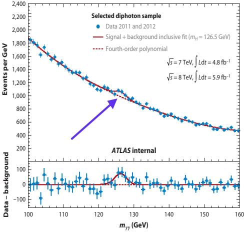
Quantification of Discovery in Astrophysics
Frequentist and Bayesian Perspectvies

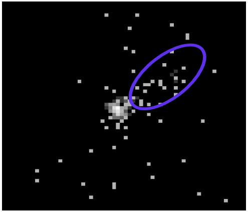
David A. van Dyk

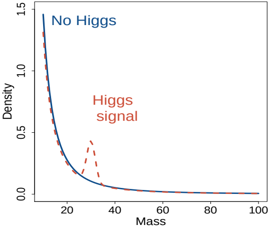
Statistics Section, Imperial College London

HEAD Meetings 2017, Sun Valley, Idaho

Searching for Structure







- **Bump Hunting:** Is there a bump? *E.g., spectral line or Higgs Boson.*
- Are circled photons due to background or a quasar jet?

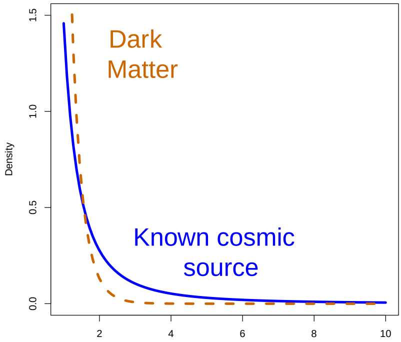
Scientific and Statistical Issues

- High-stakes science: discovery vs. estimation.
- Model selection is much harder than estimation.
- Frequentist and Bayesian methods: different conclusions.
- Is a non-partisan approach possible?

Comparing Models

Compare two different models

- Is a spectral continuum
  - a Bremsstrahlung *or*
  - a Power Law?
- Is γ-ray emmission due to
  - known cosmic sources *or*
  - dark matter?



Neutrino 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 other parameters.

Outline

Outline

Statistical Framework for Discovery

Model / Hypothesis Testing

H_0: The null hypothesis (e.g., no jet; known cosmic sources)

H_1: The alternative hypothesis (e.g., jet; dark matter)

- 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? or 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,

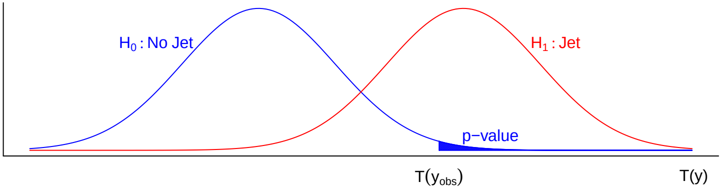
$$\text{p-value} = \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid H_0\right)$$

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

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

Likelihood under H_0

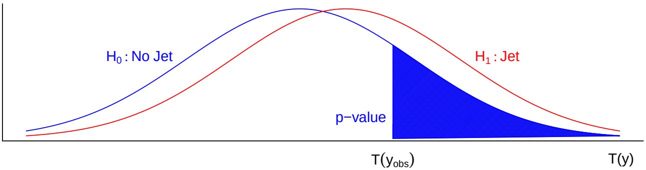
Likelihood under H_1



Computing p-values

The most common criterion is the p-value,

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



Requires distribution of $T(y)$ under H_0

- Distributions depend on unknown parameters

  (e.g., continuum / background parameters)

- Standard Theory: models nested, all parameters have values under H_0, "large" data set. *... often violated in astro/physics*
- Monte Carlo / Bootstrap infeasible with 5σ criterion.

Misuse of P-values

The most common criterion is the p-value,

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

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,

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

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 a data summary, not the 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.

Outline

A Bayesian Criterion for Discovery

When trying to detect a jet, suppose we find

$$\text{p-value} = \Pr\Big(T(y) \geq T(y_{\text{obs}}) \mid \text{No Jet}\Big) = 0.0001$$

Questions

- Can we conclude that there is probably a Jet?
- Does Pr(Data | No Jet) small imply
  Pr(No Jet | Data) is small?

Order of conditioning matters!

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

A:

B:

A Bayesian Criterion for Discovery

When trying to detect a jet, suppose we find

$$\text{p-value} = \Pr\Big(T(y) \geq T(y_{\text{obs}}) \mid \text{No Jet}\Big) = 0.0001$$

Questions

- Can we conclude that there is likely a Jet?
- Does Pr(Data | No Jet) small imply
  Pr(No Jet | 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

$$\Pr(\text{Jet} \mid \text{Data}) = \frac{\Pr(\text{Data} \mid \text{Jet}) \Pr(\text{Jet})}{\Pr(\text{Data} \mid \text{Jet}) \Pr(\text{Jet}) + \Pr(\text{Data} \mid \text{No Jet}) \Pr(\text{No Jet})}$$

Bayesian methods

- have cleaner mathematical foundations
- more directly answer scientific questions

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

- $\Pr(\text{Jet}) =$ probability of a Jet before seeing data.

Prior distributions must also be specified for model parameters.

The Problem with Priors

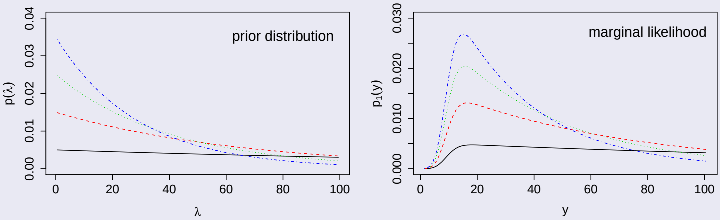
Bayesian Criteria for Discovery:

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

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

Example: (simplified) Higgs search

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



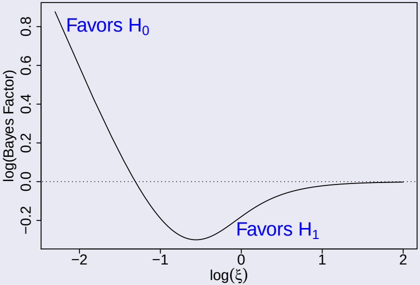
Value of $p_1(y)$ depends on prior!

Choice of Prior Matters!

Bayes Factor

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

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



Must think hard about choice of prior and report!

Frequentist vs Bayesian: Does it Matter?

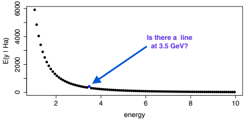
Model Testing and Model Selection

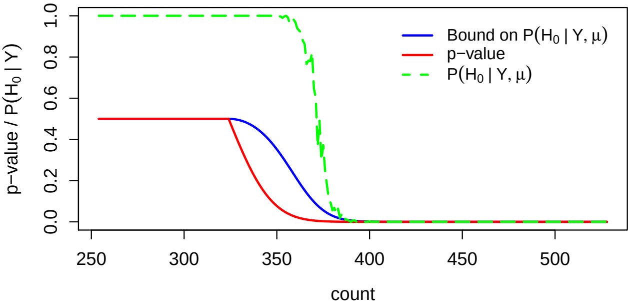
- Frequency and Bayesian methods **may not agree.**
  - Bayes automatically penalizes larger models *(Occam's Razor)*
  - and adjusts for trial factors / look elsewhere effect.
- Choice of prior distribution **is often critical.**
- **Difficult cases:** Dimension of model parameters differ.
  - Higgs search: location and intensity of bump above bkgd.
  - Added structure in image.
- 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 discovery.*

Example: Searching for a bump above background.





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

Solution: Report both.

Outline

Normal Hierarchy versus Inverted Hierarchy

Non-nested parameterized models

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

... recall $|\Delta m^2_{32}|$ is well constrained.

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., Bremsstrahlung vs. Power Law)*
  - Std theory (Wilks, Chernoff) does not apply: dist'n of LRT unknown.
  - Problem-specific theory, requires strong assumptions.
  - What about uncertainty in $|\Delta m^2_{32}|$?
  - PPP-values / parametric bootstrap, *(e.g., Protassov et al., ApJ, 2002).*

Back to Monte Carlo / Bootstrap? 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

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

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:

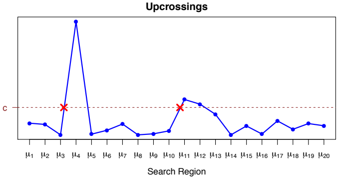
$$\text{Posterior Odds} = \frac{\Pr(\Delta m^2_{32} \leq 0 \mid y)}{\Pr(\Delta m^2_{32} > 0 \mid y)}$$

...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.

Bump Hunting: Frequency vs Bayes



Frequency Methods:

- Fixed bump location:
  standard methods apply
- Multiple testing problem.

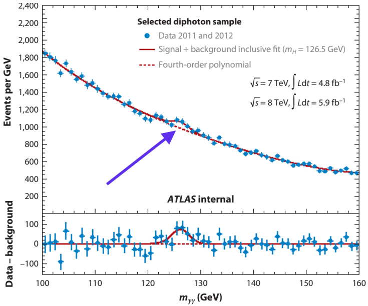
(Algeri, van Dyk et al., 2016)

Bayesian Methods:

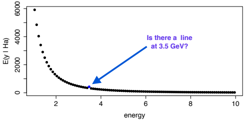
– *Prior specification is key.*

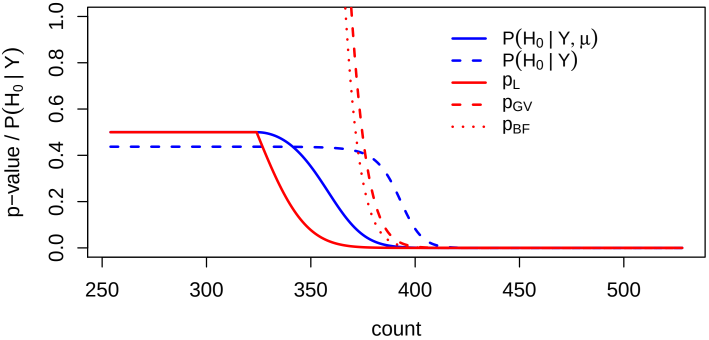
- Intensity parameter
  P-values favor H_1
  Use prior most favorable for H_1.
  Bound $\Pr(H_0 \mid \text{Data})$.
- Location:
  Prior automatically corrects
  for multiple testing.

(van Dyk and Jones, 2017+)



Bump Hunting: Frequency vs Bayes





Prior on location naturally and simply corrects for multiple testing.

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?

Be open to both Bayesian and Frequency based methods.

- Now lots of Bayesian and Frequentist methods in HEA.
- My experience with cosmologists and particle physicists.

Strategies

What is a astrophysicist 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!

References

Protassov, R., van Dyk, D., Connors, A., Kashyap, V., Siemiginowska, A. (2002).
Statistics: Handle with Care, Detecting Multiple Model Components with LRT.
The Astrophysical Journal, **571**, 545–559.

van Dyk, D. A. (2014).
The Role of Statistics in the Discovery of a Higgs Boson.
Annual Review of Statistics and Its Application, **1**, 41–59.

Stein, N. M., van Dyk, D. A., Kashyap, V. L., and Siemiginowska, A. (2015).
Detecting Unspecified Structure in Low-Count Images.
The Astrophysical Journal, **813**, 66 (15pp).

Algeri, S., Conrad, J., and van Dyk, D. A. (2016).
Comparing Non-Nested Models in the Search for New Physics.
Monthly Notices of the Royal Astronomical Society: Letters, **458 (1)**, L84-L88.

Algeri, S., van Dyk, D. A., Conrad, J., and Anderson, B. (2016).
Methods for Correcting the Look-Elsewhere Effect in Searches for New Physics.
Journal of Instrumentation, **11**, P12010.

Algeri, S. and van Dyk, D. A. (2017+).
Testing one Hypothesis Multiple Times.
In preparation.