Testing a sharp null hypothesis versus a continuous alternative: Deep issues regarding this everyday problem in high energy physics

Bob Cousins

Univ. of California, Los Angeles

STAMPS Webinar, CMU

February 12, 2021

Who am I?

* Retired physics prof, experimentalist in high energy physics
* Many searches for rare processes; since 2000 on CMS at LHC
* No formal training in statistics
* Career-long "hobby" reading statistics papers, applying some things learned to HEP

What am I doing here?

* Mikael Kuusela et al invited me! (Thanks!)

My talk today draws mainly on a paper and lecture notes, with references therein:

[1] "The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics" https://arxiv.org/abs/1310.3791 (Synthese online 2014)

[2] "Lectures on Statistics in Theory: Prelude to Statistics in Practice" https://arxiv.org/abs/1807.05996.

Outline

I. Confidence intervals and duality with NP test of the talk title
   - LR ratio ordering, advocated in HEP by Feldman and RC

II. Bayesian test of the talk title, a la Harold Jeffreys

III. Jeffreys-Lindley paradox: result of three unrelated scales

IV. Various comments on JL paradox, and my responses
   a) No good scientist believes a point null hypothesis, since "All models are wrong!", so why are we testing one?
   b) Is there an "objective" or "default" way to set the scale of prior pdf for the continuous parameter of interest?
   c) Should inference depend on sample size, for fixed *p*-value? Should we patch BF or *p*-value with $\sqrt{n}$?

Goal of the talk

I presume that most or all people in this audience will know essentially everything I say about statistics, but I say it anyway just to make sure that there are not subtle (or large) differences between the way I understand it and the way you do.

I also suspect that everyone has read or written about the Jeffreys-Lindley paradox at some point, but most or all have "moved on" to issues relevant to their daily work.

My goal today is to emphasize that these issues are at the heart of what we do in HEP, in the hope of inspiring some to reexamine them, with examples from HEP in mind.

Notation

X: "statistic" or "test statistic": the observable or any convenient or useful function of the observable(s)

x: (lower case) observed value of X

θ: parameter(s)

p(X|θ): Statistical model is probability/pdf characterizing everything that determines the probabilities/densities of X, from laws of physics to experiment setup and protocol

Confidence intervals

Given the model p(X|θ), one chooses an *ordering principle* for the sample space X, applicable to each θ: *higher rank is less extreme*.

One specifies the *fraction* of values of X that are *not* considered extreme: the *confidence level* C.L.; α = 1 – C.L.

Having obtained the observed value x, the *confidence interval* $[\theta_1,\theta_2]$ contains those values of θ for which x is *not* extreme at the chosen C.L., *given the ordering*.

Confidence Intervals and Coverage

Let θ_t be the (fixed) unknown true value of θ. Then on repeated application of the method in an ensemble of expts, (ideally)

$$P(\theta_t \in [\theta_1, \theta_2]) = \text{C.L.} = 1 - \alpha,$$

where the random variables are the endpoints of the intervals.

Nuisance parameters and discrete X are of course a nuisance.
See backup re Neyman's point that expts need not be the same.

Ordering the sample space

Three common ordering choices in 1D

(when p(X|θ) is such that higher θ implies higher average X):

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

These orderings apply only when X is 1D

Fourth ordering: Likelihood ratios

Order X using *likelihood ratio* $\mathcal{L}(X|\theta) / \mathcal{L}(X|\theta_{\text{best fit}})$, 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, θ.

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:

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

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

- **Within the *same* functional form $p(X|\theta)$, a single value $\theta = \theta_0$ (e.g., 0, SM prediction, 1, ∞) vs all other θ. The model with $\theta = \theta_0$ is *nested* within the "larger" model.**

Nested Hypothesis Testing is *very* common in HEP

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 new physics: null $\theta_0 = 0$, alternative $\theta > 0$

2) Signal strength θ of Higgs boson $\rightarrow \gamma\gamma$ *before* observation:
null $\theta_0 = 0$, alternative $\theta > 0$

3) Higgs boson $\rightarrow \gamma\gamma$ *after* observation (rejecting $\theta_0 = 0$):
null θ_0 = SM prediction, alternative is any other $\theta \neq \theta_0$

Frequentist Hypothesis Testing a la Neyman-Pearson

For null hypothesis H_0, order possible sample space 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 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 (Type I error)**

[See elsewhere for Type II error prob β]

Post-data: *p*-values and *z*-values

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 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."

In HEP, we take the p-value (which is *not* based on assumptions of normality in the model), and *simply for communication* convert it to a (one-tailed)*z-value*: the equivalent number of Gaussian σ:
E.g., for one-tailed test,

$p = 1.35 \times 10^{-3} \leftrightarrow z = 3$

$p = 2.87 \times 10^{-7} \leftrightarrow z = 5$

Nested Hypothesis Testing: Duality with Intervals

In classical/frequentist formalism, the theory of these hypo tests maps to that of confidence intervals:

Reject H_0 at α iff θ_0 is *not* in the conf. interval $[\theta_1,\theta_2]$ with C.L.=1-α

(both using the same ordering)

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

Duality in Nested Hypothesis Testing

While F-C was "in proof", Gary realized that "our" intervals were simply those obtained by "inverting" the classic 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 *good* ! Quickly included in Particle Data Group's *Review of Particle Physics*.

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

$$H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}, \tag{22.1}$$

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

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

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

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

$\hat{\hat{\boldsymbol{\theta}}}_s$ 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]

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

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

$$0 \leq l \leq 1. \tag{22.5}$$

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

$$l \leq c_\alpha, \tag{22.6}$$

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

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

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.

See backup slides for a few notes on 20+ years of experience in HEP with F-C.

Back to main theme of this talk:

Hypothesis testing of a point null vs a continuous alternative.

What do Bayesians say?

Still useful starting points (including Comments and Rejoinder): Berger & Sellke, JASA 82 (1987) 112; and:

Statistical Science
1987, Vol. 2, No. 3, 317–352

Testing Precise Hypotheses

James O. Berger and Mohan Delampady

5. WHAT SHOULD BE DONE?

First and foremost, when testing precise hypotheses, formal use of P-values should be abandoned. Almost anything will give a better indication of the evidence provided by the data against H_0.

(The key qualifier "formal" is defined in Rejoinder.)

***Forget* the duality with intervals (!).**

Traditional Bayesian Hypothesis Testing

In all editions of Harold Jeffreys's *Theory of Probability* beginning in 1930's.

Bayes's Theorem is applied to the models themselves after integrating out *all* parameters, including parameter of interest θ.

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 of interest in testing than for *same* parameter in estimation.

Various statisticians who otherwise use Bayesian methods have expressed criticisms of Jeffreys's approach. I will mention a couple alternatives in passing, but focus today on Jeffreys.

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.

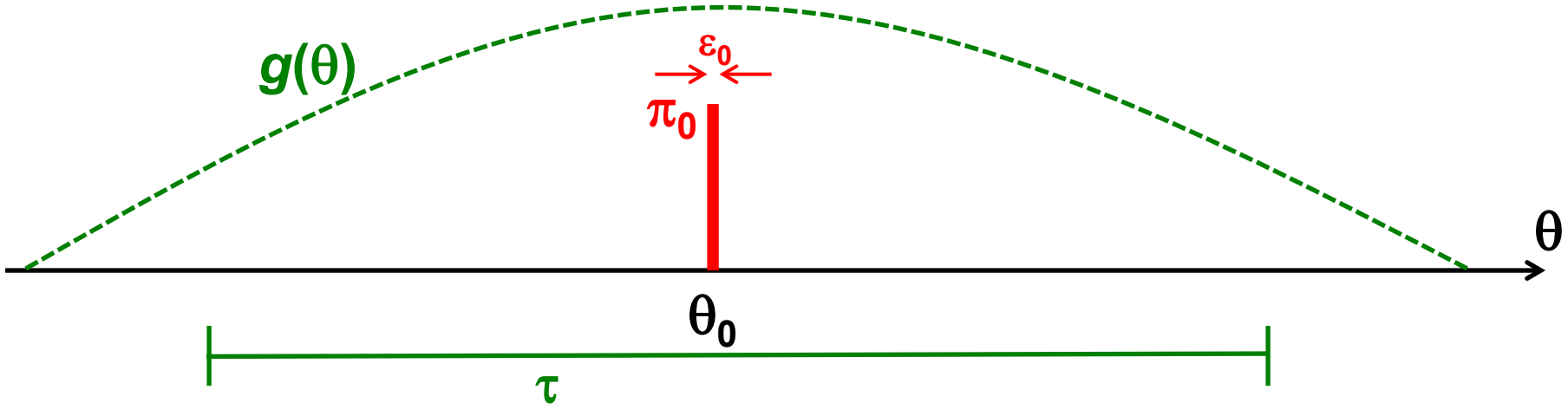
π_0 is like bit of Dirac δ-ftn ("probability mass") at $\theta=\theta_0$.

In practice can have a little width:

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

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

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



Suppose X with normal density $f(x|\theta)$ with unknown mean θ, known std dev σ, is sampled, obtaining $\{x_1, x_2, \ldots, x_n\}$.

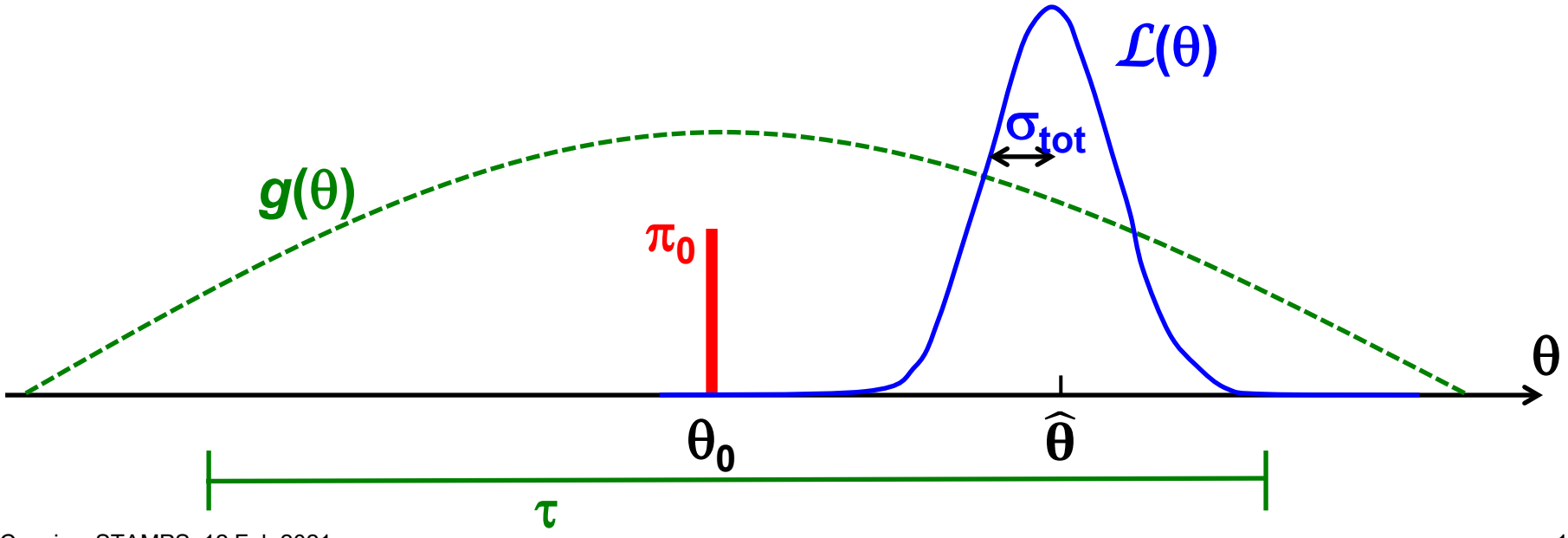
Likelihood function $\mathcal{L}(\theta)$ has maximum at $\hat{\theta} = \bar{x}$.

Std dev of $\hat{\theta}$ is $\sigma_{tot} = \sigma / \sqrt{n}$

Departure from null in "sigma":

$z = (\hat{\theta} - \theta_0) / \sigma_{tot}$.

$z \approx 5$ shown.



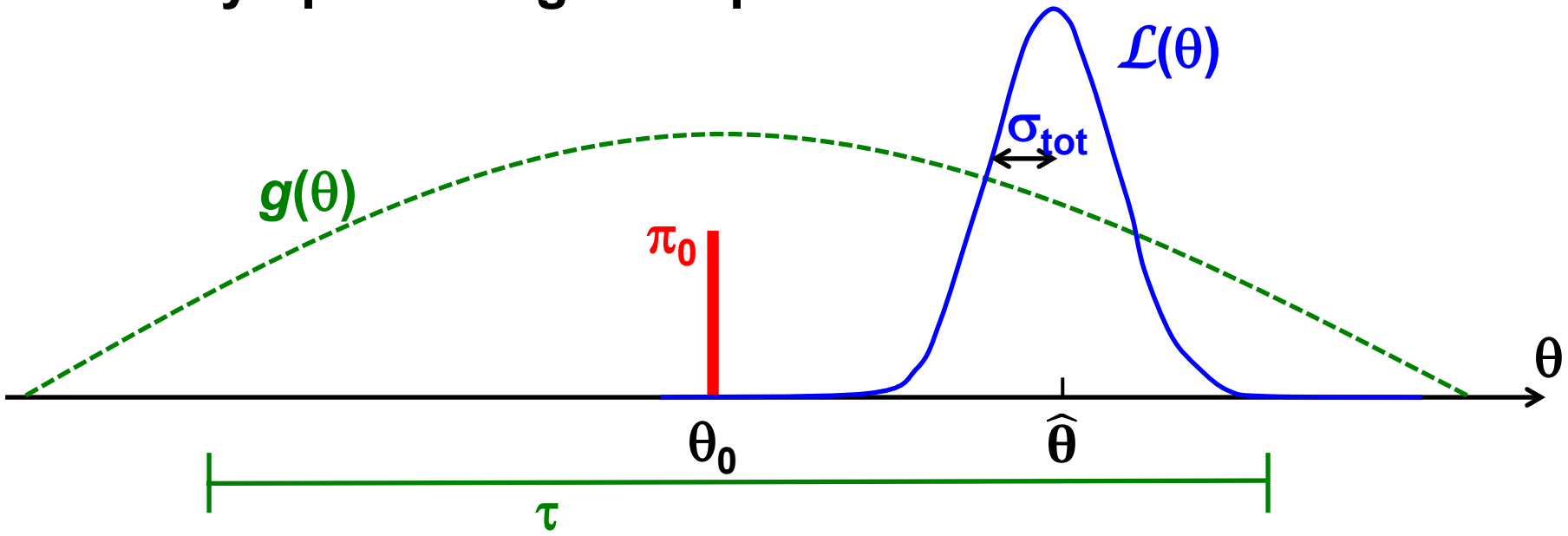
Three *independent* scales in θ: interesting when $\varepsilon_0 \ll \sigma_{tot} \ll \tau$

See backup for calculation of posterior probs of H_0, H_1, and Bayes Factor B_{01}. Key point comes from integrating out θ:

$$P(H_1|\hat{\theta}) \propto \int g(\theta)\, \mathcal{L}(\theta)\, d\theta \propto \sigma_{tot}/\tau \ .$$

"Ockham factor" σ_{tot}/τ penalizes H_1's lack of specificity re θ.

B_{01} (ratio of posterior odds to prior odds) does not depend on prior probs π_0 and π_1 , but *depends directly on scale τ of $g(\theta)$*, even in asymptotic large-sample limit.

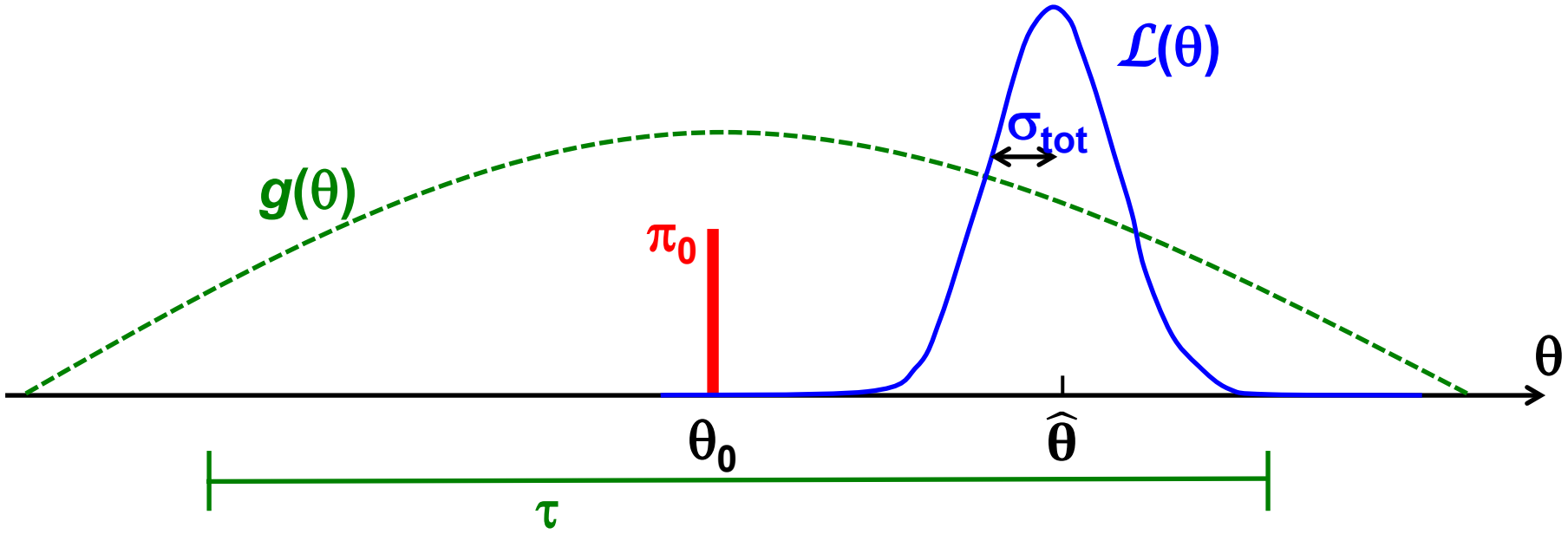


Jeffreys-Lindley paradox:

Recall $P(H_1|\hat{\theta}) \propto \sigma_{tot}/\tau$.

For arbitrarily large *z*, *p*-value for testing H_0 will be arbitrarily small (evidence against H_0)…

… and yet σ_{tot}/τ can be such that $P(H_1|\hat{\theta})$ is also arbitrarily small, with $P(H_0|\hat{\theta}) \to 1$ (strong belief in H_0).



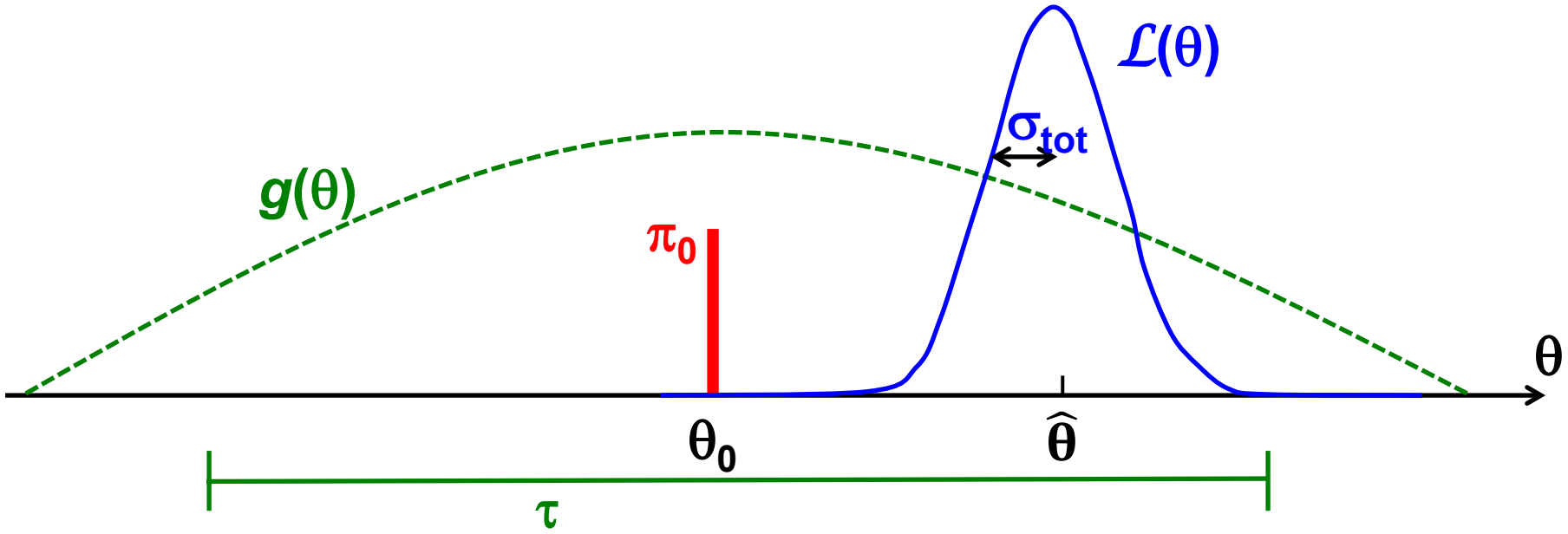
…and Bayes Factor B01 can flip likelihood ratio:

Likelihood ratio $\lambda_{01} = \mathcal{L}(\theta_0) / \mathcal{L}(\hat{\theta}) = \exp(-z^2/2) \leq 1$, *favors* H_1.

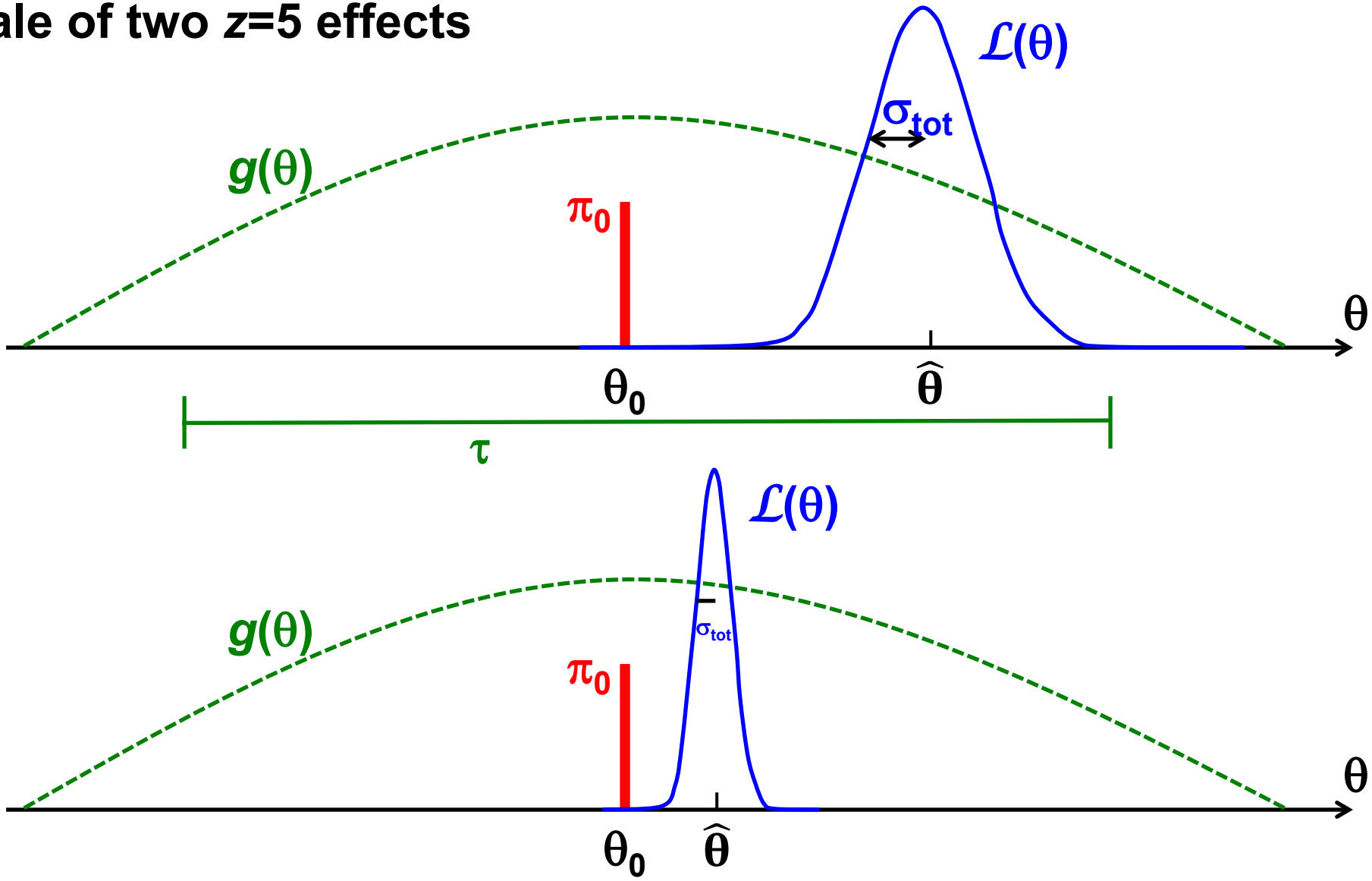
$B_{01} \propto (\tau / \sigma_{tot}) \exp(-z^2/2) = (\tau / \sigma_{tot}) \lambda_{01}$, can be >>1.

Recalling $\sigma_{tot} = \sigma / \sqrt{n}$, JL paradox is often viewed in in terms of sample size *n*: For fixed *z* and τ, $B_{01} \propto \sqrt{n}$. (Lindley 1957)

It can also be viewed as issue that B_{01} depends on the often-arbitrary scale τ. (Bartlett 1957)



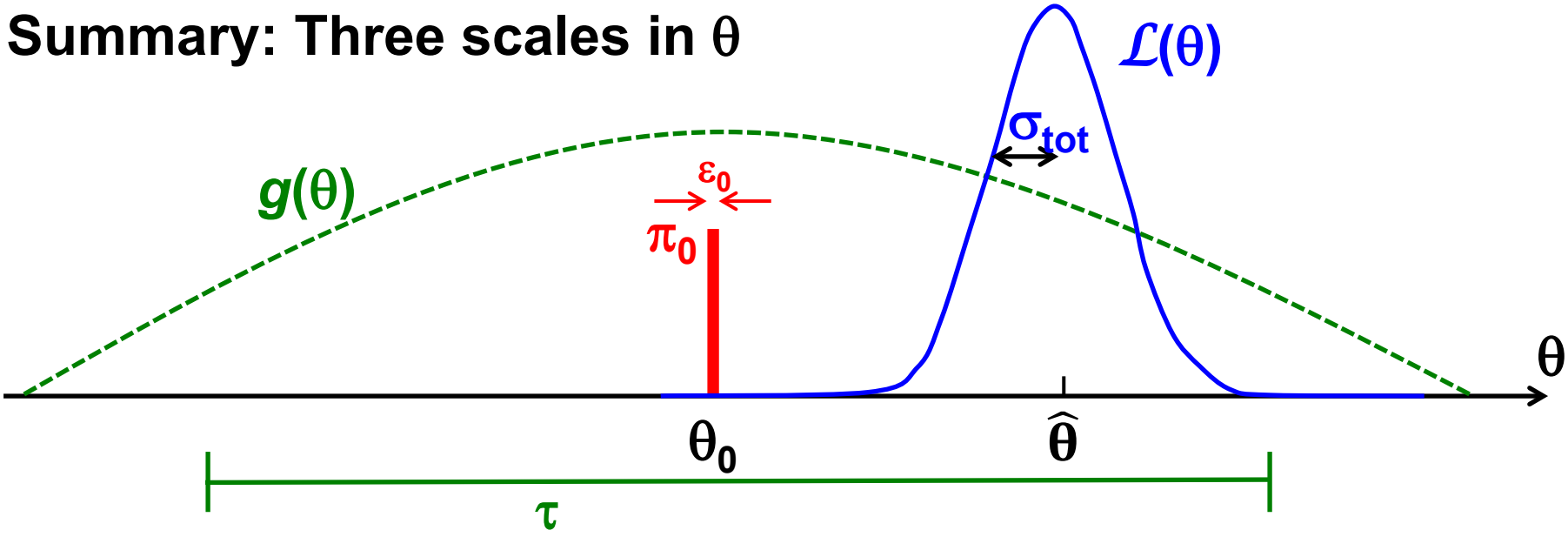
Tale of two z=5 effects



Larger τ/σ_{tot} $\Rightarrow$ Larger $B_{01} \propto (\tau / \sigma_{tot}) \exp(-z^2/2)$

Effect size in original units $(\hat{\theta} - \theta_0)$ smaller

Summary: Three scales in θ



1) ε_0 , width of "δ-ftn" with prob π_0 , scale set by H_0
2) σ_{tot} , width of $\mathcal{L}(\theta)$, total measurement uncertainty
3) τ, width of $g(\theta)$, scale set by H_1

Sufficient for JL paradox to arise if: $\varepsilon_0 << \sigma_{tot} << \tau$

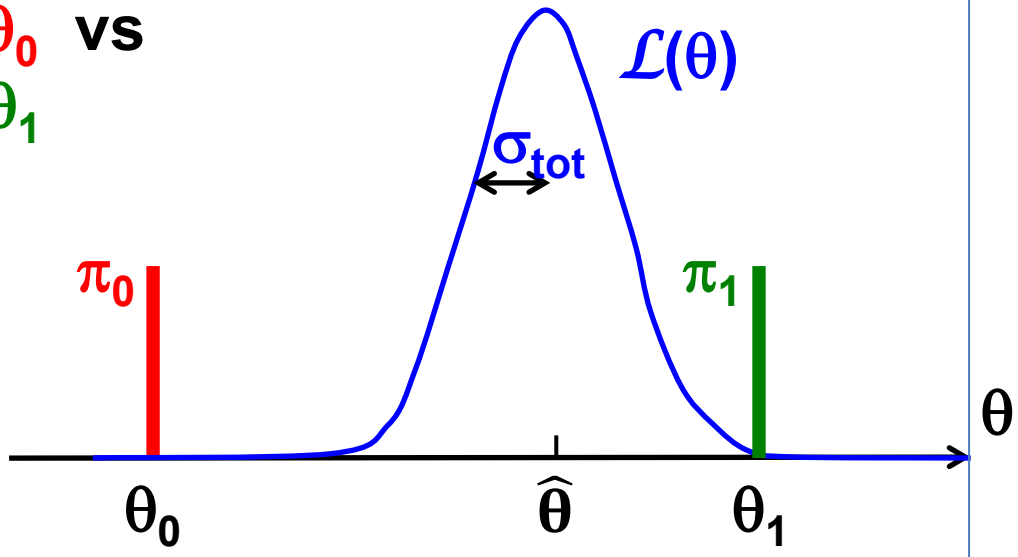
The three scales are largely independent in HEP.
In particular, τ **cannot be reliably inferred from** σ_{tot} .

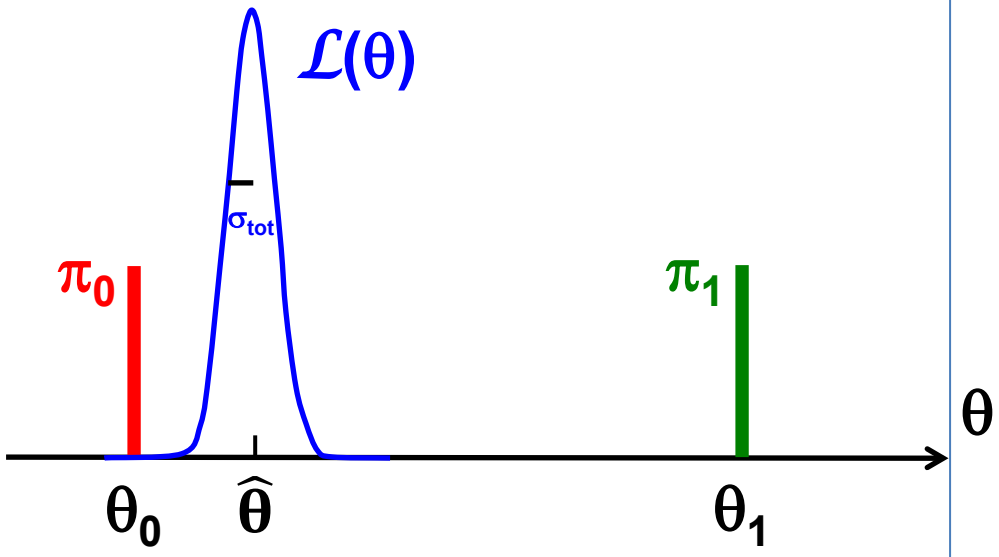
N.B. Testing simple $H_0 : \theta = \theta_0$ vs *simple* $H_1 : \theta = \theta_1$ *is completely* different from testing vs composite H_1.

With fixed z=5 for H_0 , as σ_{tot} decreases due to increasing sample size, $\hat{\theta}$ moves away from θ_1 and $\lambda = \mathcal{L}(\theta_0)/\mathcal{L}(\theta_1)$ favors H_0 !

The *p*-value for H_1 becomes smaller than *p*-value for H_0.

Arguments based on this do not apply to JL H_1, where $\lambda = \mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta})$ favors H_1 !





Every aspect of JL has been scrutinized!

Quite striking how opinions differ on where to put the "blame",e.g.

a) **No good scientist believes their point null hypothesis, since "All models are wrong!", so why are we testing a point null?**
b) **Is there an "objective" or "default" way to set the scale τ ?**
c) **Should inference depend on sample size, for fixed *p*-value? Should we patch BF or *p*-value with $\sqrt{n}$?**

[See backup slide for a couple more]

First, let's get some examples from HEP to keep in mind.

N.B. Our prototype statistical model is Poisson($N \mid \theta+b$), probability of observing *N* events passing selection criteria (out of very large number of chances, >10^{16} at LHC), for unknown signal strength θ, known mean "background" *b*.

Search for proton decay

Super-Kamiokande in Japan watched over 20 metric tonnes of water for 10-20 years, looking for a few out of the 10^{33}-10^{34} protons to decay. https://arxiv.org/abs/2010.16098

Exponential decay with ≈same mean lifetime for each proton.
Let θ be the decay rate per year into $\pi^0 e^+$, $\propto$1/(lifetime in years).

H_0: $\theta = 0$, the approximate value in Standard Model

H_1: $\theta > 0$

Super-K could reject H_0 if $\theta > 10^{-33}$ or so; that is scale of σ_{tot} .

Amazingly, SM prediction is not 0, but less than 10^{-187}.
That is scale $\varepsilon_0 << \sigma_{tot}$.

What is scale τ for the prior $g(\theta)$? Over the years, guesses from Grand Unified Theories have varied by > factor of 10000.
Certainly $\sigma_{tot} << \tau$ for the initial guesses.

Search for $K_L^0 \to \mu e$

The long-lived neutral kaon, K_L^0 , has a mean lifetime of 5×10^{-8} s, with 4 main decay modes accounting for >99.6% of the decays.

Since the 1950's, it has been of great interest to look for rare decay modes, especially those "forbidden".

In 1964, a shocking surprise was the discovery of the decay to two charged pions in about 2×10^{-3} of K_L^0 decays, which was previously (strongly) believed to be forbidden.

This Nobel-prize-winning discovery, had many implications, including the question of why the universe has (very) unequal amounts of matter and anti-matter.

In 1980's, I co-led an experiment looking for another "forbidden" K_L^0 decay, to a muon and electron (of opposite sign charge).

Let θ be the "branching fraction" B of K_L^0 that decay to μe .

- **H_0: $\theta = 0$, the presumed value of B in the Standard Model.**
- **H_1: $\theta > 0$, for which there was various speculation.**

We designed an experiment with the equivalent of $\sigma_{tot} \approx 10^{-11}$.

While $\theta = 0$ was presumed in the SM, already there was indirect evidence that neutrinos might have mass, which would allow the decay at perhaps the 10^{-17} level. That set the scale $\varepsilon_0 << \sigma_{tot}$.

What was the scale τ for the prior $g(\theta)$? When we started, it was sort of known that B < 10^{-8}. There were wildly varying guesses from speculation. Certainly $\sigma_{tot} << \tau$ for some guesses.

Discussion of proton decay and $K_L^0 \to \mu e$ examples

These experiments make use of the general principle that a nonstandard unknown particle X with very large mass m_X can cause a known particle to decay with a decay rate

$$\theta \propto \frac{1}{m_X^4} .$$

For the kaon decay experiment, the m_X probed are of order 100,000 GeV (out of reach of LHC, about 10,000 GeV).

For proton decay, the (different) m_X probed are order 10^{16} GeV!

I don't see a useful way to calculate a meaningful Bayes Factor for these examples. The scale τ and the form of the prior $g(\theta)$ are speculative at best, essentially arbitrary.

***However*, even if τ is arbitrary or unknown, the Bayesian logic that BF $\propto 1/\sigma_{tot}$ for fixed *z* remains. Thus, people claim that coherence requires sample-size dependence of inference.**

Proton decay with a mean decay rate of 10^{-32} per year would be an "effect size" of spectacular importance to physics, *even though mean lifetime of 10^{32} years is 10^{22} times age of universe!*

There is no nonzero decay rate for proton decay that is too small to be of scientific interest if observed!

The same is true for all future feasible searches for $K_L^0 \to \mu e$.

Observation of B($K_L^0 \to \mu e$) > 0 would win a Nobel Prize (once confirmed, of course), as it would point to a new force of nature mediated by a new particle X.

3rd HEP example:

At SLAC, (Prescott et al, 1978), spin-polarized electrons were scattered off nuclei to measure a *tiny* left-right asymmetry A.

Binomial model with parameter ρ. Physicists use $A = 1-2\rho$.

Experiment measured $A = -1.5 \times 10^{-4}$ with statistical and systematic uncertainties each about 10%.

Yes, sample size of scattered electrons was of order 10^{10} (!).

The result helped lead to Nobel Prize in Physics the next year for Glashow-Weinberg-Salam model (now "Standard").

Not clear that $\varepsilon_0 << \sigma_{tot} << \tau$ can be usefully defined, but it is interesting to compare with discussion in statistics literature about binomial parameters very close to 0.5 in ESP, etc.

Recall: a) No good scientist believes their point null hypothesis, since "All models are wrong!", so why are we testing it?

See the full passages (references with page numbers in my JL paper) for important context, but even out of context, these quotes give you a flavor. I am sorry that I have not updated the collection much since 2014 – I am happy to collect new quotes.

Let's start with a Subjectivist:
Lindley (2009) lauds the "triumph" of Jeffreys's "general method of significance tests, putting a concentration of prior probability on the null—no ignorance here—and evaluating the posterior probability using what we now call Bayes factors."

And an Objectivist who prefers his decision-based "reference" analysis to Bayes Factors:

Bernardo (2009): "Jeffreys…is forced to use a mixed prior which puts a lump of probability p = Pr(H_0) on the null, say $H_0 \equiv \theta = \theta_0$ and distributes the rest with a proper prior p(θ) …This has a very upsetting consequence, … for any fixed prior probability p independent of the sample size n, the procedure will wrongly accept H_0 whenever the likelihood is concentrated around a true parameter value which lies $O(n^{-1/2})$ from H_0. I find it difficult to accept a procedure which is known to produce the wrong answer under specific, but not controllable, circumstances."

Gelman and Rubin (1995) : “More generally, realistic prior distributions in social science do not have a mass of probability at zero. . . .”

Raftery (1995b): “social scientists are prepared to act as if they had prior distributions with point masses at zero. . . social scientists often entertain the possibility that an effect is small ”.

C. Robert and J. Rousseau (2009): “Down with point masses! The requirement that one uses a point mass as a prior when testing for point null hypotheses is always an embarrassment and often a cause of misunderstanding in our classrooms. [Use decision theory.]

W. Edwards, H. Lindman, J. Savage (1963): ". . . in typical applications, one of the hypotheses—the null hypothesis—is known by all concerned to be false from the outset," citing others including Berkson (1938).

Vardeman (1987): "Competent scientists do not believe their own models or theories, but rather treat them as convenient fictions. A small (or even 0) prior probability that the current theory is true is not just a device to make posterior probabilities as small as p values, it is the way good scientists think!"

Kadane (1987): "For the last 15 years or so I have been looking seriously for special cases in which I might have some serious belief in a null hypothesis. I have found only one [testing astrologer]. . . I do not expect to test a precise hypothesis as a serious statistical calculation."

Some of those statisticians have evidently not been socializing with many HEP physicists

In the literature I consulted (<2015), I encountered very few statisticians who noted, as did Zellner (2009), that physical laws such as $E = mc^2$ are point hypotheses, and
"Many other examples of sharp or precise hypotheses can be given and it is incorrect to exclude such hypotheses a priori or term them 'unrealistic'. . . ."

Already in Edwards et al (1963) there was a key point related to the situation in HEP: "Bayesians. . . must remember that the null hypothesis is a hazily defined small region rather than a point."

Indeed, ε_0 need not be 0. It just needs to be small on the scale of σ_{tot} for the math of the JL paradox to apply. There is various literature on this point.

At the heart of measurement models in HEP are well-established equations that are commonly called "laws of nature".

The current laws of elementary particle physics, which have survived several decades of intense scrutiny with only a few modifications, are collectively called the "Standard *Model*", but they have the status of "laws".

The SM has parameters such as masses of quarks and leptons, as well as "couplings" that determine reaction rates, which with few exceptions have been measured reasonably precisely.

I refer to the equations of such "laws", or alternatives considered as potential replacements for them, as "core physics models".

Alternatives to the SM are typically more generalized models in which the SM is nested.

Multiple complications arise in going from the core physics model to the full measurement model that describes the probability densities for observations such as the momentum spectra of particles emerging from proton-proton collisions.

"All models are wrong" (Box, 1976) can certainly apply to the detector simulation and calibrations, where parametric approximations such as normal or log-normal are often used.

But the pure core physics models still exist as testable hypotheses that may be regarded as point null hypotheses.

It is certainly worth trying to understand if some physical parameter θ in the larger, alternative core physics model is equal to the SM value, even if it is necessary to do so through the smoke of imperfect detector descriptions with many uninteresting and imperfectly known nuisance parameters.

Indeed, much of what distinguishes the capabilities of experimenters is how well they can do precisely that by determining the detector response through careful calibration and cross-checks, and clever expt design.

This is not just in HEP. Even (especially) with very strong prior belief in H_0 , physicists want to test H_0 as stringently as possible, e.g.

- **electric dipole moments of neutron, electron**
- **equality of gravitational and inertial masses**
- **Lorentz invariance, general relativity**

In HEP, tests of core physics models benefit from what we believe to be the world's most perfect random-sampling mechanism, namely quantum mechanics.

In each of many repetitions of a given initial state, nature *randomly* picks out a final state according to the weights given by the laws of physics and quantum mechanics.

Furthermore, a perfect incarnation of "identical" is present in the fundamental QM property that elementary particles of the same type are *indistinguishable*: the wave function squared is invariant if any two are interchanged.

There is a deeper point to be made about core physics models concerning the difference between a model being a good "approximation" in the ordinary sense of the word, and the concept of a mathematical limit.

The equations of Newtonian physics have been superseded by those of special and general relativity, but the Newtonian equations are not just approximations that did a respectable job in predicting (most) planetary orbits.

They are correct *mathematical limits* as the speed and gravitational field go to zero.

E.g., in SR, if you specify any maximum tolerance for error in kinetic energy KE due to the approximation of Newtonian mechanics, then there exists a speed v below which it will always be within that tolerance.

Equivalently, the Newtonian KE, $\frac{1}{2}$ mv^2 , is the *correct* first term in the expansion of SR KE in powers of v/c.

An analogous, deeper concept arises in the context of effective field theories (EFT).

An EFT in a sense consists of the correct first term(s) in a power series of inverse powers of some energy scale Λ that is much higher than the applicable scale of the EFT (Georgi, 1993).

It is widely believed (and hoped!) that SM will be superseded by a larger theory in which it is nested.

In that sense, we believe (hope!) that the SM is incomplete, which you can call "wrong".

It is also widely believed that the SM will be the *correct* EFT of the new theory at energies << Λ.

In that sense, we believe that the SM is *not wrong*!

My conclusion: The statement,

"All models are *wrong*, but some are *useful*",

is *not "wrong"* in HEP, but it is also *not* "*useful*"!

It is certainly wrong to use this notion to downplay the importance of testing point null hypotheses in HEP.

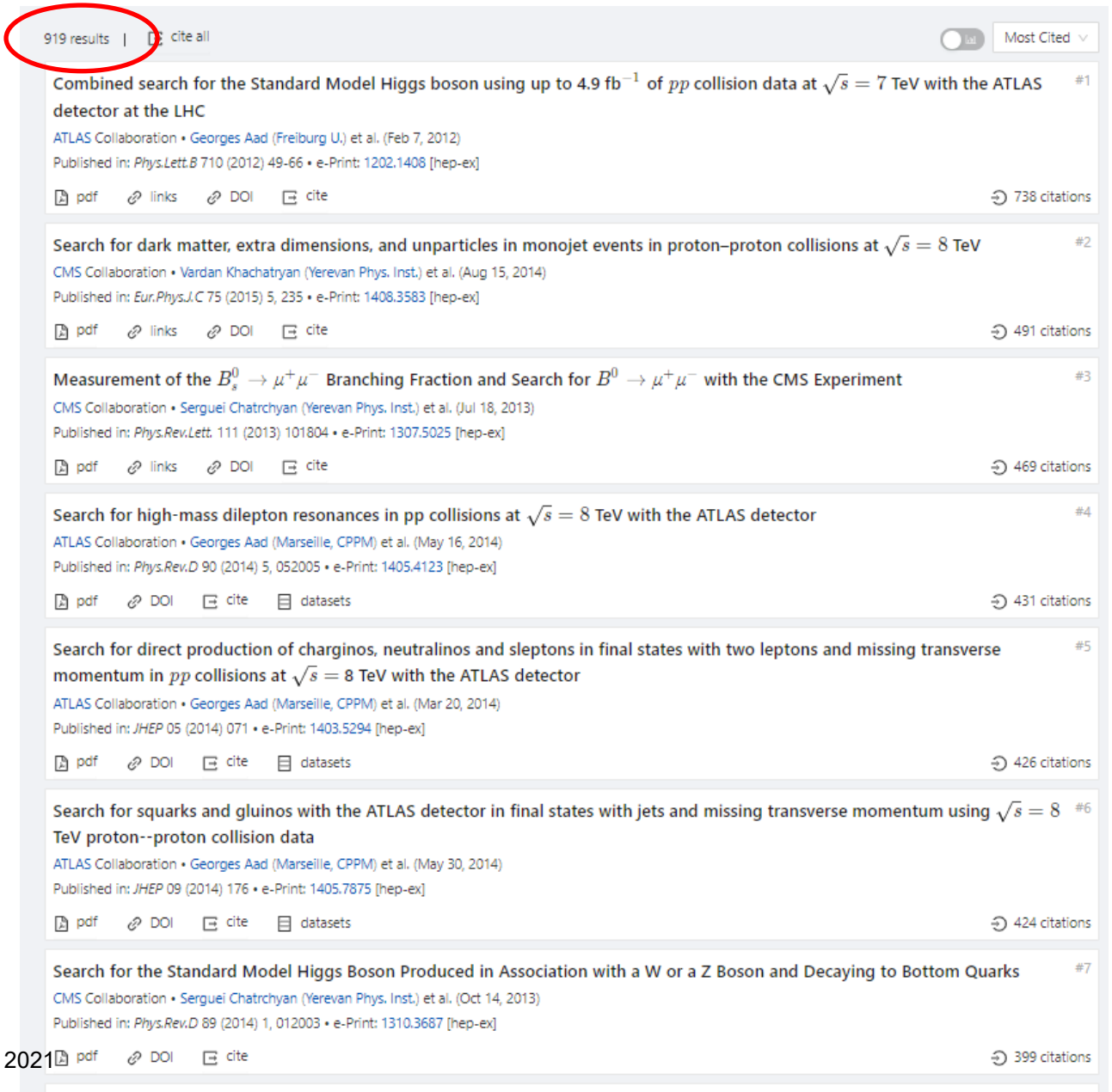
Casella and R. Berger (1987b): “Most researchers would not put 50% prior probability on H_0. The purpose of an experiment is often to disprove H_0 and researchers are not performing experiments that they believe, a priori, will fail half the time!”

Actually, in HEP, we do. Hundreds if not thousands of papers have been published on testing the Standard Model, without rejecting it. (We do not use p<0.05 !)

Inspirehep.net: papers by ATLAS or CMS with “Search for” in the title (and not “observation)

919 refereed journal articles (>100 more from LHCb), ordered by times cited.

For some of these, H_0 was Standard Model with a missing piece, indeed not believed, but not rejected in these “Search for” papers.
(See my JL paper.)



919 results | cite all

Most Cited

#1 Combined search for the Standard Model Higgs boson using up to 4.9 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector at the LHC
ATLAS Collaboration • Georges Aad (Freiburg U.) et al. (Feb 7, 2012)
Published in: *Phys.Lett.B* 710 (2012) 49-66 • e-Print: 1202.1408 [hep-ex]
pdf links DOI cite 738 citations

#2 Search for dark matter, extra dimensions, and unparticles in monojet events in proton–proton collisions at $\sqrt{s} = 8$ TeV
CMS Collaboration • Vardan Khachatryan (Yerevan Phys. Inst.) et al. (Aug 15, 2014)
Published in: *Eur.Phys.J.C* 75 (2015) 5, 235 • e-Print: 1408.3583 [hep-ex]
pdf links DOI cite 491 citations

#3 Measurement of the $B_s^0 \to \mu^+\mu^-$ Branching Fraction and Search for $B^0 \to \mu^+\mu^-$ with the CMS Experiment
CMS Collaboration • Serguei Chatrchyan (Yerevan Phys. Inst.) et al. (Jul 18, 2013)
Published in: *Phys.Rev.Lett.* 111 (2013) 101804 • e-Print: 1307.5025 [hep-ex]
pdf links DOI cite 469 citations

#4 Search for high-mass dilepton resonances in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector
ATLAS Collaboration • Georges Aad (Marseille, CPPM) et al. (May 16, 2014)
Published in: *Phys.Rev.D* 90 (2014) 5, 052005 • e-Print: 1405.4123 [hep-ex]
pdf DOI cite datasets 431 citations

#5 Search for direct production of charginos, neutralinos and sleptons in final states with two leptons and missing transverse momentum in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector
ATLAS Collaboration • Georges Aad (Marseille, CPPM) et al. (Mar 20, 2014)
Published in: *JHEP* 05 (2014) 071 • e-Print: 1403.5294 [hep-ex]
pdf DOI cite datasets 426 citations

#6 Search for squarks and gluinos with the ATLAS detector in final states with jets and missing transverse momentum using $\sqrt{s} = 8$ TeV proton--proton collision data
ATLAS Collaboration • Georges Aad (Marseille, CPPM) et al. (May 30, 2014)
Published in: *JHEP* 09 (2014) 176 • e-Print: 1405.7875 [hep-ex]
pdf DOI cite datasets 424 citations

#7 Search for the Standard Model Higgs Boson Produced in Association with a W or a Z Boson and Decaying to Bottom Quarks
CMS Collaboration • Serguei Chatrchyan (Yerevan Phys. Inst.) et al. (Oct 14, 2013)
Published in: *Phys.Rev.D* 89 (2014) 1, 012003 • e-Print: 1310.3687 [hep-ex]
pdf DOI cite 399 citations

Recall:

b) Is there an "objective" or "default" way to set the scale τ ?

c) Should inference depend on sample size, for fixed *p*-value? Should we patch BF or *p*-value with $\sqrt{n}$?

Fundamentally, the scale appears to be inescapably personal and subjective, as is the more detailed specification of $g(\theta)$.

Berger and Delampady (1987a,b): "the precise null testing situation is a prime example in which objective procedures do not exist,"…
"Testing a precise hypothesis is a situation in which there is clearly no objective Bayesian analysis and, by implication, no sensible objective analysis whatsoever."

Nonetheless, Berger and others have attempted to formulate principles for specifying default values of τ for communicating scientific results. [See backup.]

Bartlett (1957) suggests that τ might scale as $1/\sqrt{n}$, canceling the sample-size scaling in σ_{tot} and making the Bayes factor independent of n.

Cox (2006) suggests this as well, on the grounds that ". . . in most, if not all, specific applications in which a test of such a hypothesis [$\theta = \theta_0$] is thought worth doing, the only serious possibilities needing consideration are that either the null hypothesis is (very nearly) true or that some alternative within a range fairly close to θ_0 is true."

Part of Cox's argument was already given by Jeffreys (1961): ". . . the mere fact that it has been suggested that [θ] is zero corresponds to some presumption that [θ] is small."

Economist Leamer (1978) says similarly: ". . . a prior that allocates positive probability to subspaces of the parameter space but is otherwise diffuse represents a peculiar and unlikely blend of knowledge and ignorance".

N.B. This "peculiar and unlikely blend" is common in HEP.

In the search for any non-subjective *n*-independent scale τ, one option seemingly at hand is $\sigma_{tot} = \sigma/\sqrt{n}$ when n=1, i.e., the original σ that expresses std dev of single measurement.

Suggested by Jeffreys (1961), since there is nothing else in the problem that can set the scale, and was followed, e.g., in generalizations by Zellner and Siow (1980).

There are various versions of this argument (Smith and Spiegelhalter (1980); Kass and Wasserman (1995)), but IMO they do not address the common case in HEP where σ_{tot} and τ are *independent* scales!

Furthermore, the detector may have some intrinsic σ_{tot} for which *no definition of* σ or *n is evident*.

Raftery (1995a, pp. 132, 135) points out the same problem: **the "important ambiguity. . . the definition of [n], the sample size."**

He gives several examples for which he has a recommendation.

Meanwhile, for some important searches of departures from S.M. , σ_{tot} has decreased by nearly a factor 10^4 since 1980's.

(In searches with ~no background, σ_{tot} scales more like 1/n.)

I do not see a sensible way to have an Ockham factor σ_{tot}/τ for these searches.

(Looking ahead, nor do I see a convincing rationale for adjusting the *p*-value or α threshold when σ_{tot} decreases by nearly a factor 10^4 .)

See backup for paper by Jim Berger and collaborators, including “intrinsic priors”, and summaries of other approaches by others.

He gave a talk on “Bayesian Hypothesis Testing”” at PhyStat-neutrino workshop at CERN in January 2019, https://indico.cern.ch/event/735431/timetable/ , with worked examples, comparison of p-values and Bayes factors, and much food for thought.

Lindley (Bernardo (2011b) argues in cases like this that objective Bayesians can get lost in the Greek letters and lose contact with the actual context.

I too find it puzzling that one can first argue that the Ockham factor is a crucial feature of Bayesian logic that is absent from frequentist reasoning, and then resort to "objective" prior g(θ) that in effect makes the Ockham factor arbitrary.

For me, the arbitrariness of the scale τ is a big obstacle to more use of Bayes Factors in our testing problem of point null vs continuous alternative. (One could report result as function of τ.)

Kass (2009): Bayes factors for hypothesis testing "remain sensitive—to first order—to the choice of the prior on the parameter being tested."
The results are "contaminated by a constant that does not go away asymptotically." …this approach is "essentially nonexistent" in neuroscience.

Gelman approach: way too much to summarize on one slide, see article with Shalizi on next slide.

Long-time advocacy of model checking using posterior predictive p-values. (BDA textbook; Gelman and Rubin 1995,…)

Gelman, Stat. Sci. 24 (2009) 176: [In contrast to examples from physics and Jim Berger] "My own perspective as a social scientist is completely different [from those for whom Ockham's razor seems unquestionably reasonable]. I've just about never heard someone in social science object to the inclusion of a variable or an interaction in a model… In the social science problems I've seen, Ockham's razor is at best an irrelevance and at worse can lead to acceptance of models that are missing key features…."

Gelman and Shalizi (Brit. J. Math. Stat. Psych. 66 (2013) 8

"…crucial aspects of model checking and model revision, which fall outside the scope of Bayesian confirmation theory."

Different approach and justification from the "received view of Bayesian inference". "Our perspective is not new; in methods and also in philosophy we follow statisticians such as Box (1980, …"

See paper and Comments from Bayesian and frequentist perspective

In HEP, model checking is of course also crucial. For characterization of detector response, adding parameters is common if indicated by traditional methods (F-test, BIC, etc.).

But for the core physics models, adding a parameter can be a Nobel-prize-winning discovery (or an Ig Nobel embarrassment). This is the key issue for me today.

Should the interpretation of a *p*-value depend on sample size???

Let's step back from the JL paradox, and ask, what have statisticians of any flavor have said about sample size?

J. Statist. Comput. Simul., 1980, Vol. 11, pp. 307–313
0094-9655/80/1104-0307 $06.50/0

Printed in Great Britain

COMMENTS, CONJECTURES, AND CONCLUSIONS

Section Editor: *I. J. Good*

Please be succinct but lucid and interesting.

C73. THE DIMINISHING SIGNIFICANCE OF A P-VALUE AS THE SAMPLE SIZE INCREASES

Dr. Deborah Mayo raised the following question. How could one convince a very naive student, Simplissimus, that a given tail-area probability (P-value), say 1/100, is weaker evidence against the null hypothesis when the sample is larger? Although this fact is familiar in Bayesian statistics the question is how to argue it without (explicit) reference to Bayesian methods.

Bob comment: he shows it for simple vs simple, and then claims without foundation that it applies to simple vs composite.

Good (1992): “The real objection to [p-values] is not that they usually are utter nonsense, but rather that they can be highly misleading, especially if the value of [n] is not also taken into account and is large.”

He suggested a rule of thumb for taking n into account by standardizing the p-value to an effective size of n = 100, but this seems not to have attracted a following.

Richard M. Royall, "The Effect of Sample Size on the Meaning of Significance Tests," Am. Stat. 40 313 (1986)

Royall's key point: We need to distinguish between sample-size dependence of
(a) the evidence of statement "significant at p-value <0.05 or better", and
(b) the evidence from reporting the p-value itself.

In any case, Royall's discussion is about simple vs simple, in which the alterative hypothesis does not have a perfect fit to the data.

(Aside) *Statistical Evidence (*1997), p. 19 ff: Royall says law of likelihood cannot deal directly with composite hypotheses; needs to break it up into multiple simple alternatives. "It forces us to be more specific, to note and report which values greater than 0.2 are better supported than values of 0.2 or less, for example and by how much".

In addition to the likelihood/Bayesian arguments, Royall cites D. Bakan:

"The rejection of the null hypothesis when the number of cases is small speaks for a more dramatic effect…and if the p-value is the same, the probability of committing a Type I error remains the same. Thus one can be more confident with a small N than a large N." [???]

The first phrase is true, but the conclusion is a non-sequitur.

Kent R. Bailey: "The discussion of Bakan's point seems to confuse clinical significance with probability..."

Bob Wheeler: "The point that is being missed here, however, is that with small sample sizes the estimates are varying over wider ranges, and thus large observed effects do not necessarily imply large true effects…"

Deborah Mayo, *Statistical Inference as Severe Testing* (2018)

SIST, p. 239: "Rosenthal and Gaito (1963) discovered that statistical significance at a given level was often fallaciously taken as evidence of a greater discrepancy from the null hypothesis the larger the sample size n. In fact, it is indicative of less of a discrepancy from the null than if it resulted from a smaller sample size."

I clarified with Mayo that by "discrepancy", she meant the "population parametric difference from the null", i.e., $\widehat{\theta} - \theta_0$, what I called "effect size in original units" on my slide, "Tale of two *z*=5 effects". This is clear.

SIST: "*Mountains out of Molehills* (MM) *Fallacy* (large *n* problem): The fallacy of taking a rejection of H_0, just at level *P*, with larger sample size (higher power) as indicative of a greater discrepancy from H_0 than with a smaller sample size."

Summary of a few key points

The condition "$\varepsilon_0 \ll \sigma_{tot} \ll \tau$" is common in HEP.

***Tiny* effect sizes can be of *huge* scientific significance in HEP.**

Sample size scaling of inference for fixed *p*-value is problematic. Modern searches have σ_{tot} a factor of 10^4 smaller than when my career began. Really scale Ockham factor, or *p*-value, by orders of magnitude ??

Ockham factor depends on problematic τ expressing range of prior belief: how to specify for scientific communication?

Meanwhile, we continue with *p*-values, essentially *always* accompanied by confidence intervals for various effect sizes in original units.

This is what we do for a living!
The status quo of the philosophical foundations is unsettling.

Thanks to all (physicists, statisticians -- see my paper and note), including my "sponsor", the U.S. DOE Office of High Energy Physics in the Office of Science

BACKUP

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 $\overline{\theta}$ (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.

In my paper “effect size” refers to the point and interval estimates (measured values and uncertainties) of a parameter or physical quantity, typically expressed in the original units.

In HEP, point estimates and confidence intervals for model parameters are used to summarize the results of nearly all experiments, and to compare to the predictions of theory (which often have uncertainties as well).

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, sin/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 is >1D, where LR ordering a la F-C and K&S is particularly useful, and other orderings are poorly defined (metric dependent)**

Especially common in neutrino physics, also used at times in other sub-specialties (dark matter, quark mixing, LHC).

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 [2] Section 9.1 (violation of Likelihood Principle, 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. ([2] Section 12)**

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

Posterior Probabilities and Bayes Factor (BF)

Prior odds of H_0 to H_1: $\dfrac{\pi_0}{\pi_1}$

Posterior odds: $\dfrac{P(H_0|\hat{\theta})}{P(H_1|\hat{\theta})}$ **, where**

$$P(H_0|\hat{\theta}) = \frac{1}{A}\,\pi_0\,\mathcal{L}(\theta_0)$$

$$P(H_1|\hat{\theta}) = \frac{1}{A}\,\pi_1 \int g(\theta)\mathcal{L}(\theta)d\theta$$

Then $\mathrm{BF} \equiv \dfrac{P(H_0|\hat{\theta})}{P(H_1|\hat{\theta})} \Big/ \dfrac{\pi_0}{\pi_1}$ **, so priors π_0 and π_1 cancel, but the**

dependence of BF on prior g(θ) remains! (Bartlett, 1957)

$$\mathrm{BF} \propto \frac{\tau}{\sigma_{\mathrm{tot}}} \exp(-z^2/2)$$

Cf. $\lambda = \mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta}) = \exp(-z^2/2)$

Note null belief enhanced by extra factor: ratio τ/σ_{tot}

The Jeffreys-Lindley Paradox

$$\mathrm{BF} \propto \frac{\tau}{\sigma_{\mathrm{tot}}} \exp(-z^2/2)$$

$$\mathrm{BF} \propto \frac{\tau}{\sigma_{\mathrm{tot}}} \lambda$$

BF varies as τ/σ_{tot} while z is fixed. Can even reverse the inference to favor null.

Likelihood ratio $\lambda = \mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta})$ also disfavors the null (!).

The *Ockham (Occam) factor* σ_{tot}/τ penalizes H_1 for degree of freedom to choose "best-fit" θ, whereas H_0 predicts it.

For experiments obtaining the *same z*, the Bayesian answers depend on sample size (since σ_{tot} typically goes as 1/sqrt(sample size). Very different behavior!

Aside: Lindley's 1957 paper overlooked scale τ (!)

Or arbitrarily (wrongly) set it to 1 without comment.

Bartlett (1957) quickly noted that Lindley had neglected the length of the interval over which g(θ) is constant, which should appear in the numerator of the BF, and which makes the posterior probability of H_0 "much more arbitrary".

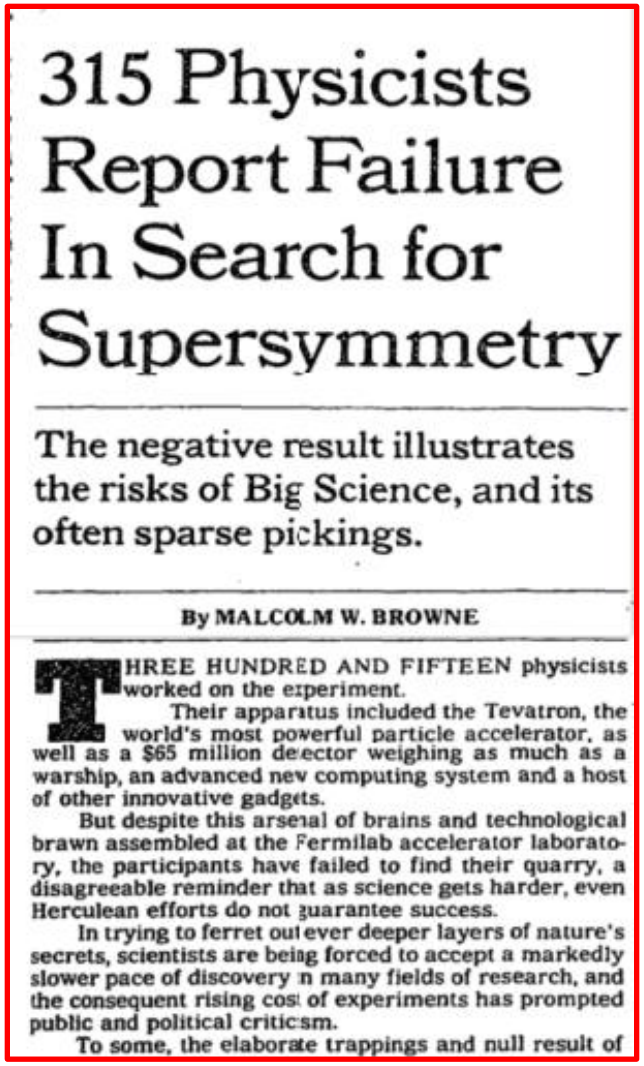
It is kind of amusing to see papers that incorrectly describe what is in Lindley's paper.

Every aspect of JL has been scrutinized!

Quite striking how opinions differ on where to put the "blame",e.g.

a) **No good scientist believes their point null hypothesis, since "All models are wrong!", so why are we testing a point null?**
b) **Is there an "objective" or "default" way to set the scale τ ?**
c) **Should inference depend on sample size, for fixed *p*-value? Should we patch BF or *p*-value with $\sqrt{n}$?**
d) **(From a Bayesian) Jeffreys's method should be replaced by a method based on decision theory that gives different results.**
e) **Frequentist tail probabilities are obviously wrong way to make inference; p-values dramatically overstate evidence against H_0.**

Restricting the allowed parameter space of extensions of the SM is considered a very worthy enterprise in HEP, but there have been some infamous headlines, e.g., NYT Jan. 5, 1993:



315 Physicists Report Failure In Search for Supersymmetry

The negative result illustrates the risks of Big Science, and its often sparse pickings.

By MALCOLM W. BROWNE

THREE HUNDRED AND FIFTEEN physicists worked on the experiment.

Their apparatus included the Tevatron, the world's most powerful particle accelerator, as well as a $65 million detector weighing as much as a warship, an advanced new computing system and a host of other innovative gadgets.

But despite this arsenal of brains and technological brawn assembled at the Fermilab accelerator laboratory, the participants have failed to find their quarry, a disagreeable reminder that as science gets harder, even Herculean efforts do not guarantee success.

In trying to ferret out ever deeper layers of nature's secrets, scientists are being forced to accept a markedly slower pace of discovery in many fields of research, and the consequent rising cost of experiments has prompted public and political criticism.

To some, the elaborate trappings and null result of

Since we do not know the scale τ of "new physics", we keep looking, with smaller and smaller σ_{tot} .

The LHC is currently being upgraded to give us >10 times larger sample sizes for many searches.

The excitement of, or belief in, a new discovery will not be diminished by the small effect size in absolute units, if the p-value is small enough! (and of course if confirmed…)

Summary of searches for proton and neutron decay, https://pdg.lbl.gov/2020/listings/rpp2020-list-p.pdf , p. 12

Lower limits at 90% C.L. on τ , the inverse of decay rate into the once-favored final state of positron and pion.

$\tau(N \to e^+\pi)$ τ_1

LIMIT (10^{30} years)	PARTICLE	CL%	EVTS	BKGD EST	DOCUMENT ID		TECN
>16000	***p***	**90**	**0**	**0.61**	ABE	17	SKAM
> 5300	***n***	**90**	**0**	**0.41**	ABE	17D	SKAM
• • • We do not use the following data for averages, fits, limits, etc. • • •							
> 2000	*n*	90	0	0.27	NISHINO	12	SKAM
> 8200	*p*	90	0	0.3	NISHINO	09	SKAM
> 540	*p*	90	0	0.2	MCGREW	99	IMB3
> 158	*n*	90	3	5	MCGREW	99	IMB3
> 1600	*p*	90	0	0.1	SHIOZAWA	98	SKAM
> 70	*p*	90	0	0.5	BERGER	91	FREJ
> 70	*n*	90	0	$\leq$ 0.1	BERGER	91	FREJ
> 550	*p*	90	0	0.7	[1] BECKER-SZ...	90	IMB3
> 260	*p*	90	0	<0.04	HIRATA	89C	KAMI
> 130	*n*	90	0	<0.2	HIRATA	89C	KAMI
> 310	*p*	90	0	0.6	SEIDEL	88	IMB
> 100	*n*	90	0	1.6	SEIDEL	88	IMB
> 1.3	*n*	90	0		BARTELT	87	SOUD
> 1.3	*p*	90	0		BARTELT	87	SOUD
> 250	*p*	90	0	0.3	HAINES	86	IMB
> 31	*n*	90	8	9	HAINES	86	IMB
> 64	*p*	90	0	<0.4	ARISAKA	85	KAMI
> 26	*n*	90	0	<0.7	ARISAKA	85	KAMI
> 82	*p* (free)	90	0	0.2	BLEWITT	85	IMB
> 250	*p*	90	0	0.2	BLEWITT	85	IMB
> 25	*n*	90	4	4	PARK	85	IMB
> 15	*p*, *n*	90	0		BATTISTONI	84	NUSX
> 0.5	*p*	90	1	0.3	[2] BARTELT	83	SOUD
> 0.5	*n*	90	1	0.3	[2] BARTELT	83	SOUD
> 5.8	*p*	90	2		[3] KRISHNA...	82	KOLR
> 5.8	*n*	90	2		[3] KRISHNA...	82	KOLR
> 0.1	*n*	90			[4] GURR	67	CNTR

Jim Berger and collaborators:

Berger and Pericchi (2001) review more general possibilities based on use of the information in a small subset of the data, and for one method claim that “this is the first general approach to the construction of conventional priors in nested models.”

Berger (2008, 2011) applied one of these so-called “intrinsic priors” to a pedagogical example and its generalization from HEP. I am not aware of anyone in HEP who has pursued these suggestions.

Meanwhile, Bayarri, Berger, Forte, and Garca-Donato (2012) have reconsidered the issue and formulated principles resulting “. . . in a new model selection objective prior with a number of compelling properties.”

Jim gave a talk on “Bayesian Hypothesis Testing”” at PhyStat-neutrino workshop at CERN in January 2019, https://indico.cern.ch/event/735431/timetable/ , with worked examples, comparison of p-values and Bayes factors, and much food for thought.

Textbook by Lee (2004) “Although it seems reasonable that [τ] should be chosen proportional to [σ], there does not seem to be any convincing argument for choosing this to have any particular value. . . .”

Andrews (1994) also explores the consequences of τ shrinking with sample size, but these ideas seem not to have led to a standard.

Robert (1993) considers π_1 that increases with τ, but this seems not to have been pursued further.

Rosenthal and Gaito: “Ss completed a questionnaire requesting them to rate their degree of belief or confidence in a variety of *p* levels: once assuming an *n* of 10, and then assuming an *n* of 100.”

…The finding that sizeable proportions of Ss had greater confidence in levels of significance based on an *n* of 100 suggests that most investigators not only use the probability of Type I errors as a criterion but consider (at least intuitively) Type II errors as well. If Type I errors only were of concern, the same degree of belief should be attached to the p for both *ns*. However, for a given *p* level the Type II error decreases with increasing *n*; thus the power of the test increases in that Power = 1-p (Type II errors), Objectively speaking, one would expect greater belief attached to the larger sample size at every p level because for any p level, all samples of any size have the same probability of a Type I error but the probability of a Type II error is less for the larger ns. Thus, the statistical test using larger ns has a greater tendency to detect real differences if they are present.

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) !

How many sigmas?
I.e., choice of threshold α for rejecting H_0

Neyman and Pearson (1933a)

"These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second.

"...The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator."

How many sigmas? (cont.)

Lehmann (2005 and earlier editions)

"The choice of a level of significance α is usually somewhat arbitrary...the choice should also take in consideration the power that the test will achieve against the alternatives of interest...."

[But if power is a function of the unknown θ, that's a problem for frequentists, who generally cannot put a prior on θ and take a weighted average of the power.]

How many sigmas? (cont.)

Lehmann (cont.)

"Another consideration that may enter into the specification of a significance level is the attitude toward the hypothesis before the experiment is performed.

"If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low."

How many sigmas? (cont.)

Kendall and Stuart and successors (Stuart 1999)

"...unless we have supplemental information in the form of the *costs* (in money or other common terms) of the two types of error, and costs of observations, we cannot obtain an optimal combination of α, β, and *n* for any given problem."

How many sigmas? (cont.)

And, in HEP, the tradition started in Alvarez group, largely motivated by multiple trials factors:

Arthur Rosenfeld (1968) "To the theorist or phenomenologist the moral is simple: wait for nearly 5σ effects. For the experimental group who have spent a year of their time and perhaps a million dollars, the problem is harder...go ahead and publish...but they should realize that any bump less than about 5σ calls only for a repeat of the experiment."

Allan Franklin studied the modern history: