Statistical Quantification of Discovery in Neutrino Physics

David A. van Dyk

Statistics Section, Imperial College London

PhyStat-nu, Fermilab, 2016

Statistical Discovery in Neutrino Physics

I am a statistician, not a neutrino physicists...

- I collaborate with astrophysicists, solar physicists, and particle physicists on statistical methodology.
- First contact with neutrino physics: PhyStat-ν *...3 months ago*

Today:

- Summarize a number of statistical issues that pertain to discovery in neutrino physics *... as discussed in PhyStat-ν, Tokyo*
- Illustrate how they play out in three examples.

Outline

Outline

Motivating Problems

Mass Hierarchy

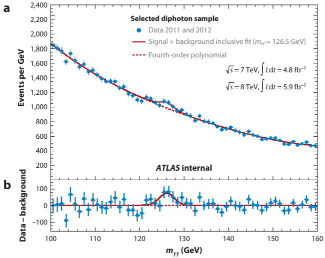
- normal ($\Delta m^2_{32} > 0$) *vs* inverted hierarchy ($\Delta m^2_{32} < 0$)
- $|\Delta m^2_{32}|$ well constrained, degeneracy of sign with θ_{23} or $\delta_{\rm CP}$.

CP-violation

- Is there evidence to counter $\delta_{\rm CP} \in \{0, \pi\}$?
- Current data is limited.

Bump Hunting (e.g., Higgs serach)

- no bump *vs* bump
- location of bump unknown
- What is the bump location if there is no bump?



Outline

Statistical Framework for Discovery

Model / Hypothesis Testing

H_0: The null hypothesis (e.g., no CP-violoation, $\delta_{\mathrm{CP}} = 0$)

H_1: The alternative hypothesis (e.g., CP-violation)

- Without further evidence, H_0 is presumed true.
- "Deciding" on H_1 means scientific discovery: new physics.
- Model Selection: No presumed model. (normal/inverted hierarchy)

Appropriate Statistical Approach Depends on:

- Is H_0 the *presumed* model? *are there more than 2 possible models?*
- Is H_0 a special case of H_1, "nested models"
- Parameters: (i) Unknown values under H_0?
  (ii) No "true value" under H_0?, (iii) Boundary concerns.
- Bayesian vs. Frequentist methods

Statistical Criterion for Discovery

The most common criterion is the p-value,

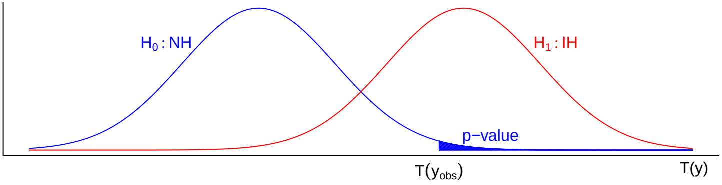
$$\text{p-value} = \Pr\Big(T(y) \geq T(y_{\text{obs}}) \mid H_0\Big)$$

- $T(\cdot)$ is a *Test Statistic*, e.g., $\Delta\chi^2$ or likelihood ratio statistic

Likelihood under H_0

$$\text{Likelihood Ratio Test} = -2\log\frac{\max_\theta p_0(y \mid \theta)}{\max_\theta p_1(y \mid \theta)}$$

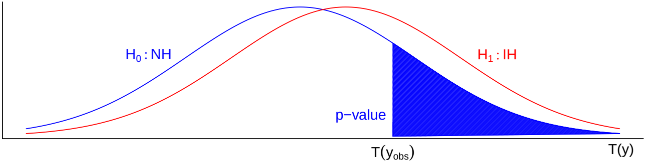
Likelihood under H_1



Computing p-values

The most common criterion is the p-value,

$$\text{p-value} = \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid H_0\right)$$



Requires distribution of $T(y)$ under H_0

- Distributions depend on unknown parameters *(e.g., δ_{CP}, θ_{23})*
- Standard Theory: models nested, all parameters have values under H_0, "large" data set. *... often violated in physics*
- Monte Carlo toys infeasible with 5σ criterion.

Misuse of P-values

The most common criterion is the p-value,

$$\text{p-value} = \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid H_0\right) \text{ with } T = \text{test statistic}$$

But....

Misuse of P-values

The most common criterion is the p-value,

$$\text{p-value} = \Pr\Big(T(y) \geq T(y_{\text{obs}}) \mid H_0 \Big) \quad \text{with } T = \text{test statistic}$$

But....


nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive

Archive > Volume 519 > Issue 7541 > Research Highlights: Social Selection

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

Misuse of P-values

The most common criterion is the p-value,

$$\text{p-value} = \Pr\Big(T(y) \geq T(y_{\text{obs}}) \mid H_0\Big) \quad \text{with } T = \text{test statistic}$$

But....

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive

Archive > Volume 519 > Issue 7541 > Research Highlights: Social Selection

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For

Archive > Volume 531 > Issue 7593 > News > Article

NATURE | NEWS

Statisticians issue warning over misuse of *P* values

Policy statement aims to halt missteps in the quest for certainty.

Monya Baker

07 March 2016

(ASA Statement on Statistical Significance and P-values)
February 5, 2016

The Problem with P-values

The misuse of P-values:

- **Do not measure relative likelihood of hypotheses.**
- Large p-values do not validate H_0.
- May depend on bits of H_0 that are of no interest.
- **Single filter** for publication / judging quality of research.
- **Should be viewed as <u>a</u> data summary, not <u>the</u> summary**

Reviewers, Editors, and Readers want a simple black-and-white rule: $p < 0.05$, or $> 5\sigma$.

But, statistics is about quantifying uncertainty, not expressing certainty.

A Bayesian Criterion for Discovery

To determine mass hierarchy, suppose we find

$$\text{p-value} = \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid \text{NH}\right) = 0.0001$$

Questions

- Can we conclude NH is unlikely?
- Does Pr(data | NH) small imply Pr(NH | data) is small?

Order of conditioning matters!

Consider $\Pr(A \mid B)$ and $\Pr(B \mid A)$ with

A: A person is a woman.

B: A person is pregnant.

Bayesian Methods

Bayes Theorem

$$\Pr(\mathrm{NH} \mid \mathrm{data}) = \frac{\Pr(\mathrm{data} \mid \mathrm{NH}) \Pr(\mathrm{NH})}{\Pr(\mathrm{data} \mid \mathrm{NH}) \Pr(\mathrm{NH}) + \Pr(\mathrm{data} \mid \mathrm{IH}) \Pr(\mathrm{IH})}$$

Bayesian methods

- have cleaner mathematical foundations
- more directly answer scientific questions

*... but they depend on **prior distributions***

- $\Pr(\mathrm{NH}) =$ probability of NH before seeing data.

Prior distributions must also be specified for model parameters.

The Problem with Priors

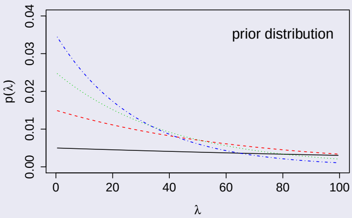
Bayesian Criteria for Discovery:

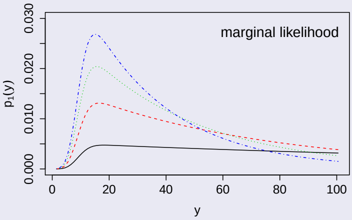
$$\text{Bayes Factor} = \frac{p_0(y)}{p_1(y)} \text{ with } p_i(y) = \int p_i(y|\theta)p_i(\theta)d\theta.$$

$$\Pr(H_0 \mid y) = \frac{p_0(y)\pi_0}{p_0(y)\pi_0 + p_1(y)\pi_1} = \frac{\pi_0}{\pi_0 + \pi_1(\text{Bayes Factor})^{-1}}$$

Example: (simplified) Higgs search

Likelihood: $y|\lambda \sim \text{Poisson}(10+\lambda)$ Test: $\lambda = 0$ vs $\lambda > 0$





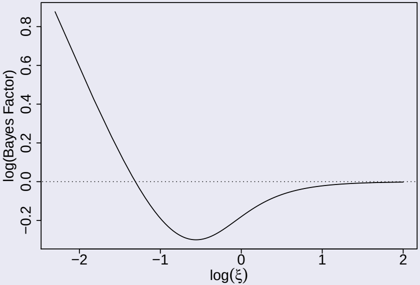
Value of $p_1(y)$ depends on prior!

Choice of Prior Matters!

Bayes Factor

$$H_0: \quad y \sim \mathrm{Poisson}(10).$$
$$H_1: \quad y \sim \mathrm{Poisson}(10+\lambda).$$
$$\text{with } \lambda \sim \exp(\xi)$$

- Observe $y = 15$
- log(Bayes Factor)



Must think hard about choice of prior and report!

Frequentist vs Bayesian: Does it Matter?

Model Testing and Model Selection

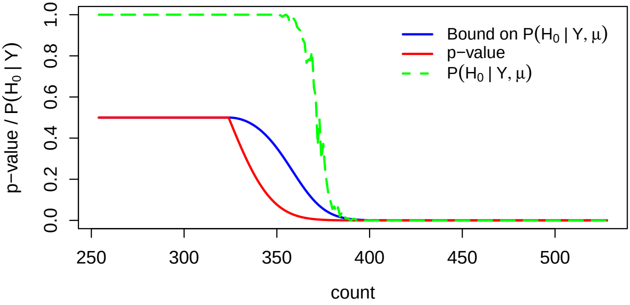
- Frequency and Bayesian methods **may not agree.**
  - Bayes automatically penalizes larger models *(Occam's Razor)*
  - and adjusts for trial factors / look elsewhere effect.
  - Choice of prior distribution **is often critical.**
- **Problem cases:** Dimension of model parameters differ.
  - CP-violation: $H_0 : \delta_{\text{CP}} \in \{0, \pi\}$ vs. $H_1 : \delta_{\text{CP}} \notin \{0, \pi\}$.
  - Higgs search: location and intensity of bump above bkgd.
- Anti-conservative: p-value $\ll \Pr(H_0 \mid y)$.
- *Remember:*

  p-value and $\Pr(H_0 \mid y)$ *quantify different things!*

Interpreting p-value as $\Pr(H_0 \mid y)$ *may significantly overstate evidence for new physics.*

Example: Searching for a bump above background.

E.g., in toy version of Higgs search with known mass...



.... but researchers interpret p-value as $\Pr(H_0 \mid y)$*.*

Solution: Report both.

5σ Discovery Threshold

5σ is required for "discovery"

- High profile false discoveries led to conservative threshold
- Treat Higgs mass as known (multiple-testing)
- "What would you have done had you had different data"
- **Calibration, systematic errors, and model misspecification**
- Of course **cranking up to** 5σ **does not address these issues**

"In particle physics, this criterion has become a convention ... but should not be interpreted literally [1]*."*

At PhyStat-nu (Tokyo)....

Cousins: *Two* 3.5σ *results are better than one* 5σ *result.*

van Dyk: *Calibrated* 3.5σ *result better than uncalibrated* 5σ*.*

[1] Glossary in the *Science* review of the 2012 CMS and ATLAS discoveries.

Outline

Normal Hierarchy versus Inverted Hierarchy

Non-nested parameterized models

$$\begin{aligned} H_0 &: \text{normal hierarchy} \qquad \text{i.e., } \Delta m^2_{32} \le 0 \\ H_1 &: \text{inverted hierarchy} \qquad \text{i.e., } \Delta m^2_{32} > 0 \end{aligned}$$

Computing a p-value using LRT

- Non-nested models. If no unknown parameters in either model:
  - LRT follows a Gaussian distribution under H_0 or H_1.
- With unknown parameters (e.g., Δm^2_{32}, δ_{CP}, θ_{23}):
  - Std theory (Wilks, Chernoff) does not apply: dist'n of LRT unknown.
  - What is null distribution of $\hat{\delta}$ when fitting H_1?
  - Some results, but strong assumptions *(Blennow, et al. arXiv:1311.1822)*
    Apply to reactor neutrino experiments, not accelerator experiments involving δ_{CP} (Emilo Ciuffoli).
  - Low power owing to degeneracy.
  - What about uncertainty in $|\Delta m^2_{32}|$?

Are we back to Monte Carlo (toys)? at 5σ??

Is There an Easier Solution?

Two paradigms for statistical inference:

Likelihood: inference based on $p(y \mid \theta)$. *... and LRT, p-value, etc.*

Bayesian: inference based on $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$.

Model Fitting

- Specify one model, fit parameters, estimate uncertainty.
- Frequency and Bayesian methods tend to agree.
- Choice of prior distribution is often not critical.

Some "model selection" can be accomplished via model fitting, e.g., confidence intervals.

Normal versus Inverted Hierarchy: Easier Way?

Non-nested parameterized models

$$H_0 : \text{normal hierarchy} \qquad \text{i.e., } \Delta m^2_{32} \leq 0$$
$$H_1 : \text{inverted hierarchy} \qquad \text{i.e., } \Delta m^2_{32} > 0$$

Is there an easier solution??

Why not just compute $\Pr(H_0 \mid y) = \Pr(\Delta m^2_{32} \leq 0 \mid y)$?

In this case Bayes Criterion is particularly easy:

$$\text{Posterior Odds} = \frac{\Pr(\Delta m^2_{32} \leq 0 \mid y)}{\Pr(\Delta m^2_{32} > 0 \mid y)}$$

...model fitting with Δm^2_{32} a free parameter.

One model and one prior, easy to compute, not sensitive to prior... what's not to like?

Bayesian solution is easier in this case.

CP-violation

Test: $H_0 : \delta_{\mathrm{CP}} \in \{0, \pi\}$ versus $H_1 : \delta_{\mathrm{CP}} \notin \{0, \pi\}$

p-value

- Standard theory (Wilks, Chernoff) applies...
  but insufficient data for asymptotics.
- Monte Carlo (toys) required to assess p-value.
- More data required! (For 5σ??)

Posterior Odds or Bayes Factor (JOHANNES BERGSTRÖM)

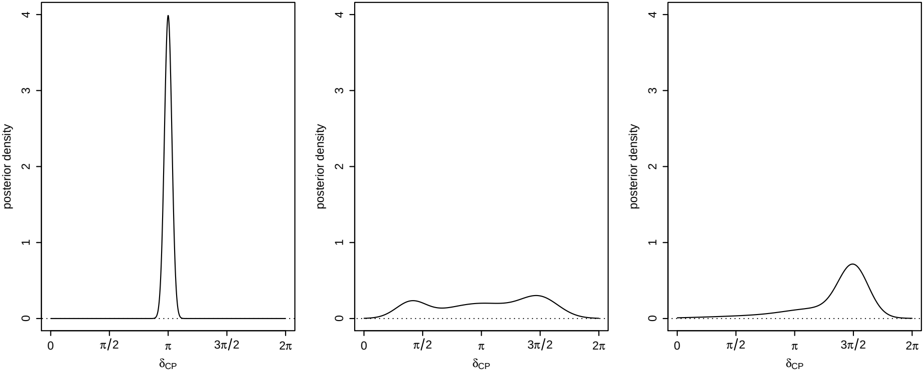
- Sensitive to prior on δ, but finite support.

Again, Bayesian solution is easier (with limited data).

Still Easier:

- Report a confidence/credible interval for δ_{CP}.
- Employ model fitting rather than model selection.

Assessing CP-violation via Model Fitting



Is data consistent with $\delta_{\mathrm{CP}} \in \{0, \pi\}$??

Higgs Search: Is a Bayes Factor Possible?

Basic Model:

$$f(y_i|\theta) = (1-\lambda)f_0(y_i|\alpha) + \lambda f_1(y_i|\mu)$$
$$= \text{background} + \text{Higgs}$$

P-values are anti-conservative.
What about $\Pr(H_0 \mid y)$*?*

Challenge: Setting priors on λ and μ.

- *Prior on* α: Luckily, $\Pr(H_0 \mid y)$ is not sensitive to this prior.

Lower Bound on Bayesian evidence for H_0

- P-values tend to favor H_1 more strongly than $\Pr(H_0 \mid y)$.

[At least when H_0 is "precise".]

- *Prior on* λ: Use a parameterized prior, $\lambda \sim p(\lambda \mid \beta)$,

$$\bar{p}_1(y \mid \mu) = \sup_{\beta} \int p_1(y \mid \lambda, \mu) p(\lambda \mid \beta) \mathrm{d}\lambda$$

$$\Pr(H_0 \mid y, \mu) = \frac{\pi_0 p_0(y)}{\pi_0 p_0(y) + \pi_1 p_1(y \mid \mu)} \geq \frac{\pi_0 p_0(y)}{\pi_0 p_0(y) + \pi_1 \bar{p}_1(y \mid \mu)}$$

Prior on μ

...or more generally, parameters unidentified under H_0

Local $p(H_0|y)$: $\inf_\mu p(H_0 \mid y, \mu)$
Global $p(H_0|y)$: properly average over $p(\mu)$

Like global p-value, averaging over $p(\mu)$ penalizes wide search

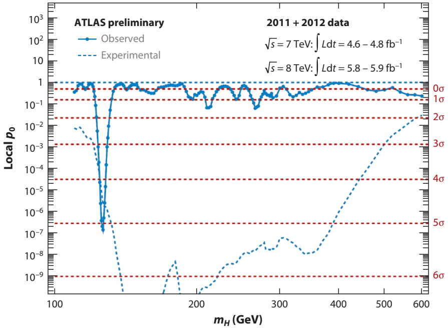
$$
\begin{aligned}
p_1(y) &= \int p_1(y \mid \mu) p(\mu) \mathrm{d}\mu \leq \sup_\mu p_1(y \mid \mu) \\
\Pr(H_0 \mid y) &= \frac{\pi_0 p_0(y)}{\pi_0 p_0(y) + \pi_1 p_1(y)} \geq \frac{\pi_0 p_0(y)}{\pi_0 p_0(y) + \pi_1 \sup_\mu p_1(y \mid \mu)} \\
&= \inf_\mu p(H_0 \mid y, \mu)
\end{aligned}
$$

- Simplest choice of $p(\mu)$ is uniform over the search region.
- Results in a "Bonferroni like correction" to local $p(H_0|y)$.

Is there a better choice??

[skip?]

Choice of Prior on μ



Sensitivity of detector varies

- Do we want to search thoroughly everywhere?
- E.g., BF unlikely to favor H_1 for $\mu > 500$.
- Good choice:

$$\text{detection prior} \propto p(\text{Detection} \mid \mu)p(\mu) \propto p(\mu \mid \text{Detection}).$$

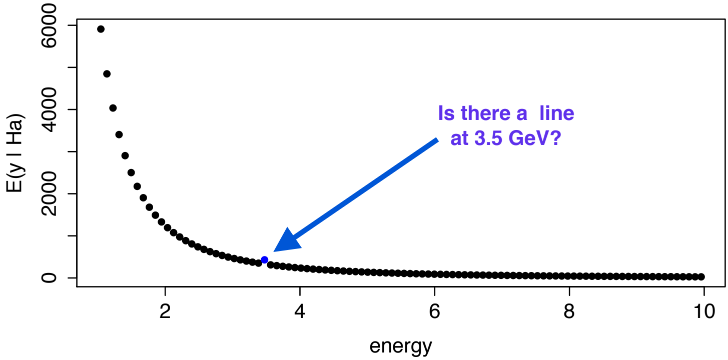
Example: Are P-values Biased in Favor H_1?

Model:

$y_i \overset{\text{indep}}{\sim} \textsc{Poisson}\Big(f_0(\alpha, i) + \lambda f_1(\mu, i)\Big)$

Test: $H_0 : \lambda = 0$ vs $H_0 : \lambda > 0$

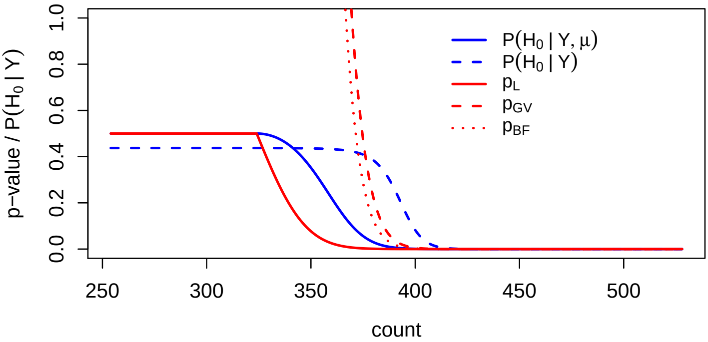
- f_0 = power law
- $f_1 = \mathcal{I}\{i = \mu\}$
- 100 bins



Natural Bayesian correction for multiple testing

- Varying the count in the line bin (3.5 GeV).
- The expected count in this bin under H_0: 330.

Compare local/global p-value (red); local/global Bayes (blue), p-value vs Bayes.



Prior on μ naturally and simply corrects for the “look elsewhere effect”

Outline

Frequentist or Bayesian?

Do you have to choose??

- Bayes prescribes methodology.
- Frequentists evaluate methods.
- Frequency evaluation of Bayesian methods.
- Model fitting: often little difference in fits and errors.
- Why not control rate of false detection
  and assess probability of new physics?
- Why throw away half of your tool box?

I'm impressed with the openness of neutrino researchers to both Bayesian and Frequency based methods.

- Lots of Bayesian and Frequentist proposals at PhyStat-ν.
- My experience with cosmologists and particle physicists.

Strategies

What is a physicists to do?

- Controlling false discovery is critical in physical sciences.
- Comparing p-values with a predetermined significant level can control false discovery.... *if used with care, e.g., no cherry picking!*
- When confronted with small p-values researchers *...even statisticians!!...* may believe H_0 is unlikely.
- Bayesian solutions can better quantify likelihood of H_0 / H_1.
- **Solution:** Compute both *global* p-value *and* Bayes Factor.

But be Careful...

1. *quantification of p-values in non-standard problems*
2. *choice and validation of prior distributions*

remain challenging!