The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics

Bob Cousins

Univ. of California, Los Angeles

Workshop on Evidence, Discovery, Proof: Measuring the Higgs Particle

Univ. of South Carolina (via video)

April 25, 2014

Based on http://arxiv.org/abs/1310.3791 , which has any references not given here.

Also some slides adapted from my 2009 Hadron Collider Physics Summer School lectures, http://indico.cern.ch/getFile.py/access?contribId=1&resId=0&materialId=slides&confId=44587

The New York Times

Wednesday, July 4, 2012 Last Update: 6:54 AM ET

Discovery of New Particle Could Redefine Physical World

By DENNIS OVERBYE

21 minutes ago

The discovery by physicists at CERN's Large Hadron Collider, if confirmed to be the Higgs boson particle, could lead to a new understanding of how the universe began.

- The Lede Blog: What in the World Is a Higgs Boson? 4:16 AM ET

Fabrice Coffrini/Agence France-Presse — Getty Images

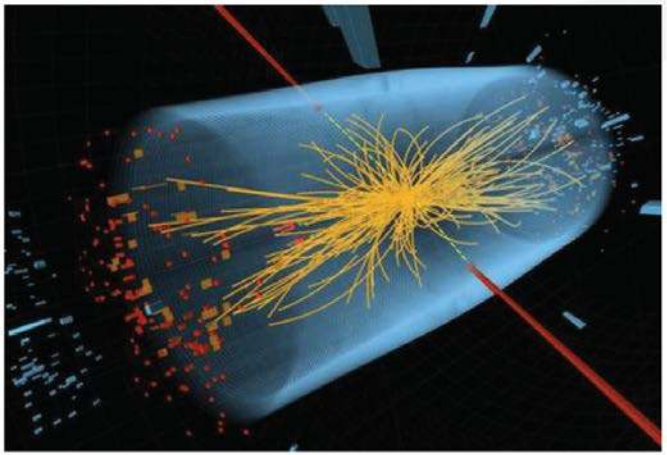
CERN officials held a press conference near Geneva on Wednesday.

LA NEWS DEL GIORNO | CANALE: Politica 10:21 – Roma, 4 lug 2012

T - +

Il Bosone di Higgs esiste, oggi l'annuncio del Cern a Ginevra

Tanti indizi per il "Santo Graal" della fisica quantistica teorizzato nel 1964. E' l'ultima particella ancora da scoprire

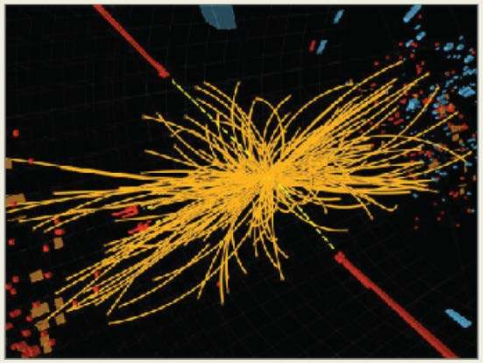


Roma, 4 lug. (TMNews) - L'enigma relativo all'esistenza del "bosone di Higgs", il "Santo Graal" della fisica delle particelle elementari, potrebbe essere oramai vicino alla soluzione: la conferenza stampa in programma oggi al Cern potrebbe dissipare gli ultimi dubbi.

LENTA·RU вторник, Прогресс

издание Rambler Media Group

04.07.2012, 12:13:02 Версия для печати | PDA

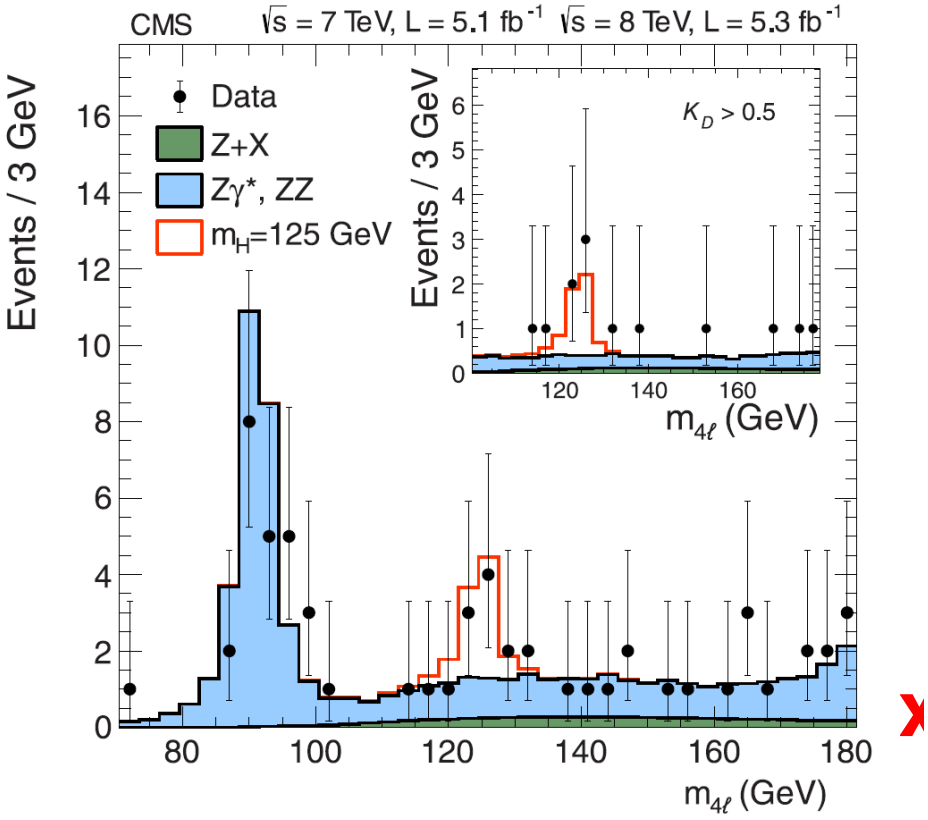


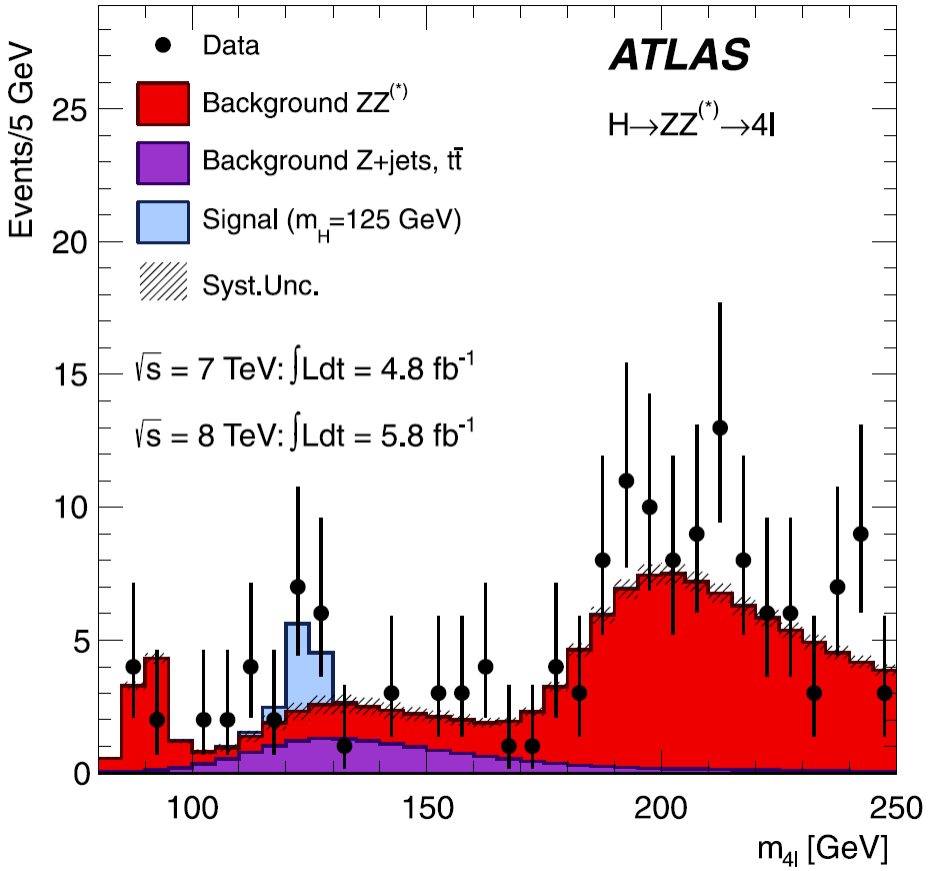
Изображение с сайта CERN

Физики обнаружили претендента на роль бозона Хиггса

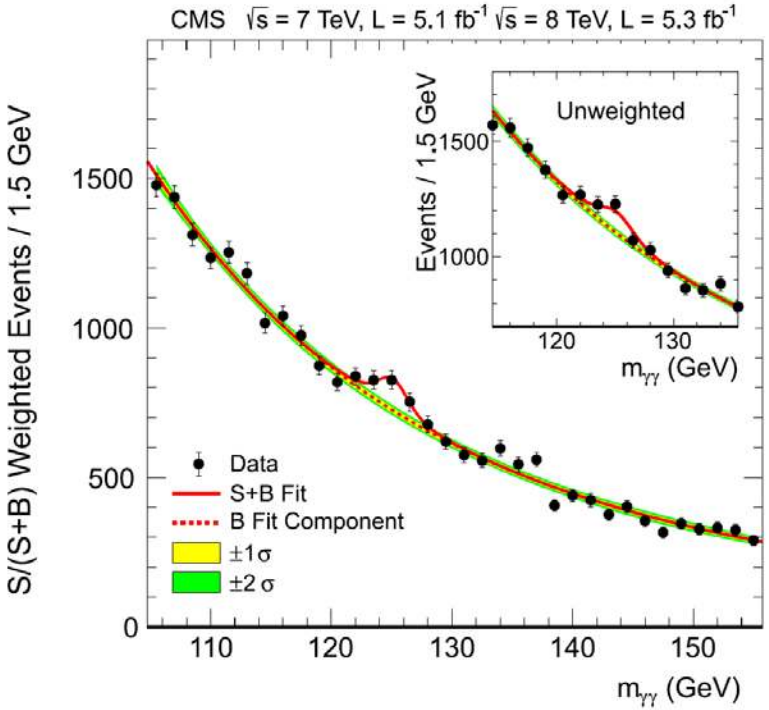
Physicists discover a candidate for the boson Higgs

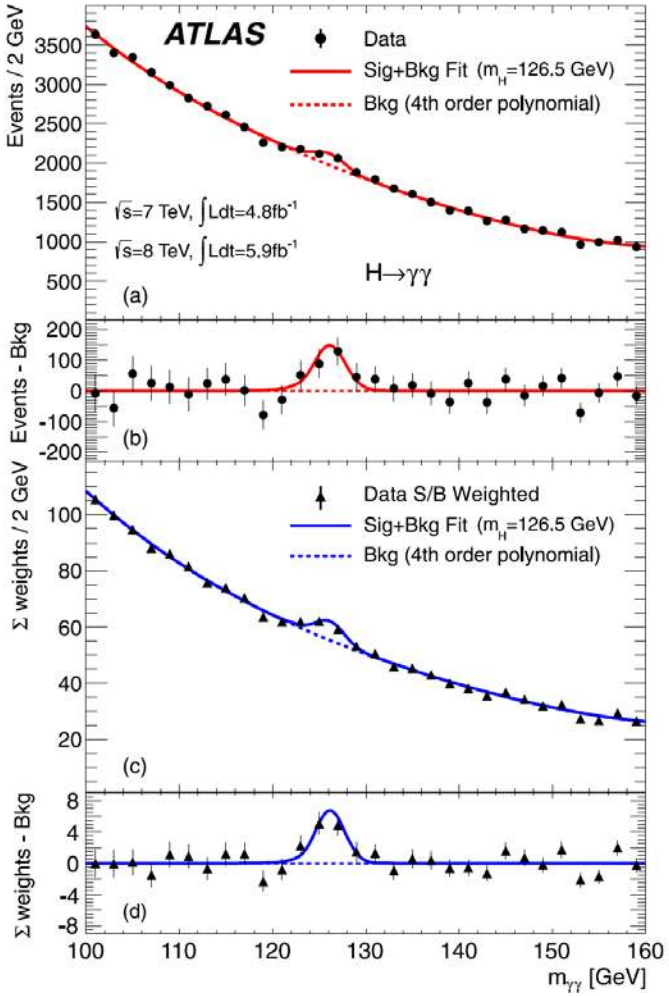
July 4: Each expt shows invariant mass spectrum for four-lepton events.





July 4: γγ invariant mass spectra
CMS, ATLAS both see excess near 125 GeV



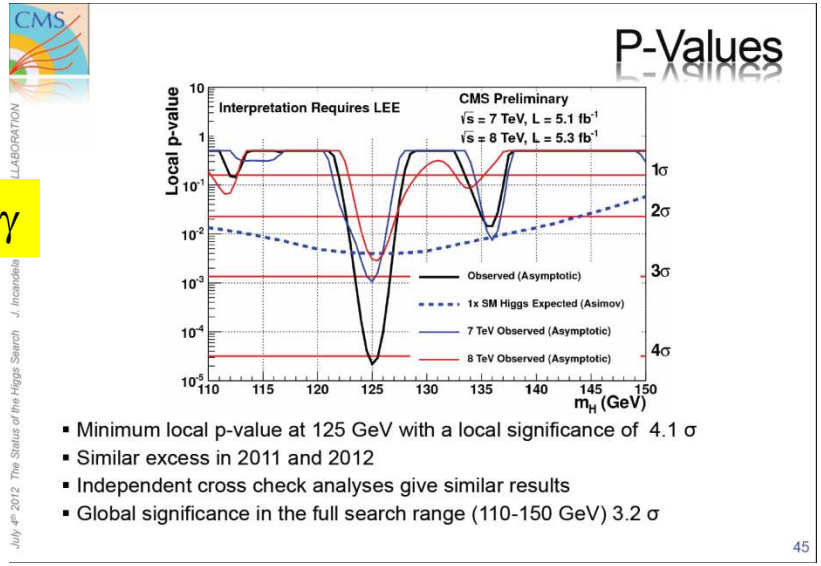


CMS and ATLAS each conclude "observation"of a new particle.

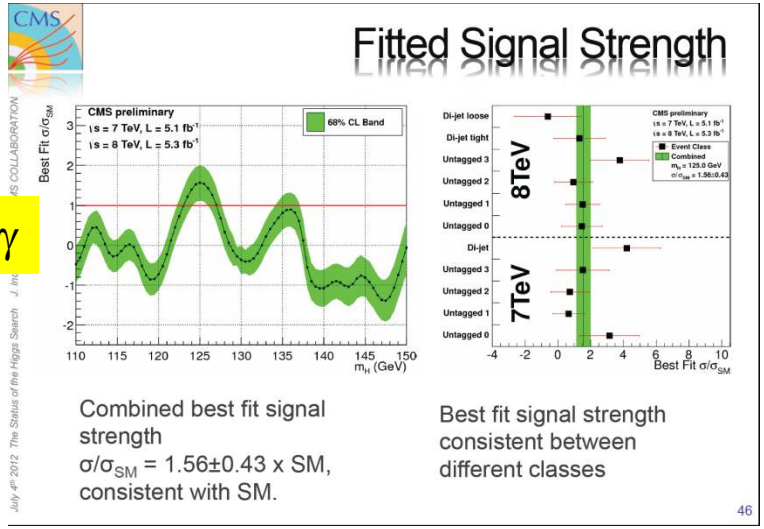
For many of us, it passed the "interocular traumatic test."

July 4 '12: CMS and ATLAS presented p-values *and* effect sizes. E.g., CMS:

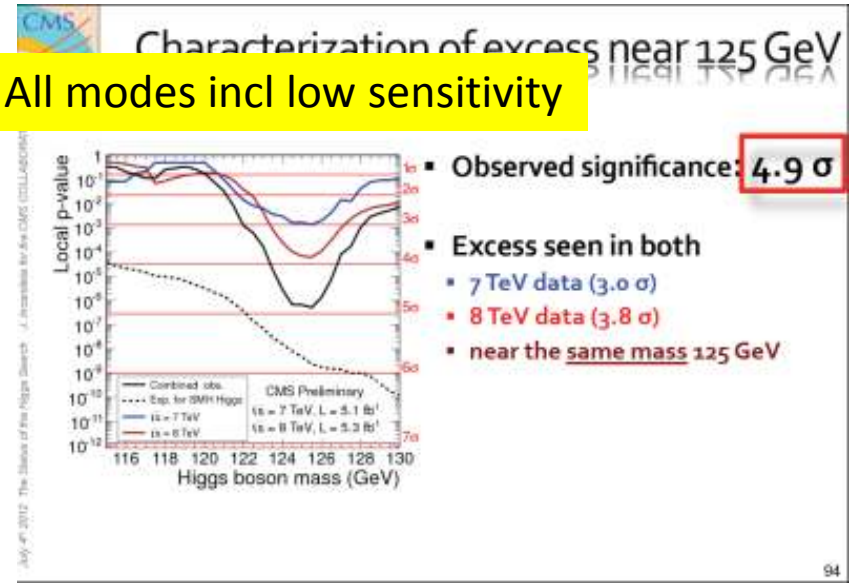
γγ

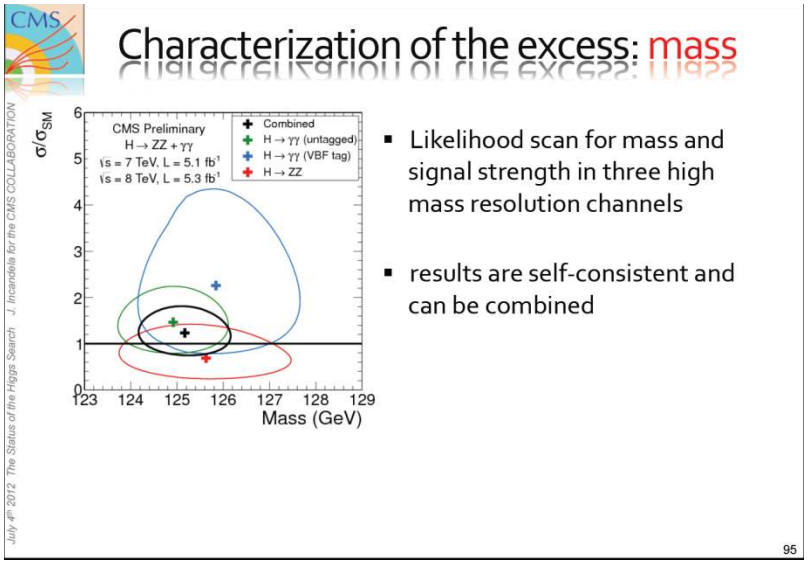


γγ



All modes incl low sensitivity





http://indico.cern.ch/event/197461/

Effect sizes: Cross section in units of SM cross section. Also mass.

Publications: Physics Letters B, vol. 216 (2012)

ATLAS: "The significance of an excess in the data is first quantified with the local p_0, the probability that the background can produce a fluctuation greater than or equal to the excess observed in data. The equivalent formulation in terms of number of standard deviations, Z_l, is referred to as the local significance.

CMS: "The probability for a background fluctuation to be at least as large as the observed maximum excess is termed the local *p*-value, and that for an excess *anywhere* in a specified mass range the global *p*-value"... Both the local and global *p*-values can be expressed as a corresponding number of standard deviations using the one-sided Gaussian tail convention."

As one involved, I think that both experiments did an excellent job in scientific analysis, both in the talks and the papers.

But p-values remain controversial...

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.

Google on "criticism of p-values", or "criticism of null hypothesis statistical testing (NHST)", for a plethora of articles.

What's it all about?

A STATISTICAL PARADOX (1957)

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.

Big issue: Bayesian methods use only the probability of obtaining the data actually observed (Likelihood Principle), while tail probabilities such as *p*-values also use probability of obtaining data *more unlikely* than that observed.

Oft-quoted Jeffreys: "*What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.*"

Raftery (1995}) "...there is no justification for the step in the derivation of the *p*-value where "probability density for data as extreme as that observed" is replaced with "probability for data as extreme, *or more extreme*".

...but this is not the only issue: Bayesians disagree with each other.

"Estimation" for parameter θ

Physicist: "measured value" or "best-fit value" of θ
Psychologist: "effect size" (in original units)
Statistician: "*point estimate*" of θ

Physicist: "uncertainty", "confidence interval", less often "credible interval" for θ
Statistician: "*interval estimate*" for θ (confidence interval, credible interval, ...)

In *estimation*, there is already a lot to say about frequentist vs Bayesians methods.
But for many problems (fixed dimensionality...), *dependence on the Bayesian prior goes away asymptotically (large n)*.

My talk is *not* about differences in estimation, but about something much more disturbing (at least to me).

Hypothesis testing (aka Model Selection)
E.g.: Is θ equal to some particular value θ_0 ?

Frequentist approach **is usually some mix of rival Fisher and Neyman-Pearson methods and terminology.**

Closely related to interval estimation (1-to-1 map).

Bayesian **hypothesis testing is** ***separate from estimation*****:**

It's so separate **that recommended "objective" priors for θ are different for estimation and testing (!).**

Bayesians compute probabilities that hypotheses are true ***a la Theory of Probability*** **by Harold Jeffreys (1939, 1961).**

This talk is about ***hypothesis testing*****, and Jeffreys-Lindley paradox contrasting frequentist and Bayesian results.**

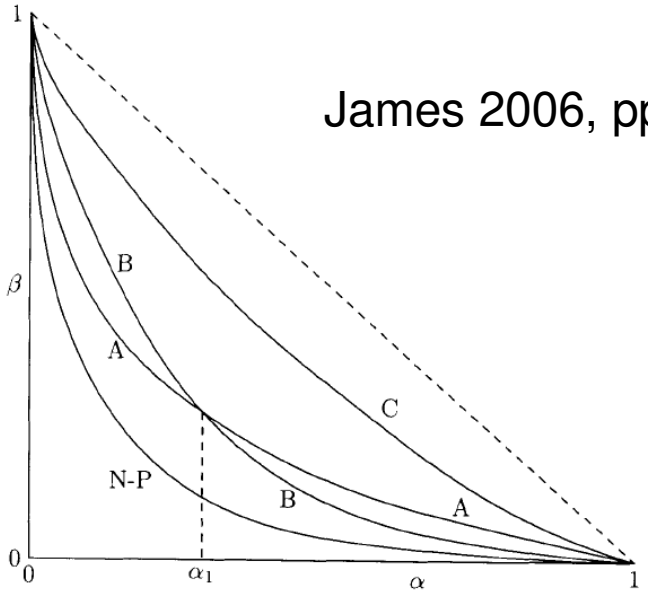
The dependence of the Bayesian result on the prior for θ ***remains*** **asymptotically.**

Classical Hypothesis Testing

- **In Neyman-Pearson hypothesis testing (James 2006), frame discussion in terms of null hypothesis H_0 and an alternative H_1. (E.g., H_0 = S.M., H_1 = CMSSM). Consider repeated tests on independent samples.**
  - α: **probability (under H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error)**
  - β: **probability (under H_1) of accepting H_0 when it is false, i.e., not claiming a discovery when there is one (Type II error)**
  - θ: **parameters in the hypotheses**
- **Common for H_0 to be *nested* in H_1 , i.e. H_0 corresponds to particular parameter values θ_0 (e.g. zero or ∞) in H_1.**
- **Competing analysis methods can be compared by looking at graphs of β vs α at various θ, and at graphs of *power* $1-\beta$ vs θ at various α (power function).**

Classical Hypothesis Testing (cont.)

James 2006, pp. 258, 262



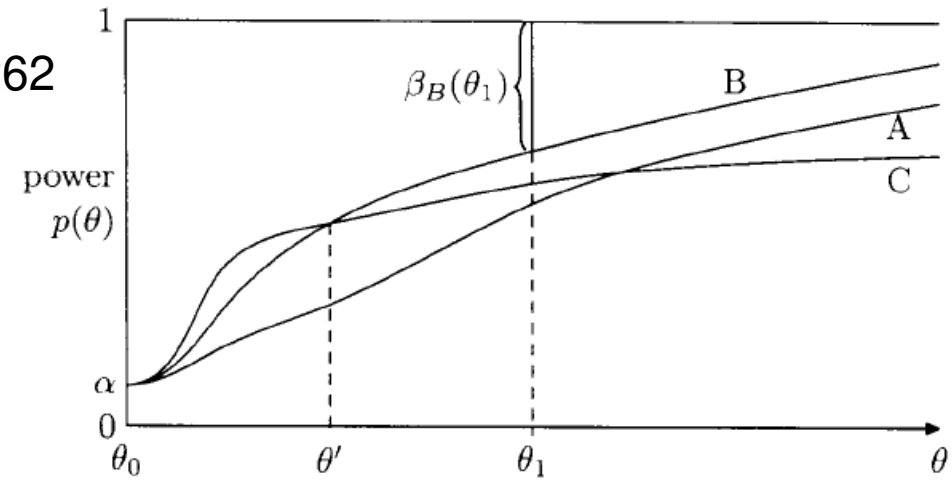


Fig. 10.3. Power functions of tests A, B, and C at significance level α. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.

Where to live on the β vs α curve is a *long* discussion. (Even longer when considered as *n* events increases, so curve moves toward origin.)
***Decision* on whether 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)**

A one-size-fits-all criterion of α corresponding to some fixed threshhold like 5σ is without foundation in the frequentist stat literature.

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. "

Lehmann (2005) "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.]

Lehmann (2005) "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."

Choice of threshold α for rejecting H_0 (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."

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

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."

Classical Hypothesis Testing: p-values and Z-values

In N-P theory, α is *specified in advance*. Suppose after obtaining data, you notice that with α=0.05 previously specified, you reject H_0, but with α=0.01 previously specified, you accept H_0. In fact, you determine that with the data set in hand, H_0 would be rejected for $\alpha \geq 0.023$. This interesting critical value has a name:

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

Typically that critical value of α was *not* specified in advance, so p-values do *not* correspond to Type I error rates of the experiments which report them.

Interpretation of p-values is a long, contentious story – beware! Fisher had a very different philosophical view of them than N-P. In HEP, converted to Z-value, equivalent number of Gaussian σ. E.g., for one-tailed test, p=1.35E-3 is Z=3; p=2.87E-7 is Z=5.

Aside on Confidence Intervals

"Confidence intervals", and this phrase to describe them, were invented by Jerzy Neyman in 1934-37.

Neyman described a way to construct a set of confidence intervals. It is really ingenious – perhaps a bit *too* ingenious given how often confidence intervals are misinterpreted.

In particular, the confidence level does *not* tell you "how confident you are that the unknown true value is in the interval" – only a *subjective* Bayesian credible interval has that property!

Confidence Intervals and Coverage

Recall: in math, one defines a *vector space* as a *set* with certain properties, and then
The definition of a *vector* is "an element of a vector space".
(A vector is not defined in isolation.)

Similarly, whether constructed in practice by Neyman's construction or some other technique,

The definition of a *confidence interval* is "an element of a confidence set*",

where the *confidence set* is a set of intervals defined to have the property of frequentist *coverage* under repeated sampling.

*** Also called *family* of intervals, or (when graphed) confidence *band* or confidence *belt.* (Set is also used by some authors to mean one interval.)**

Confidence Intervals and Coverage (cont.)

Let the unknown true value of θ be θ_t .

In repeated experiments, the confidence intervals obtained will have different endpoints $[\theta_1, \theta_2]$, since the endpoints are functions of the randomly sampled *x*.

The fraction C.L. = 1 - α of intervals obtained by Neyman's construction will contain ("cover") the fixed but unknown θ_t :

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

The random variables in this equation are θ_1 and θ_2, and not θ_t.

Coverage is a property of the *set of intervals*, not of an individual interval.

It is not necessary that all experiments have the same θ_t , or even measure the same quantity.

Classical Hypothesis Testing: Duality

"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" – Stuart 1999, p. 175.

Classic example of *Bayesian* hypothesis testing
Test H_0: $\theta=\theta_0$ vs H_1: $\theta\neq\theta_0$

Concept: Calculate posterior probability of H_0.

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

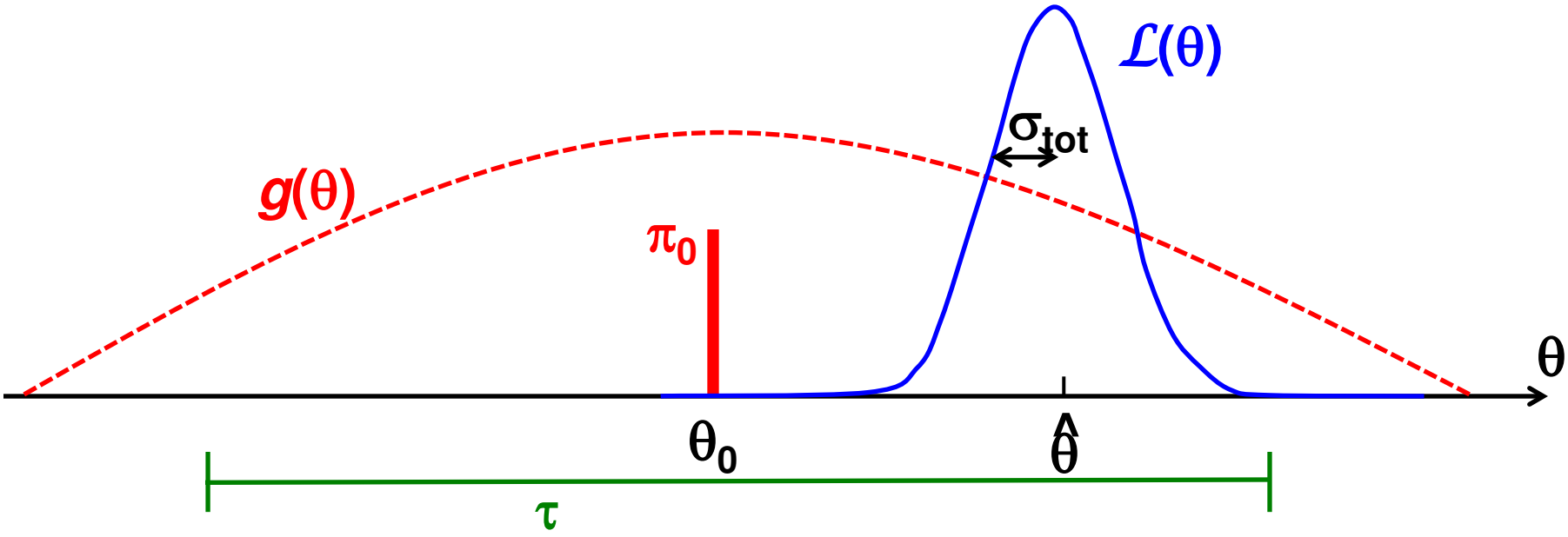
Conditional on H_1 being true, we also need prior probability *density* for θ: $g(\theta)$. [Lebesgue measure]

[Conceptual issue: π_0 is like a bit of Dirac δ-ftn at $\theta=\theta_0$. Called "probability mass" or "counting measure".]

So suppose X having density $f(x|\theta)$ is observed.

Consider case: $f(x|\theta)$ is normal with mean θ, rms σ, and sample is $\{x_1, x_2, \ldots, x_n\}$.

Max Lik. Est. for θ is $\hat{\theta} = \overline{x}$ **having rms** $\sigma_{\rm tot} \equiv \sigma/\sqrt{n}$



$$\hat{\theta} = \overline{x}$$

$$\sigma_{\rm tot} \equiv \sigma / \sqrt{n}$$

Posterior Probs from Bayes's rule: prior × $\mathcal{L}$

$$P(H_0|\hat{\theta}) = \frac{1}{A}\,\pi_0\,\mathcal{L}(\theta_0) = \frac{1}{A}\,\pi_0\,\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}}\exp\left\{-(\hat{\theta}-\theta_0)^2/2\sigma^2_{\rm tot}\right\}$$

$$P(H_1|\hat{\theta}) = \frac{1}{A}\,\pi_1\int g(\theta)\mathcal{L}(\theta)d\theta = \frac{1}{A}\,\pi_1\int g(\theta)\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}}\exp\left\{-(\hat{\theta}-\theta)^2/2\sigma^2_{\rm tot}\right\}d\theta$$

A = normalization constant so that sum of above two is unity.

Posterior Probs from Bayes's rule: prior × $\mathcal{L}$

$$P(H_0|\hat{\theta}) = \frac{1}{A}\,\pi_0\,\mathcal{L}(\theta_0) = \frac{1}{A}\,\pi_0\,\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}}\exp\left\{-(\hat{\theta}-\theta_0)^2/2\sigma_{\rm tot}^2\right\}$$

$$P(H_1|\hat{\theta}) = \frac{1}{A}\,\pi_1\int g(\theta)\mathcal{L}(\theta)d\theta = \frac{1}{A}\,\pi_1\int g(\theta)\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}}\exp\left\{-(\hat{\theta}-\theta)^2/2\sigma_{\rm tot}^2\right\}d\theta$$

A = normalization constant so that sum of above two is unity.

Let τ be scale that characterizes range of θ for which prior *g* is relatively large.

Consider case $\sigma_{\rm tot} << \tau$.

$$P(H_1|\hat{\theta}) \approx \frac{1}{A}\,\pi_1\,g(\hat{\theta})\ .$$

Posterior Probs from Bayes's rule: prior × $\mathcal{L}$

$$P(H_0|\hat{\theta}) = \frac{1}{A}\,\pi_0\,\mathcal{L}(\theta_0) = \frac{1}{A}\,\pi_0\,\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}}\exp\left\{-(\hat{\theta}-\theta_0)^2/2\sigma_{\rm tot}^2\right\}$$

$$P(H_1|\hat{\theta}) = \frac{1}{A}\,\pi_1\int g(\theta)\mathcal{L}(\theta)d\theta = \frac{1}{A}\,\pi_1\int g(\theta)\frac{1}{\sqrt{2\pi}\sigma_{\rm tot}}\exp\left\{-(\hat{\theta}-\theta)^2/2\sigma_{\rm tot}^2\right\}d\theta$$

A = normalization constant so that sum of above two is unity.

Let τ be scale that characterizes range of θ for which prior *g* is relatively large.

Consider case $\sigma_{tot} << \tau$.

$P(H_1|\hat{\theta}) \approx \frac{1}{A}\,\pi_1\,g(\hat{\theta})$. *Bayes Factor* = (Posterior odds *favoring* H_0)/Prior odds :

$$\mathrm{BF} = \frac{P(H_0|\hat{\theta})}{P(H_1|\hat{\theta})}\Big/\frac{\pi_0}{\pi_1} = \frac{1}{\sqrt{2\pi}\sigma_{\rm tot}g(\hat{\theta})}\exp\left\{-(\hat{\theta}-\theta_0)^2/2\sigma_{\rm tot}^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_{\rm tot}g(\hat{\theta})}\exp(-z^2/2),$$

where $z = (\hat{\theta}-\theta_0)/\sigma_{\rm tot} = \sqrt{n}(\hat{\theta}-\theta_0)/\sigma$ is usual discrepancy in units of σ .

$$\mathrm{BF} = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{tot}} g(\hat{\theta})} \exp(-z^2/2)$$

Key point: g(θ) is normalized, so $g(\hat{\theta}) \propto 1/\tau$. Hence very generally,

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

Proportionality constant depends on the form of g, specifically on $g(\hat{\theta})$.

Meanwhile, classical likelihood ratio comparing null value and the MLE value:

$$\begin{aligned}\lambda &= \mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta}) \\ &= \exp\left\{(\hat{\theta}-\theta_0)^2/2\sigma_{\mathrm{tot}}^2\right\} \Big/ \exp\left\{(\hat{\theta}-\hat{\theta})^2/2\sigma_{\mathrm{tot}}^2\right\} \\ &= \exp(-z^2/2)\end{aligned}$$

So also very generally,

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

The Jeffreys-Lindley Paradox

Recall *z* is the "number of sigma" of the effect.

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

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

The factor σ_{tot}/τ is called the Ockham (Occam) factor, penalizing H_1 in the likelihood ratio for having the degree of freedom to choose "best-fit" θ, whereas H_0 predicts it.

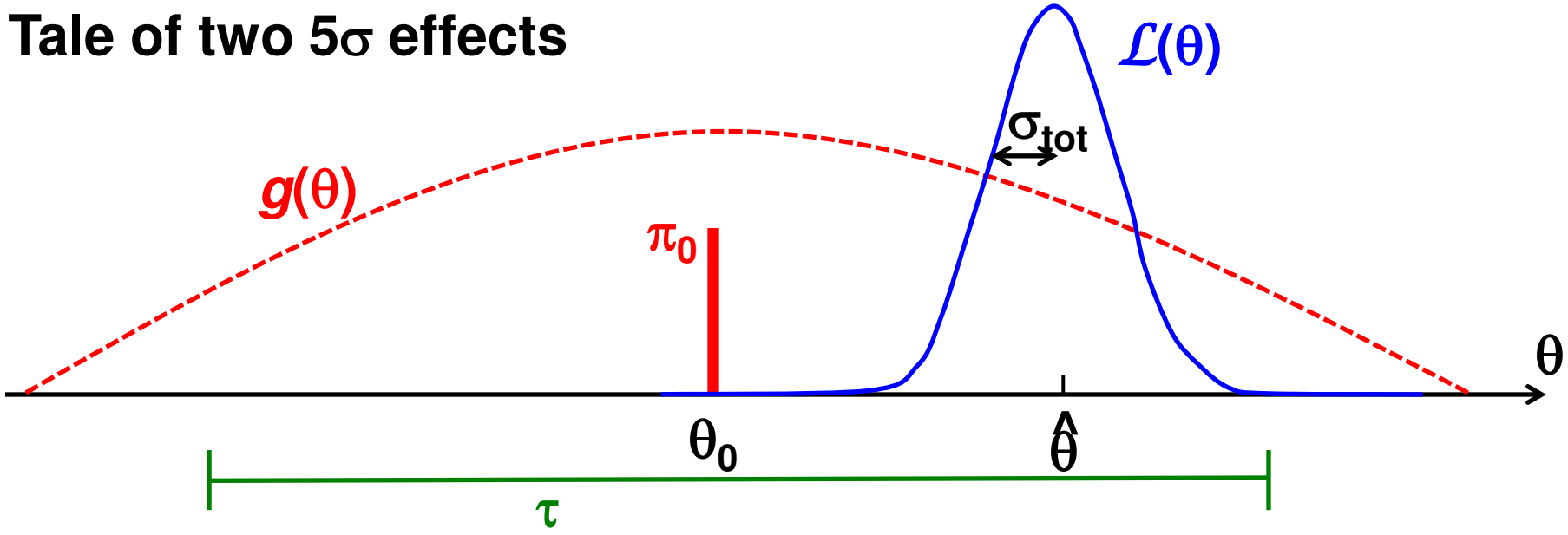
For experiments having the *same z* (say 5σ effect), the Bayes Factors can be dramatically different.

For 5σ results with small enough σ_{tot}, BF can strongly favor H_0 !

(And meanwhile the likelihood ratio also favors the alternative H_1.)

Recalling $\sigma_{\mathrm{tot}} \equiv \sigma/\sqrt{n}$, the JL paradox is often viewed as an issue of sample size *n*: the interpretation of "5σ" should depend on *n*, according to the BF. From this point of view, the paradox arises for large n.

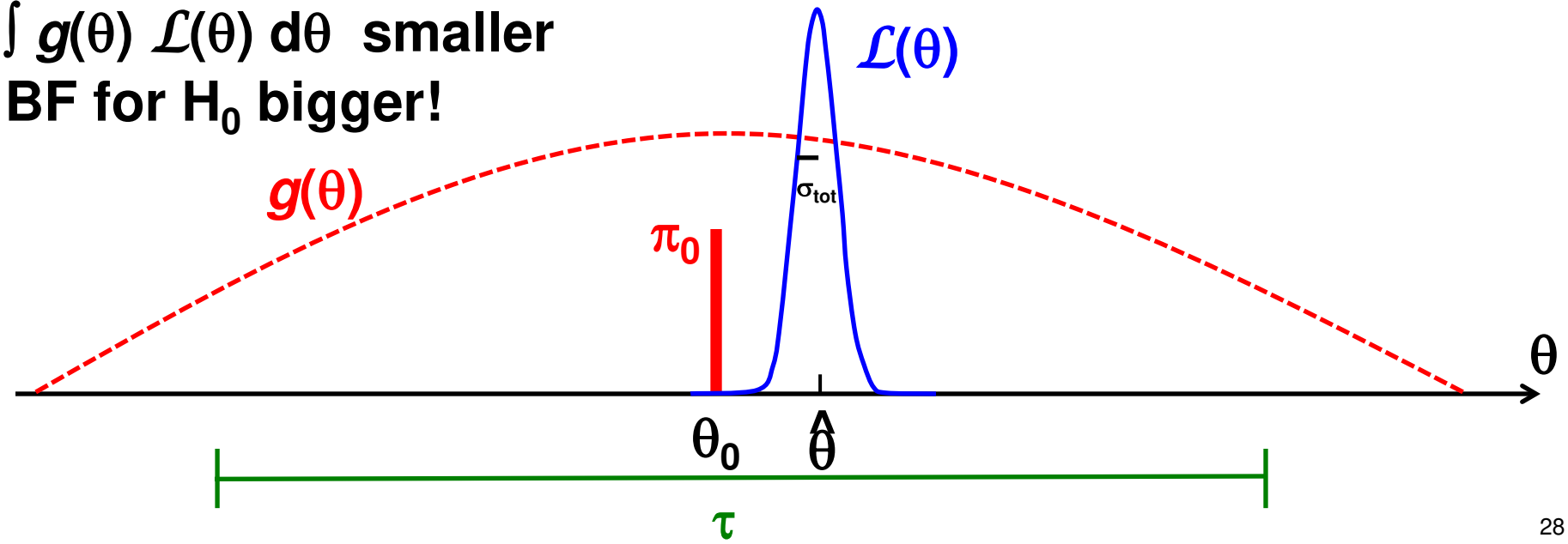
Tale of two 5σ effects



$\pi_0\,\mathcal{L}(\theta_0)$ unchanged

$\int g(\theta)\,\mathcal{L}(\theta)\,d\theta$ smaller

BF for H_0 bigger!



Jeffreys (1939, 1961) curiously downplayed the discrepancy:

"In spite of the difference in principle between my tests and those based on the [p-values],
and the omission of the latter to give the increase in 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 a value out of a long range and should allow for selection,
it appears that there is not much difference in the practical recommendations."

Lindley (1957) highlighted how bad it could be, difference in scaling in sample-size *n*.
Bartlett (1957) noted Lindley's oversight of the role of τ: makes result "much more arbitrary" (remains asymptotically).

Every assumption in JL has been scrutinized!

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

1) The δ-function for H_0 makes no sense.
2) Very small σ_{tot} means we should not be trusting the model $f(x|\theta)$, so what's the fuss about? No good scientist believes their null hypothesis!(And "All models are wrong!")
3) (From a Bayesian) Jeffrey's method should be replaced by a method based on decision theory that gives different results.
4) Frequentist tail probabilities are obviously wrong way to make inference; p-values dramatically overstate evidence against H_0.
5) Paradox goes away if the prior depends on *n* in compensating way.
6) Etc., etc.

A sampling of quotes in remaining slides

The δ-function for H_0

Lindley (2009) lauds the "triumph" of Jeffreys "...putting a concentration of prior probability on the null---no ignorance here---and evaluating the posterior probability using what we now call Bayes factors."

Bernardo (2009) "Jeffreys intends to obtain a posterior probability for a precise null hypothesis, and, to do this, he is forced to use a mixed prior which puts a lump of probability $p=Pr(H_0)$ on the null...This has a very upsetting consequence, usually known as Lindley's paradox...I find it difficult to accept a procedure which is known to produce the wrong answer under specific, but not controllable, circumstances."

Zellner 2009) 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'... ."

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

The δ-function for H_0 (cont.)

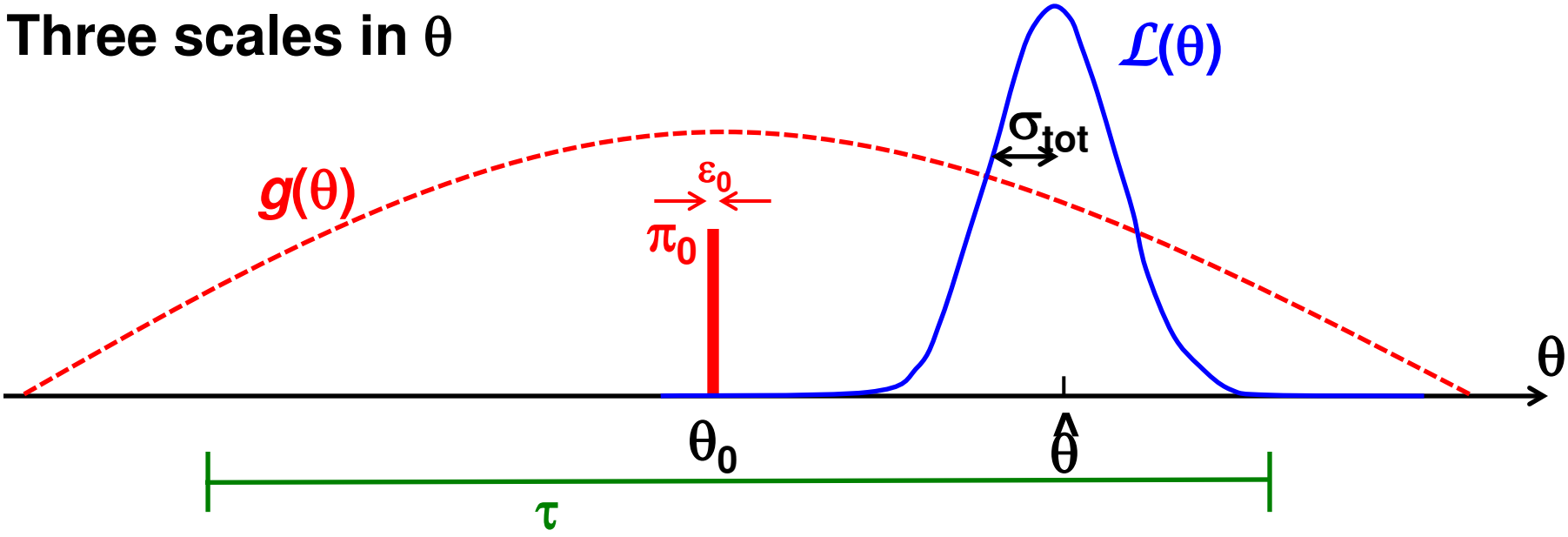
Raftery(1995) "social scientists are prepared to act *as if* they had prior distributions with point masses at zero...social scientists often entertain the possibility that an effect is *small* ".

C. Robert and J. Rousseau (2011), "*Down with point masses!* " ... "What matters in pointwise hypothesis testing is not whether or not $\theta=\theta_0$ holds but what the consequences of a wrong decision are."

But the key point was made in 1963 by Edwards, Lindman, Savage: "Bayesians...must remember that the null hypothesis is a hazily defined small region rather than a point."

Thus a key issue is the size of this "hazily define region"; I call it ε_0 .

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

JL paradox arises if: $\varepsilon_0 << \sigma_{tot} << \tau$

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

Is there an "objective" (default) prior $g(\theta)$?

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

A common idea, dating to Jeffreys: use a prior with the same amount of information as a dataset with n=1. Also generalizations.
Kass and Wasserman (1995) refer to this as "unit information prior".

Raftery(1995) points out the problem: the "important ambiguity...the definition of [n], the sample size."

It's a frontier research topic: Berger and Pericchi (2001) "this is the first general approach to the construction of conventional priors in nested models."

Bayarri et al (including Berger) (2012) "...a new model selection objective prior with a number of compelling properties."

Does anyone believe their null hypotheses?

Edwards (1963) "...in typical applications, one of the hypotheses---the null hypothesis---is known by all concerned to be false from the outset"

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!"

Casella and R. Berger (1987c) "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!"

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."

This is all rather amusing to a high energy physicist. See paper for discussion.

The Bayesian alternative of José Bernardo

Bernardo refers unapprovingly to point null hypotheses in an "objective" framework, and to the use begun by Jeffreys of two "*radically different*" types of priors for estimation and for hypothesis testing.

The JL paradox "clearly poses a very serious problem to Bayes factors, in that, under certain conditions, they may lead to misleading answers. Whether you call this a paradox or a disagreement, the fact that the Bayes factor for the null may be arbitrarily large for sufficiently large *n*, *however relatively unlikely the data may be under* H_0 is, to say the least, deeply disturbing..."

He has an alternative based on Bayesian decision theory that I hope is investigated further.

Conclusion

Lots of food for thought. The conditions for the JL paradox are common in HEP, so it should not be ignored.

In my paper I discuss what it means for a high energy physicist to "believe" our models. (The core physics model is a mathematical limit, or an effective field theory, for the more complete one.)

In searches for physics beyond the SM, belief in null (SM) is high. In search for first observation of physics *within* the SM (as in Higgs search) our null hypothesis is rather artificial (all of SM except some piece).

Meanwhile we continue with p-values, nearly always accompanied by confidence intervals for the various effect sizes (as was the case on July 4, 2012). I opine in my paper that the combination of information given to consumers in papers such as the Higgs discovery even allows for an informal estimate of a Bayes factor.

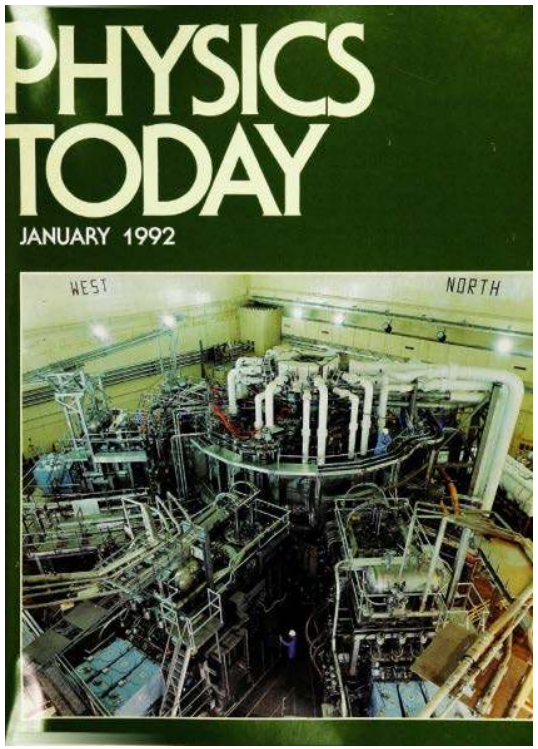
But clearly more work (both foundational and practical) on the JL paradox is called for.

BACKUP

REFERENCE FRAME

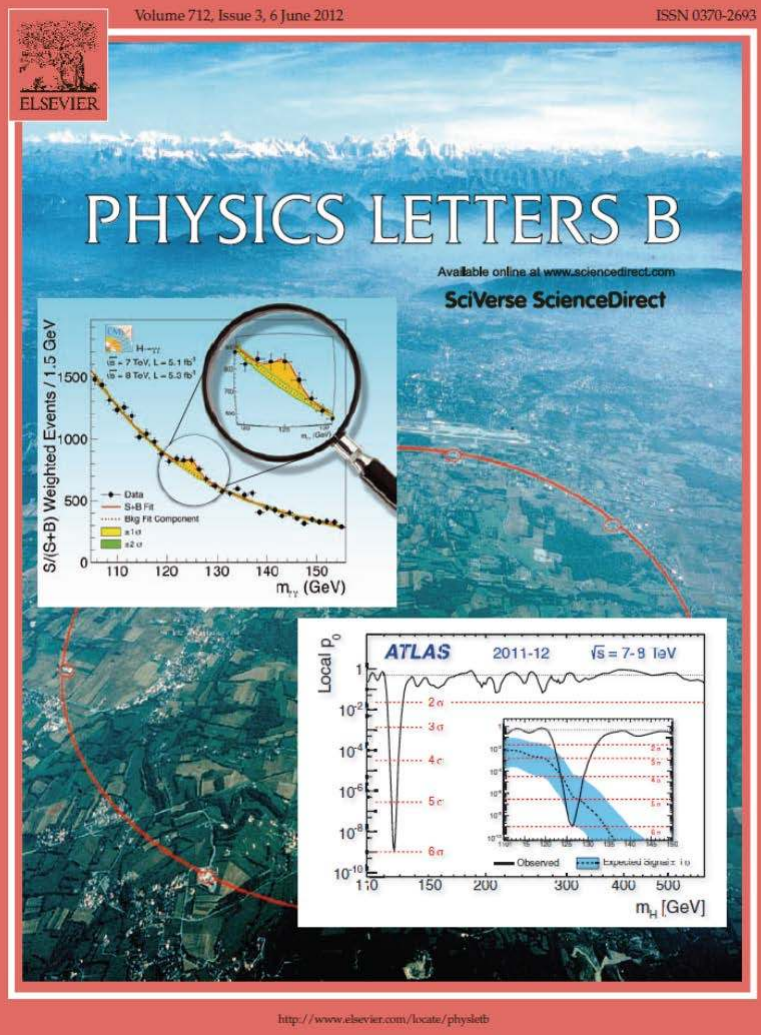
THE REVEREND THOMAS BAYES, NEEDLES IN HAYSTACKS, AND THE FIFTH FORCE

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



June: Peer-reviewed papers:

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC.

"…compatible with the production and decay of the Standard Model Higgs boson."

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

"..consistent, within uncertainties, with expectations for the standard model Higgs boson."

Definition of "Probability"

- **Abstract mathematical probability P can be defined in terms of sets and axioms that P obeys. If the axioms are true for P, then P obeys Bayes' Theorem (see next slide)**

  P(B|A) = P(A|B) P(B) / P(A).

- **Two established* incarnations of P are:**

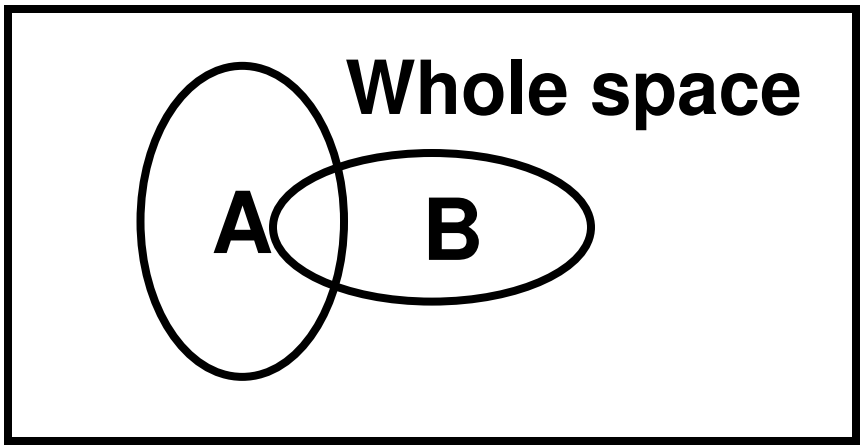
  ***1) Frequentist P*: limiting frequency in ensemble of imagined repeated samples (as usually taught in Q.M.).**
  P(constant of nature) and P(SUSY is true) do not exist (in a useful way) for this definition of P (at least in one universe).

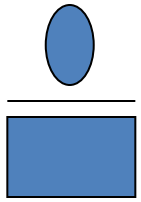
  ***2) (Subjective) Bayesian P*: subjective (personalistic) *degree of belief.* (de Finetti, Savage)**
  P(constant of nature) and P(SUSY is true) exist for You.
  Shown to be basis for coherent personal decision-making.

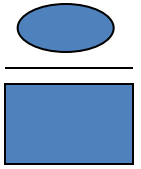
- ***It is important to be able to work with either definition of P, and to know which one you are using!***

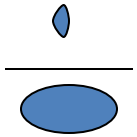
***Of course they are still argued about, but to less practical effect, I think.**

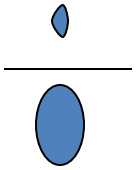
P, Conditional P, and Derivation of Bayes' Theorem in Pictures

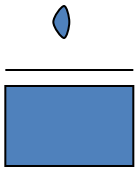


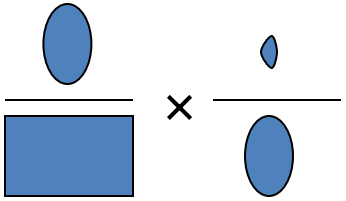
$P(A) =$ 

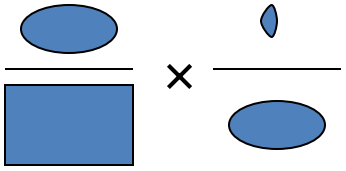
$P(B) =$ 

$P(A|B) =$ 

$P(B|A) =$ 

$P(A \cap B) =$ 

$P(A) \times P(B|A) =$  $= P(A \cap B)$

$P(B) \times P(A|B) =$  $= P(A \cap B)$

$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$

What is the "Whole Space"?

- **Note that for probabilities to be well-defined, the "whole space" needs to be defined, which can be hard for both frequentists and Bayesians!.**
- **Thus the "whole space" itself is more properly thought of as a conditional space, conditional on the assumptions going into the model (Poisson process, whether or not total number of events was fixed, etc.).**
- **Furthermore, it is widely accepted that restricting the "whole space" to a relevant subspace can sometimes improve the quality of statistical inference – see the discussion of "Conditioning" in later slides.**
- **I will not clutter the notation with explicit mention of the assumptions defining the "whole space", but some prefer to do so – in any case, please keep them in mind.**

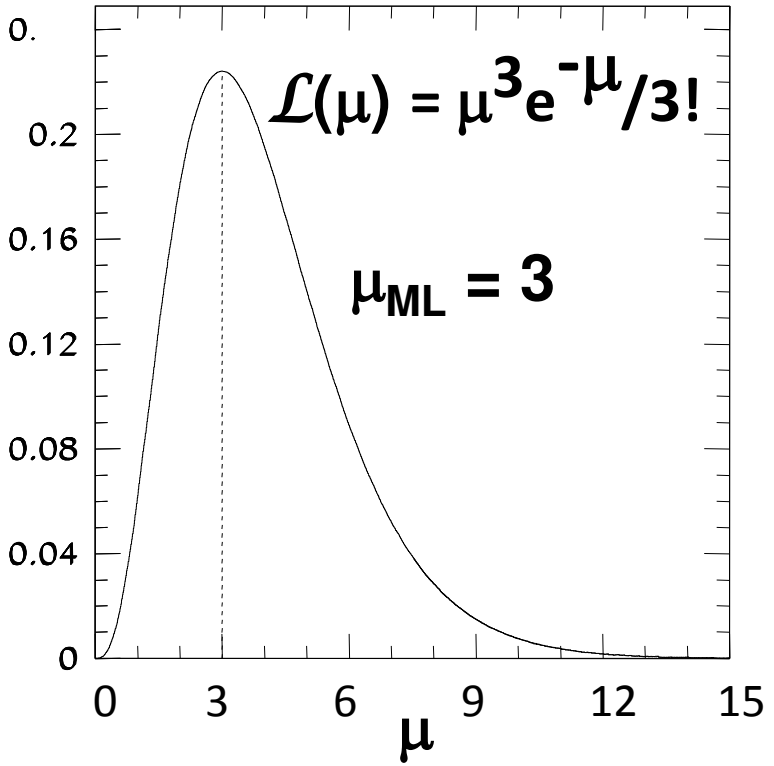
Probability, Probability Density, and Likelihood

- Poisson *probability* $P(n|\mu) = \mu^n \exp(-\mu)/n!$
- Gaussian *probability density function* (pdf) $p(x|\mu,\sigma)$: $p(x|\mu,\sigma)dx$ is differential of probability dP.
- In Poisson case, suppose n=3 is observed. Substituting n=3 into $P(n|\mu)$ yields the *likelihood function* $\mathcal{L}(\mu) = \mu^3 \exp(-\mu)/3!$

It is tempting to consider area under $\mathcal{L}$, but $\mathcal{L}(\mu)$ is *not* a probability density in μ:

Area under $\mathcal{L}$ is meaningless.

Likelihood *Ratios* $\mathcal{L}(\mu_1)/\mathcal{L}(\mu_2)$ are useful and frequently used.



Change of variable x, change of parameter θ

- **For pdf $p(x|\theta)$ and (1-to-1) change of variable from x to y(x):**

  $$p(y(x)|\theta) = p(x|\theta) / |dy/dx|.$$

  Jacobian modifies probability *density*, guaranties that

  $$P(y(x_1) < y < y(x_2)) = P(x_1 < x < x_2),$$ **i.e., that**

  ***Probabilities* are invariant under change of variable x.**

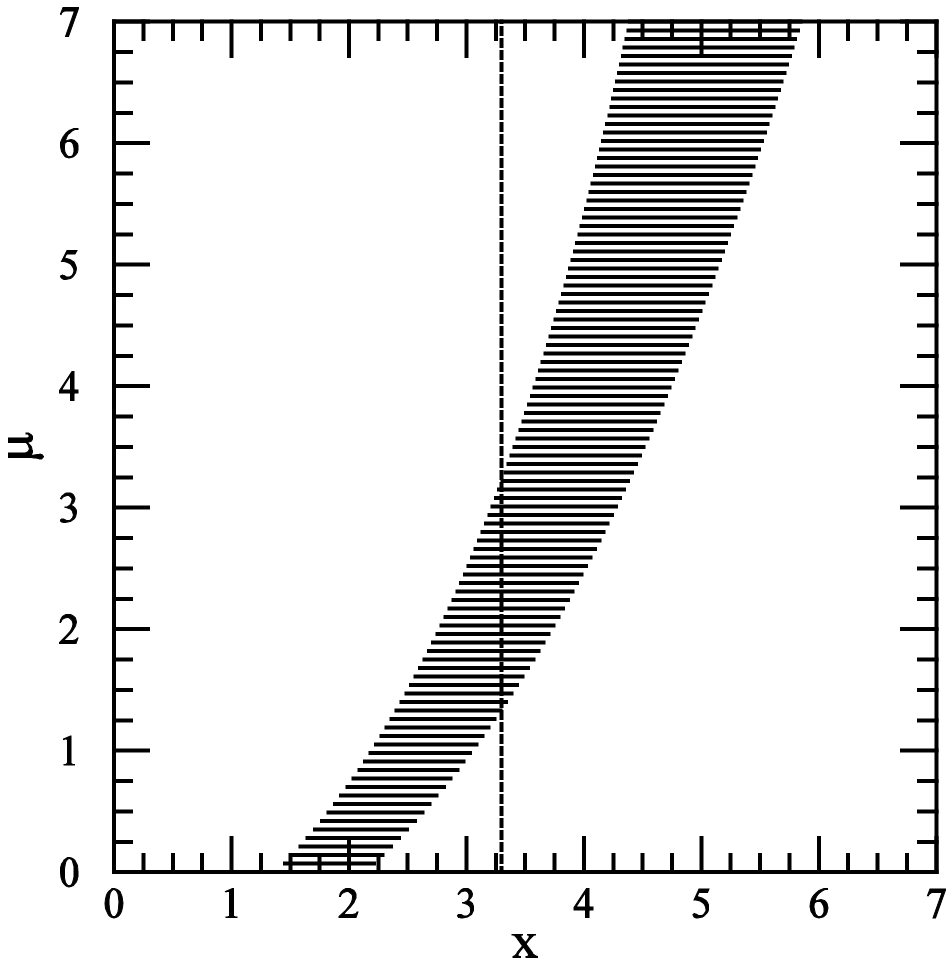
  - **Mode of probability *density* is *not* invariant (so, e.g., criterion of maximum probability density is ill-defined).**
  - **Likelihood *ratio* is invariant under change of variable x. (Jacobian in denominator cancels that in numerator).**
- **For likelihood $\mathcal{L}(\theta)$ and reparametrization from θ to u(θ):**

  $$\mathcal{L}(\theta) = \mathcal{L}(u(\theta))$$ **(!).**

  - **Likelihood $\mathcal{L}(\theta)$ is invariant under reparametrization of parameter θ (reinforcing fact that $\mathcal{L}$ is *not* a pdf in θ).**

Neyman's Confidence Interval construction

Given $p(x|\mu)$ from a model:
For each value of μ, one draws a horizontal *acceptance interval* $[x_1,x_2]$ such that $p(x \in [x_1,x_2] \mid \mu) = 1 - \alpha$. (Ordering principle needed to well-define.)
Upon performing an experiment to measure x and obtaining the value x_0, one draws the vertical line through x_0.
The vertical *confidence interval* $[\mu_1, \mu_2]$ with Confidence Level C.L. = $1 - \alpha$ is the union of all values of μ for which the corresponding *acceptance interval* is intercepted by the vertical line.

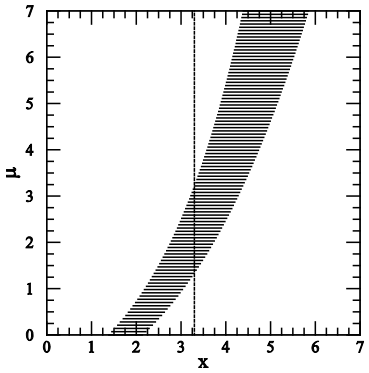


Note: x and μ need not have the same range, units, or (in generalization to higher dimensions) dimensionaliity!

Figure from G. Feldman, R Cousins, Phys Rev D57 3873 (1998)

Aside on the note regarding x and μ

Note: x and μ need not have the same range, units, or (in generalization to higher dimensions) dimensionaliity!



I actually think it is *much easier* to avoid confusion when x and μ are qualitatively different.
Louis Lyons give the example where x is the flux of solar neutrinos and μ is the temperature at the center of the sun;
I like examples where x and μ have different dimensions.

After studying examples such as those, one learns that in the Gaussian "measurement" of a mass μ which obtains the value x, it is crucial to distinguish between the data x, which can be negative, and the mass μ, for which negative values do not exist in the model. (I.e., for which $P(x/\mu)$ does not exist)

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.

Classical Hypothesis Testing: Neyman-Pearson Lemma

IX. *On the Problem of the most Efficient Tests of Statistical Hypotheses.*

By J. NEYMAN, *Nencki Institute, Soc. Sci. Lit. Varsoviensis, and Lecturer at the Central College of Agriculture, Warsaw, and* E. S. PEARSON, *Department of Applied Statistics, University College, London.*

(*Communicated by* K. PEARSON, *F.R.S.*)

(Received August 31, 1932.—Read November 10, 1932.)

Phil. Transactions of the Royal Society of London. Vol. 231, (1933), pp. 289-337

If Type I error probability α is specified in a test of *simple* hypothesis H_0 against *simple* hypothesis H_1, then the Type II error probability β is minimized by using as the test statistic the *likelihood ratio*
$\lambda = \mathcal{L}(x|\ H_0) / \mathcal{L}(x|\ H_1)$, and rejecting H_0 if $\lambda \leq k_\alpha$

The "lemma" applies only to a very special case: no nuisance parameters, not even undetermined parameters of interest!
But it has inspired many generalizations, and likelihood ratios are a oft-used component of both frequentist and Bayesian methods.

Conceptual proof in Second lecture of Kyle Cranmer, February 2009
http://indico.cern.ch/categoryDisplay.py?categId=72 . See also Stuart 1999, p. 176