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

Bob Cousins
Univ. of California, Los Angeles (UCLA)

PHYSSTAT-$\nu$ 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 \textit{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 $\nu$ 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 $0^-$
For MH, interesting complication of non-trivial nuisance parameters: phase $\delta_{CP}$, angle $\theta_{23}$

I concentrate on the simplest (but still rich!) case of simple vs simple testing. Although the $\nu$ community seems to have its confusion about that sorted out now, I thought it might be worth a “tutorial” on $\chi^2$, likelihood ratios.
Likelihood ratios, Central Limit Thm, $\chi^2$, and all that

$N$ quantities to measure. $i=1,N$
Simple $H_A$: true values are $\{f_{A,i}\}$
Simple $H_B$: true values are $\{f_{B,i}\}$
Measurements $\{d_i\}, i=1,N$ with Gaussian rms $\sigma_i$.

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

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

*N.B: No mention yet of Wilks, \(\chi^2\), 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 \(\nu\) community calls the above “\(\Delta\chi^2\)”, where, individually under \(H_A\) and \(H_B\):

\[
\chi^2(A) = \sum_{i=1}^{N} \frac{(d_i - f_{A,i})^2}{\sigma_i^2} \quad \chi^2(B) = \sum_{i=1}^{N} \frac{(d_i - f_{B,i})^2}{\sigma_i^2}
\]

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.

Bob Cousins, PhyStat-nu Fermilab 2016
$\mathcal{L}(H_{\text{sat}}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2}}$
Repeat the above with binned Poisson data

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

Simple \( H_A \): true Poisson means are \( \{f_{A,i}\} \)

Simple \( H_B \): true Poisson means are \( \{f_{B,i}\} \)

\( H_{\text{sat}} \): \( f_{\text{sat},i} = n_i \).

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

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

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

\( -2\ln \lambda_{A,\text{sat}} = -2\ln \frac{\mathcal{L}(H_A)}{\mathcal{L}(H_{\text{sat}})} \quad \rightarrow \chi^2, N \text{ DOF, similarly for } -2\ln \lambda_{B,\text{sat}} \)
GOF test based on Poisson LR $-2\ln\lambda_{A,\text{sat}}$ with saturated model was subject of my first foray into statistics literature...


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

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

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

\[-2\ln \lambda_{A,\text{sat}} \rightarrow \chi^2, \text{ in this case 10 DOF}\]

\[-2\ln \lambda_{B,\text{sat}} \text{ very similar}\]

\[-2\ln \lambda_{A,B} \rightarrow \text{Gaussian}\]
What about binned Poisson data?

Now $N$ is number of events (not bins).

Let $\theta$ be vector of observable (energy, angles, etc.) with pdf $p(\theta|H_A)$.

$$L(H_A) = \prod_{i=1}^{N} p(\theta_i|H_A) \quad -2 \ln(H_A) = \sum_{i=1}^{N} -2\ln(p(\theta_i|H_A))$$

similarly for $L(H_B)$.

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

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

$\Rightarrow$ A reason why I prefer direct $-2\ln\lambda_{A,B}$ approach to "$\Delta\chi^2$" approach.
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 \( \theta_{CS} \) in Collins-Soper frame.

spin-1 \( Z' \)
mass 1.5 TeV

spin-2 \( G^* \)
mass 1.5 TeV
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 $-2\ln \lambda$'s from events:

Gaussian by CLT
mean $= \text{event mean} \times 50$
RMS $= \text{event RMS} \times \sqrt{50}$

(An earlier paper had erroneously assumed that $-2\ln \lambda$ was $\chi^2$.)

(Peak separation) / RMS scales beautifully with $\sqrt{N}$. 
Above is all “pre-data” characterization of the test

How to characterize post-data?

In N-P theory, \( \alpha \) is specified in advance.

Suppose after obtaining data, you notice that with \( \alpha = 0.05 \) previously specified, you reject \( H_0 \), but with \( \alpha = 0.01 \) 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 \emph{p-value} is the smallest value of \( \alpha \) 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.

Interpreting p-values and Z-values

It is crucial to realize that that value of $\alpha$ was typically \textit{not} specified in advance, so p-values do \textit{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 $0^+ \text{ vs. } 0^-$

Paper reported (fixing typo here):

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

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

3) p-value = 0.7 for $0^+$

4) $\text{CL}_s = \frac{(0.72\%)}{(1-0.7)} = 2.4\%$, “a more conservative value for judging whether the observed data are compatible with $0^-$”

FIG. 3 (color online). Expected distribution of $-2\ln\frac{\mathcal{L}_{0^-}}{\mathcal{L}_{0^+}}$ under the pure pseudoscalar and pure scalar hypotheses (histograms). The arrow indicates the value determined from the observed data.
Test of point null vs point alternative, two Gaussians with same $\sigma$, peak separation $\Delta \mu$.

At a glance can see that contours of constant $\lambda_{01}$ 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 $\nu$ flavors in 1989: 3 light $\nu$ 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 $N_\nu$ 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, $N_\nu < 3.9$, 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 $\alpha$. 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 $\alpha$, with relevant set of C.L.’s).

I encourage continuous $\alpha$ approach as part of toolkit. But…Eligio Lisi has explained to me that $\alpha$ is highly correlated with $\Delta m^2$, 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 $\delta$ 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 $\delta$ is treated as a nuisance parameter in the MH determination, then great care is needed.

Providing the MH results as a function of $\delta$ (same $\delta$ in numerator and denominator of LR) would seem to be mandatory, before attempting to “eliminate” $\delta$ by profiling or marginalizing..
Addition of Nuisance Parameter $\delta$ to MH Test

Then I would try:

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

b) Marginalizing: Takes average weighted by prior for $\delta$ of numerator $L$; weighted average of denominator $L$ over $\delta$, 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” $\delta_{CP}$ 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.”


*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:

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 $\alpha \leftrightarrow \text{Is } \theta_0 \text{ in conf. int. for } \theta \text{ with C.L.} = 1 - \alpha$

“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>
<thead>
<tr>
<th>Property of test</th>
<th>Property of corresponding confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size $= \alpha$</td>
<td>Confidence coefficient $= 1 - \alpha$</td>
</tr>
<tr>
<td>Power $= \text{probability of rejecting a}$</td>
<td>Probability of not covering a false value of $\theta = 1 - \beta$</td>
</tr>
<tr>
<td>Most powerful</td>
<td>Uniformly most accurate</td>
</tr>
<tr>
<td>Equal-tails test $\alpha_1 = \alpha_2 = \frac{1}{2} \alpha$</td>
<td>Central interval</td>
</tr>
</tbody>
</table>

\[
\{ \text{Unbiased} \quad 1 - \beta \geq \alpha \} \quad \leftrightarrow 
\]

Bob Cousins, PhyStat-nu Fermilab 2016
Classical Hypothesis Testing (cont.)

“Test for \( \theta = \theta_0 \)” \( \leftrightarrow \) “Is \( \theta_0 \) in confidence interval for \( \theta \)”

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.
Duality: p-values from F-C

Post-data, find the threshold C.L. for which $\theta$ is on the boundary of the F-C confidence interval/region.

I.e., for C.L. lower than that threshold, $\theta$ 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 $\theta$.

In contrast to frequentist method, one does not find a Bayesian credible interval for $\theta$ and test if $\theta$ 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 $\theta=\theta_0$ (or equivalent).
Brings in a whole new set of issues not present in estimation!
(“Can of worms” – Jim Berger)
Dependence on the prior for $\theta$ 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 $\rho$ (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 $\theta \neq \theta_0$: $\delta_{CP}$

$\delta_{CP}$ is bounded similarly to binomial $\rho$, 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:

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 $\text{CL}_S$ (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
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, $\pi \mu$ 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 $\nu$ oscillation searches at NOMAD at CERN
CMS at LHC since 2000 (Deputy Spokes, 2007-2009)

Many committees, some relevant to $\nu$’s, especially:
Lehmann review of NUMI project (1998)
Fermilab PAC 1999-2003
P5 panel 2013-14 ⇒ crash course on all $\nu$ experiments in world
(and their proposed statistical methods)
### Summary of Three Ways to Make Intervals

<table>
<thead>
<tr>
<th></th>
<th>Bayesian Credible</th>
<th>Frequentist Confidence</th>
<th>Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requires prior pdf?</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Obeys likelihood principle?</td>
<td>Yes (exception re Jeffrey’s prior)</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Random variable in “μ ∈ µ µ”</td>
<td>µ</td>
<td>µ µ</td>
<td>µ µ</td>
</tr>
<tr>
<td>Coverage guaranteed?</td>
<td>No</td>
<td>Yes (but over-coverage…)</td>
<td>No</td>
</tr>
<tr>
<td>Provides P(parameter</td>
<td>data)?</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

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”.

Bob Cousins, PhyStat-nu Fermilab 2016
68% intervals by various methods for mean $\mu$ of Poisson process with $n=3$ observed

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

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)
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
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 $H_0 = \text{S.M.}$, and an alternative $H_1 = \text{mSUGRA}$, etc.

- $\alpha$: probability (under $H_0$) of rejecting $H_0$ when it is true, i.e., false discovery claim (Type I error)
- $\beta$: 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)
- $\theta$: parameters in the hypotheses

Competing analysis methods can be compared by looking at graphs of $\beta$ vs $\alpha$ at various $\theta$, and at graphs of $\beta$ vs $\theta$ at various $\alpha$ (power function).
Where to live on the $\beta$ vs $\alpha$ 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 $\alpha$ corresponding to $5\sigma$ 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.)

If Type I error probability \( \alpha \) is specified in a test of hypothesis \( H_0 \) against hypothesis \( H_1 \), then the Type II error probability \( \beta \) is minimized by using as the test statistic the

\[
\lambda = \frac{\mathcal{L}(x|H_0)}{\mathcal{L}(x|H_1)}, \quad \text{and rejecting } H_0 \text{ if } \lambda \leq k_\alpha
\]

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
Tale of two $=5$ effects

$\sigma_{\text{tot}}/\tau$ 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 robustness 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 view is that it could also be given more emphasis in astro/cosmo – some exemplary papers certainly exist!

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.”
“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.”
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 $\mu_u$ such that posterior probability $P(\mu > \mu_u) = 0.1$.

2. Likelihood ratio method for approximate 90% C.L. U.L.:
   find $\mu_u$ such that $L(\mu_u) / L(3)$ has prescribed value.

3. Frequentist one-sided 90% C.L. upper limit:
   find $\mu_u$ 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 $P(n \mid \mu)$ 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!
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:
   \[ \text{UL} = \max(0, 0) + 1.64 \sigma \]

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

3) 1997: Alex Read et al. (LEP)
   \[ \text{CLS} \]

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

5) 2010: Power Constrained Limits;
   Cowan, Cranmer, Gross, Vitells (ATLAS):
   \[ \text{UL} = \max(0, \max(x, x_{PCL}) + 1.64 \sigma) \]
References Cited in Talk Slides


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


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.


