Dangers of Frequentist Estimates of the “False Alarm Probability” and Other Pretenders

Eric B. Ford
Penn State Astronomy & Astrophysics Dept.
Center for Astrostatistics
Center for Exoplanets & Habitable Worlds
Institute for CyberScience

July 19, 2016
2016 Sagan Workshop

Dangers of Frequentist Estimates of the “False Alarm Probability” and Other Pretenders

Eric B. Ford

drawing extensively from Tom Loredo
http://astrostatistics.psu.edu/RLectures/cast16-PValueNote.pdf
http://www2.stat.duke.edu/~berger/p-values.html

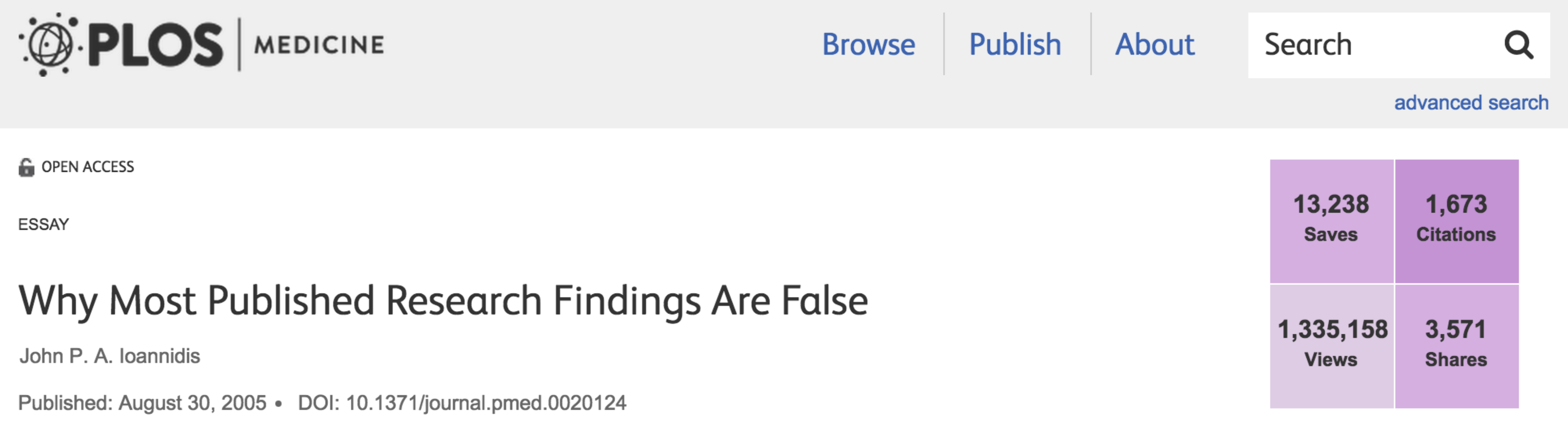
July 19, 2016
2016 Sagan Workshop

Goals: Improve Quality of Your Science

- Recognize ad hoc and/or unjustified statistical methodology
- Recognize misleading language
- Increase awareness of better methods for model comparison
- Increase reliability of scientific conclusions

Why Astronomy? Exoplanets?

- Learn about nature
- Train future generation of scientists, engineers, entrepreneurs,…, policy makers.
- Engage public with the power of science



ESSAY

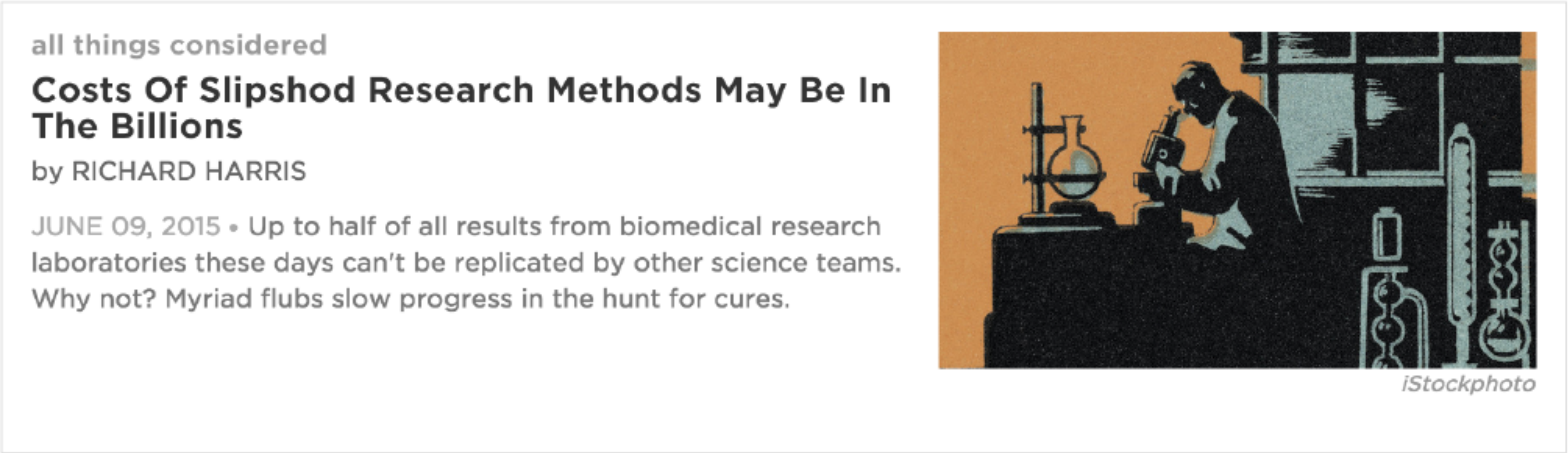
Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

13,238 Saves	1,673 Citations
1,335,158 Views	3,571 Shares

- *"Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true."*



all things considered

Costs Of Slipshod Research Methods May Be In The Billions

by RICHARD HARRIS

JUNE 09, 2015 • Up to half of all results from biomedical research laboratories these days can't be replicated by other science teams. Why not? Myriad flubs slow progress in the hunt for cures.

iStockphoto

Laboratory research seeking new medical treatments and cures is fraught with pitfalls: Researchers can inadvertently use bad ingredients, design the experiment poorly, or **conduct inadequate data analysis**. Scientists working on ways to reduce these sorts of problems have put a staggering price tag on research that isn't easy to reproduce: $28 billion a year.

Aim for Reproducible Research

- Reproducibility is fundamental tenet of the scientific process
- Reproducible Research has several aspects
  - Precise & accurate descriptions on methods
  - Inadequate attention to experimental design
  - Availability of data, materials, instruments, algorithms, codes, etc.
  - Validity of statistical inference
- Credibility of science depends on it

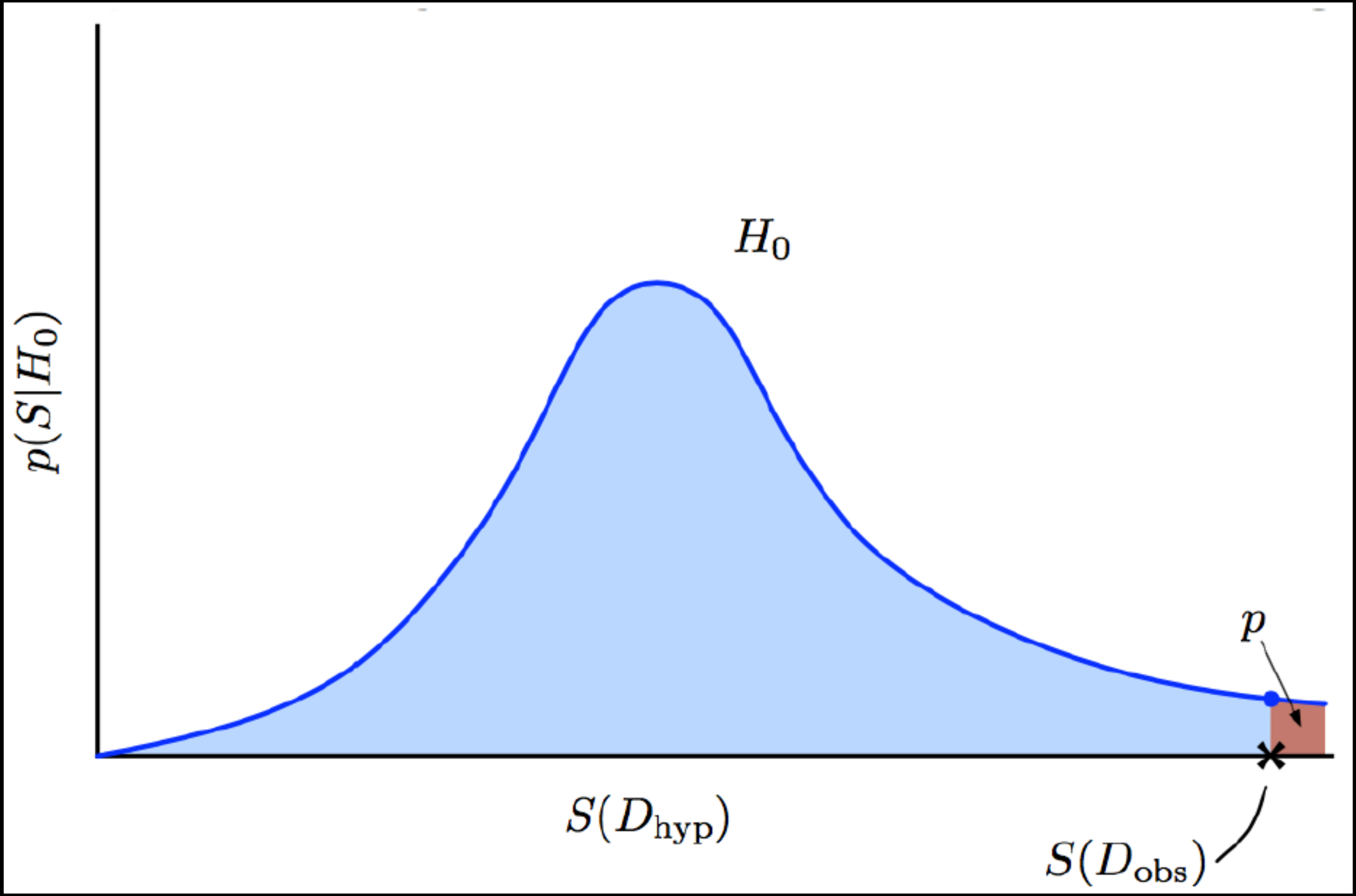
Reproducibility & Exoplanets

Would the field of exoplanet emerge unscathed if subject to a similar level of scrutiny as devoted to life sciences?

What does it mean to “Discover a Planet”?

- *Frequentist* approach:
  Reject the null hypothesis that a simpler model without the planet could reasonably explain your data
- *Bayesian* approach:
  The Evidence for a model with the planet is significantly greater than the alternative models.

What is a p-value?



T. Loredo

A valid interpretation of a p-value

“In a world with absolutely no exoplanets, with a threshold set so we wrongly claim to detect planets 100*p% of the time, this data would be judged a detection, and it would be the data providing the weakest evidence for a planet in that world.”

- That’s a mouthful, so we say “p-value”.
- But’s less precise language makes it too easy to misinterpret p-value.
- E.g., “False alarm rate” is very misleading.

T. Loredo

Misuses of *p*-values in Exoplanet Literature

- “This detection has a signal-to-noise ratio of [X.X] with an empirically estimated upper limit on false alarm probability of 1.0%.”
- “...the false alarm probability for this signal is rather high at a few percent”
- “This signal has a false alarm probability of <4% and is consistent with a planet of minimum mass...”
- “We find a false-alarm probability $<10^{-4}$ that the RV oscillations attributed to [STAR]b and [STAR]c are spurious effects of noise and activity.”

T. Loredo

What’s wrong with stating p-value as a False Alarm Probability?

"*This signal, with $S(D)=S_{obs}$, has a FAP of p…*“

- *p* is not a property of **this** signal; rather, it's the size of the **ensemble** of possible null-generated signals with $S(D)>S_{\text{obs}}$.
- *Every one* of those signals is a false alarm: each one has a FAP=1 in the context producing the *p*-value!
- For any signal to have FAP≠1, alternatives to the null must sometimes act; the FAP will depend on how often they do (and what they are)

T. Loredo

Why report a p-value?

- *The main virtual of a p*-value is that *p* is uniformly distributed under the null hypothesis.

- Provide “*p*-value” to give a measure of how “surprisingness” under the null hypothesis (but not how surprising in a world with alternatives)

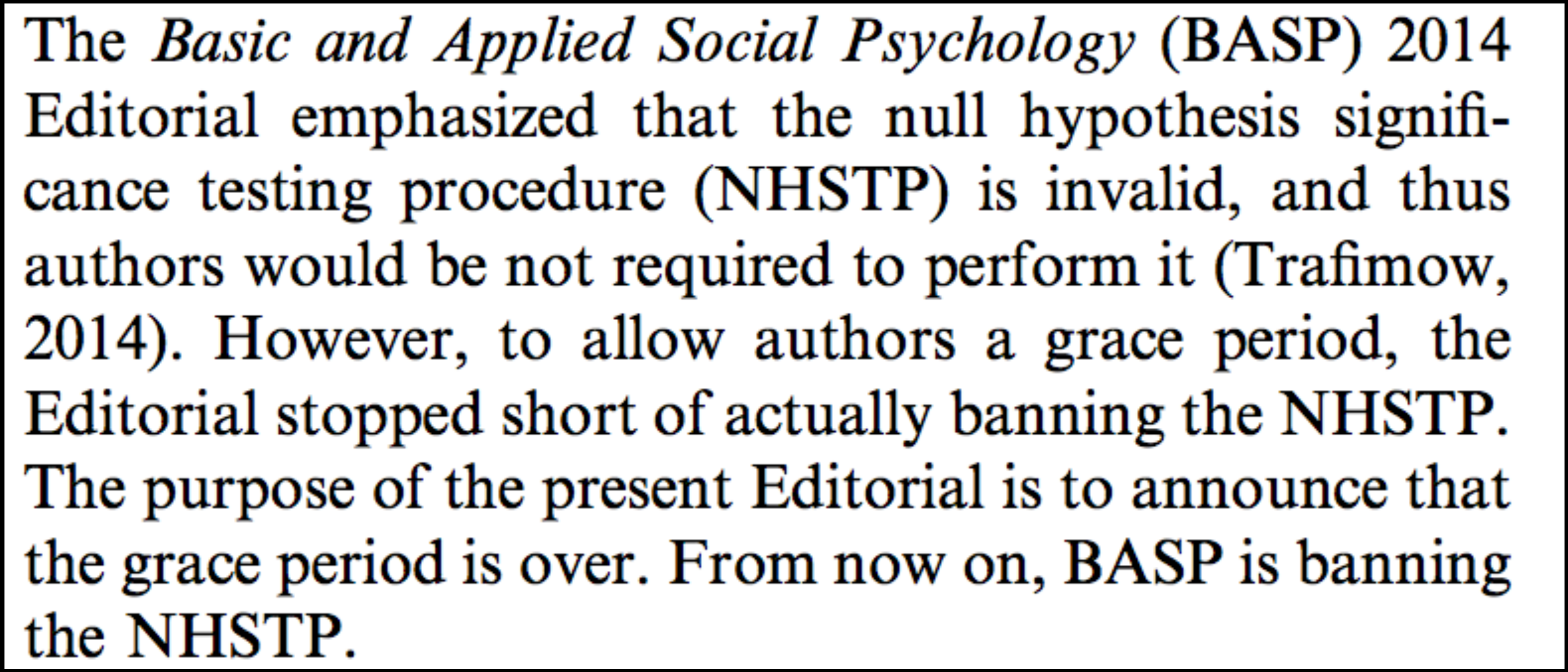
T. Loredo

Why need to be careful about *p*-value?

- A *p*-value is not an easily interpretable measure of the weight of evidence against the null hypothesis.
- It does not measure how often the null will be wrongly rejected among similar data sets
- A naive false alarm interpretation typically overestimates the evidence
- For fixed *p*-value, the weight of the evidence decreases with increasing sample size

T. Loredo

Some Journals are Taking this Seriously



The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

Basic & Applied Social Psychology: 37: 1-2, 2015

Entries to the literature

- "402 Citations Questioning the Indiscriminate Use of Null Hypothesis Significance Tests in Observational Studies" (Thompson 2001) [web site]

- *The significance test controversy: a reader* (ed. Morrison & Henkel 1970, 2006) [Google Books]

- "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" (Berger 2003 with discussion; 2001 Fisher Lecture), *Statistical Science*, **18**, 1–32

- "Odds Are, It's Wrong: Science fails to face the shortcomings of statistics" (By Tom Siegfried 2010) [*Science News*, March 2010]

- "Scientific method: Statistical errors" (By Regina Nuzzo 2014) [*Nature* news feature, Feb 2014]

- "The ASA's statement on p-values: context, process, and purpose" [*The American Statistician*, March 2016]

Demonstration that *p*-value is not a FAP

Model: $x_i = \mu + \epsilon_i$, $(i = 1 \text{ to } n)$ $\quad \epsilon_i \sim N(0, \sigma^2)$

Null hypothesis, H_0: $\quad \mu = \mu_0 = 0$

Test statistic:

$$t(x) = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$$

p-value:

$$\begin{aligned} p(t|H_0) &= \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \\ p\text{-value} &= P(t \geq t_{\rm obs}) \end{aligned}$$

T. Loredo based on Berger (2003)

For this simple model, we can compute critical *p*-values analytically

(but this is a very bad idea for most real problems)

t	p-value
1	0.317
2	0.046
3	0.003

$$t(x) = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$$

$p = .05 \rightarrow$ "significant"

$p = .01 \rightarrow$ "highly significant"

maybe for life or social scientists:

$p = 10^{-3}$ "significant to typical(?) astronomers"

$p = 10^{-12}$ "significant to typical(?) physicists"

T. Loredo based on Berger (2003)

Collect the p-values from a large number of tests in situations where the truth eventually became known, and determine how often H_0 is true at various p-value levels.

- Suppose that, overall, H_0 was true about half of the time.
- Focus on the subset with $t \approx 2$ (say, $[1.95, 2.05]$ so $p \in [.04, .05]$, so that H_0 was rejected at the 0.05 level.
- Find out how many times in that subset H_0 turned out to be true.
- Do the same for other significance levels.

T. Loredo based on Berger (2003)

A Monte Carlo experiment

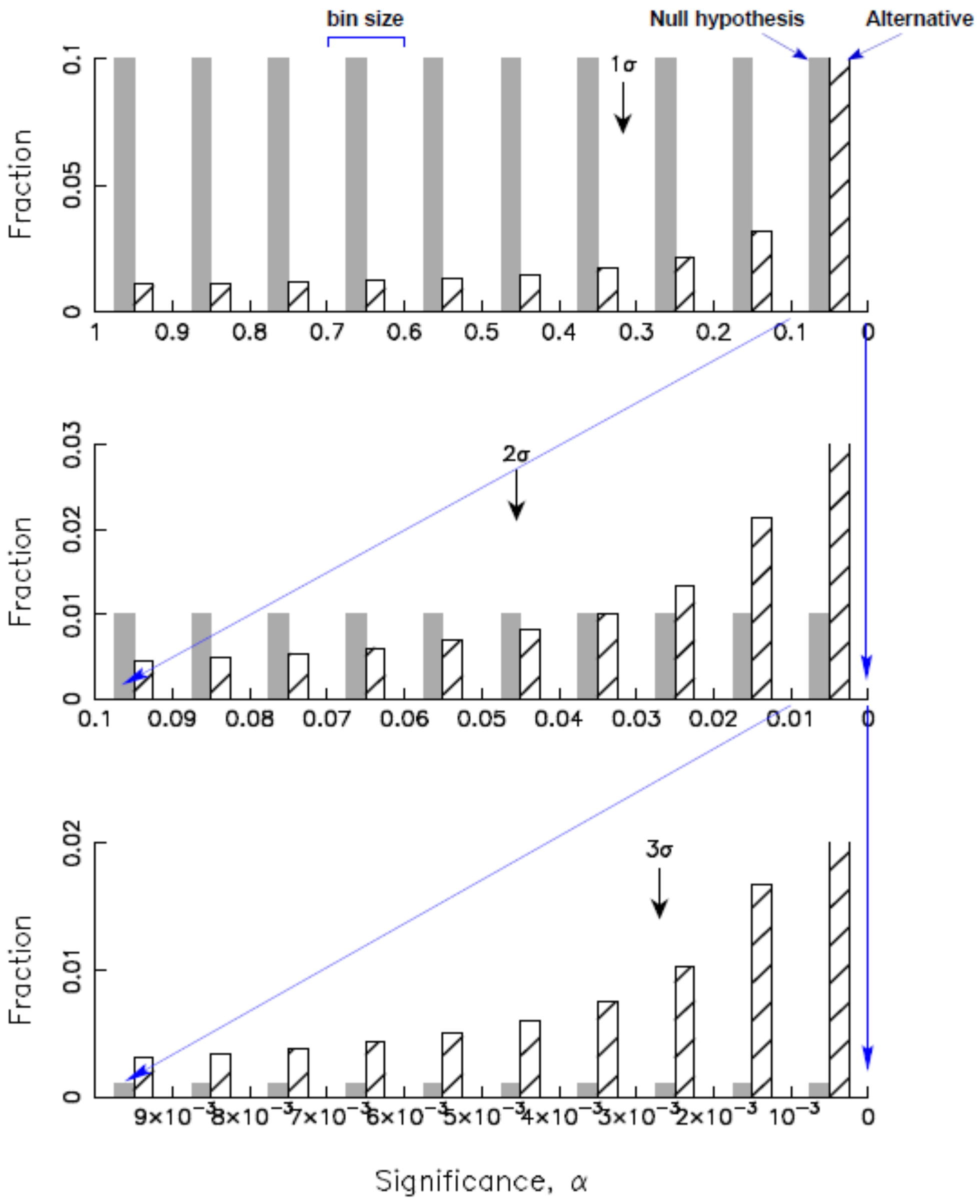
- Choose $\mu = 0$ OR $\mu \sim N(0, 4\sigma^2)$ with a fair coin flip*
- Simulate n data, $x_i \sim N(\mu, \sigma^2)$ (use $n = 20$, 200, 2000)
- Calculate $t_{\rm obs} = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$ and $p(t_{\rm obs}) = P(t > t_{\rm obs}|\mu = 0)$
- Bin $p(t)$ separately for each hypothesis; repeat

Compare how often the two hypotheses produce data with a 1–, 2–, or 3–σ effect.

*A neutral assumption that gives alternatives a “fair” chance and may *over*estimate the evidence against H_0 in real settings where the null is more prevalent

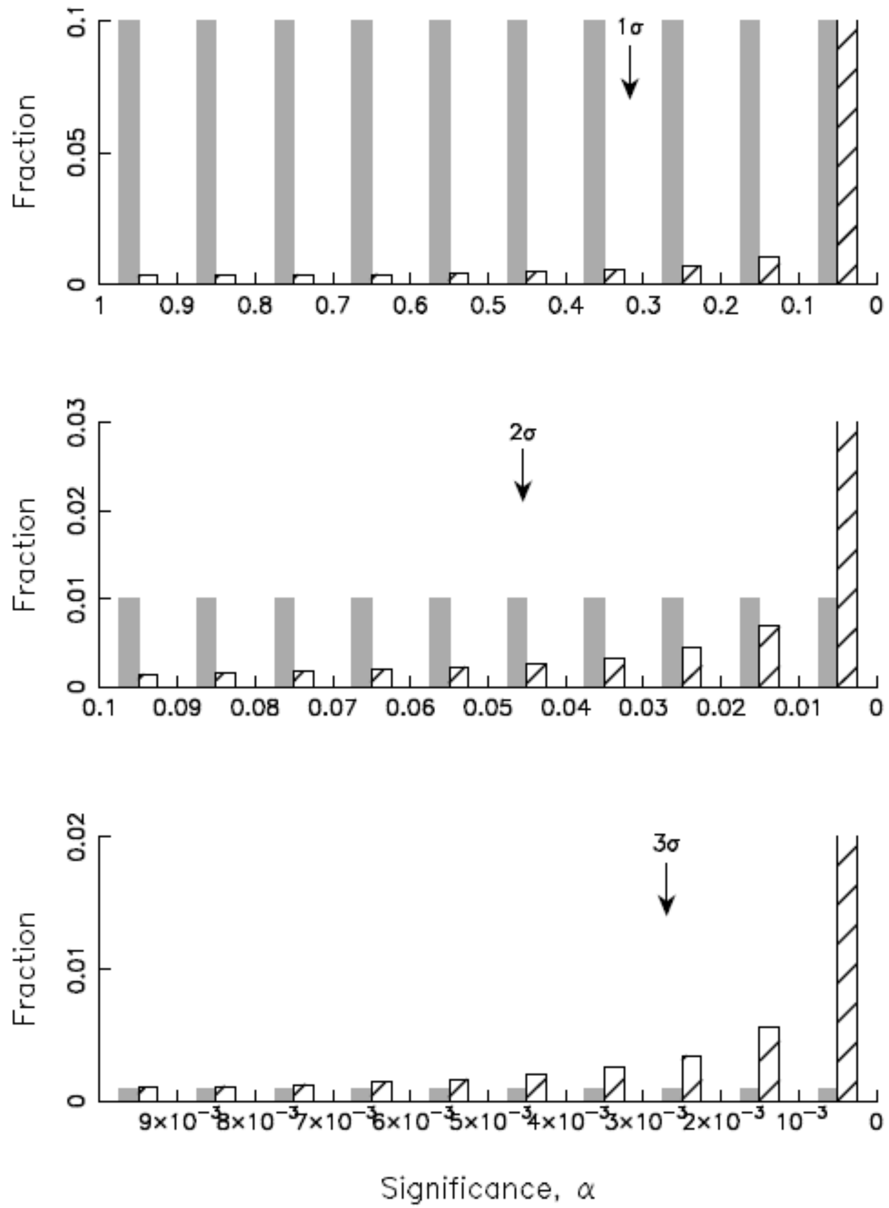
T. Loredo based on Berger (2003)

Significance Level Frequencies, $n = 20$



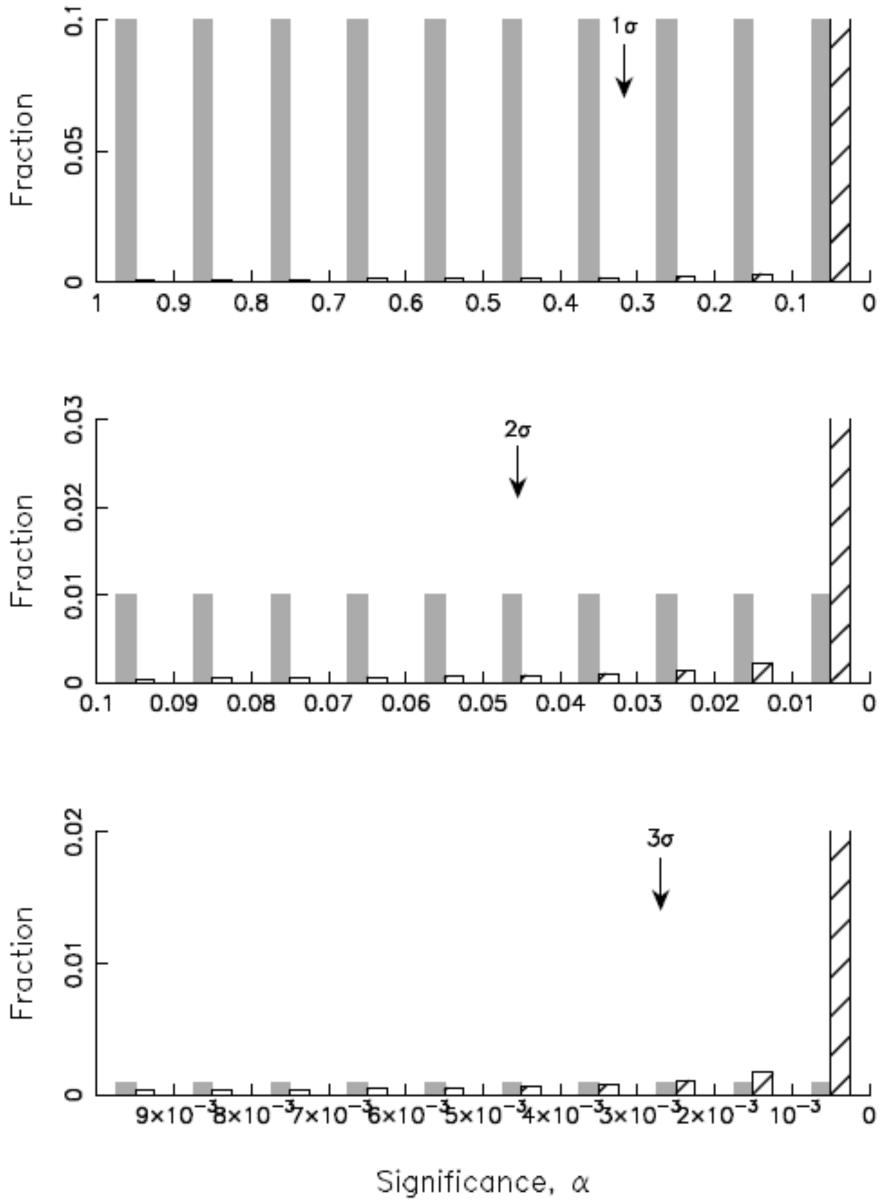
T. Loredo based on Berger (2003)

Significance Level Frequencies, $n = 200$



T. Loredo based on Berger (2003)

Significance Level Frequencies, $n = 2000$



T. Loredo based on Berger (2003)

A p-value is not an easily interpretable measure of the weight of evidence against the null.

- It does not measure how often the null will be wrongly rejected among similar data sets
- A naive false alarm interpretation typically overestimates the evidence
- For fixed *p*-value, the weight of the evidence decreases with increasing sample size

T. Loredo based on Berger (2003)

What about another μ prior?

- For data sets with H_0 rejected at $p \approx 0.05$, H_0 will be true *at least* 23% of the time (and typically close to 50%). (Edwards et al. 1963; Berger and Selke 1987)
- At $p \approx 0.01$, H_0 will be true *at least* 7% of the time (and typically close to 15%).

What about a different "true" null frequency?

- If the null is initially true 90% of the time (as has been estimated in some disciplines), for data producing $p \approx 0.05$, the null is true at least 72% of the time, and typically over 90%.

In addition . . .

- At a fixed p, the proportion of the time H_0 is falsely rejected grows as $\sqrt{n}$. (Jeffreys 1939; Lindley 1957)
- Similar results hold generically; e.g., for χ^2. (Delampady & Berger 1990)

T. Loredo based on Berger (2003)

What does it mean to “Discover a Planet”?

- *Frequentist* approach:
  Reject the null hypothesis that a simpler model without the planet could reasonably explain your data
- *Bayesian* approach:
  The Evidence for a model with the planet is significantly greater than the alternative models.

Why did the Frequentist Approach Work?

- Large planetary RV amplitudes
- Stellar activity need to explain RVs would cause other readily recognizable spectral signatures
- RV surveys focused on quiet FGK stars

Astronomers prioritized a track record of minimal false alarms

- When any doubt… collect more data

Why change?

- Goals shifting to planets with small RV amplitudes
- Unknown if stellar activity needed to explain RVs of low-mass planets is otherwise recognizable
- Prime targets selected for properties other than low stellar activity
- Increasing amount of telescope time required for low mass planet detections

What does it mean to “Discover a Planet”?

- *Frequentist* approach:
  Reject the null hypothesis that a simpler model without the planet could reasonably explain your data
- *Bayesian* approach:
  The Evidence for a model with the planet is significantly greater than the alternative models.

Bayesian Model Comparison

The probability of a radial velocity dataset {*d*} being generated from some model *M* parameterized by {*θ*} is given by…

$$p(d|\mathcal{M}) = \int p(\theta|\mathcal{M})p(d|\theta,\mathcal{M})d\theta$$

Evidence (or Fully marginalized likelihood)

Prior

Likelihood

To choose between two competing models M_1 and M_2, take the ratio of the evidences…

$$\text{Bayes Factor} = \frac{p(d|\mathcal{M}_2)}{p(d|\mathcal{M}_1)}$$

Bayesian view of false-alarm rate

$$B \equiv \frac{p(\{x_i\}|H_1)}{p(\{x_i\}|H_0)} = \frac{p(p_{\rm obs}|H_1)}{p(p_{\rm obs}|H_0)}$$

→ B here is just the ratio calculated in the Monte Carlo!

Why is p-value a poor measure of the weight of evidence?

- We should be *comparing hypotheses*, not trying to identify rare/surprising events—an observation surprising under the null motivates rejection only if it is not surprising under reasonable alternatives

- Comparison should use the *actual data*, not merely membership of the data in some larger set. A p-value conditions on incomplete information.

T. Loredo based on Berger (2003)

Features of Bayesian Approach

- Answer questions that you want to ask
- Rigorous basis for:
  - Quantifying parameter uncertainties
  - Comparing evidence for competing models (quantitative “Occam’s razor”)
  - Making principled decisions (via utility function)
- Makes assumptions explicit (encoded in model, priors and likelihood)

Interpreting the Bayesian Evidence

In a world with exactly N possible models, where our knowledge prior to taking data *d* is specified by p(M_i) that model i is the correct model and the Bayesian evidence for each model is calculated to be p($d|M_i$), our knowledge of the relative probability of each model after taking data *d* is given by

$$p(M_i|d) = p(d|M_i) \,/\, \Sigma_i \, p(M_i)\, p(d|M_i)$$

Interpreting the Bayesian Evidence

In a world with *exactly* N possible models, where our knowledge prior to taking data *d* is specified by p(M_i) that model i is the correct model and the *Bayesian evidence* for each model is *calculated* to be p($d|M_i$), our knowledge of the relative probability of each model after taking data *d* is given by

$$\mathrm{p}(M_i|d) = \mathrm{p}(d|M_i) \,/\, \Sigma_i\, \mathrm{p}(M_i)\, \mathrm{p}(d|M_i)$$

Limitations of Bayesian Approach

- Requires computing multi-dimensional integrals
- Can be computationally expensive:
  - Complex models can be expensive to evaluate
  - Parameter estimation for models with many parameters require performing high-dimensional integrals
  - Comparing evidence for competing models requires even more integrals
  - Computing expected utility requires even more integrals

Potential Estimates of Bayes Factor

- Laplace (or WKB) Approximation
  - If target density is nearly Gaussian
- Importance Sampling
  - If you have a good approximation to the target density
- Thermodynamic Integration
  - If integrals computed accurately/efficiently
- Nested Sampling
  - If it converges to correct answer

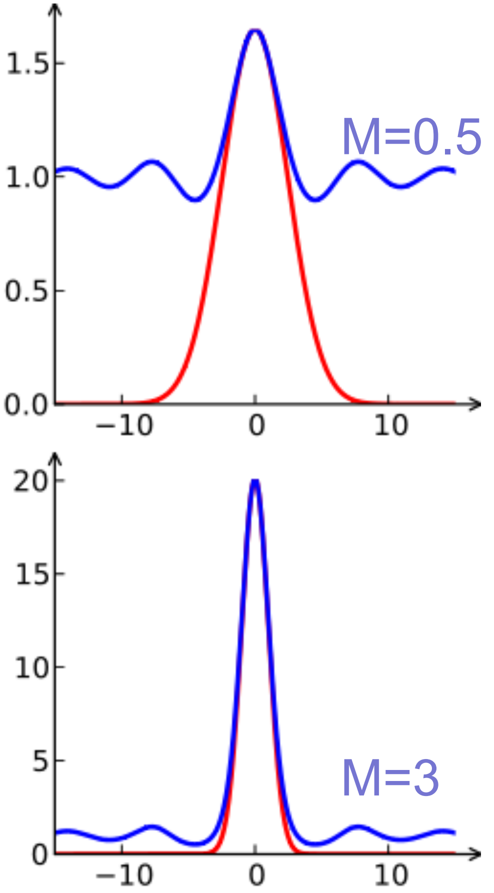
Laplace (WKB) Approximation

- 1-D:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) \\ + \frac{1}{2} f''(x_0)(x - x_0)^2 + O\left((x - x_0)^3\right)$$

$$\int_a^b e^{Mf(x)}\,dx \approx e^{Mf(x_0)} \int_a^b e^{-M|f''(x_0)|(x-x_0)^2/2}\,dx$$

$$\approx \sqrt{\frac{2\pi}{M|f''(x_0)|}} e^{Mf(x_0)} \text{ as } M \to \infty$$



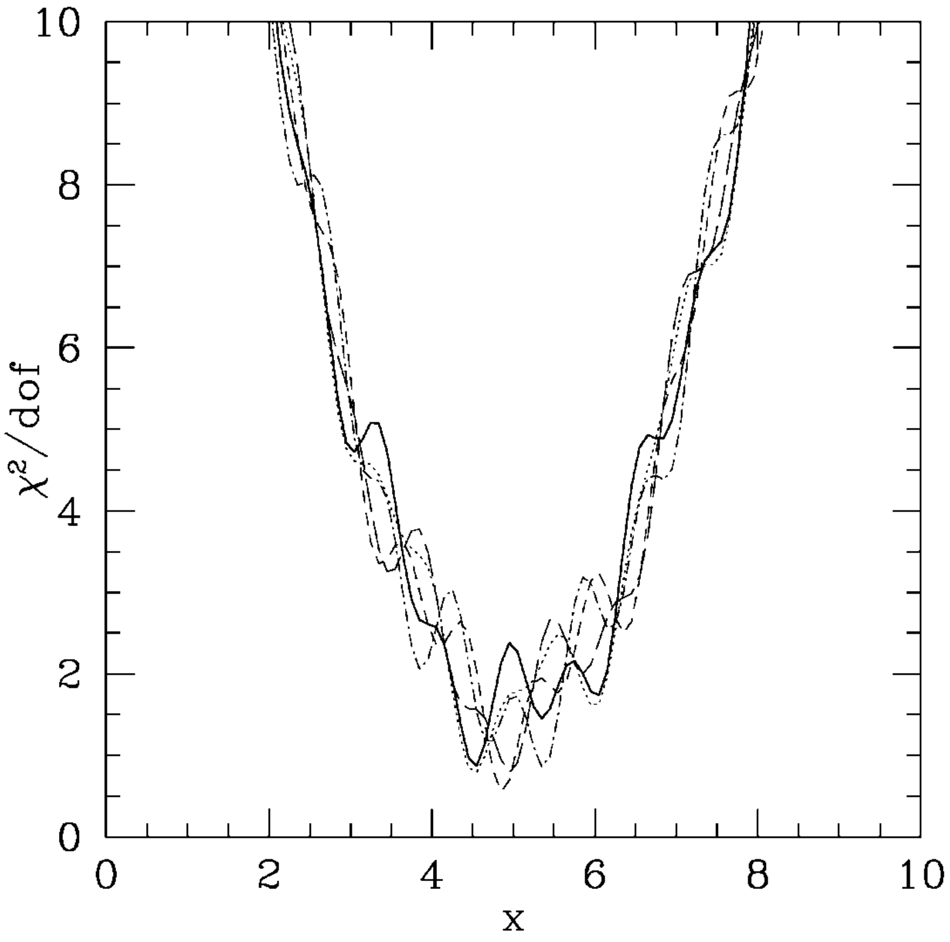
- In multiple dimensions:

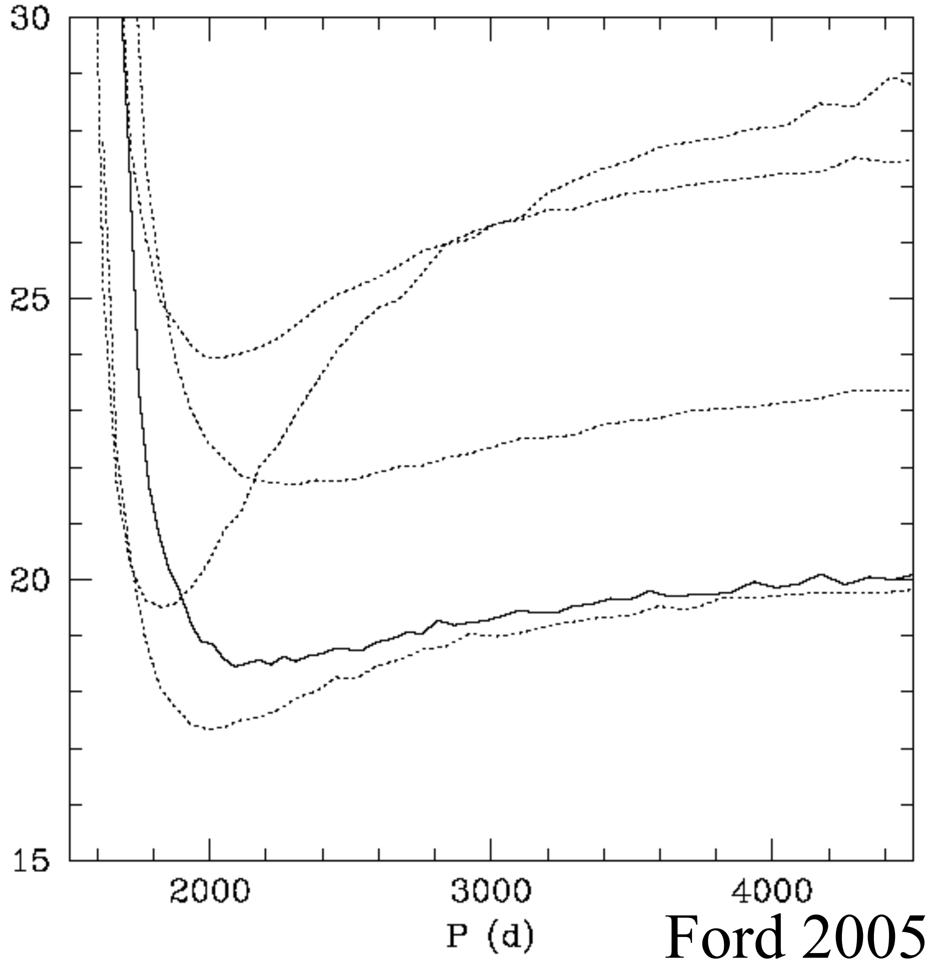
$$\int e^{Mf(\mathbf{x})}\,d\mathbf{x} \approx \qquad \approx \left(\frac{2\pi}{M}\right)^{d/2} |-H(f)(\mathbf{x}_0)|^{-1/2} e^{Mf(\mathbf{x}_0)} \text{ as } M \to \infty$$

$$\mathbf{H}_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Wikipedia

Examples of Problematic χ^2 Surfaces





Ford 2005

The Pretenders

- False alarm rate
- Likelihood ratio test / *F*-test
- Penalized likelihood
  - Bayesian Information Criterion (BIC)
  - Akaike Information Criterion (AIC)
  - …

Bayesian Information Criterion

- BIC = – 2 ln($\max_\theta L(\theta)$) + k ln(n) – k ln(2π)
  k = number of parameters to be estimated
  n = number of data points

- BIC is *not* Bayesian, as it ignores:
  - Prior over models
  - Prior over θ
  - Steepness or shape of likelihood near $\theta_{bf} = \mathrm{argmax}_\theta L(\theta)$
  - Other modes

Akaike Information Criterion

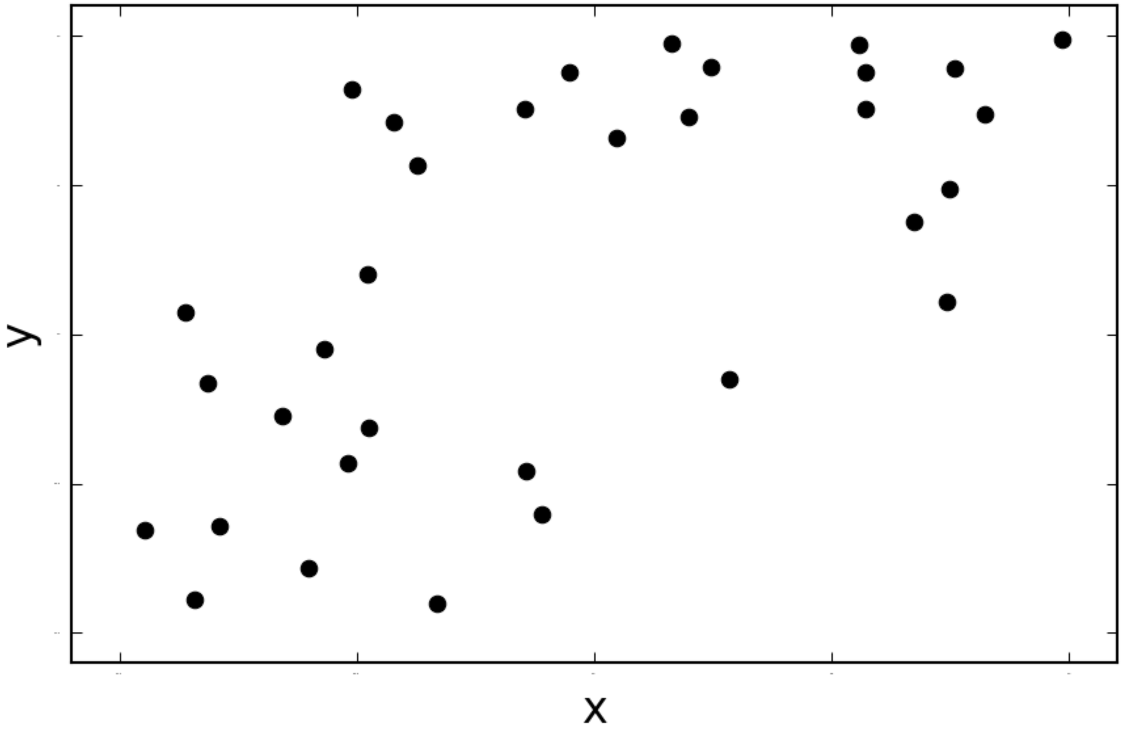
- $\text{AIC} = -\ln(\max_{\theta} L(\theta)) + 2k$
  k = number of parameters to estimated
  n = number of data points

- Often better than BIC for prediction or high-dimensional problems when true model isn't among models considered
- AIC "Corrected" for finite sample size
  $\text{AICc} = \text{AIC} + 2k(k+1)/(n-k-1)$

(If univariate, linear, normal residuals)

“Why not use maximum likelihood estimates?”

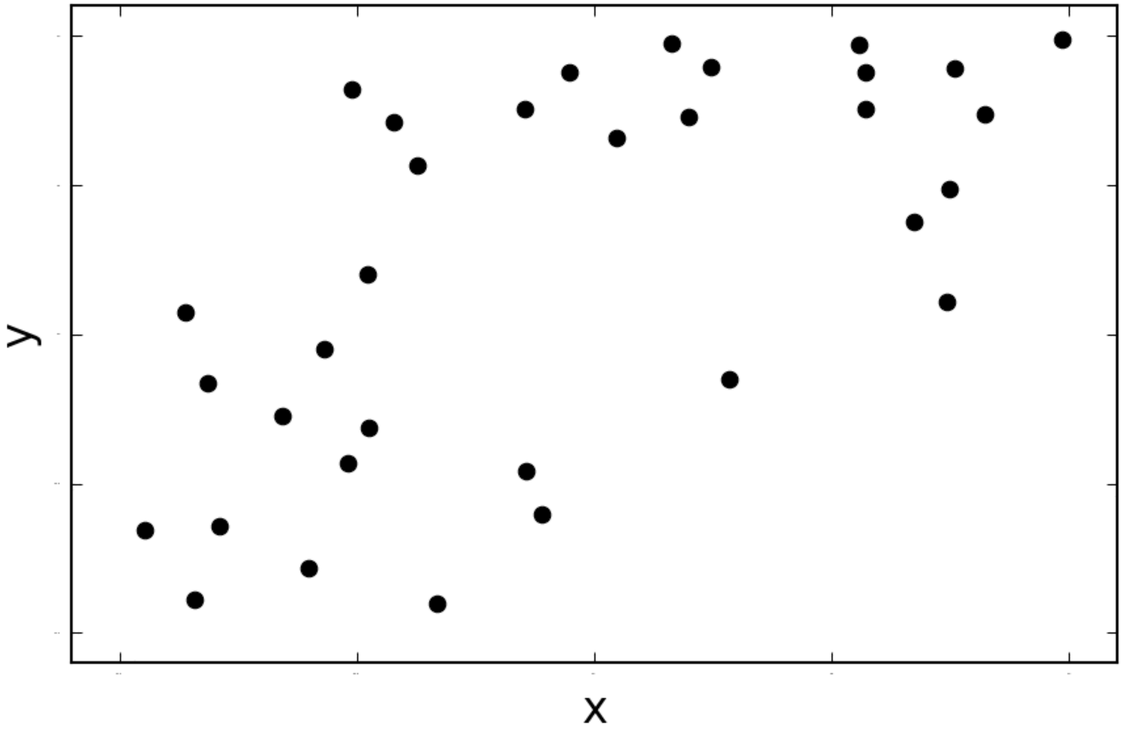
1. You can, but…
2. There’s a 3-parameter model can fit any 2-D scatterplot **exactly**:



B. Nelson based on D. Hogg

“Why not use maximum likelihood estimates?”

1. You can, but…
2. There’s a 3-parameter model can fit any 2-D scatterplot **exactly**: $y = A\cos(kx + \phi)$

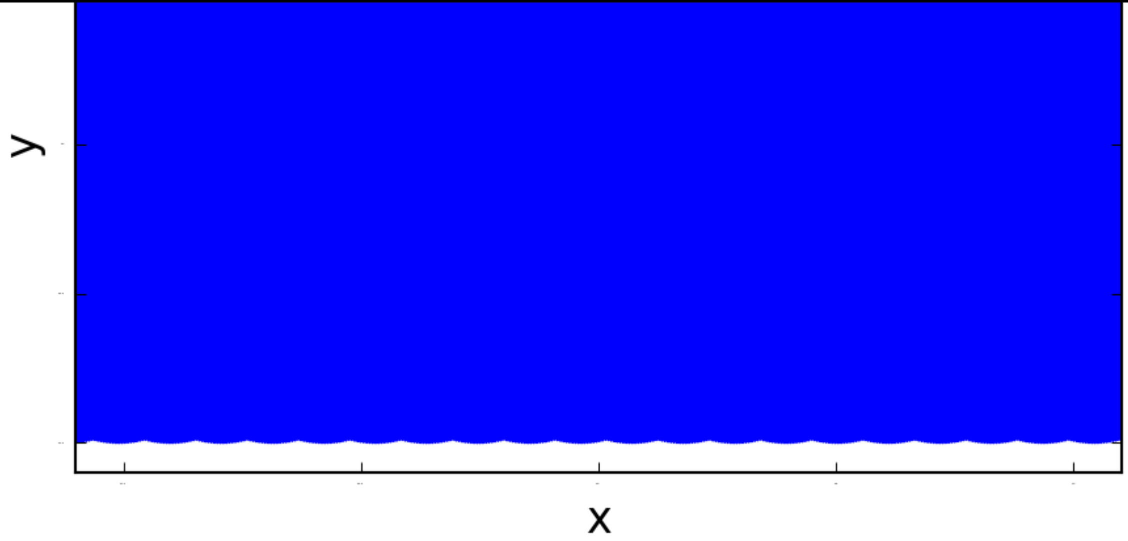


B. Nelson based on D. Hogg

“Why not use maximum likelihood estimates?”

1. You can, but…
2. There’s a 3-parameter model can fit any 2-D

Troll-tip: $y = A\cos(kx + \phi)$ will fit any scatterplot exactly. Use it if someone asks you to compute a MLE.



B. Nelson based on D. Hogg

What is the value of pretenders?

- Rough qualitative assessment of whether something is surprising
- Is it worth your time to perform a statistically meaningful calculation?
- Advantage: Can be computed quickly (i.e., maximize rather than marginalize)
- Disadvantages:
  - Arbitrary, often misleading
  - Asymptotic limits rarely relevant for real problems/data

- Search for extraordinary evidence

Best Practices

"All models are wrong; some models are useful."

– George Box 1979

Analyze datasets generated with physical model before you analyze astronomical data

- Great starting point
- Validate, understand & improve your statistical algorithm here

e.g., Nelson+ 2014; Jontof-Hutter+ 2015; Rajpaul+ 2015

Best Practices

“All models are wrong; some models are useful.”
– George Box 1979

When planning and analyzing astronomical observations, keep in mind that:

- Characterizing stellar activity requires large, well-sampled datasets
- Quantifying rare events requires large survey size
- Avoid ad-hoc revisions to model based on data for the target in question.

Best Practices

- Apply principled & tested algorithms
  - Validate stellar activity & statistical models with simulated data
  - Verify stellar activity & statistical models with astronomical observations
- Test for non-convergence of iterative algorithms
- Test sensitivity of results/conclusions to choice of priors & likelihoods (e.g., stellar activity model)

Conclusions

- Strive to perform statistically valid and reproducible research
- Be skeptical of any claim based on inappropriate statistical methodology
  - Don’t use *p*-values to claim detections of planets, atmospheric features, …
  - Never pretend a *p*-value is a false alarm probability
- Learn fundamentals of how to perform Bayesian model comparison
- Use practical approximations to Bayesian evidence when appropriate for your problem

SAMSI Program on Statistical, Mathematical and Computational Methods for Astronomy

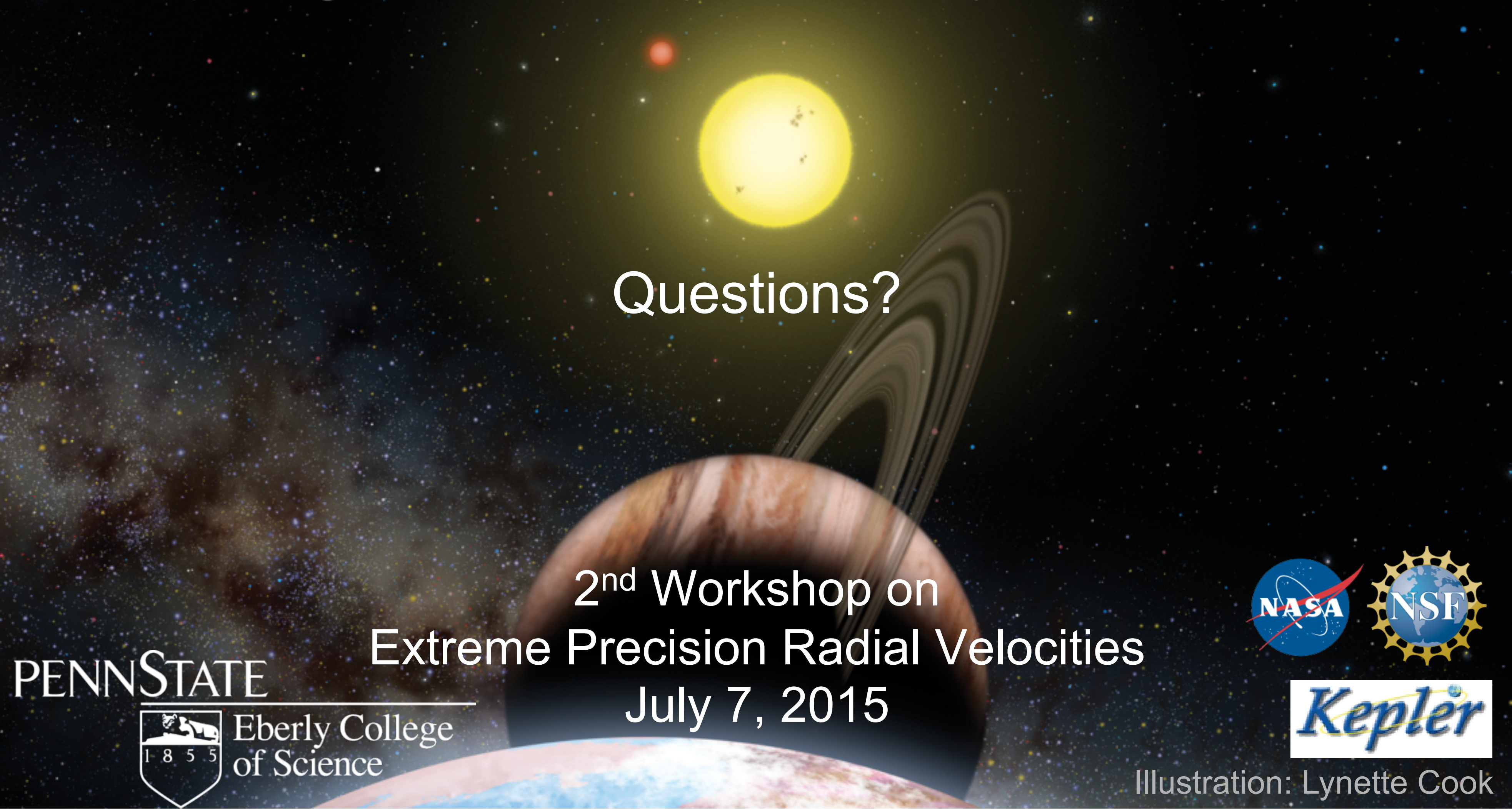
- Opening workshop Aug 22-26, 2016
- Planned working groups (Fall 2016/Spring 2017):
  - Uncertainty Quantification and Reduced Order Modeling in Gravitation, Astrophysics, and Cosmology
  - Synoptic Time Domain Surveys
  - Time Series Analysis for Exoplanets & Gravitational Waves: Beyond Stationary Gaussian Processes
  - Population Modeling & Signal Separation for Exoplanets & Gravitational Waves
  - Statistics, computation, and modeling in cosmology

DEPARTMENT OF ASTRONOMY AND ASTROPHYSICS

CENTER FOR

Exoplanets & Habitable Worlds

Seeking to discover habitable planets and life beyond the Solar System.



Questions?

2nd Workshop on
Extreme Precision Radial Velocities
July 7, 2015

PENNSTATE
1855
Eberly College
of Science

Illustration: Lynette Cook