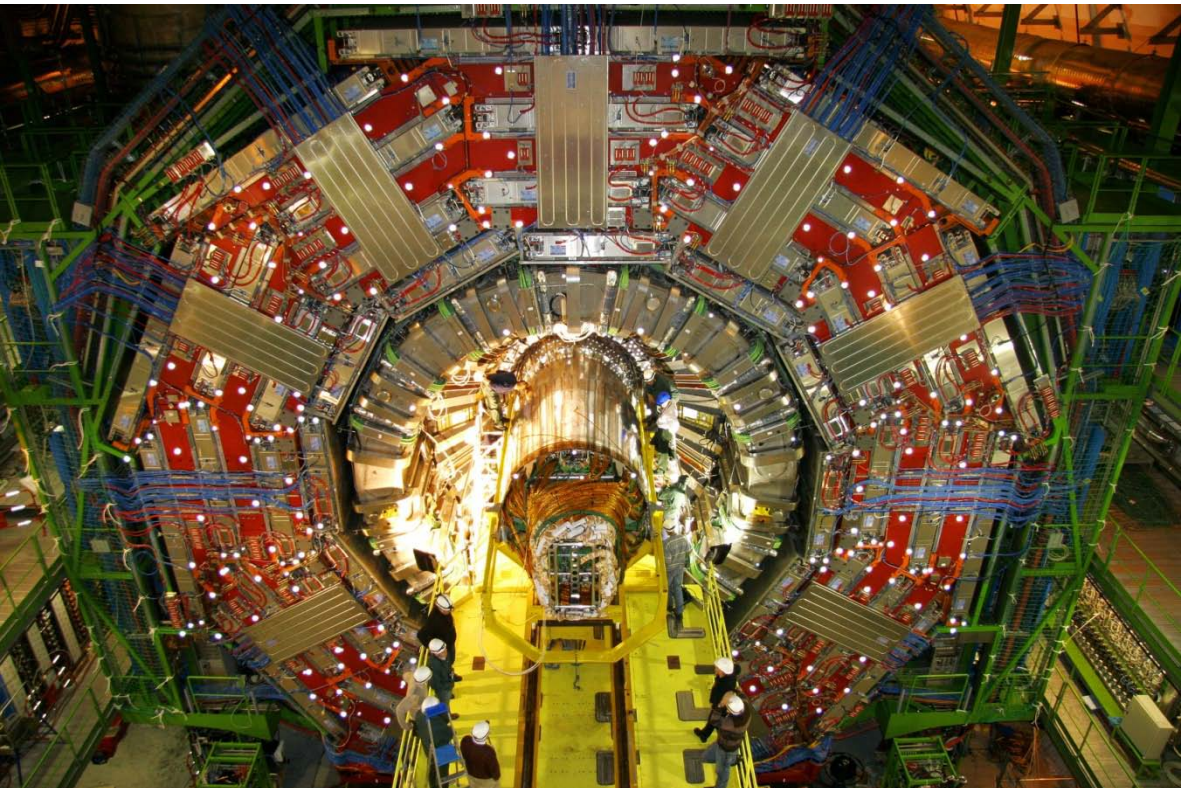
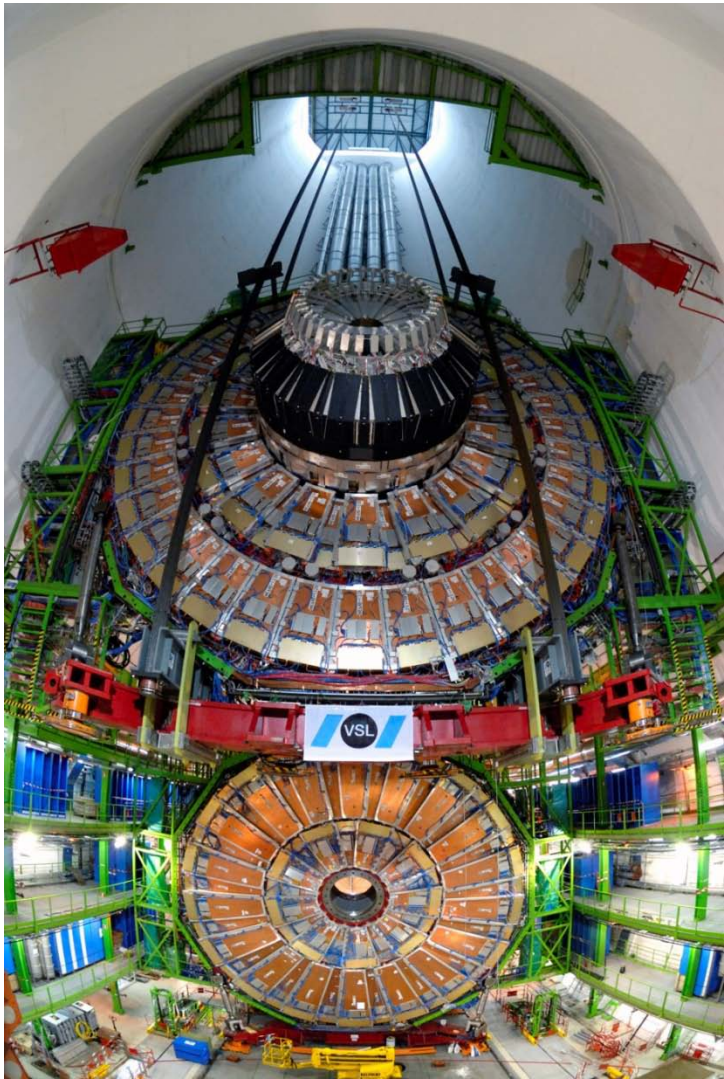
Can High Energy Physicists Learn Something Useful from the Jeffreys-Lindley Paradox?

Bob Cousins, UCLA

Banff

July 13, 2010





IS HINCHLIFFE'S RULE TRUE? *

Boris Peon

Abstract

Hinchliffe has asserted that whenever the title of a paper is a question with a yes/no answer, the answer is always no. This paper demonstrates that Hinchliffe's assertion is false, but only if it is true.

*Accepted for publication in *Annals of Gnosis*.

Statistics of Discovery: Theorists' View

PHYSICAL REVIEW D, VOLUME 62, 015009

Measurements of masses in supergravity models at CERN LHC

Henri Bachacou Ian Hinchliffe Frank E. Paige

I. INTRODUCTION

If supersymmetric (SUSY) particles exist at the TeV mass scale, they will be produced at the CERN Large Hadron Collider (LHC) with large rates, so discovery of their existence will be straightforward.

But Maybe It Won’t Be Easy.
We’ll want Efficient Techniques.

- As mentioned yesterday, somehow *fixed* 5 sigma (alpha=2.87E-7) has become (even more than) the rule of thumb for discovery in HEP.
- Partly it is a very crude attempt to account for the LEE.
- Partly it is to account for unknown systematic errors.
- We have to do better at the LHC. But what?
- Two obvious candidates to consider:
  - N-P test with more intelligent (and variable) choice of alpha.
  - Bayesian Model selection.
- What does the statistics literature say about each, and about comparisons?

- **We all know that p-value is neither probability nor odds of H given data.**
- **Still, one frequently finds comparisons of the two (especially by Bayesians).**
- **It should not be surprising that the numbers are different.**
- **But can they be calibrated with respect to each other with "rules of thumb"?**
- **Jim Berger and others, before and after, have many examples and arguments to say *no*.**
- **The most disturbing thing to me is that for fixed values of alpha (such as the 5-sigma criterion), the scaling of "the answer" with sample size is different!**
- **Physicists like to look at limiting cases. That takes us to the Jeffreys-Lindley paradox: large sample limit.**

The setup:

- **Null model is $\theta = \theta_0$, where for definiteness I take $\theta_0 = 0$.**
- **Alternative is $\theta > 0$.**

Classical Hypothesis Testing (cont.)

"Test for $\theta=\theta_0$" $\leftrightarrow$ "Is θ_0 in confidence interval for θ"

Table 20.1 Relationships between hypothesis testing and interval estimation

Property of test	Property of corresponding confidence interval
Size $= \alpha$	Confidence coefficient $= 1 - \alpha$
Power = probability of rejecting a false value of $\theta = 1 - \beta$	Probability of not covering a false value of $\theta = 1 - \beta$
Most powerful	Uniformly most accurate
$\longleftarrow$	$\left\{ \begin{array}{c} \textit{Unbiased} \\ 1 - \beta \geq \alpha \end{array} \right\}$ $\longrightarrow$
Equal-tails test $\alpha_1 = \alpha_2 = \frac{1}{2}\alpha$	Central interval

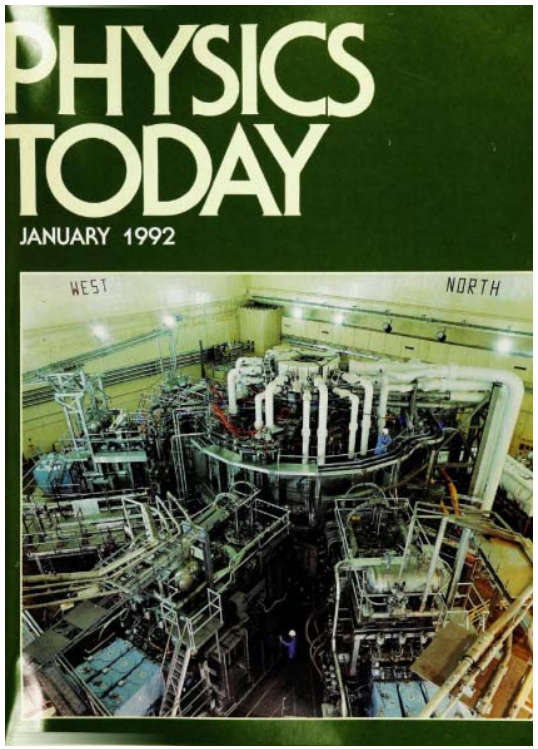
"There is thus no need to derive optimum properties separately for tests and for intervals; there is a one-to-one correspondence between the problems as in the dictionary in Table 20.1" – Stuart99, p. 175.

- **So for frequentist test, choose your favorite test and proceed. In 1D, people tend to focus exclusively on inherently 1-sided test.**
- **This does not generalize to two new parameters of interest, and also has some bad properties. Some of this is ameliorated by "unified approach advocated by F-C. But use one-sided for now.**
- **Consider a set of experiments, all giving (exactly) 5-sigma effect, but some having much better resolution on θ than others.**
- **Useful to think of the "better and better" experiments as simply having same intrinsic measurement apparatus, with larger and larger sample size n.**
- **Does "5-sigma" give the necessary and sufficient information to convey to a consumer?**

REFERENCE FRAME

THE REVEREND THOMAS BAYES, NEEDLES IN HAYSTACKS, AND THE FIFTH FORCE

Philip W. Anderson

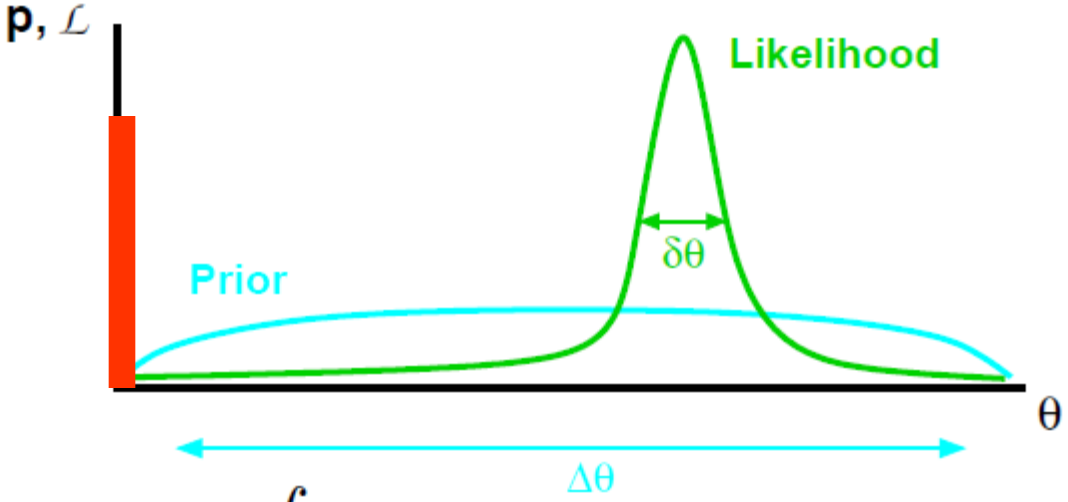


These statistics are the correct way to do inductive reasoning from necessarily imperfect experimental data.

Let us take the "fifth force." If we assume from the outset that there *is* a fifth force and we need only measure its magnitude, we are assigning the bin with zero range and zero magnitude an infinitesimal probability to begin with. Actually, we should be assigning this bin, which is the null hypothesis we want to test, some *finite a priori* probability—like $\frac{1}{2}$—and sharing out the remaining $\frac{1}{2}$ among all the other strengths and ranges. We then ask the question, Does a given set of statistical measurements increase or decrease this share of the probability? It turns out that when one adopts this point of view, it often takes a *much larger* deviation of the result from zero to begin to decrease the null hypothesis's share than it would in the conventional approach. The formulas are complicated, but there are a couple of rules of thumb that give some ideas of the necessary factor. For a large number N of statistically independent measurements, the probability of the null hypothesis must increase by a factor of something like $N^{1/2}$. (For a rough idea of where this

Recall from Tom Loredo Yesterday
(I have added "atom" of probability at null.)

The Occam Factor



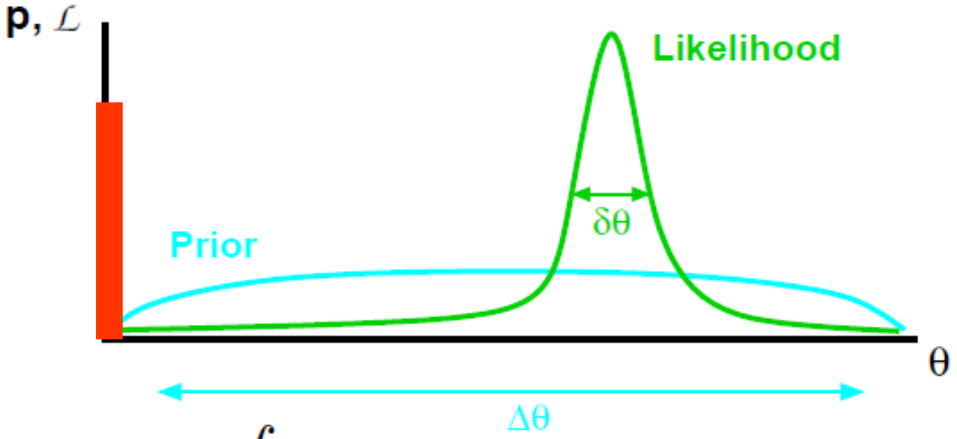
$$p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i$$

$$\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i}$$

$$= \text{Maximum Likelihood} \times \text{Occam Factor}$$

As sample size increases, $\delta\theta$ decreases as $1/\sqrt{n}$

The Occam Factor



$$p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i$$

$$\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i}$$

$$= \text{Maximum Likelihood} \times \text{Occam Factor}$$

$\Rightarrow$ Given ε and fixed p-value, there exists an n for which posterior P in favor of alternative is < ε. (Still assuming null has a fixed prior p.)

Lindley, 1957

A STATISTICAL PARADOX

By D. V. LINDLEY

Statistical Laboratory, University of Cambridge

An example is produced to show that, if H is a simple hypothesis and x the result of an experiment, the following two phenomena can occur simultaneously:

(i) a significance test for H reveals that x is significant at, say, the 5 % level;

(ii) the posterior probability of H, given x, is, for quite small prior probabilities of H, as high as 95 %.

Clearly the common-sense interpretations of (i) and (ii) are in direct conflict. The phenomenon is fairly general with significance tests and casts doubts on the meaning of a significance level in some circumstances.

...

The paradox is not, in essentials, new, although few statisticians are aware of it. The difference between the two approaches has been noted before by Jeffreys (see, in particular, 1948, Appendix), who is the originator of significance tests based on Bayes's theorem and a concentration of prior probability on the null value. But Jeffreys is concerned to emphasize the similarity between his tests and those due to Fisher and the discrepancies are not emphasized.

Lindley's Example (corrected by Bartlett)

Let $(x_1, x_2, x_3, \ldots, x_n)$ be a random sample from a normal of mean θ and known variance σ^2.

Let the probability that $\theta = \theta_0$, the value on the null hypothesis, be *c*.

Suppose the remainder of the probability is distributed uniformly over some interval *I* containing θ_0. [Suppose mean of x well within *I*].

Then the posterior odds for $\theta = \theta_0$ are:

$$\frac{c}{1-c}\left[\frac{I}{\sigma}\sqrt{\left(\frac{n}{2\pi}\right)}\exp\left\{\frac{-n(\bar{x}-\theta_0)^2}{2\sigma^2}\right\}\right]$$

Note that variance of mean of x is σ^2/n.

Now consider the scaling with n in the situations with same
Z = (mean-of-x $-$ θ_0) / ($\sigma/\sqrt{n}$).

The odds in favor of θ_0 increase without bound as $\sqrt{n}$.

Bartlett, 1957

A comment on D. V. Lindley's statistical paradox

By M. S. BARTLETT
University of Manchester

reasonable to choose the sample size n analogously, making $\sqrt{n}$ proportional to $1/I$. If we write $\sqrt{n} = A\sigma/I$, we obtain

$$\frac{\bar{c}}{1-\bar{c}} = \frac{c}{1-c}\left\{\frac{A}{\sqrt{(2\pi)}}e^{-\frac{1}{2}\lambda^2}\right\}, \qquad (2)$$

where $\lambda = \sqrt{n}(\bar{x}-\theta_0)/\sigma$; in (2) there is a constant relation between $\bar{c}$ and λ for fixed c.

Editor's Note (Kendall)

In regard to Prof. Bartlett's final point, it may be useful to observe that some procedure of the type he suggests is implicit in the idea of the asymptotic relative efficiency of a test. In general, the power of a test against a specific alternative tends to unity with increasing sample size. To compare two tests asymptotically at a fixed significance level, it is necessary to allow the alternative to approach the null hypothesis as the sample size increases.

M.G.K.

Jeffreys, 3rd Edition, Appendix B

(p. 434 ff) It is interesting to compare the results with those based on the customary use of the *P* integral…

In spite of the difference in principle between my tests and those based on the *P* integrals, and the omission of the latter to give the increase of the critical values for large *n*, dictated essentially by the fact that in testing a small departure found from a large number of observations we are selecting out a value out of a long range and should allow for selection, it appears that there is not much difference in the practical recommendations…

At large numbers of observations there is a difference, since the tests based on the integral would sometimes assert significances at departures that would actually give $K>1$. But these will be very rare…

If an estimate gives $K>1$ and $P<0.01$, internal correlation should be suspected and tested…

Ward Edwards, Harold Lindman, Leonard J. Savage, 1963

VOL. 70, NO. 3 MAY 1963

PSYCHOLOGICAL REVIEW

BAYESIAN STATISTICAL INFERENCE FOR PSYCHOLOGICAL RESEARCH [1]

WARD EDWARDS, HAROLD LINDMAN, AND LEONARD J. SAVAGE

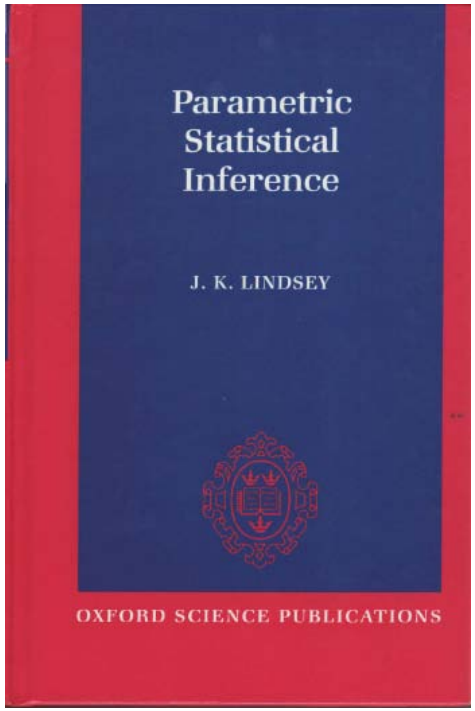
University of Michigan

Experiment #:	1	2	3	4
n	50	100	400	10 000
y	32	60	220	5098
$\hat{\pi}$	0.64	0.60	0.55	0.51
Bayes factor	0.82	1.09	2.17	11.69
Normed likelihood	0.14	0.13	0.13	0.15

More important, Experiments 3 and 4, which would lead a classical statistician to reject the null hypothesis, leave the Bayesian who happens to have a roughly uniform prior, more confident of the null hypothesis than he was to start with. And Experiment 4 should reassure even a rather skeptical person about the truth of the null hypothesis. Here, then, is a blunt practical contradiction between conclusions produced by classical and

Translated into modern language/notation by J.K. Lindsey, p. 356, who also notes that Bayes factor is not monotonic with *n*.

J.K. Lindsey, 1996



2. Consider again the normal distribution with known variance, but let us now compare the model H_0 with $\mu_0 = 0$ to H_A with μ_A unspecified. We shall here take a proper uniform prior for μ_A: $dP(\mu_A) = 1/(2c)$ on $(-c, c)$ for c large. Then, the Bayes factor for comparing the two hypotheses is

$$\frac{2c\sqrt{\frac{n}{\sigma^2}}\exp\left[-\frac{n\bar{y}_\bullet^2}{2\sigma^2}\right]}{\int_{\sqrt{n}(\bar{y}_\bullet - c)/\sigma}^{\sqrt{n}(\bar{y}_\bullet + c)/\sigma}\exp[-u^2]du}$$

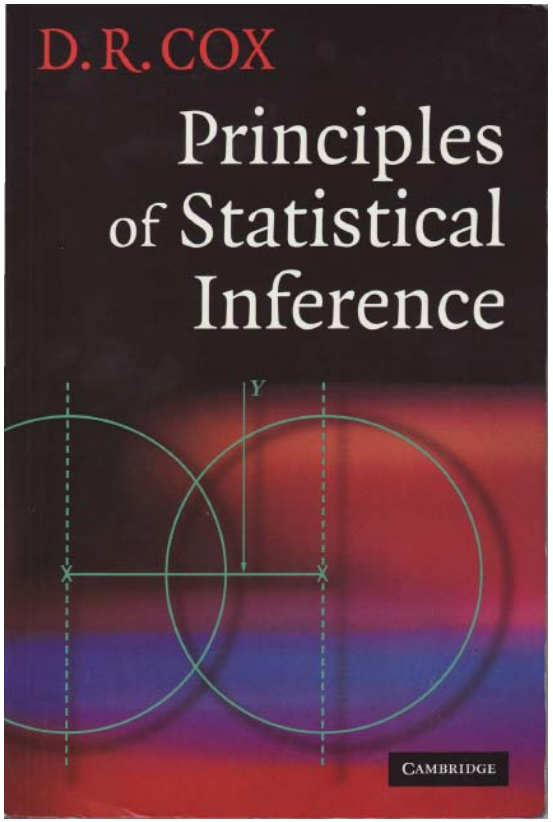
For a given frequentist significance level, we would fix $\sqrt{n}\bar{y}_\bullet/\sigma$. The denominator rapidly approaches unity as c increases. Thus, this Bayes factor becomes arbitrarily large as $c \to \infty$ or $n \to \infty$, for any fixed value of the significance level, increasingly favouring H_0 over H_A.

p. 356: In the light of such results, some Bayesians have argued that sharp hypotheses are unreasonable…

Others have held that the improper prior causes the problem and that Bayesian decision-making should be limited to informative priors. Example 2 above demonstrates that the problem does not lie here.

Thus, this paradox appears to imply that, if one already has enough prior information to place a point mass (counting measure) on one hypothesis, but not on other possible individual models, so that the latter set has Lebesgue measure, then empirical data are not necessary. In this sense, the prior probability of a sharp hypothesis should always be zero (Novick, 1969). But this evidently causes problems for model selection.

David Cox, 2006



p. 106: …we now consider where there is an atom of probability π_0 at a null hypothesis $\theta = \theta_0$, and the remaining prior probability is distributed over nonnull values. It is tempting to write this latter part in the form $(1 - \pi_0)p_0(\theta)$, where $p_0(\theta)$ is some smooth density not depending on n. This is however, often to invite misinterpretation, because in most, if not all, specific applications in which a test of a hypothesis is thought worth doing, the only serious possibilities of such a hypothesis 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. This suggest that the remaining part of the prior density should usually be taken of the form $q\{(\theta - \theta_0)\sqrt{n}\}\sqrt{n}$, where q(.) is some fixed probability density function…

Thus as $n \to \infty$ the posterior odds are asymptotically a fixed function of the test statistic… the relationship between the significance level and the posterior odds is independent of n.
[Bartlett 1957 same idea: ext has $\sqrt{n}$ prop to 1/scale.]

Christian Robert, 1993

The fundamental argument underlying our reevaluation of the Jeffreys-Lindley paradox is that the prior probability ρ_0 of the null hypothesis H_0 should depend on the prior variance under the alternative hypothesis H_1, σ^2. …

The dependency of ρ_0 on σ^2 thus avoids the undesirable convergence to 1 and provides an estimator which can be considered as a noninformative answer and a Bayesian counterpart to the p-value…

…behavior seems to be quite unreasonable… [more discussion]

O'Hagan (K&S vol. 2B), 1994

Section 7.43: Difficulties arise with nested models if $f_1(\varphi)$ is specified to represent very weak prior information about φ …

7.46 Sensitivity will also be a problem even when prior information about φ is not particularly weak but the data are strong… Any perturbation of the prior distribution $f_1(\varphi)$ that alters its value $f_1(\varphi\text{-hat})$ at $\varphi = \varphi\text{-hat}$ will result in a proportionally identical change in the Bayes factor…

This is in direct contrast to the argument in 3.26 which says that as the amount of data increases the prior information is overwhelmed by the data and becomes irrelevant. [3.26 was about posterior for continuous parameter.]

Richard M. Royall, 1986

Conclusion on direction of Sample-Size dependence depends on whether one works with p-value, or with index of whether fix threshold was passed!

Berger and Sellke, 1987, with comments

"The main reason for the substantial difference between the magnitude of p and the evidence against H_0…is essentially one of conditioning. The actual vector of observations is x, and Pr(H_0|x) and l_x depend only on the evidence from the actual data observed. To calculate a P value, however, one effectively replaces x by the "knowledge" that X is in A = {y: T(y) >= T(x)} and then calculates p = $Pr_{\theta=\theta_0}(A)$… Common sense supports the distinction between x and A…

I.J.Good: "…I proposed "standardizing" a tail-area probability P to sample size 100, by replacing P by min(1/2, n^{1/2}P/10) (Good 1982b)

Varadaman: finds priors with "spike at θ_0" "completely unappealing". Uses "a physical constant light the speed of light" as example (!).

Casella and R. Berger: "A good frequentist would always report the probabilities of both Type I and Type II error, and Morris shows us that reporting the sample size along with the p value is somewhat equivalent to this; we thoroughly agree with him.

B&S rejoinder: "…nonconstancy in interpretation of P value: as the sample size increases, a given P value provides less and less real evidence against the null… We remain unconvinced that p-values have any merit."

Glenn Shafer, 1982 and comments

Morris DeGroot comment: I do not agree with the notion expressed by Shafer, and by many others before him, that a diffuse prior represents ignorance about, θ or with the statement, "the more ignorant we are, the more diffuse" the prior distribution should be. Indeed, a diffuse prior distribution, represented by a normal distribution, indicates not that I am totally ignorant about θ_0 but that I am quite certain the $|\theta|$ is large. I doubt that the concept of total ignorance about θ has any precise meaning…

In summary, diffuse prior distributions…are never appropriate for tests of significance.

I.J. Good comment: The Neyman-Pearson theory of errors of the first and second kind also shows that a given tail-area probability has less power when N is increased" (see e.g., Leamer 1978, Good 1980)

Aitkin, 1991

Smith and Spiegelhalter (1980) proposed the use of the prior N(mu_1,sigma^2/n) for mu_2, in the context of local alternatives to the null…
This Bayes factor is not a function of n and therefore does not suffer from the Lindley paradox, it requires very large values of z for evidence against M_1…

Bernard comment: In this spirit I wish that we could agree on regularly quoting the observed likelihood ratio L(H_1|y)/L(H_0)|y) in addition to attained mid-P values on hypothesis H_0, and power on a specified alternative H_1.
With Aitkin's example…such a practice would produce the correct conclusion – that model 1 fits much better than model 2, but neither model 1 nor model 2 fits at all well.

Cox comment: I agree with Professor Bernard that we must distinguish between:
(a) The assessment of the relative fit of two models, M_1 and m_2, assuming provisionally that one of the models is "true" , and
(b) Analysis of the adequacy of M_1 looking for departures in the direction of M_2, and vice versa.
In (b), the conclusion may be that the fit of both, one, or neither model is adequate.

Jose Bernardo and Raul Rueda, 2002

…If θ is a continuous parameter, this *forces* the use of a *non-regular* (not absolutely continuous) `sharp' prior concentrating a positive probability mass on θ_0. One unappealing aspect of this non-regular prior structure, noted by Lindley (1957) and generally known as *Lindley's paradox*, is that for any fixed value of the pertinent test statistic, the Bayes factor increases as sqrt(n) with the sample size; hence, with large samples, "evidence" in favor of H_0 may be over-whelming with data sets that are both implausible under H_0 and quite likely under alternative θ_0 values, such as (say) the MLE θ-hat.

The Bayes factor approach to hypothesis testing in a continuous parameter setting…analyzes how such *very strong* beliefs about the value of θ should be modified by the data…
Bayes factors should *not* be used unless this strong prior formulation is an appropriate assumption.

[Italics in original: they evidently have strong beliefs about this!]

Berger and Delampady, 1987, and comments

Do such objective Bayesian answers always exist, however? The answer is no, and the precise null testing situation is a prime example in which objective procedures so not exist...

Unfortunately, the choice of the scale factor, tau, for *g* has a large effect on the answer…
A dramatic effect o the Bayesian and likelihood answer. Furthermore, letting $\tau^2 \rightarrow \infty$ so that g becomes "non-informative" is ridiculous, since then P(H_0|x)→ 1. Thus, a Bayesian must, at a minimum, subjectively specify τ^2 , and there is no default value that "lets the data speak for itself".

…it becomes ridiculous to argue that we can intuitively learn to properly calibrate P-values…
First and foremost, when testing precise hypotheses, formal use of P-values should be abandoned.

Cox comment: In summary, it seems to me that the paper is a valuable and thought-provoking one, but that the conclusion that P-values have no role at all is wrong

What about using the p-value as the "test statistic" and proceeding with Bayesian analysis?

Physicists interpretation of "number of sigma" combined with prior belief is an informal attempt at something like this, I think.

Various papers try this:, are generally unhappy with results:

James Dickey, 1977. (Also has idea of prior densities locally a power of mu)
Berger and Mortera, 1991
Johnstone and Lindley, 1995

I think that Jim Berger will discuss conditioning aspects, including his work on Conditional Frequentist tests and "unification" with Bayesian approach

E.g., Berger, Brown, Wolpert 1994 and Berger, Boukai, Wang, 1997.
From the latter:

A final comment on this issue is that precise hypothesis testing should not be done by forming a traditional confidence interval (frequentist or Bayesian) and simply checking whether or not the precise hypothesis is compatible with the confidence interval. A confidence interval is usually of considerable importance in determining where the unknown parameter (say) is likely to be, given that the alternative hypothesis is true, but it is not useful in determining whether or not a precise null hypothesis is true. For discussion of this point, see Berger and Delampady (1987).

Bounds on Bayes Factors

Edwards, Lindman, Savage, 1963
p. 225 "…classical procedures quite typically are, from a Bayesian point of view, far too ready to reject null hypotheses."

At least for Bayesian, however, no procedure for testing a sharp null is likely to be appropriate unless the null hypothesis deserves special attention.

Idea of bound on Bayes factors developed by Jim Berger and Collaborators.

Using Bayes Factor as test statistic in frequentist calculation

Idea promoted as part of "Bayes frequentist compromise" by I.J. Good (1957, 1982, 1992).

In fact I think we do essentially this in HEP! We often use likelihood ratios as the test statistics, and often integrate out nuisance parameters rather than profile them.

Of course, many Bayesians do not want to compromise and see this as inheriting ills of frequentism.

"The real objection to P values is not that they are utter nonsense, but that they can be highly misleading, especially if the value of N is not taken into account… replace P by $P_{stan} = \min(1/2, P\sqrt{N} / 10)$.

Comic relief: Good 1992 ends with crackpot numerology on fine structure constant a la Eddington! Seems not to know u and d quark have different masses (in 1992!). Is this how physicists look when talk about statistics? Yikes.

In “modern” use of N-P testing, how should alpha be chosen?

E.L. Lehmann, Testing Statistical Hypotheses, 1959

The choice of a level of significance alpha will usually be somewhat arbitrary…
In fact, when choosing a level of significance one should also consider the power that the test will achieve against various alternatives. If the power is too low one may wish to use much higher values of alpha…

Another consideration that frequently enters 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 be most unlikely to occur if H were true.

In “modern” use of N-P testing, how should alpha be chosen? (cont.)

E.L. Lehmann, 1993:

> It is interesting to note that unlike Fisher, Neyman and Pearson (1933a, p. 296) did not recommend a standard level but suggested that “how the balance [between the two kinds of error] should be struck must be left to the investigator,” and (1933b, p. 497) “we attempt to adjust the balance between the risks P_{I} and P_{II} to meet the type of problem before us.”

E.L. Lehmann, 1993:

1. Both Neyman–Pearson and Fisher would give at most lukewarm support to standard significance levels such as 5% or 1%. Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice. Neyman–Pearson, in their original formulation of 1933, recommended a balance between the two kinds of error (i.e., between level and power). For a disucssion of how to achieve such a balance, see, for example, Sanathanan (1974). Both level and power should of course be considered conditionally whenever conditioning is deemed appropriate. Unfortu-

Journal of the American Statistical Association, Vol. 69, No. 346 (Jun., 1974)

Critical Power Function and Decision Making

LALITHA SANATHANAN*

It is generally recognized that when deciding on a significance level for a test, its power must also be taken into account. Determination of the optimal significance level, however, is not a straightforward task. It is pointed out here that by use of critical significance level and power we can achieve the usual objective of making a reject-or-accept decision without explicitly determining the optimal significance level.

arXiv:1005.1891v1

Trial factors or the look elsewhere effect in high energy physics

Eilam Gross, Ofer Vitells

Weizmann Institute of Science, Rehovot 76100, Israel

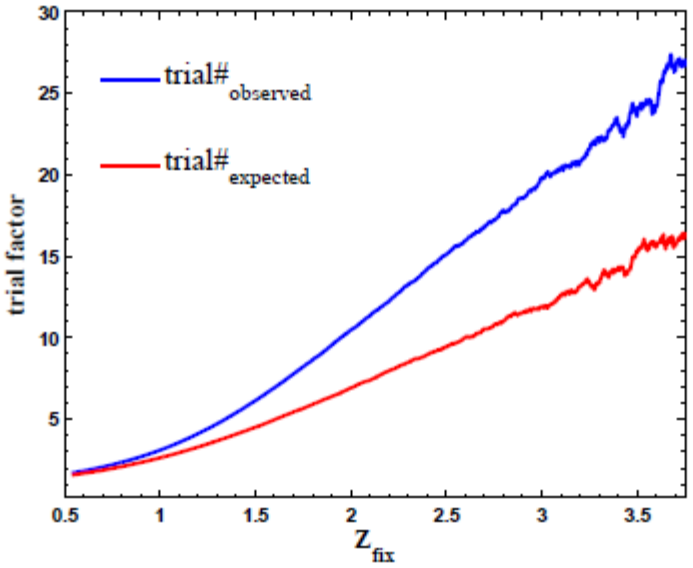
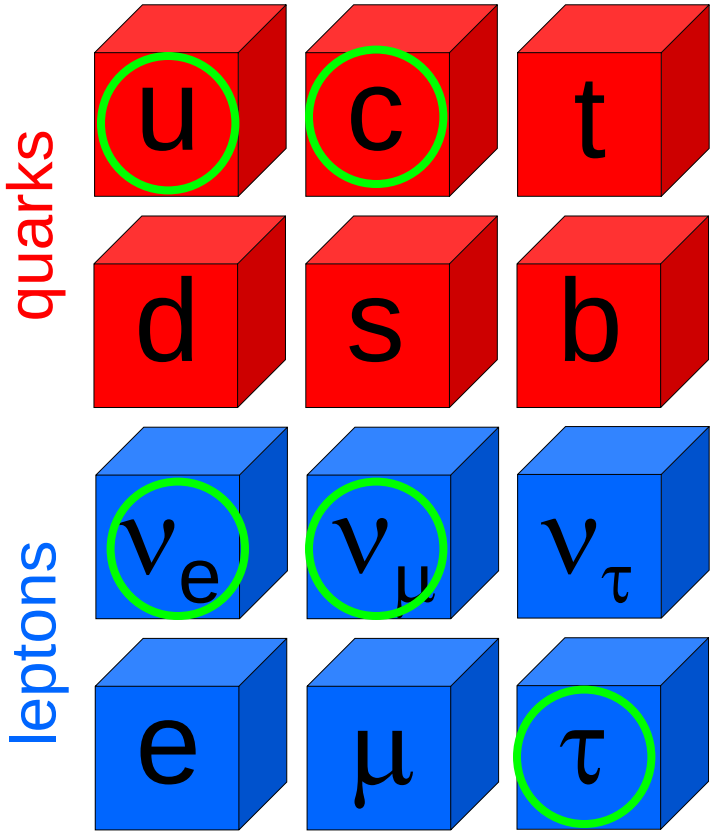


Figure 6: The trial factors as a function of the fixed mass significance Z_{fix}.

$$trial\#_{observed} \simeq \frac{1}{3}\frac{range}{resolution}Z_{fix}$$

!!! √n

Claim: **In most, if not all, of HEP's "Nobel" level discoveries, there is a new particle or new "discrete" symmetry in nature.**



Circles: Nobel Prizes

Also Nobels for two antiparticles: positron, anti-proton, and force carriers: W, Z bosons

Discovery of b quark

1 AUGUST 1977

Observation of a Dimuon Resonance at 9.5 GeV in 400-GeV Proton-Nucleus Collisions

S. W. Herb, D. C. Hom, L. M. Lederman, J. C. Sens,[a] H. D. Snyder, and J. K. Yoh
Columbia University, New York, New York 10027

and

J. A. Appel, B. C. Brown, C. N. Brown, W. R. Innes, K. Ueno, and T. Yamanouchi
Fermi National Accelerator Laboratory, Batavia, Illinois 60510

and

A. S. Ito, H. Jöstlein, D. M. Kaplan, and R. D. Kephart
State University of New York at Stony Brook, Stony Brook, New York 11974
(Received 1 July 1977)

[5]The errors quoted on the magnitude of the continuum and resonance cross sections and the resonance masses are statistical only. Systematic normalization effects are probably less than 25% and do not affect the conclusions drawn here. Systematic errors on the mass calibration are probably less than 1%.

The discovery was all in the shape.

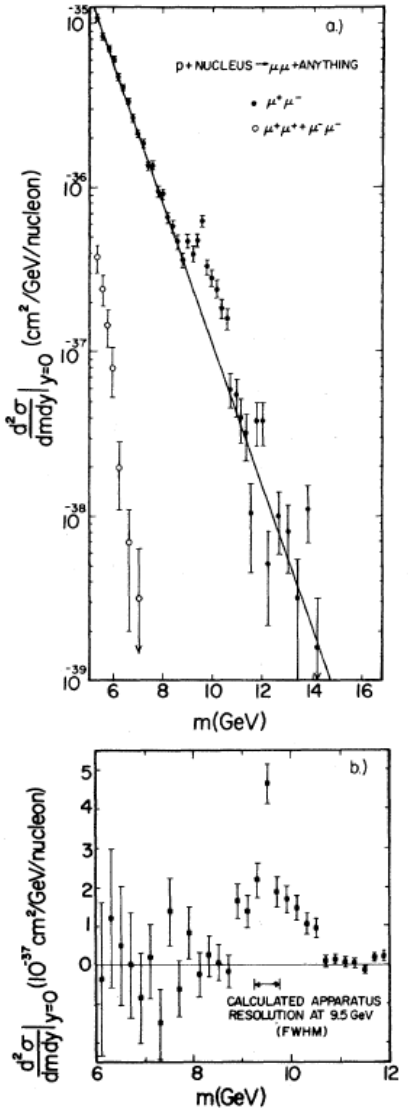


FIG. 3. (a) Measured dimuon production cross sections as a function of the invariant mass of the muon pair. The solid line is the continuum fit outlined in the text. The equal-sign–dimuon cross section is also shown. (b) The same cross sections as in (a) with the smooth exponential continuum fit subtracted in order to reveal the 9–10-GeV region in more detail.

…The Upsilon was Preceded by the Oops-Leon

Observation of High-Mass Dilepton Pairs in Hadron Collisions at 400 GeV

D. C. Hom, L. M. Lederman, H. P. Paar, H. D. Snyder, J. M. Weiss, and J. K. Yoh
*Columbia University, New York, New York 10027**

and

J. A. Appel, B. C. Brown, C. N. Brown, W. R. Innes, and T. Yamanouchi
Fermi National Accelerator Laboratory, Batavia, Illinois 60510†

and

D. M. Kaplan
*State University of New York at Stony Brook, Stony Brook, New York 11794**
(Received 28 January 1976)

We report preliminary results on the production of electron-positron pairs in the mass range 2.5 to 20 GeV in 400-GeV p-Be interactions. 27 high-mass events are observed in the mass range 5.5–10.0 GeV corresponding to $\sigma = (1.2 \pm 0.5) \times 10^{-35}$ cm^2 per nucleon. Clustering of 12 of these events between 5.8 and 6.2 GeV suggests that the data contain a new resonance at 6 GeV.

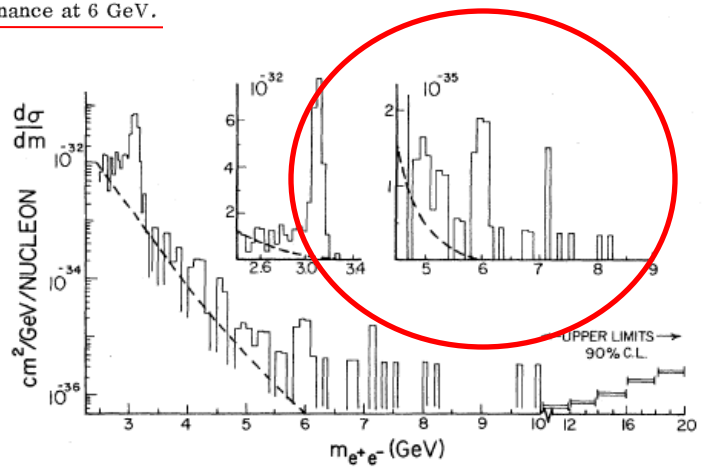
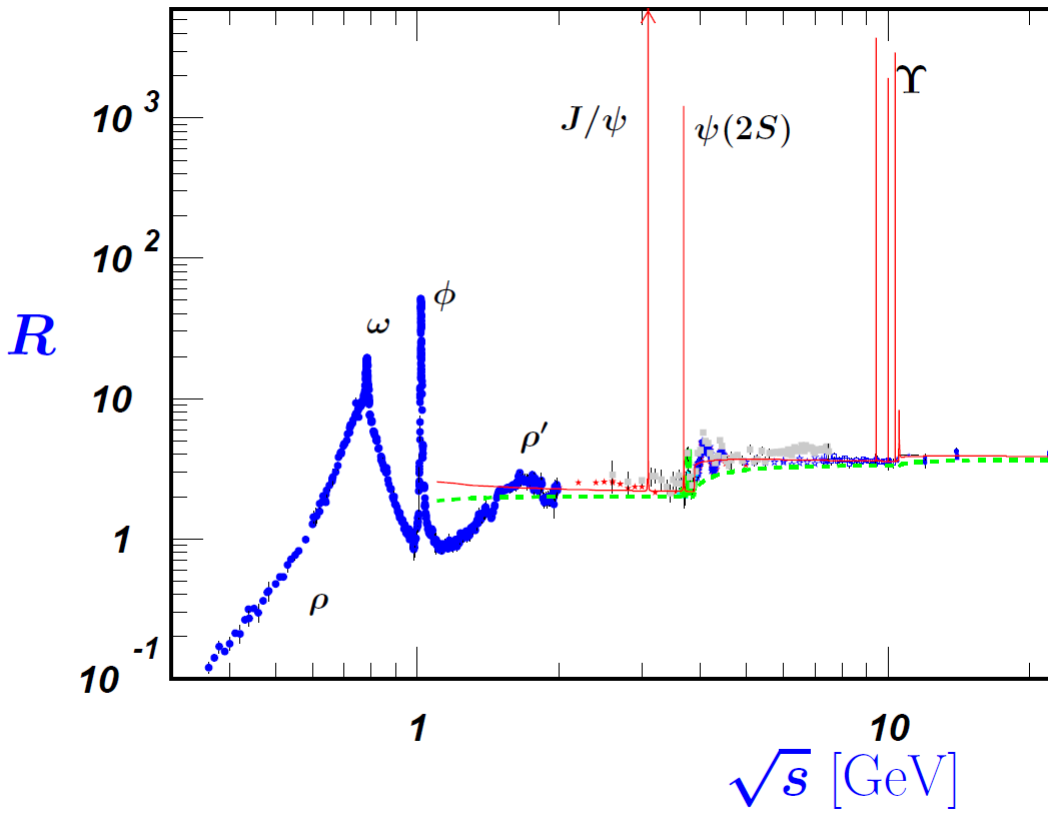


FIG. 2. Electron-positron mass spectrum: $d\sigma/dm$ per nucleon versus the effective mass. A linear A dependence is assumed. Note bin-width changes.

The events near 6 GeV correspond to a total cross section of $\sigma B = (5.2 \pm 2.0) \times 10^{-36}$ cm^2 per nucleon under the assumptions of Eq. (3) and of a linear A dependence.[7] We have studied the probability for a clustering of events as is observed here to result from a fluctuation in a smooth distribution, e.g., Eq. (3). To avoid the difficult problems involved in the statistical theory associated with small numbers of events per resolution bin, a Monte Carlo method was used. Histograms were generated by throwing events according to a variety of smooth distributions, modulated by the mass acceptance, over the mass range 5.0 to 10.0 GeV. Clusters of events as observed occurring anywhere from 5.5 to 10.0 GeV appeared less than 2% of the time.[8] Thus the statistical case for a narrow (< 100 MeV) resonance is strong although we are aware of the need for confirmation. These data, at a level of

Toy M.C. of look-elsewhere effect!
LEE-corrected p was 0.02.

Ratio of production of muons and quarks in e+e- collisions



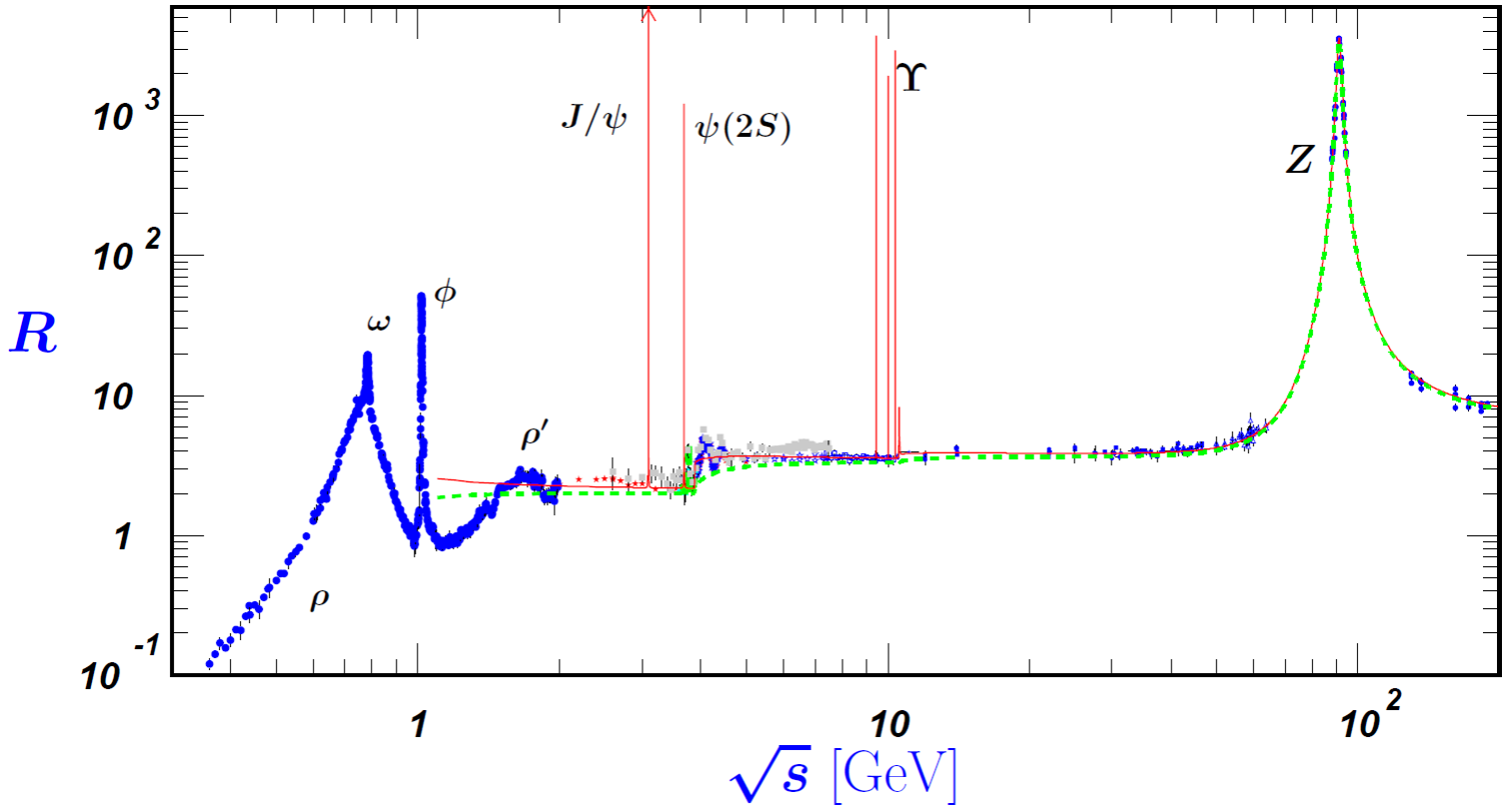
To the trained eye, the different horizontal levels correspond to small *integers*:

Electric Charge of produced quarks in units of e/3:

(n=1,2; e = charge of proton) , and

Number of types of charges in the strong force (3 *colors !*)

Continuation of the plot is fun as well!



The first evidence of the new "discrete" addition to the fundamental theory can be a small effect change in an experimental observable

- **Cronin & Fitch Nobel Prize for discovering that in 1 out of 500 decays of the long-live neutral kaon, there were two pions rather than three in the final state.**
- **Implication was that equations of physics were not time-reversal invariant!**

EVIDENCE FOR THE 2π DECAY OF THE K_2^0 MESON*†

J. H. Christenson, J. W. Cronin,‡ V. L. Fitch,‡ and R. Turlay§
Princeton University, Princeton, New Jersey
(Received 10 July 1964)

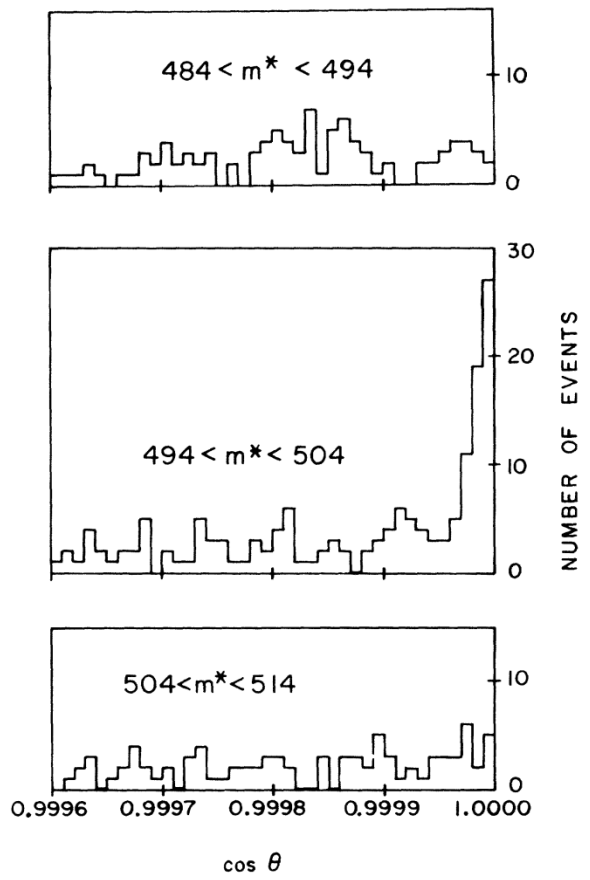


FIG. 3. Angular distribution in three mass ranges for events with $\cos\theta > 0.9995$.

[$m^* = m_{\pi\pi}$; $\cos\theta=1 \leftrightarrow$ MET=0]

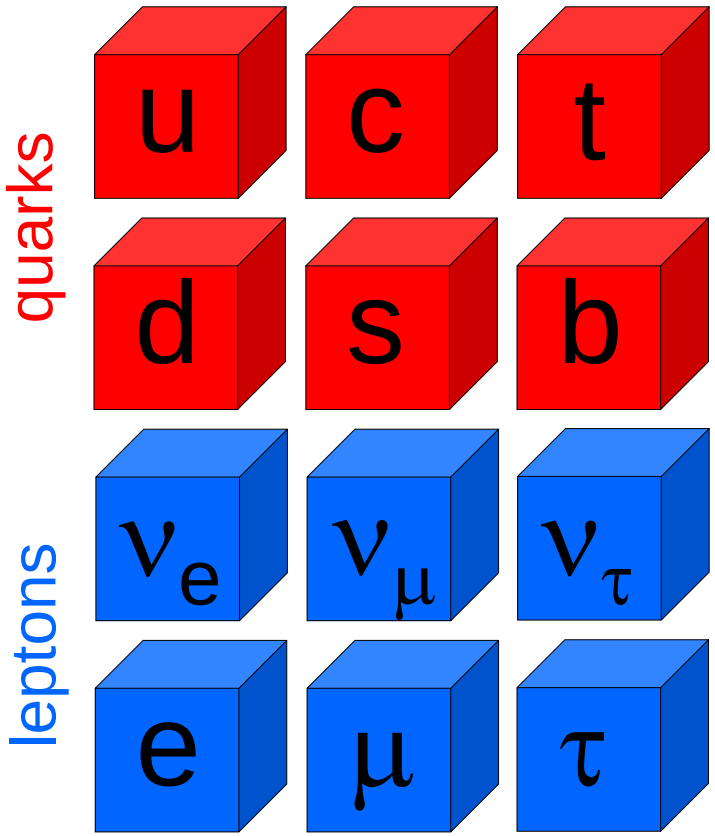
More generally, the existence of a new force-carrying boson can show up as a *tiny* change in the interaction rate of order

$1/M^2$ or $1/M^4$, where

***M* is mass of the new particle (and absolute rate involves a coupling constant as well)**

Ranges of *M* currently accessible are ~10^3 GeV at LHC, ~10^5 GeV rare decays, ~10^{12} GeV from neutrinos, ~10^{15} GeV from proton decay.

What are the smallest building blocks of matter?

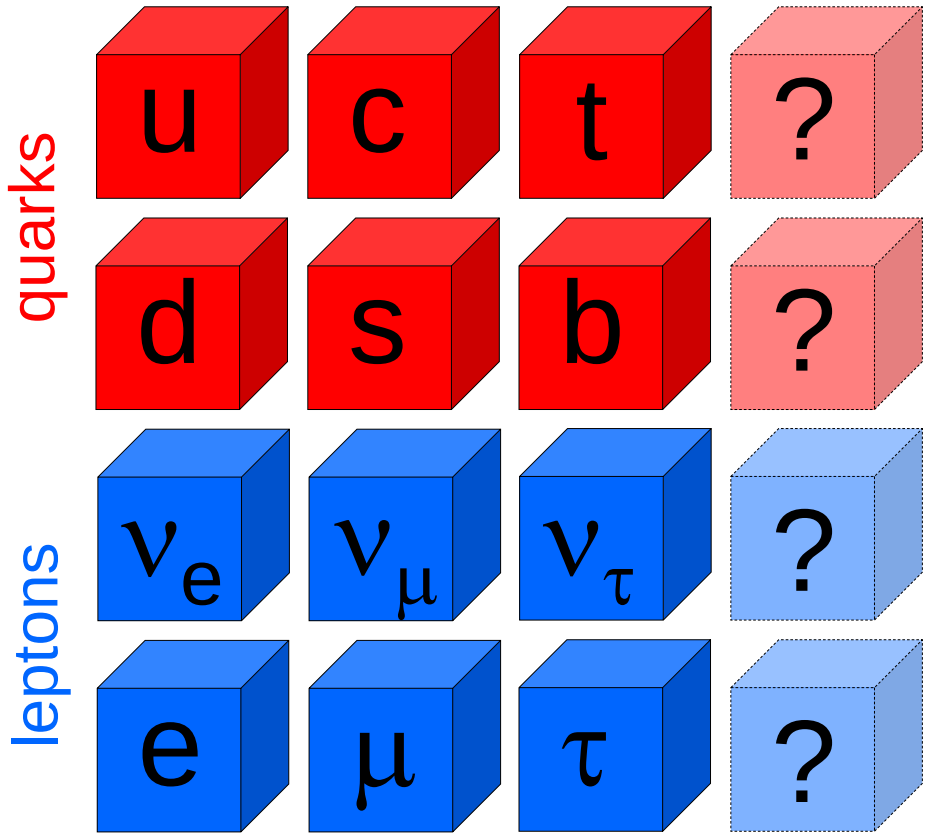


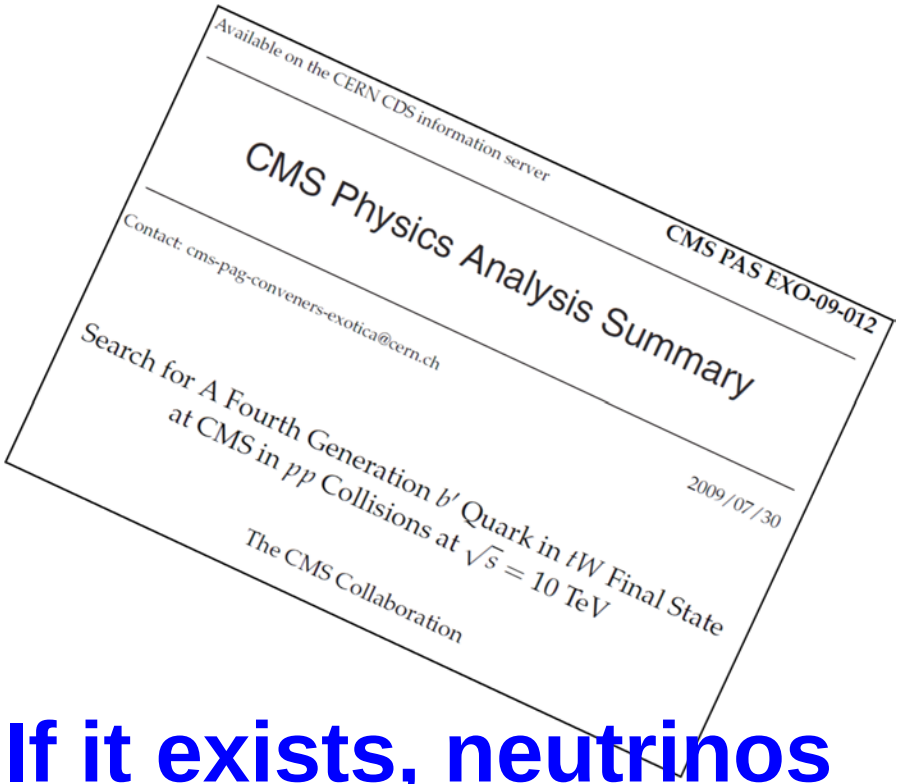
LHC experiments will look for signs of yet smaller building blocks.

(So far, no sign from experiment *or theory*.)

And is there yet another column???

LHC experiments will look for it.




Available on the CERN CDS information server

CMS PAS EXO-09-012

CMS Physics Analysis Summary

Contact: cms-pag-conveners-exotica@cern.ch

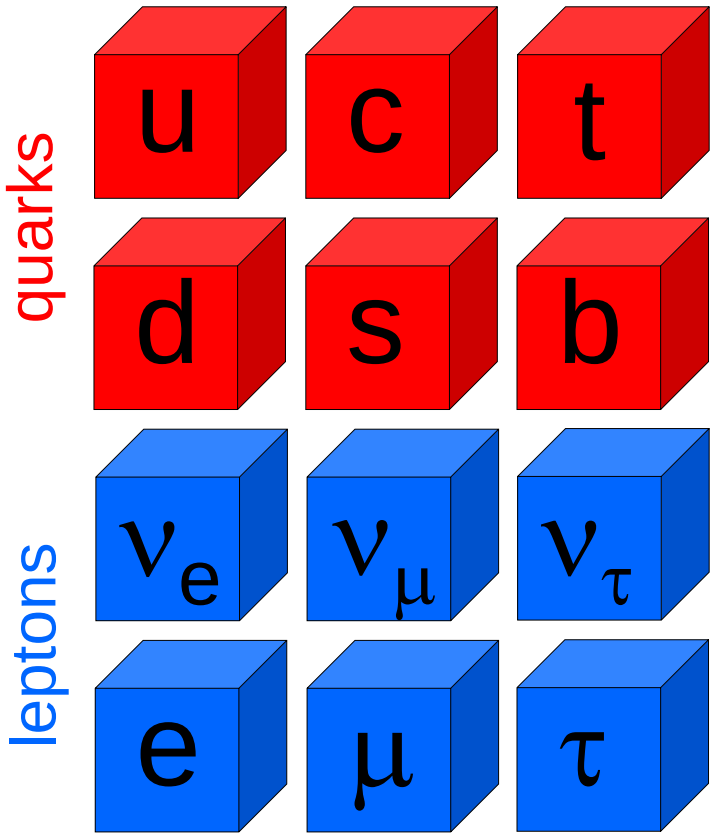
2009/07/30

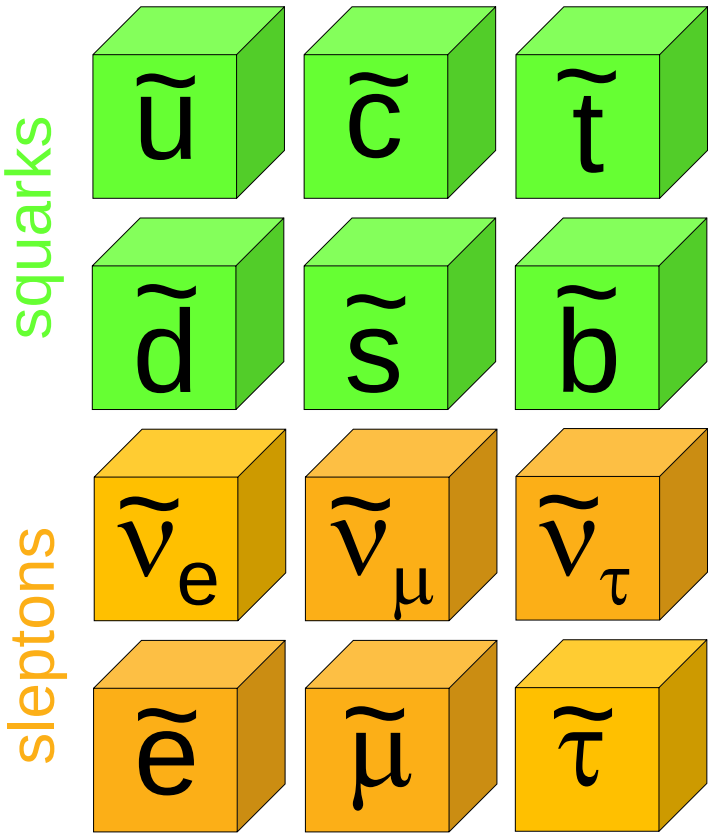
Search for A Fourth Generation b' Quark in tW Final State at CMS in pp Collisions at $\sqrt{s} = 10$ TeV

The CMS Collaboration

If it exists, neutrinos are very massive; that downgrades many people's prior.

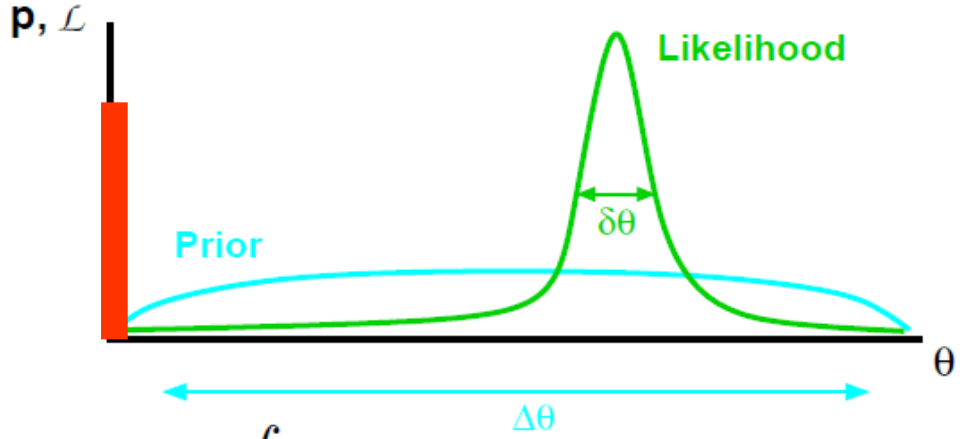
Supersymmetry: Double the whole table with a new type of matter!?





Heavy versions of every quark and lepton!

The Occam Factor



$$p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i$$

$$\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i}$$

$$= \text{Maximum Likelihood} \times \text{Occam Factor}$$

$\Rightarrow$ Given ε and fixed p-value, there exists an n for which posterior P in favor of alternative is < ε. (Still assuming null has a fixed prior p.)

Does this Occam Factor Really Correspond to the Way "Good" Physicists Adjust Beliefs?

Personally, I doubt it.

All kinds of issues, beginning with the obvious ones: To make any sense of it at all, one needs both a cut-off scale and a metric in the parameter space.

For most exploratory experiments I can think of, these metrics just don't exist in a relevant way.

Recent example with Higgs seems to be a counter-example (no one's belief has been modified even though favored region of parameter space is "excluded".

Some recent examples: neutrino oscillations, Minimal Supersymmetry, axion. All had parts of parameter space ruled out.

Axion

- **In 1977, Peccei and Quinn proposed a new symmetry to "explain" why the strong interaction does not have a term violating time-reversal invariance.**
- **Weinberg and Wilzcek independently pointed out that this symmetry would be spontaneously broken, and there would be a new particle: the axion.**
- **Mass of the axion depends depends on the completely unknown scale, from say, below 1 GeV to 10,000,000,000,000,000,000 GeV.**
- **Axion is also a viable "cold dark matter" candidate – and some people think that its supersymmetric partner would be an even better one!**

Annu. Rev. Nucl. Part. Sci. 2006. 56:293–326
doi: 10.1146/annurev.nucl.56.080805.140513

First published online as a Review in Advance on July 5, 2006

SEARCHES FOR ASTROPHYSICAL AND COSMOLOGICAL AXIONS*

Stephen J. Asztalos,[1] Leslie J Rosenberg,[1] Karl van Bibber,[1]
Pierre Sikivie,[2,3] and Konstantin Zioutas[4]

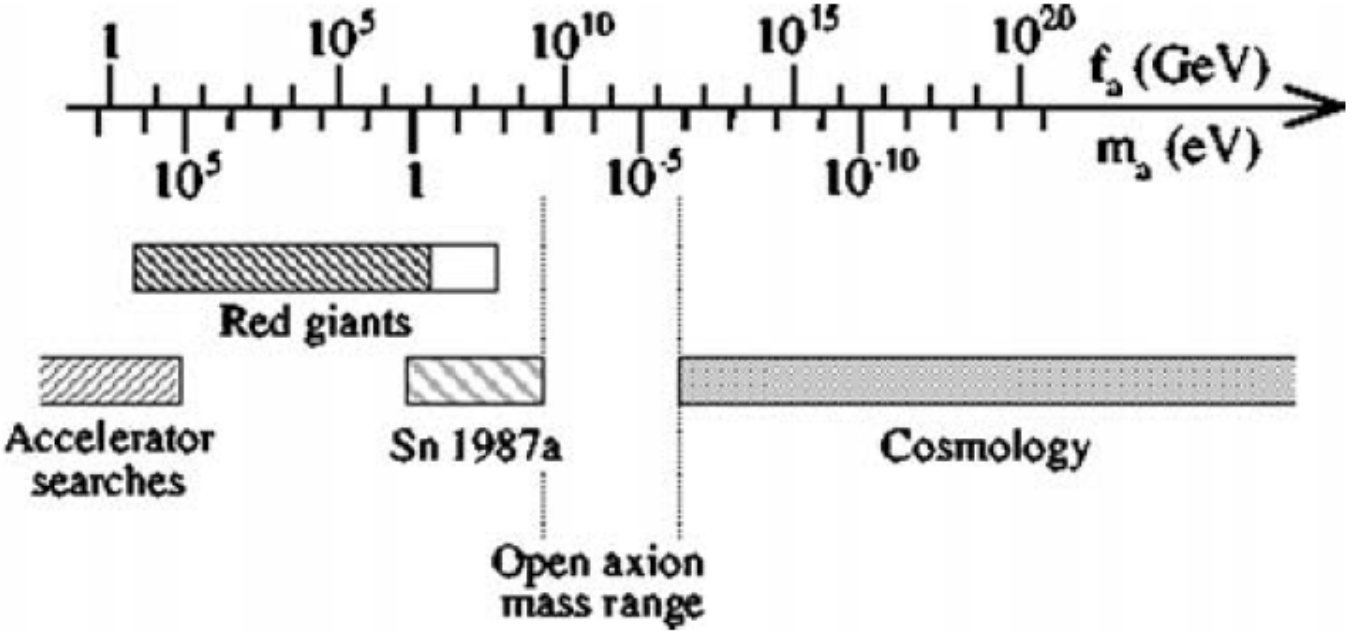


Figure 2 The allowed axion mass range, bounded from below by the requirement that axions should not overclose the universe and from above by accelerator searches and stellar evolution.

Did “real” theorists’ belief get updated by this reduction in parameter space?

- **Using what metric??**
- **This is not necessarily a failure of Bayesian theory in principle, but rather just a statement that in this example, I think it is useless.**

What about the claims in the literature (e.g., Cox) that there is some intrinsic connection between what experiments that get designed and done, and the σ^2/n of the experiment ?

- **Typical modern HEP experiments make many many measurements beyond those which motivated the design, sample size etc. So not really a connection.**

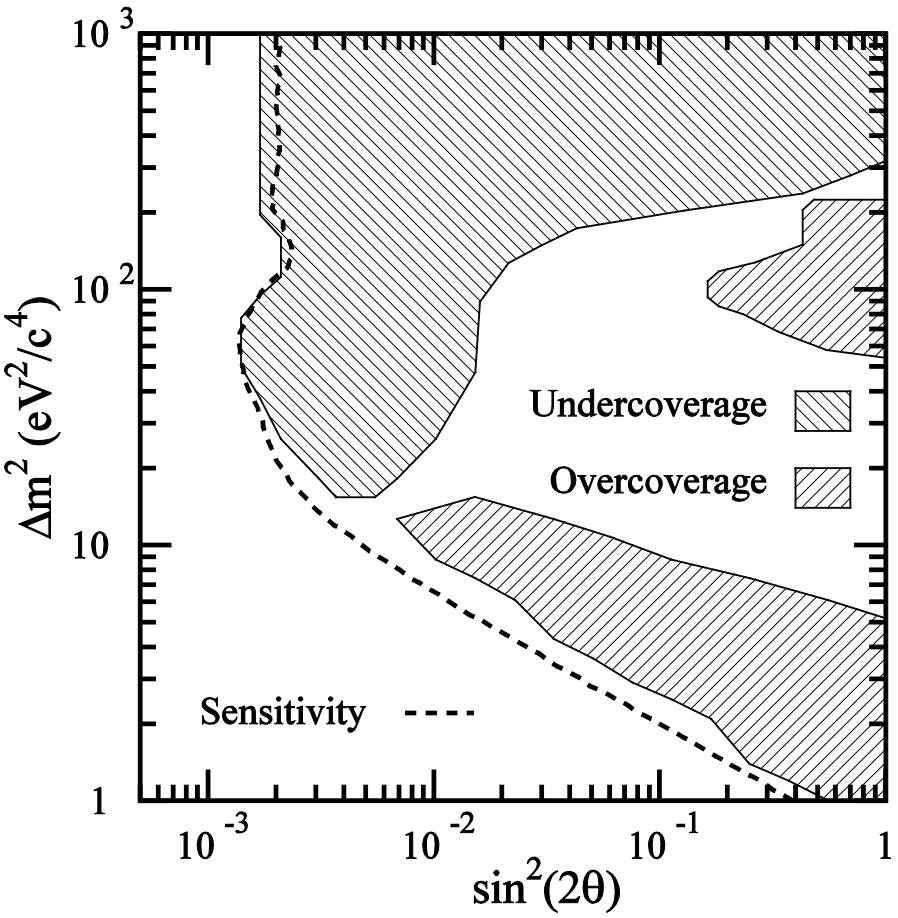
- **For more specialized experiments, there *is* an informal rule of thumb: worth doing "fishing expedition" for a rare process if you can gain a factor of 10 in rareness. So there may indeed be a number that can be used, traced back not to any belief, but to the number of fingers we have. And our preference for multiplication makes the metric flat in the log.**

- **Back to the frequentist p-value approach. What to use if the alternative model has two parameters of interest?**

- **And what does it say about 1 parameter of interest?**

From Feldman and Cousins, 1998

Neutrino Oscillations: Null hypothesis is (0,0)



The global $\Delta\ln(L)$ using "book" values of critical values over-covers in some places and undercovers in others. The reason as that the effective dimensionality of the number of degrees of freedom changes continuously across the plot!

So the notion is to calibrate the critical value as a function of point in parameter space.

The big conceptual point is that, for exact coverage, this table of critical values has to be organized by unknown *true* value!

And that is "all" F-C does!

Classical Hypothesis Testing (cont.)

"Test for $\theta=\theta_0$" $\leftrightarrow$ "Is θ_0 in confidence interval for θ"

Using the likelihood ratio hypothesis test, this correspondence is the basis of intervals/regions we advocated in Phys. Rev. D57 3873 (1998):

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

While paper was "in proof", Gary realized that the method (including nuisance parameters) was all on 1¼ pages of "Kendall and Stuart" ! →

We thought this was *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

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

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

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

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 $\hat{\hat{\boldsymbol{\theta}}}_s$ in (22.2). Now consider the likelihood ratio[2]

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

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,

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

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 $\tau(\hat{\hat{\boldsymbol{\theta}}})$ – cf. **18.3**. Thus the LR statistic is invariant under reparametrization.

Bayesians, Frequentists, and Physicists

Bradley Efron
Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA

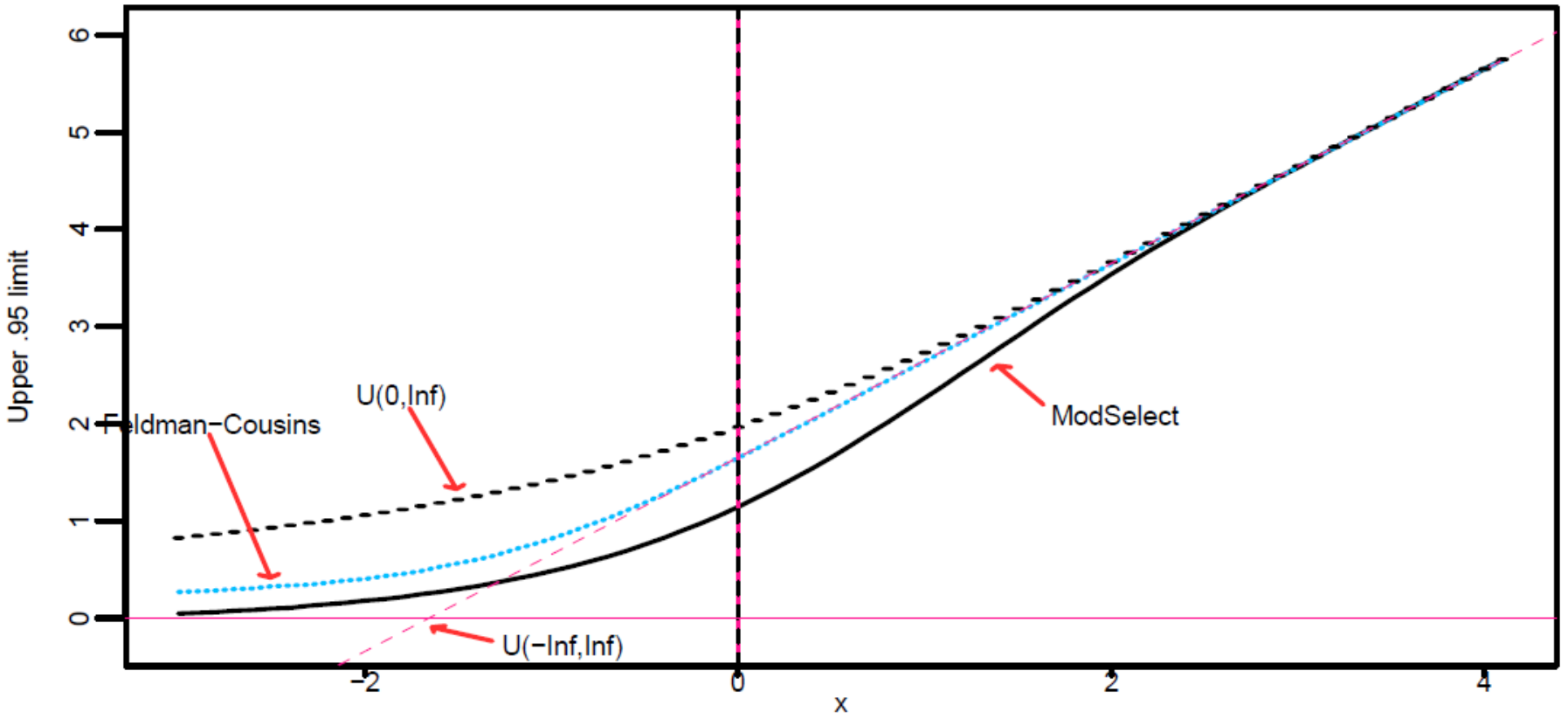


Figure 2: Four possible 95% upper bounds for μ having observed x in model (7,8): standard frequentist and Bayesian bounds $x + 1.645$ (labelled U(-Inf, Inf) in figure); Bayesian bound given a uniform prior on $[0, \infty]$ (labelled U(0, Inf)); Model Selection bound.

As time permits

- **"Objective" Bayes priors depend on the model" means that they are derived from *measurement* model, not the physics model.**
  - **Jeffreys's Rule gives flat prior for a parameter measured with Gaussian/Normal measurement uncertainty**
  - **But thinking about charged of electron charge itself, Jeffreys concluded prior should be 1/q.**
- **This is why "objective" priors have a connection to coverage (Welch and Peers, etc.)**
- **Lots of nice properties for estimation, but for model checking, as Jim said, need subjective prior.**

Conclusions

- **My answer to the title question violates Hinchliffe's Rule (if it is true).**
- **The result of Eilam and Ofer may be the way to see that the √n behavior in the LEE-corrected p-value (!)**
- **Even within the frequentist paradigm, our test reporting is not complete enough.**
- **Likelihood-ratio test a la K&S may alleviate some issues with testing, as it did with confidence intervals.**
- **More than ever, I think we need to provide the consumer with the results of different ways of testing.**

Beyond that, Jim said it best

I shall resist the temptation of saying more, because model selection is a can of worms for both objectivists and subjectivists.

J. Berger , “The Case for Objective Bayesian Analysis,” Bayesian Analysis 1.

But he was incomplete

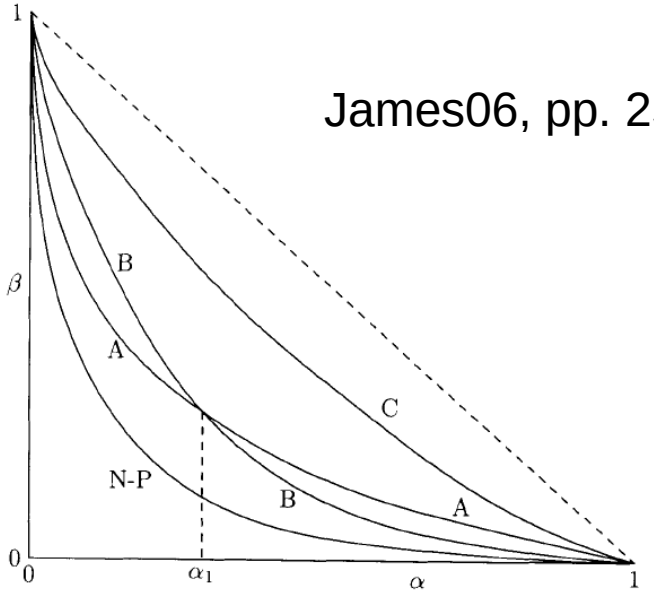
I shall resist the temptation of saying more, because model selection is a can of worms for both objectivists and subjectivists.

…and frequentists.

Backup

Classical Hypothesis Testing (cont.)

James06, pp. 258, 262



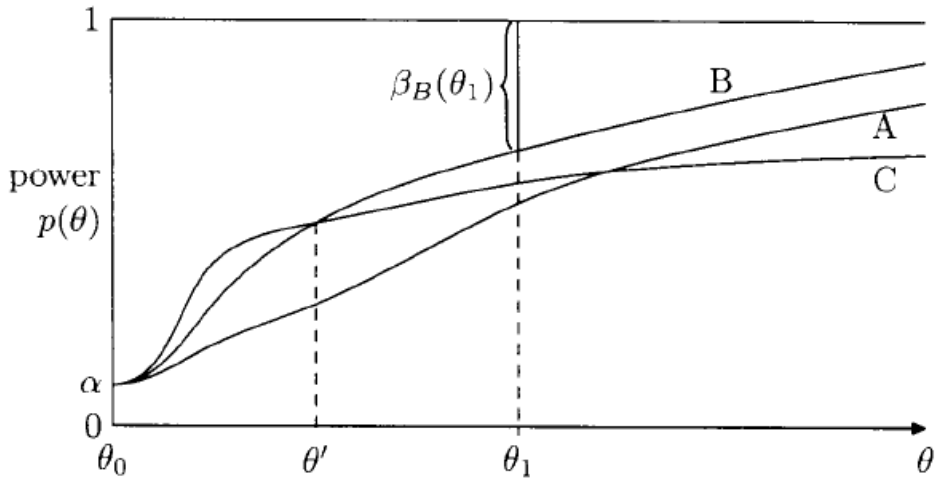


Fig. 10.3. Power functions of tests A, B, and C at significance level α. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.

Where to live on the β vs α curve is a *long* discussion. (Even longer when considered as number of events increases, so curve moves toward origin.) *Decision* on whether or not to declare discovery requires two more inputs:

1) **Prior belief in H_0 vs H_1**
2) **Cost of Type I error (false discovery claim) vs cost of Type II error (missed discovery)**

I argue in HEP that a one-size-fits-all criterion of α corresponding to 5σ is without foundation, but it is a common convention.