Reflections on Statistical Data Analysis in Neutrino Experiments since NOMAD and F-C

Bob Cousins

Univ. of California, Los Angeles (UCLA)

PHYSTAT-ν Workshop on Statistical Issues in Experimental Neutrino Physics, Fermilab

September 21, 2016

Notes added after talk: mistake fixed on slide 18.

Work partially supported by U.S. Dept. of Energy Award DE-SC000993

Neutrino Mass Hierarchy

Choosing between two *simple* hypotheses is the prototype problem in classic Neyman-Pearson theory of hypothesis testing ("simple" = no fit parameters).

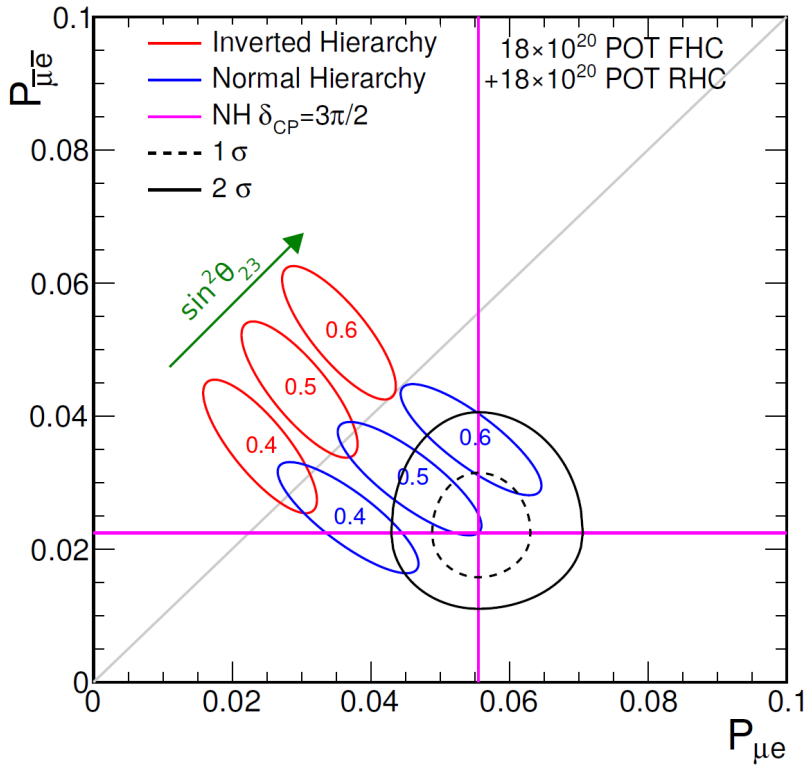
But rare in HEP. We do have (almost) simple cases, e.g.,

- Number of light ν flavors (e.g., 3 vs 4 in late 1980's)
- Spin 1 vs spin 2 for new resonance
- Higgs spin-parity (assuming spin 0) either 0^+ or 0^-

Neutrino Mass Hierarchy

For MH, interesting complication of non-trivial nuisance parameters:

phase δ_{CP} , angle θ_{23}



Backhouse talk

I concentrate on the simplest (but still rich!) case of simple vs simple testing.

Although the ν community seems to have its confusion about that sorted out now, I thought it might be worth a "tutorial" on χ^2, likelihood ratios.

Likelihood ratios, Central Limit Thm, χ^2 , and all that

N quantities to measure. $i=1,N$

Simple $\mathrm{H_A}$: true values are $\{ f_{\mathrm{A},i} \}$

Simple $\mathrm{H_B}$: true values are $\{f_{\mathrm{B},i}\}$

Measurements $\{d_i\}$, $i=1,N$ with Gaussian rms σ_i .

$$\mathcal{L}(\mathrm{H_A}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2}}$$

$$-2\ln\lambda_{AB} = -2\ln\mathcal{L}(\mathrm{H_A}) + 2\ln\mathcal{L}(\mathrm{H_B}) = \sum_{i=1}^{N} \frac{(d_i - f_{\mathrm{A},i})^2}{\sigma_i^{\,2}} - \frac{(d_i - f_{\mathrm{B},i})^2}{\sigma_i^{\,2}}$$

By **Central Limit Theorem**, $-2\ln\lambda_{AB} \rightarrow$ **Gaussian**, independent of true H.
(Can let the σ_i depend on H as well!)

N.B: No mention yet of Wilks, χ^2, DOF. Just CLT, if we are in asymptopia or dominated by the certain term when re-expressed in certain way.

(A more high-powered discussion can invoke non-central chisquare; see Blennow et al, JHEP 1403 (2014) 028)

The ν community calls the above "$\Delta\chi^2$", where, individually under $\mathrm{H_A}$ and $\mathrm{H_B}$:

$$\chi^2(\mathrm{A}) = \sum_{i=1}^{N} \frac{(d_i - f_{\mathrm{A},i})^2}{\sigma_i^{\,2}} \qquad \chi^2(\mathrm{B}) = \sum_{i=1}^{N} \frac{(d_i - f_{\mathrm{B},i})^2}{\sigma_i^{\,2}}$$

Since this is a bit of a long way around in my opinion, it is instructive to take a closer look, viewing these also as likelihood ratios.

$$\mathcal{L}(\mathrm{H}_{\mathrm{sat}}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2}}$$

Repeat the above with binned Poisson data

Observed bin contents $\{n_i\}$, $i=1,N$.

Simple $\mathrm{H_A}$: true Poisson means are $\{f_{\mathrm{A},i}\}$

Simple $\mathrm{H_B}$: true Poisson means are $\{f_{\mathrm{B},i}\}$

$\mathrm{H_{sat}}$: $f_{\mathrm{sat},i} \equiv n_i$.

$$\mathcal{L}(\mathrm{H_A}) = \prod_{i=1}^{N} f_{\mathrm{A},i}{}^{n_i} \frac{e^{-f_{\mathrm{A},i}}}{n_i!}$$, similarly for $\mathcal{L}(\mathrm{H_B})$.

$$\mathcal{L}(\mathrm{H_{sat}}) = \prod_{i=1}^{N} n_i{}^{n_i} \frac{e^{-n_i}}{n_i!}$$

Once again, $-2\ln\lambda_{\mathrm{A,B}} = -2\ln\frac{\mathcal{L}(\mathrm{H_A})}{\mathcal{L}(\mathrm{H_B})}$ $\rightarrow$ Gaussian by CLT.

$-2\ln\lambda_{\mathrm{A,sat}} = -2\ln\frac{\mathcal{L}(\mathrm{H_A})}{\mathcal{L}(\mathrm{H_{sat}})}$ $\rightarrow \chi^2$, N DOF, similarly for $-2\ln\lambda_{\mathrm{B,sat}}$

GOF test based on Poisson LR $-2\ln\lambda_{A,sat}$ with saturated model was subject of my first foray into statistics literature...

Nuclear Instruments and Methods in Physics Research 221 (1984) 437–442

CLARIFICATION OF THE USE OF CHI-SQUARE AND LIKELIHOOD FUNCTIONS IN FITS TO HISTOGRAMS

Steve BAKER and Robert D. COUSINS

Department of Physics, University of California, Los Angeles, California 90024, USA

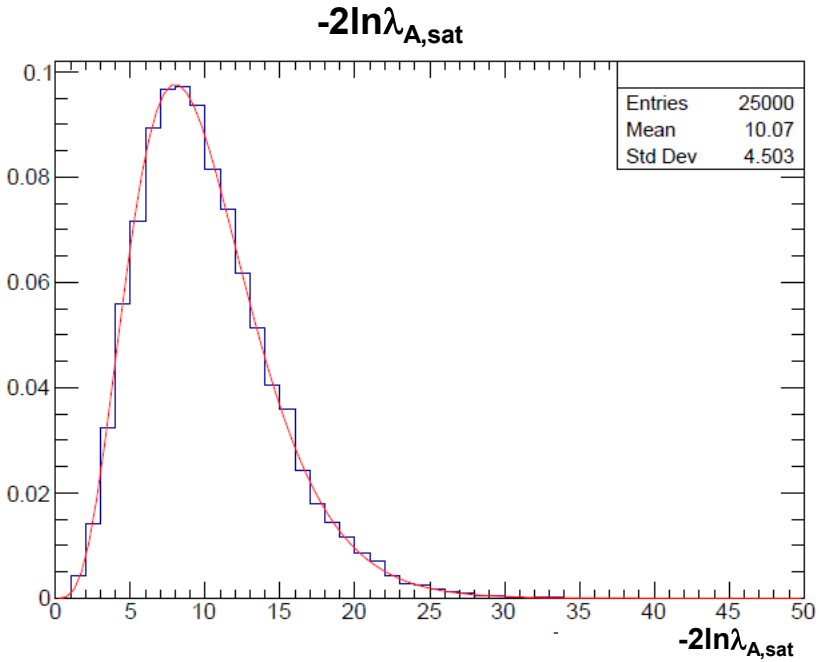
$$\chi^2_{\lambda,p} = 2\sum y_i - n_i + n_i \ln(n_i/y_i)$$

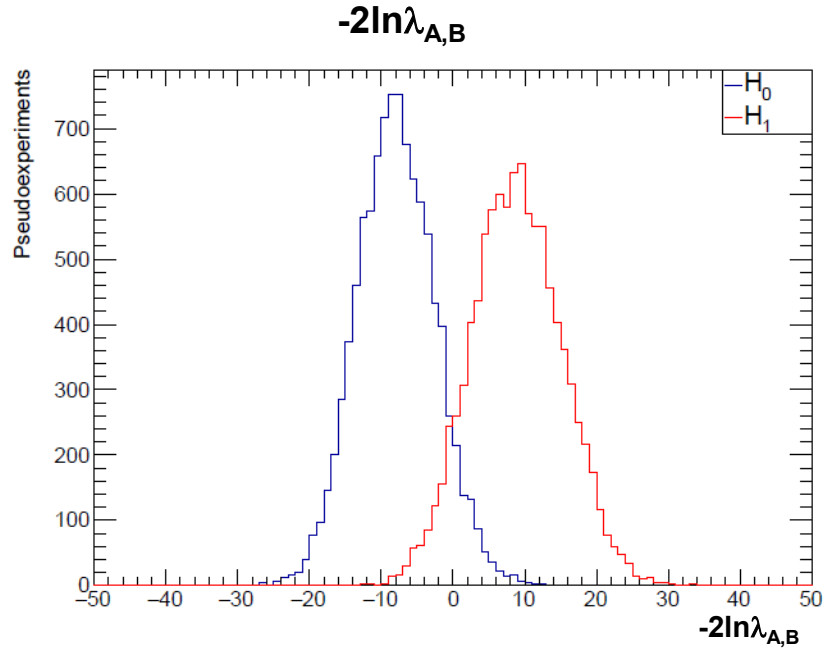
[20] Some early and later studies include: J. Neyman and E.S. Pearson, Biometrika 20A (1928) 263; S.S. Wilks, Ann. Math. Stat. 6 (1935) 190; S.S. Wilks, Ann. Math. Stat. 9 (1938) 60; R.A. Fisher, Biometrics 6 (1950) 17; J. Berkson, Biometrics 28 (1972) 443.

A recent worked MC example for binned Poisson data is in

"Should unfolded histograms be used to test hypotheses?"

Cousins, May, Sun http://arxiv.org/abs/1607.07038





$-2\ln\lambda_{A,sat}$ → χ^2, in this case 10 DOF

$-2\ln\lambda_{B,sat}$ very similar

$-2\ln\lambda_{A,B}$ → Gaussian

What about binned Poisson data?

Now N is number of events (not bins).

Let $\boldsymbol{\theta}$ be vector of observable (energy, angles, etc.) with pdf $p(\boldsymbol{\theta}|H_{\mathrm{A}})$.

$$\mathcal{L}(\mathrm{H_A}) = \prod_{i=1}^{N} p(\boldsymbol{\theta}_i|H_{\mathrm{A}}) \quad -2\ln(\mathrm{H_A}) = \sum_{i=1}^{N} -2\ln(p(\boldsymbol{\theta}_i|\mathrm{H_A}))$$

similarly for $\mathcal{L}(\mathrm{H_B})$.

Once again, $-2\ln\lambda_{\mathrm{A,B}} = -2\ln\dfrac{\mathcal{L}(\mathrm{H_A})}{\mathcal{L}(\mathrm{H_B})}$ $\rightarrow$ Gaussian by CLT.

However, there is no natural analog to the saturated model and hence the individual GOF tests are ~arbitrary, and $-2\ln\lambda_{\mathrm{A,B}}$ is not equivalent to a "$\Delta\chi^2$".

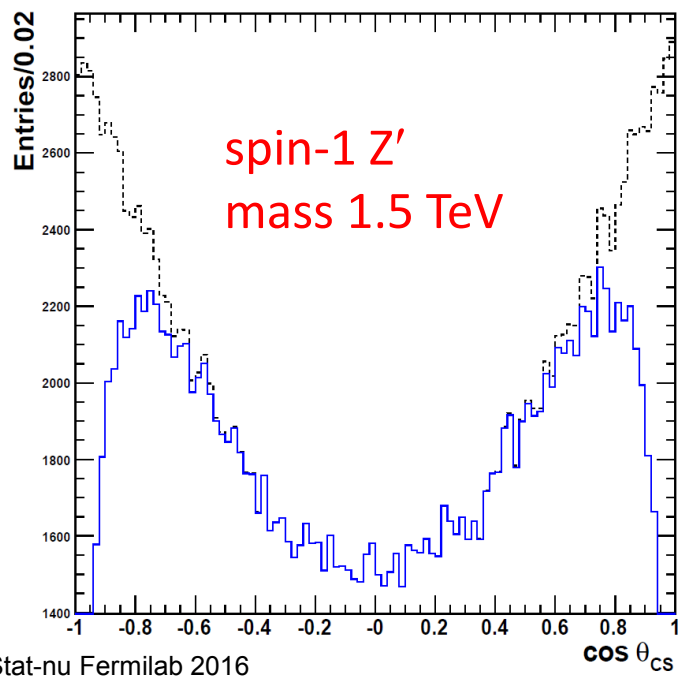
$\Rightarrow$ A reason why I prefer direct $-2\ln\lambda_{\mathrm{A,B}}$ approach to "$\Delta\chi^2$" approach.

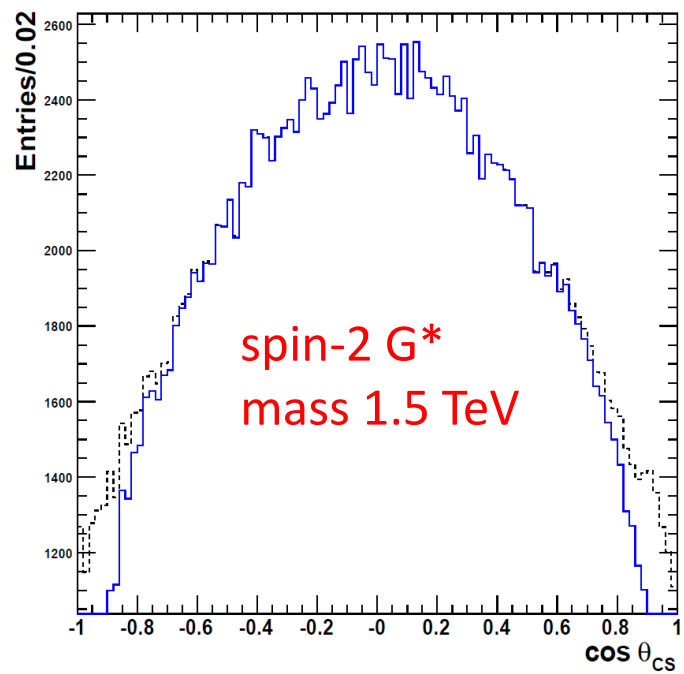
Published by Institute of Physics Publishing for SISSA
Received: *September 8, 2005*

Spin discrimination of new heavy resonances at the LHC

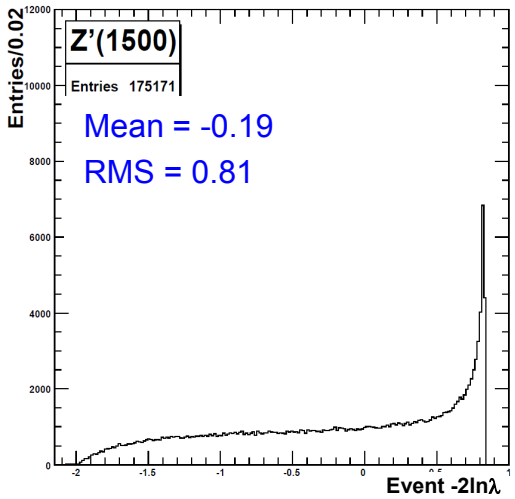
Robert Cousins, Jason Mumford, Jordan Tucker and Viatcheslav Valuev

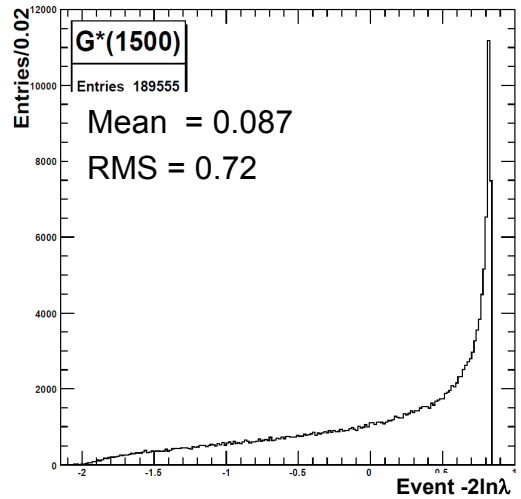
Preparing for LHC, we imagined new dilepton resonance.
H_A: spin-1 Z′, or
H_B: spin-2 graviton G*
Discriminating variable: quark-muon angle θ_{CS} in Collins-Soper frame.



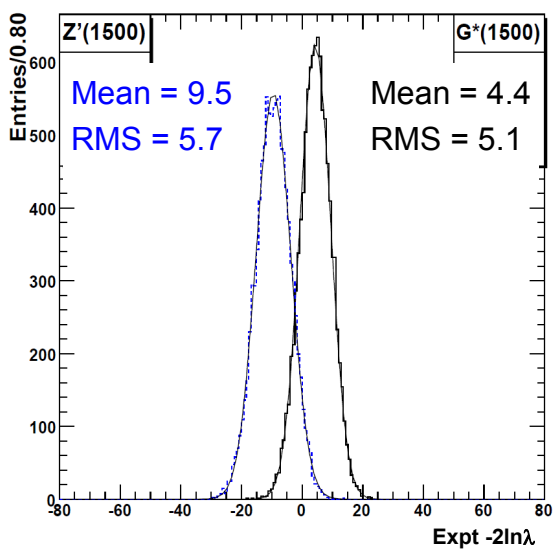


Histograms of $-2\ln\lambda_{A,B} = -2\ln\frac{\mathcal{L}(H_A)}{\mathcal{L}(H_B)}$ for individual MC events





Each MC experiment is 50 samples from above. Add -2lnλ's from events:



Gaussian by CLT

mean = event mean × 50

RMS = event RMS × √50

(An earlier paper had erroneously assumed that -2lnλ was χ^2 .)

(Peak separation) / RMS scales beautifully with √N.

Above is all "pre-data" characterization of the test

How to characterize *post-data*?

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

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

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

Numerically (if not philosophically) the same as usual "value obtained or more extreme" due to Fisher.

Large literature bashing p-values. I defend HEP: http://arxiv.org/abs/1310.3791

Interpreting p-values and Z-values

It is crucial to realize that that value of α was typically *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!

In HEP, typically converted to Z-value, equivalent number of Gaussian sigma.

At LHC, we had recent case that forced us to think about post-data interpretation of (nearly) simple vs simple test.

Early CMS Higgs spin-parity test of 0^+ vs. 0^-

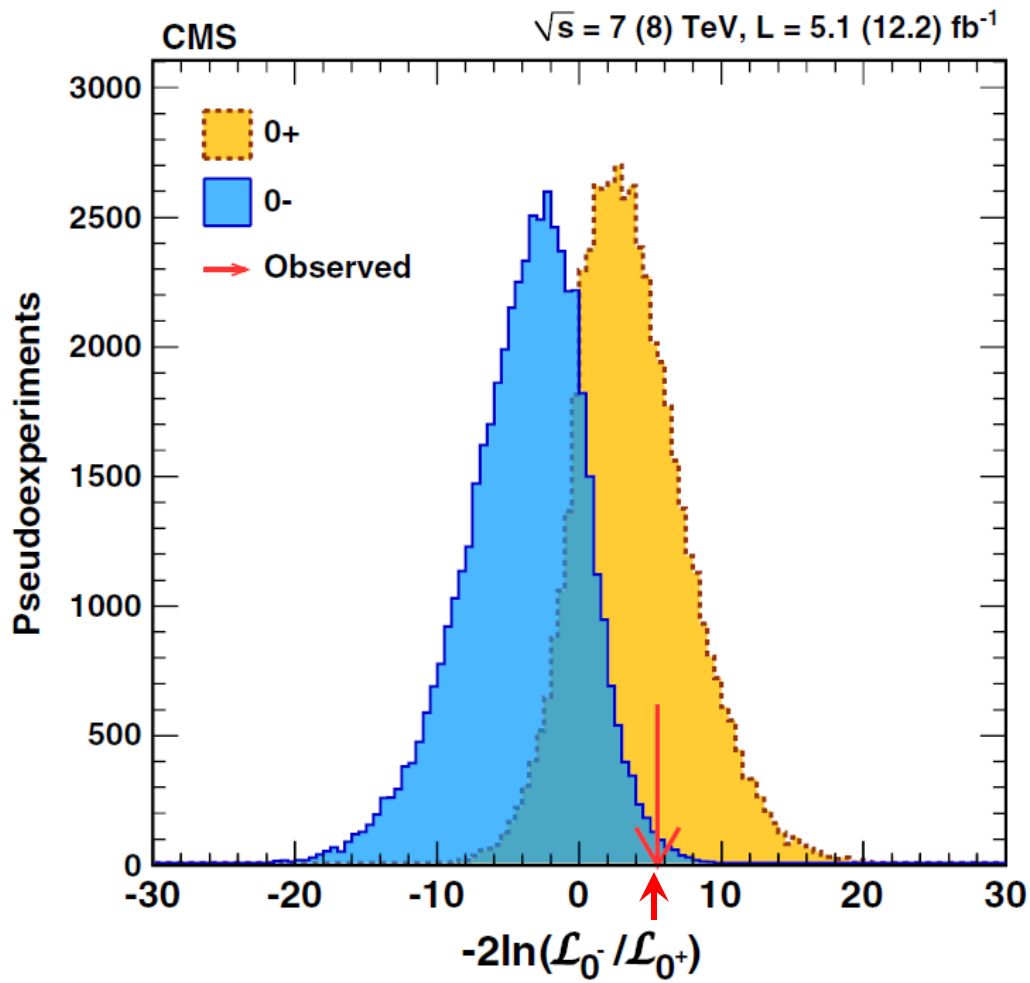


FIG. 3 (color online). Expected distribution of $-2\ln\mathcal{L}_{0^-}/\mathcal{L}_{0^+}$ under the pure pseudoscalar and pure scalar hypotheses (histograms). The arrow indicates the value determined from the observed data.

Paper reported (fixing typo here):

1) $-2\ln(\mathcal{L}_{0-}/\mathcal{L}_{0+})$
$= 5.5$ favoring 0^+

2) p-value = 0.72% for 0^-

3) p-value = 0.7 for 0+

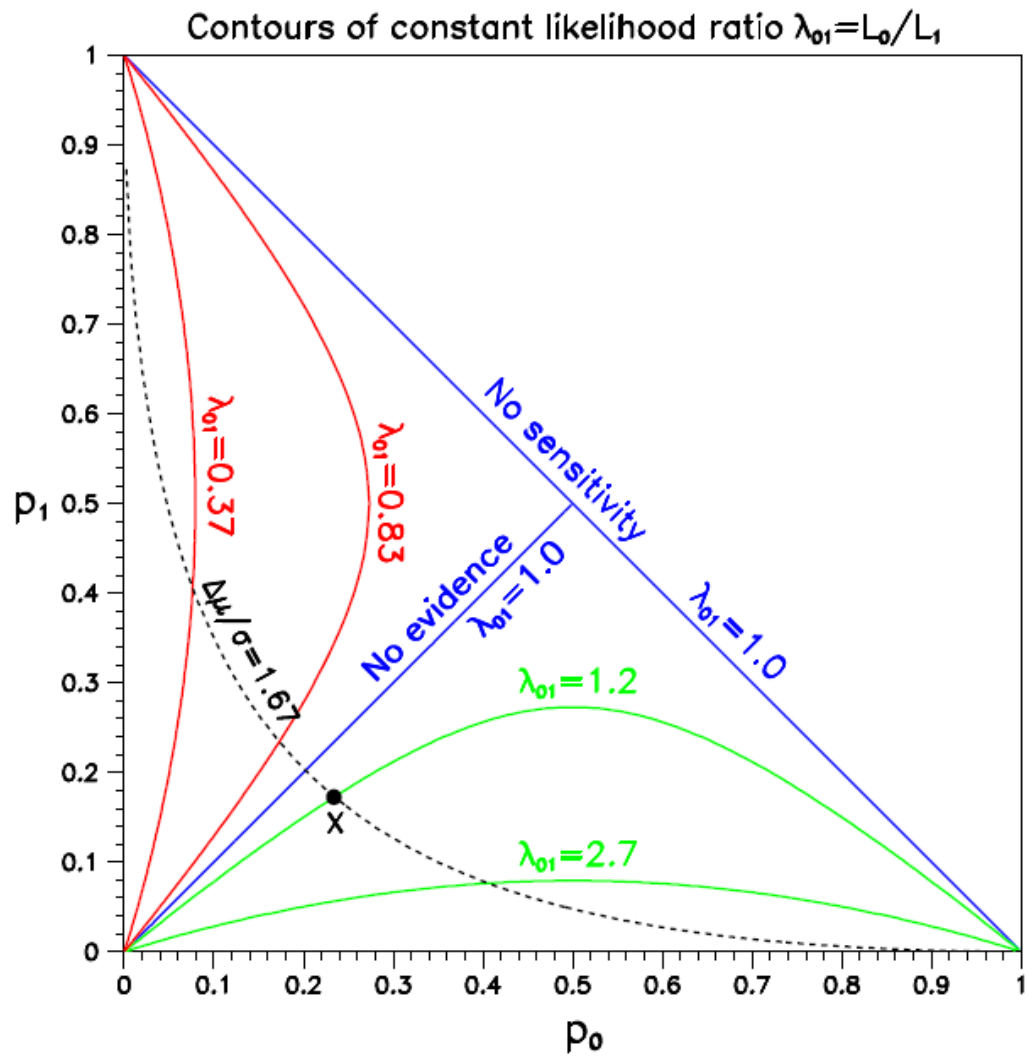
4) CL_s = (0.72%) / (1–0.7) = 2.4%, "a more conservative value for judging whether the observed data are compatible with 0^- "

Luc Demortier and Louis Lyons, http://arxiv.org/abs/1408.6123
"Testing Hypotheses in Particle Physics: Plots of p_0 versus p_1"

Test of point null vs point alternative, two Gaussians with same σ, peak separation $\Delta\mu$.

At a glance can see that contours of constant λ_{01} are completely different topology from contours of e.g. p_0.

(For rest of plot, you will have to read their paper or stare at it for a long time.)



Number of light ν flavors in 1989: 3 light ν s known

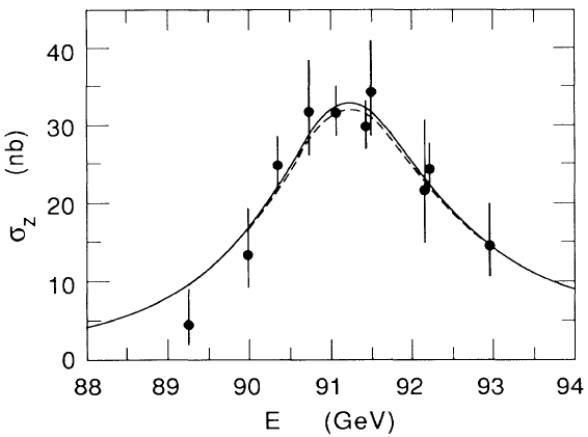
Crucial to test $_{\nu}$=3 vs $_{\nu}$ (or more?) in Z decay.

Mark II collab at SLAC SLC, facing imminent competition from LEP.

Rather than treating $_{\nu}$ 3 and $_{\nu}$ 4 has "point hypotheses", they treated N_{ν} as a *continuous* parameter estimated with standard techniques, obtaining $N_{\nu} = 2.8 \pm 0.6$ from resonance parameters of Z.

"The 95%-C.L. limit, N_{ν}<3.9, excludes to this level the presence of a fourth massless neutrino species within the standard-model framework."

(*Several* interesting discussion points, Including benefit of downward fluctuation!)



PRL 63, 2173 (1989)

Continuous Mass Hierarchy variable?

The +1 and -1 for MH appear in the equations as simply that: arithmetic signs.

Various authors (e.g., Capozzi, Lisi, and Marrone, PRD 89 013001) have suggested replacing ± 1 with (unbounded) continuous variable α.

Reminiscent of continuous "number of light neutrino species" (which recall had BSM physics interpretation).

In frequentist treatment, I think it is mostly a matter of presentation, since results from discrete way map to continuous way, and vice versa (particularly if F-C construction is used for confidence interval for α, with relevant set of C.L.'s).

I encourage continuous α approach as part of toolkit.

But…Eligio Lisi has explained to me that α is highly correlated with Δm^2, and contributes to increase its overall uncertainty. This leads to the undesired result that power is lost due to consideration of unphysical (or at least non-SM) values of MH. Ugh.

NOTE added after talk: I mis-stated Eligio's point above at the time of the talk; I believe that it is now repaired. -BC

Addition of Nuisance Parameter δ to MH Test

Small variation of nuisance parameters seems not to upset the formalism, and some relevant examples with toys still give nicely Gaussian distribution of LR test statistic. However the situation can become harder – see talk by Sara Algeri at Tokyo.

If the CP phase δ is treated as a nuisance parameter in the MH determination, then great care is needed.

Providing the MH results as a function of δ (same δ in numerator and denominator of LR) would seem to be mandatory, before attempting to "eliminate" δ by profiling or marginalizing..

Addition of Nuisance Parameter δ to MH Test

Then I would try:

a) Profiling: for each value of MH, find best case δ. So numerator and denominator in LR would generally be evaluated at different values of δ. Info complementary to giving MH as ftn of same δ in num and denom.

b) Marginalizing: Takes average weighted by prior for δ of numerator $\mathcal{L}$; weighted average of denominator $\mathcal{L}$ over δ, then ratio. Scary! Obviously mandatory to study prior dependence, freq. coverage.

Profiling or Marginalizing Nuisance Parameters

Although profiling nuisance params is the “native” treatment in a frequentist context, it comes with no performance guarantees at small sample sizes.

Marginalizing might even do better from *frequentist* point of view!

Classic pathological case has integrable singularity in data pdf, and hence singularity in likelihood – Tom Loredo likes analogy of temperature vs heat.

But something about "eliminating" δ_{CP} reminds me of the quote by "likelihoodist" A.W.F. Edwards:

"Let me say at once that I can see no reason why it should always be possible to eliminate nuisance parameters. Indeed, one of the many objections to Bayesian inference is that it always permits this elimination."

(commenting on J.D. Kalbfleisch and J.D. Sprott, J. Roy. Stat. Soc. Series B 32, 175 (1970). See my paper Oxford05.)

For further reading:

For PhyStat 2005, I wrote, "Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature". Small compared to:

Luc Demortier, "P Values: What They Are and How to Use Them" http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf (174 pages!)

Testing point null(s) alternative

Prototype: "Test for continuous $\theta=\theta_0$ vs $\theta\neq\theta_0$"

Example before us at this conference: Is there CP violation in PMNS matrix ?

H_0: $\delta=0$ or $\delta=\pi$, vs

H_1: $\delta \subset$ open intervals $(0,\pi) \cup (\pi,2\pi)$

Jargon (e.g. in Bayesian framework)

Discrete parameter values with non-zero *probability*:

Counting measure, probability mass, Dirac Delta-ftn in density

Continuous parameter values with non-zero *probability density*,

probability for any single value is zero: *Lebesgue measure*.

Bayesian and frequentist frameworks treat mix of counting and Lebesgue measures in testing *completely differently*.

In general the asymptotic convergence we are used to in estimation does not happen.

Classical Hypothesis Testing: Duality

Test $\theta=\theta_0$ at $\alpha \leftrightarrow$ Is θ_0 in conf. int. for θ with C.L. = 1- α

"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.

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

Classical Hypothesis Testing (cont.)

"Test for $\theta=\theta_0$" ↔ "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 PRD 57 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

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

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.

Duality: p-values from F-C

Post-data, find the threshold C.L. for which θ is on the boundary of the F-C confidence interval/region.
I.e., for C.L. lower than that threshold, θ is not in the confidence region.

(Strangely, this post-data C.L. seems not to have a special name to distinguish from pre-data C.L.)

The F-C p-value is then 1 minus this C.L.
(Since pre-data alpha is 1−C.L.)

But this just takes us full circle in the duality: this “F-C p-value” is just the p-value for the LR test in Kendall and Stuart!

test for continuous $\theta=\theta_0$ vs $\theta\neq\theta_0$

Bayesian approach does not use the duality!!!

Point and interval estimation for a parameter are *separate* from hypothesis testing.
Also known as *model selection*: the lower-dimensional model with $\theta=\theta_0$ vs the model with one more dimension, in θ.

In contrast to frequentist method, one does *not* find a Bayesian credible interval for θ and test if θ is in it.

The historical standard for Bayesian model selection is due to Harold Jeffreys, calculating posterior probability of each model.

Requires counting measure for null: bit of Dirac delta function in prior density at $\theta=\theta$ (or equivalent).
Brings in a whole new set of issues not present in estimation!
("Can of worms" – Jim Berger)

Bayesian test for continuous $\theta=\theta_0$ vs $\theta\neq\theta_0$ (cont.)

Dependence on the prior for θ does not go away asymptotically: the *Bayes Factor* (ratio of posterior odds to prior odds) depends on $\pi(\theta)$!

Improper priors lead to zero or infinity: if made finite by cut-off, answer depends on the cutoff.

All the Ockham's razor praise you hear about Bayesian model selection depends on cutoff if it's an unbounded parameter (e.g. Poisson mean): fine if you are carefully subjective – beware of default priors!

Even for bounded binomial parameter ρ ($0\leq \rho \leq 1$), testing e.g. $\rho = 0.5$ vs $\rho \neq 0.5$ has issues (though Bayesians like this example for bashing p-values because at least there is no prior cutoff disaster)
$\Rightarrow$ Jeffreys used different priors for estimation and testing.

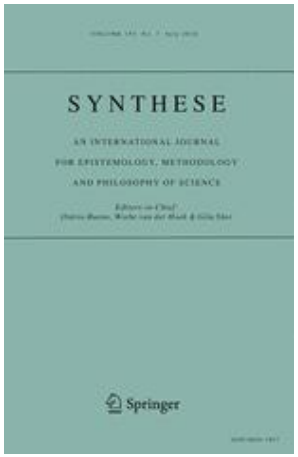
Bayesian test for continuous $\theta=\theta_0$ vs $\theta\neq\theta_0$: δ_{CP}

δ_{CP} is bounded similarly to binomial ρ, so at least no cut-off issue.
But still involves contentious frequentist vs Bayesian testing issues.

This is deep stuff – physicists exploring use of using this (myself included) need to talk to Bayesian statisticians, read literature.

My attempt to analyze and explain HEP p-value practice, with many references to Bayesian testing literature:

"The Jeffreys–Lindley paradox and discovery criteria in high energy physics",
Synthese (2014), DOI 10.1007/s11229-014-0525-z
http://arxiv.org/abs/1310.3791 (editing error fixed).



My advocacy for >10 years:

- Have in place tools to allow computation of results using a variety of recipes, for problems up to intermediate complexity:
  - Bayesian with analysis of sensitivity to prior
  - Profile likelihood ratio (Minuit MINOS)
  - Frequentist construction (incl F-C) with approximate treatment of nuisance parameters
  - Other "favorites" such as LEP's CL_S (which is an HEP invention)
- The community can then demand that a result shown with one's preferred method also be shown with the other methods, *and sampling properties studied*.
- When the methods all agree, we are in asymptopic nirvana.
- When the methods disagree, we learn something! E.g.:
  - The results are answers to different questions (but...)
  - Bayesian methods can have poor frequentist properties
  - Frequentist methods can badly violate likelihood principle

BACKUP

These are mostly from my stock slides on things that I “wish everyone knew”.

See also my summary talk at Tokyo PhyStat-nu, which include examples of pseudo-Bayes detection.

Who am I (other than Gary Feldman's co-author)?

Ph.D. Thesis: Fermilab E533, 1978-80, π μ atoms
Forward charm production in pp collisions at CERN ISR
Rare kaon decays at BNL
H dibaryon search at BNL
Sabbatical year at Harvard working on CDF
1994-1999 ν oscillation searches at NOMAD at CERN
CMS at LHC since 2000 (Deputy Spokes, 2007-2009)

Many committees, some relevant to ν's, especially:
Lehmann review of NUMI project (1998)
Fermilab PAC 1999-2003
P5 panel 2013-14 $\Rightarrow$ crash course on all ν experiments in world (and their proposed statistical methods)

Summary of Three Ways to Make Intervals

	Bayesian Credible	Frequentist Confidence	Likelihood Ratio
Requires prior pdf?	Yes	No	No
Obeys likelihood principle?	Yes (exception re Jeffreys prior)	No	Yes
Random variable in " $\mu \in \mu \quad \mu$	μ	$\mu \quad \mu$	$\mu \quad \mu$
Coverage guaranteed?	No	Yes (but over-coverage…)	No
Provides P(parameter|data)?	Yes	No	No

Frequentist intervals map to frequentist hypothesis tests, as discussed above.
Bayesian equivalent of hypothesis testing is called and is a whole other "can of worms".

68% intervals by various methods for mean μ of Poisson process with n=3 observed

Method	Prior	Interval	Length	Coverage?
rms deviation	–	(1.27, 4.73)	3.46	no
Bayesian central$_{\pm\sqrt{}}$	1	(2.09, **5.92**)	3.83	no
Bayesian shortest	1	(1.55, 5.15)	3.60	no
Bayesian central	$1/\mu$	(**1.37**, 4.64)	3.27	no
Bayesian shortest	$1/\mu$	(0.86, 3.85)	2.99	no
Likelihood ratio	–	(1.58, 5.08)	3.50	no
Frequentist central	–	(**1.37, 5.92**)	4.55	yes
Frequentist shortest	–	(1.29, 5.25)	3.96	yes
Frequentist LR ordering	–	(1.10, 5.30)	4.20	yes

For the Jeffreys prior ($1/\sqrt{\mu}$), Bayesian central interval is (1.72, 5.27).
Fastest approach to correct coverage as n increases (Welch Peers, 1963)

Numerical coindence of *upper endpoint* of intervals led to *flat prior on Poisson mean in early HEP Bayesian analyses*: probability matching prior for upper limits!

Adapted from Cousins05 and
R. Cousins, Am. J. Phys. 63 398 (1995)

Sir David Cox at PhyStat-LHC 2007

Five faces of Bayesian statistics

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

From my Summary Talk: his list, as augmented in HEP

Six faces of Bayesian statistics

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

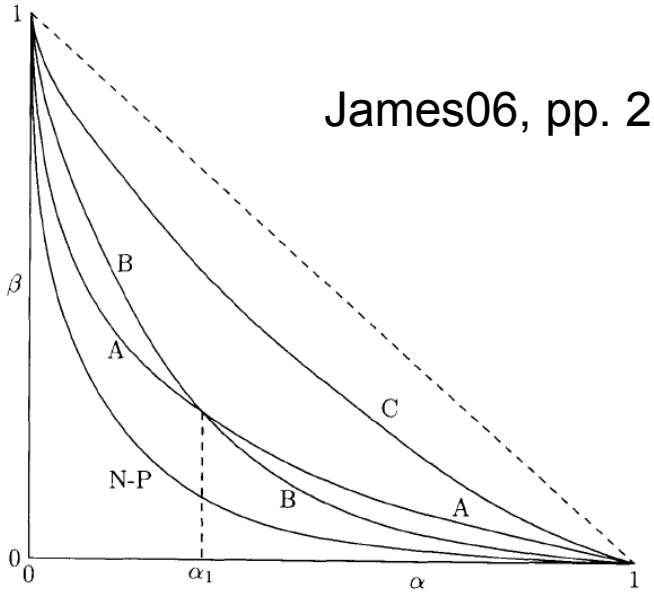
Mini-review: Classical (N-P) Hypothesis Testing

In Neyman-Pearson hypothesis testing (James06), frame discussion in terms of null hypothesis H_0 = S.M., and an alternative H_1 = mSUGRA, etc.

- **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**

Competing analysis methods can be compared by looking at graphs of β vs α at various θ, and at graphs of β vs θ at various α (power function).

Classical Hypothesis Testing (cont.)



James06, pp. 258, 262

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

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 5σ is without foundation.

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 hypothesis H_0 against hypothesis H_1, then the Type II error probability β is minimized by using as the test statistic the

$\lambda = \mathcal{L}(x|\, H_0) / \mathcal{L}(x|\, H_1)$, 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 Stuart99, p. 176

Conditioning*

- An "ancillary statistic" (see literature for precise math definition) is a function of your data which carries information about the precision of your measurement of the parameter of interest, but no info about parameter's value.
- The classic example is a branching ratio measurement in which the total number of events N can fluctuate if the expt design is to run for a fixed length of time. Then N is an ancillary statistic.
- You perform an experiment and obtain N total events, and then do a toy M.C. of repetitions of the experiment. Do you let N fluctuate, or do you fix it to the value observed?
- It may seem that the toy M.C. should include your procedure, including fluctuations in N.
- But there are strong arguments, going back to Fisher, that inference should be based on probabilities

!

*See Reid95 for a review; my post http://arxiv.org/abs/1109.2023 has discussion in controversial non-ancillary case of bounded Gaussian mean problem.

Conditioning (cont.)

- The 1958 thought expt of David R. Cox focused the issue:
  - Your procedure for weighing an object consists of flipping a coin to decide whether to use a weighing machine with a 10% error or one with a 1% error; and then measuring the weight. (Coin flip result is ancillary stat.)
  - Then "surely" the error you quote for your measurement should reflect which weighing machine you actually used, and not the average error of the "whole space" of all measurements!
  - But classical most powerful Neyman-Pearson hypothesis test uses the whole space!
- In more complicated situations, ancillary statistics do not exist, and it is not at all clear how to restrict the "whole space" to the relevant part for frequentist coverage.
- In methods obeying the likelihood principle, in effect one conditions on the exact data obtained, giving up the frequentist coverage criterion for the guarantee of relevance.

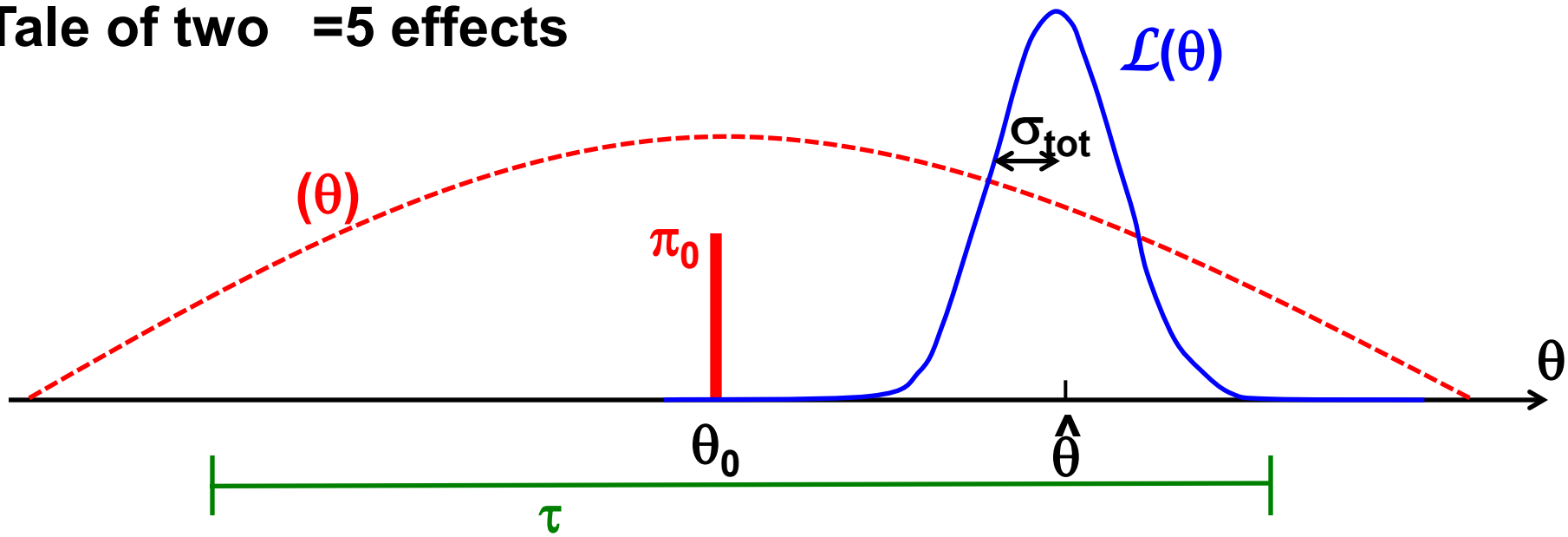
Conditioning (cont.)

- At past PhyStats and CosmoStats, Jim Berger said that one of his "pet peeves" was people ignoring frequentist conditioning – improving inference by noting where one's own sample is in the unconditional sample space.
- Fisher introduced the concept of "recognizable subsets" in the sample space.
- It's interesting that D.R. Cox does not mind null intervals (essentially viewing them as bad fit), but others view downward fluctuations as a "recognizable subset".
- From this point of view, there IS a problem with null-set confidence intervals!
- So the situation is significantly more interesting than saying that all frequentists care about is the "unconditional ensemble", and not about your particular data.
- Feldman-Cousins intervals seem to survive an important test failed by standard intervals: Buehler's betting game"

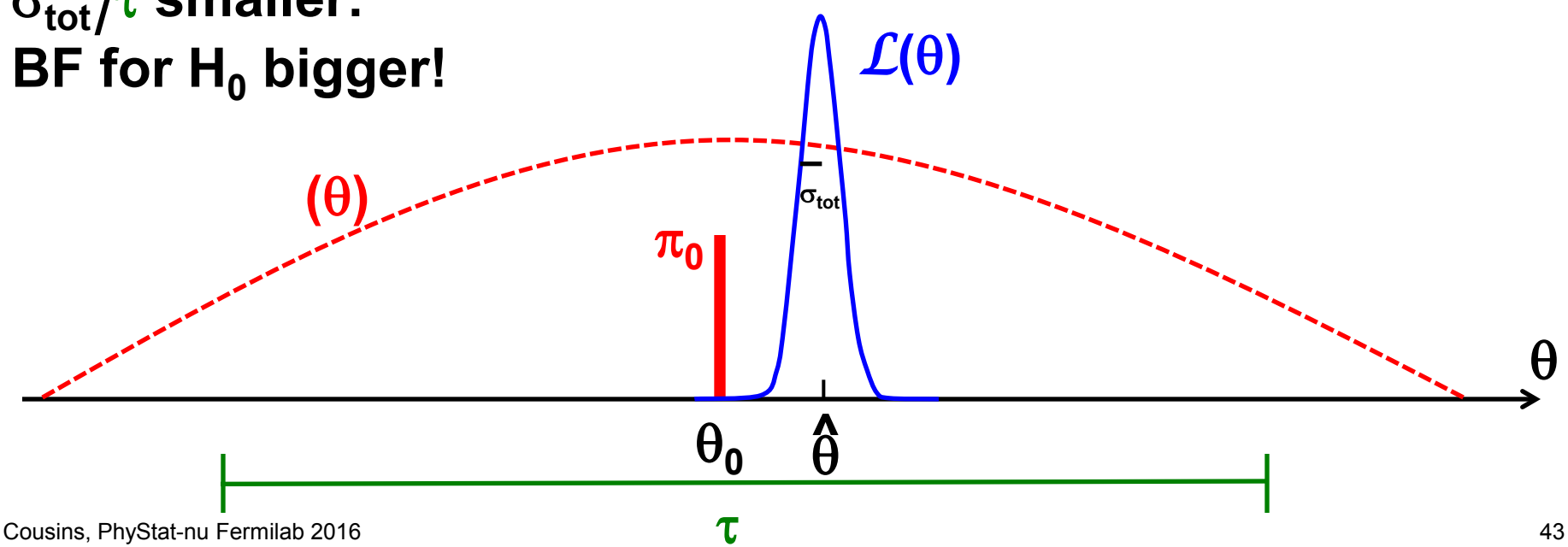
See http://arxiv.org/abs/1109.2023

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

Tale of two =5 effects



σ_{tot}/τ smaller:
BF for H_0 bigger!

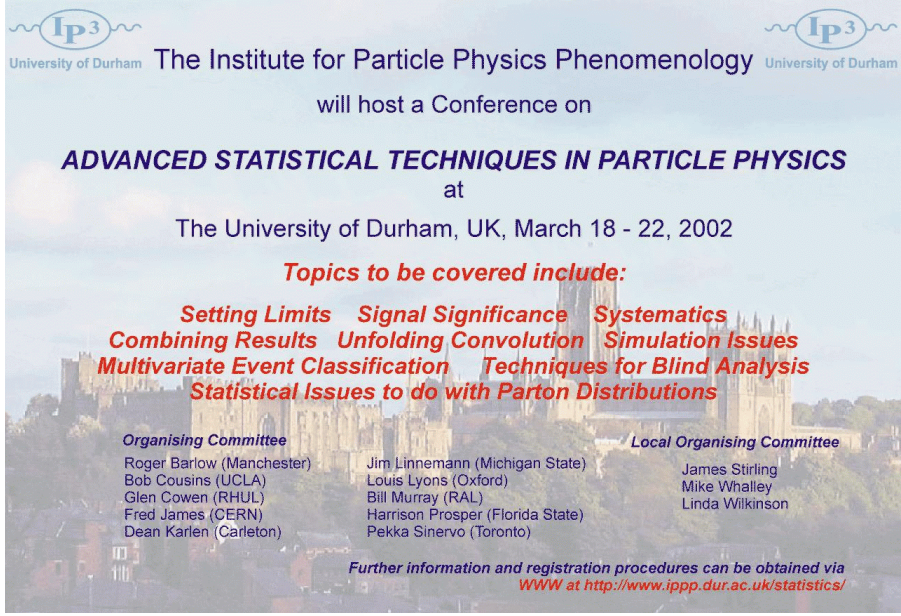


Sensitivity Analysis

- **Since a Bayesian result depends on the prior probabilities, which are either personalistic or with elements of arbitrariness, it is widely recommended by Bayesian statisticians to study the of the result to varying the prior.**
- **I express dismay a lot in HEP at how little emphasis this is given by Bayesian advocates in HEP. My is that it could also be given more emphasis in astro/cosmo – some exemplary papers certainly exist!**

- **An "objective Bayesian's" point of view: "Non-subjective Bayesian analysis is just a part -- an important part, I believe – of a healthy analysis to the prior choice…" – J.M. Bernardo, J. Roy. Stat. Soc., Ser. B 41 113 (1979)**

Sensitivity analysis: A subjective Bayesian's point of view:



WHY BE A BAYESIAN?

Michael Goldstein
Dept. of Mathematical Sciences, University of Durham,England

From the Proceedings: "…Again, different individuals may react differently, and the sensitivity analysis for the effect of the prior on the posterior is the analysis of the scientific community..."

From his transparencies:

"Sensitivity Analysis is at the heart of scientific Bayesianism."

Bayesians, Frequentists, and Physicists

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

"Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions."

Workshop Homepage

Workshop on Confidence Limits

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

Jim Berger:

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

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

What should be the view today:

Objective Bayesian analysis is the best frequentist tool around.

U.L. in Poisson Process, n=3 observed: 3 ways

1. **Bayesian upper limit at 90% credibility: find μ_u such that posterior probability $P(\mu > \mu_u) = 0.1$.**
2. **Likelihood ratio method for approximate 90% C.L. U.L.: find μ_u such that $\mathcal{L}(\mu_u) / \mathcal{L}(3)$ has prescribed value.**
3. **Frequentist one-sided 90% C.L. upper limit: find μ_u such that $P(n \leq 3 \mid \mu_u) = 0.1$.**

Deep foundational issues

- **Only #3 has guaranteed ensemble properties (though issues arise with systematics.) Good ?!?**
- **Only #3 uses $P(n|\mu)$ for $n \neq$ observed value. Bad?!? (Violates likelihood principle)**

These issues will not (should not) be resolved in HEP: aim to have software for reporting all 3 answers, and sensitivity to prior. Make all available to consumer.

Likelihood Principle

- As noted above, in both Bayesian methods and likelihood-ratio based methods, the probability (density) for obtaining the
  is used (via the likelihood function),

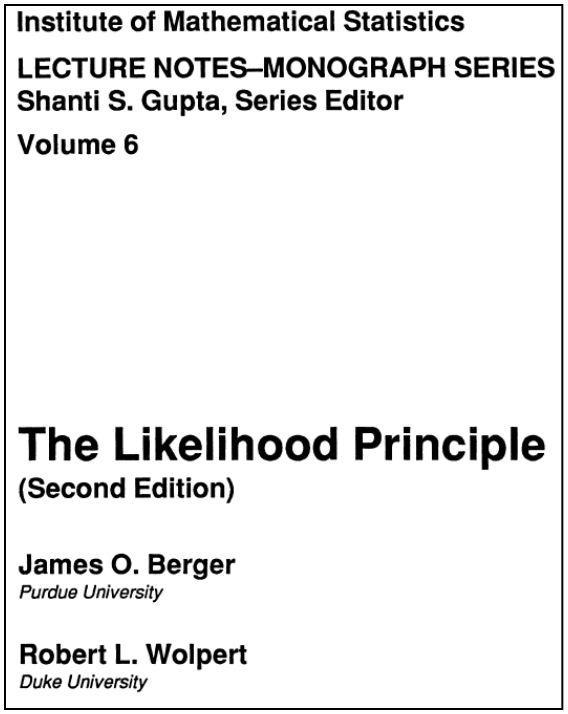
- In contrast, in typical frequentist calculations (e.g., a p-value which is the probability of obtaining a value as extreme or
  than that observed), one uses probabilities of data .
- This difference is captured by the : If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.
- L.P. is built in to Bayesian inference (except e.g., when Jeffreys prior leads to violation).
-
- Although practical experience indicates that the L.P. may be too restrictive, it is useful to keep in mind. When frequentist results "make no sense" or "are unphysical", in my experience the underlying reason can be traced to a bad violation of the L.P.

*There are various versions of the L.P., strong and weak forms, etc. See Stuart99 and book by Berger and Wolpert.

Likelihood Principle Discussion

We will not resolve this issue, but should be aware of it.

- **See book by Berger & Wolpert, but be prepared for the "Stopping Rule Principle" to set your head spinning.**
- **When frequentist intervals and limits badly violate the L.P., use great caution in interpreting them!**
- **And when Bayesian inferences badly violate the frequentist coverage, again use great caution!**

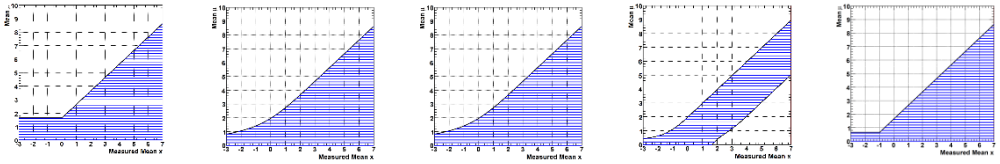

Institute of Mathematical Statistics
LECTURE NOTES–MONOGRAPH SERIES
Shanti S. Gupta, Series Editor
Volume 6

The Likelihood Principle
(Second Edition)

James O. Berger
Purdue University

Robert L. Wolpert
Duke University

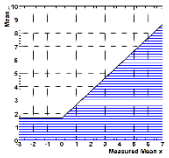
Bayes, Fisher, Neyman, Neutrino Masses, and the LHC



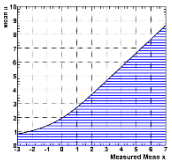
Bob Cousins
Univ. of California, Los Angeles

Virtual Talk
12 September 2011

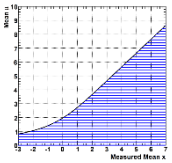
Five methods used for bounded Gaussian mean problem



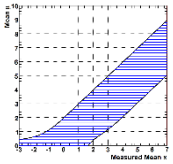
1) 1960's and beyond:
   UL = max(, 0) + 1.64σ



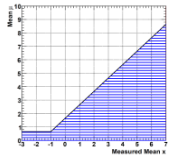
2) 1979 "PDG" (real 1986 PDG) and beyond:
   Bayesian with uniform prior



3) 1997: Alex Read et al. (LEP)
   CL_S



4) 1997: Feldman and Cousins (NOMAD)
   Unified Approach



5) 2010: Power Constrained Limits;
   Cowan, Cranmer, Gross, Vitells (ATLAS):
   UL = max(0, max(x, x_{PCL}) + 1.64σ)

References Cited in Talk Slides

Cousins05: Robert Cousins, "Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature", PhyStat05: Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, 12-15 Sept. 2005.

James06: Frederick James, Statistical Methods in Experimental Physics, World Scientific, 2006.

Reid95: N. Reid, "The Roles of Conditioning in Inference", Statistical Science 10 138 (1995).

Stuart99: A. Stuart, K. Ord, S. Arnold, Kendall's Advanced Theory of Statistics, Vol. 2A, 6th edition, 1999; and earlier editions by Kendall and Stuart.

Recommended reading

Books: Among the many books available, I usually recommend the following progression, reading the first three cover-to-cover, and consulting the last two as needed:

1) Philip R. Bevington and D.Keith Robinson, Data Reduction and Error Analysis for the Physical Sciences (Quick read for undergrad-level review)

2) Glen Cowan, Statistical Data Analysis (Solid foundation for HEP)

3) Frederick James, Statistical Methods in Experimental Physics, World Scientific, 2006. (This is the second edition of the influential 1971 book by Eadie et al., has more advanced theory, many examples)

4) A. Stuart, K. Ord, S. Arnold, Kendall's Advanced Theory of Statistics, Vol. 2A, 6th edition, 1999; and earlier editions of this "Kendall and Stuart" series. More modern books include:

5) George Casella and Roger L. Berger, Statistical Inference, 2nd Ed., 2002

PhyStat conference series: Beginning with Confidence Limits Workshops in 2000, links at http://phystat-lhc.web.cern.ch/phystat-lhc/ and http://www.physics.ox.ac.uk/phystat05/

By now there are many many web pages with lists of statistics references – Google on your favorite topic.

My Bayesian reading list is the set of citations in my Comment, PRL 101 029101 (2008), especially refs 2, 8, 9, 10, 11 (and 7 for model selection)

Memorable Quotes Therein from Jim Berger

I call such analyses *pseudo-Bayes* because, while they utilize Bayesian machinery, they do not carry with them any of the guarantees of good performance that come with either true subjective analysis (with a very extensive elicitation effort) or (well-studied) objective Bayesian analysis. I will briefly discuss the problem with each of these pseudo-Bayes procedures.

***

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

"The Case for Objective Bayesian Analysis," Bayesian Analysis 1. See pp. 397, 459.

Must-Read for HEP/Astro/Cosmo (incl discussion!)

Robert E. Kass and Larry Wasserman, "The Selection of Prior Distributions by Formal Rules," J. Amer. Stat. Assoc. 91 1343 (1996).

Telba Z. Irony and Nozer D. Singpurwalla, "Non-informative priors do not exist: A dialogue with Jose M. Bernardo," J. Statistical Planning and Inference 65 159 (1997).

J.O. Berger and L.R. Pericchi, "Objective Bayesian Methods for Model Selection: Introduction and Comparison," in Model Selection, Inst. of Mathematical Statistics Lecture Notes-Monograph Series, ed. P. Lahiri, vol 38 (2001) pp .135-207

James Berger, "The Case for Objective Bayesian Analysis," Bayesian Analysis 1 385 (2006)

Michael Goldstein, "Subjective Bayesian Analysis: Principles and Practice," Bayesian Analysis 1 403 (2006)