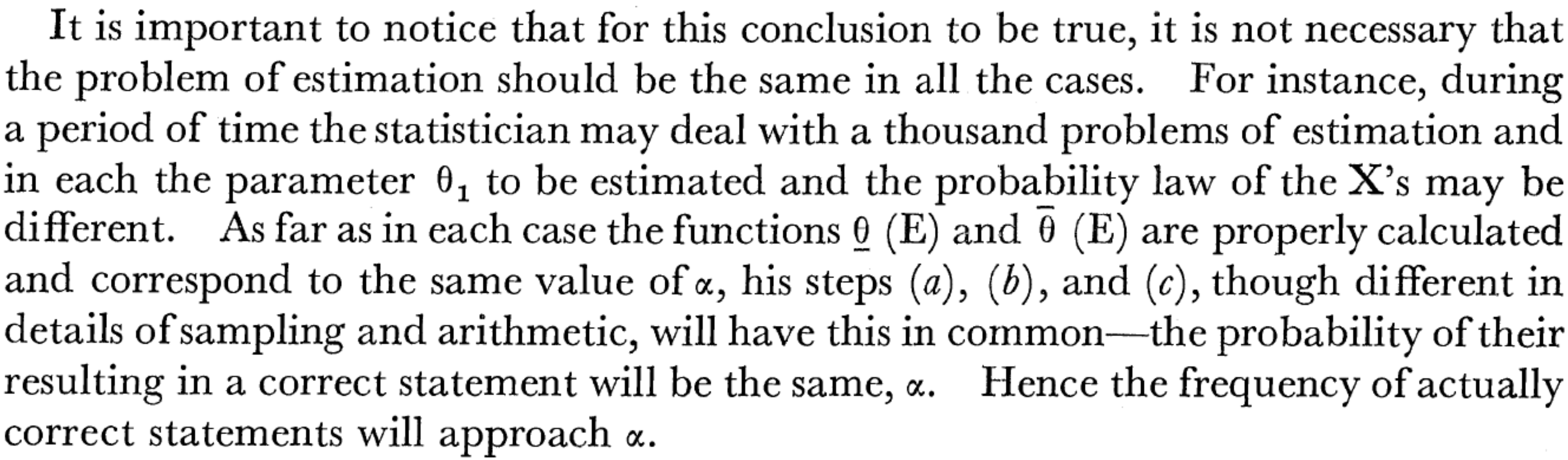
As a function of assumed H_0 and H_1 and any parameters, what is the distribution of the Bayes Factors that they will report: what is the probability, given each model $p(x|H_i,\mu)$ (with whatever definition of p they use), that their procedure will obtain various values of the Bayes Factor (or posterior probabilities).

This is also a crucial diagnostic to report to the consumer, especially if attempts at "noninformative" priors are used!

(enlightening for seeing relationship between Bayes Factors and p-values)

Coverage: The experiments in the ensemble do not have to be the same.

Neyman pointed this out in his 1937 paper (in which his α is the modern 1 - α):



It is important to notice that for this conclusion to be true, it is not necessary that the problem of estimation should be the same in all the cases. For instance, during a period of time the statistician may deal with a thousand problems of estimation and in each the parameter θ_1 to be estimated and the probability law of the X's may be different. As far as in each case the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are properly calculated and correspond to the same value of α, his steps (*a*), (*b*), and (*c*), though different in details of sampling and arithmetic, will have this in common—the probability of their resulting in a correct statement will be the same, α. Hence the frequency of actually correct statements will approach α.

Above is all “pre-data” characterization of the test

How to characterize *post-data*?

p-values and Z-values

In N-P theory, α is *specified in advance*.

Suppose after obtaining data, you notice that with α=0.05 previously specified, you reject H_0, but with α=0.01 previously specified, you accept H_0.

In fact, you determine that with the data set in hand, H_0 would be rejected for $\alpha \geq 0.023$. This interesting value has a name:

After data are obtained, the *p-value* is the smallest value of α for which H_0 would be rejected, *had it been specified in advance*.

This is numerically (if not philosophically) the same as definition used e.g. by Fisher and often taught: “*p-value* is probability under H_0 of obtaining x as extreme *or more extreme* than observed x_0.”

Interpreting p-values and Z-values

It is crucial to realize that that value of α (0.023 in the example) was typically *not* specified in advance, so p-values do *not* correspond to Type I error probs of experiments reporting them.

In HEP, p-value typically converted to Z-value (unfortunately commonly called "the significance S"), equivalent number of Gaussian sigma.*

E.g.., for one-tailed test, p = 2.87E-7 is Z = 5.

Interpreting p-values and Z-values (cont.)

Interpretation of p-values (and hence Z-values) is a long, contentious story – beware!

Widely bashed. I give some reasons why in https://arxiv.org/abs/1807.05996 .

I defend their use in HEP. See https://arxiv.org/abs/1310.3791.)

Whatever they are, p-values are *not* the probability that H_0 is true!

- **They are calculated *assuming that* H_0 *is true*, so they can hardly tell you the probability that H_0 is true!**
- **Calculation of "probability that H_0 is true" requires prior(s)!**

**Please help educate press officers and journalists!
(and physicists) !**

Whatever you call non-subjective priors, they do *not* represent ignorance!

Dennis V. Lindley Stat. Sci 5 85 (1990), "the mistake is to think of them [Jeffreys priors or Bernardo/Berger's reference priors] as representing ignorance"

This Lindley quote is emphasized by Christian Robert, *The Bayesian Choice*, (2007) p. 29.

Jose Bernardo: "[With non-subjective priors,] The contribution of the data in constructing the posterior of interest should be "dominant". Note that this does not mean that a non-subjective prior is a mathematical description of "ignorance". Any prior reflects some form of knowledge."

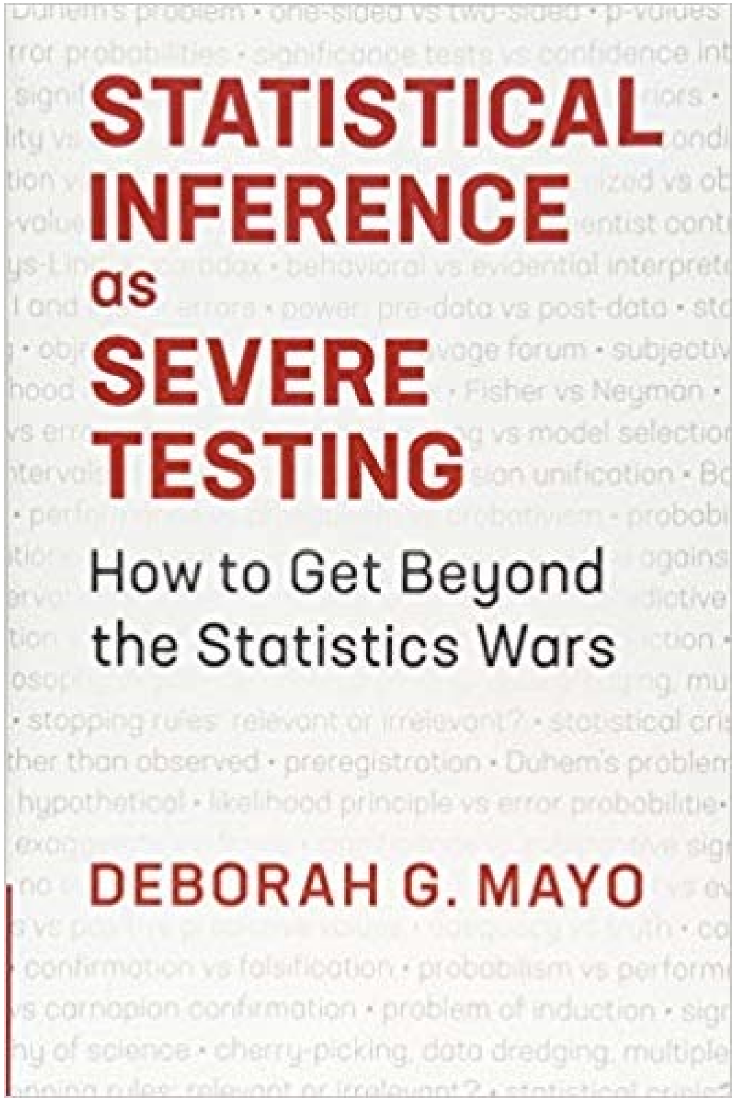
Nonetheless, Berger (1985, p. 90) argues that Bayesian analysis with noninformative priors (older name for objective priors) such as Jeffreys and Barnardo/Berger "is the single most powerful method of statistical analysis, in the sense of being the *ad hoc* method most likely to yield a sensible answer for a given investment of effort".

Recent book exploring Bayesian-frequentist divide

Much interesting history and up to-date discussion of both theory and practice, including, for example, internal debates among Bayesians. Very well-referenced.

Mayo has long advocated “error statistics”, and in particular the concept of how *severely* a hypothesis has been tested in a test that “passes”.

See my note, https://arxiv.org/abs/2002.09713 , “Connections between statistical practice in elementary particle physics and the severity concept as discussed in Mayo's Statistical Inference as Severe Testing”



STATISTICAL INFERENCE as SEVERE TESTING

How to Get Beyond the Statistics Wars

DEBORAH G. MAYO