Quantifying the Strength of Evidence for a Planet

p-values, FAPs, Bayes factors, and all that

Tom Loredo
Cornell Center for Astrophysics and Planetary Science
&
Carl Sagan Institute

http://hosting.astro.cornell.edu/~loredo/

Sagan Exoplanet Workshop — 23 July 2018

Sagan on scientific method

Science is more than a body of knowledge;
it is a way of thinking.
The method of science, as stodgy and grumpy as it may seem,
is far more important than the findings of science.
—Carl Sagan

The weather forecaster

Joint Frequencies of
Actual & Predicted Ithaca Weather

	Actual	
Prediction	Rain	Sun
Rain	**1/4**	**1/2**
Sun	**0**	**1/4**

The weather forecaster

Joint Frequencies of
Actual & Predicted Pasadena Weather

	Actual	
Prediction	Swelter	Sun
Swelter	$\mathbf{1/4}$	$\mathbf{1/2}$
Sun	$\mathbf{0}$	$\mathbf{1/4}$

The weather forecaster

Joint Frequencies of
Actual & Predicted Weather

	Actual		
Prediction	Rain	Sun	
Rain	**1/4**	**1/2**	3/4
Sun	**0**	**1/4**	1/4
	1/4	3/4	

Forecaster is right only 50% of the time

Observer notes a prediction of 'Sun' *every day* would be right 75% of the time, and applies for the forecaster's job

Should the observer get the job?

	Actual	
Prediction	Rain	Sun
Rain	1/4	1/2
Sun	0	1/4

Forecaster: You'll never be in an unpredicted rain

Observer: You'll be in an unpredicted rain 1 day out of 4

Lessons (Jaynes 1976)

- The value of an inference often lies in its usefulness in *the individual case* at hand
- Long run performance is not an adequate criterion for assessing the usefulness of individual case inferences
- When long run performance is deemed important, it needs to be separately evaluated

Exoplanet discovery questions

- **Single-host planet detection:**
  Is there a planet orbiting this observed star?
  (Or: How many planets are orbiting this observed star?)
- **Planet demographics:**
  How many of these N observed stars host a planet?
  (Or: What is the planet multiplicity distribution?)
  These may be generalized to infer the *underlying* planet prevalence

These two questions are inextricably related

Two styles of answers

Quantify uncertainty about planet detection using *probability*

But there are two competing understandings of what "probability" means, and thus how to use it:

- **Frequentist ($\mathcal{F}$):** $P =$ how often a procedure will be right or wrong in the long run
- **Bayesian ($\mathcal{B}$)**: $P =$ measure of strength of evidence for/against rival hypotheses explaining the case at hand

Exoplanet discovery answers (our agenda)

- **Single-host planet detection:**
  - $\mathcal{F}$: Null hypothesis ("no planet") significance testing via *maximum* likelihood ratio:
    - Fixed-threshold Type I (false alarm) & II (false no-alarm) error probabilities
    - p-values [Note: *p-value $\neq$ FAP!*]
  - $\mathcal{B}$: Posterior probability (or odds) for a planet via *marginal* likelihood
- **Planet demographics:**
  - $\mathcal{B}$: Hierarchical Bayesian modeling—learning priors from populations
  - $\mathcal{F}$: Adaptive thresholding via p-value distribution (e.g., controlling false discovery *rate* — FDR; out-of-scope!)
- **Feedback**: Single-host inference $\leftrightarrow$ demographic inference

Plan

1. **Frequentist and Bayesian parameter estimation**

2. **Frequentist and Bayesian model assessment**

3. **Demographics and detection**

Plan

1. **Frequentist and Bayesian parameter estimation**

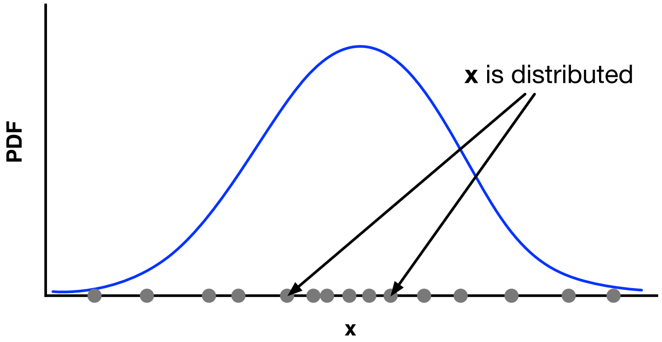
2. Frequentist and Bayesian model assessment

3. Demographics and detection

Interpreting probability densities (PDFs)

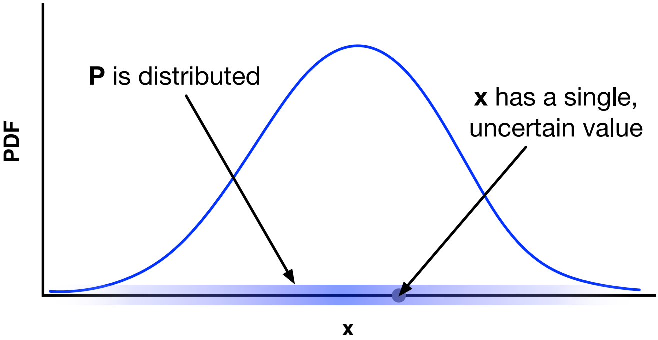
Frequentist

Probabilities are always (limiting) rates/proportions/frequencies that *quantify variability* across a sequence of independent trials/replications. $p(x)$ describes how the *values of x* would be distributed among infinitely many replications:



Bayesian

Probability *quantifies uncertainty* in a single-case inductive inference. $p(x)$ describes how *probability* is distributed over the possible values x might have taken in the single case before us:



"The 250-year debate between Bayesians and frequentists is unusual among philosophical arguments in actually having *important practical consequences*. . . ."

—Brad Efron, ASA President (2005)

Bayes's theorem:

$$\begin{aligned} p(A, B) &= p(A)p(B|A) \\ &= p(B)p(A|B) \\ \to p(A|B) &= p(A)\frac{p(B|A)}{p(B)}, \quad \text{Bayes's th.} \end{aligned}$$

$\mathcal{F}$: BT is only valid when A and B refer to *events*—statements about about the same underlying "random" outcomes (outcomes varying across replicated trials)

E.g.: $A \equiv i$, the number of dots shown on a fairly rolled die
$B \equiv i \in \mathcal{P}$, the number of dots is a prime number (2, 3, 5)

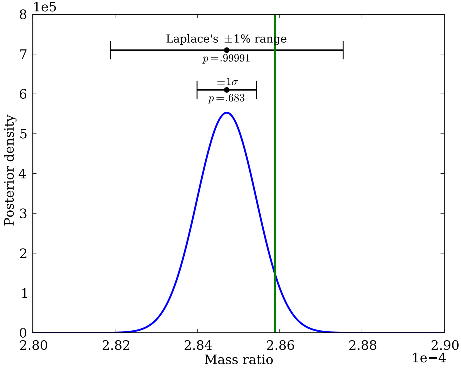
$$p(i|\mathcal{P}) = \frac{1}{6} \times \frac{[\![i \in \mathcal{P}]\!]}{\frac{1}{2}} = \frac{1}{3} \text{ if } i = 2, 3, \text{ or } 5; \text{ else } 0$$

$\mathcal{B}$: BT is potentially valid for *any propositions*, so long as we can assign the required strength-of-argument probabilities. In particular, for B = specification of data, D, and A = choice of one of rival hypotheses H_i explaining the data,

$$p(H_i|D) = p(H_i)\frac{p(D|H_i)}{p(D)} = \text{prior} \times \frac{\text{likelihood for } H_i}{\text{marginal likelihood}}$$

E.g.: Laplace (ca. 1818) computed the posterior PDF for $M_{Sat}/M_{\odot}$:

"Applying to them my formulae of probability I find that it is a bet of 11,000 against one that the error of this result is not 1/100 of its value..."



Inference With Parametric Models

Models M_i ($i = 1$ to N), each with parameters θ_i, each imply a *sampling dist'n* (conditional predictive dist'n for possible data):

$$p(D|\theta_i, M_i)$$

The θ_i dependence when we fix attention on the *observed* data is the *likelihood function*:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\rm obs}|\theta_i, M_i)$$

We may be uncertain about i (model uncertainty) or θ_i (parameter uncertainty); a hypothesis would specify these

Sometimes I'll drop the cumbersome subscript: $D = D_{\rm obs}$; D often refers to *hypothetical* data in $\mathcal{F}$ calculations

A model with *no* free parameters is a *simple hypothesis*; otherwise it is a *compound or composite hypothesis*

Additive Gaussian noise models

Setup

Data $D = \{d_i\}$ are noisy measurements of an underlying signal $f(t;\theta)$ at N sample points $\{t_i\}$. Let $f_i(\theta) \equiv f(t_i;\theta)$:

$$d_i = f_i(\theta) + \epsilon_i, \qquad \epsilon_i \sim \mathtt{Norm}(0, \sigma_i^2), \text{ indep.}$$

We seek to learn θ, or to compare different signal or noise hypotheses (model choice, M). Note: To a statistician, "model" means *everything* needed to make predictions—both the signal and noise hypotheses.

Likelihood

$$\begin{aligned} p(D|\theta, M) &= \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d_i - f_i(\theta)}{\sigma_i}\right)^2\right] \\ &\propto \exp\left[-\frac{1}{2}\sum_i \left(\frac{d_i - f_i(\theta)}{\sigma_i}\right)^2\right] \\ &= \exp\left[-\frac{\chi^2(\theta)}{2}\right] \end{aligned}$$

Posterior

For prior density $\pi(\theta)$ (perhaps uniform...),

$$p(\theta|D, M) \propto \pi(\theta) \exp\left[-\frac{\chi^2(\theta)}{2}\right]$$

The normalization constant (*marginal likelihood*) is

$$p(D|M) = \int d\theta \pi(\theta) \exp\left[-\frac{\chi^2(\theta)}{2}\right]$$

If you have a least-squares or χ^2 code:

- Treat $\chi^2(\theta)$ as $-2 \log \mathcal{L}(\theta)$
- Bayesian inference amounts to exploration and *numerical integration* (by quadrature or Monte Carlo) of $\pi(\theta) e^{-\chi^2(\theta)/2}$

A Simple (?) confidence region

Problem

Estimate the location (mean, μ) of a Gaussian distribution from a set of N IID samples $D = \{x_i\}$. Report a region summarizing the uncertainty.

Here assume std dev'n σ is *known*; we are uncertain only about μ

Model

The *sampling distribution* for *any* set $\{x_i\}$ is

$$
\begin{aligned}
p(\{x_i\}|\mu) &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\
&\propto e^{-\chi^2(\mu)/2}
\end{aligned}
$$

This gives the *likelihood function*, $\mathcal{L}(\mu)$ if we set $\{x_i\}$ to the *observed values*. The *log* likelihood is a parabola here.

Classes of variables—the two spaces

- μ is the unknown we seek to estimate—the *parameter*. The *parameter space* is the space of possible values of μ—here the real line (perhaps bounded). *Hypothesis space* is a more general term.
- A particular set of N data values $D = \{x_i\}$ is a *sample*. The *sample space* is the N-dimensional space of possible samples.

Standard inferences

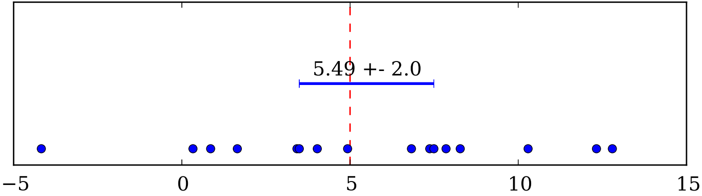
Let $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

- "Standard error" (rms error) is $\sigma/\sqrt{N}$
- "1σ" interval: $\bar{x} \pm \sigma/\sqrt{N}$ with conf. level CL $= 68.3\%$
- "2σ" interval: $\bar{x} \pm 2\sigma/\sqrt{N}$ with CL $= 95.4\%$

Some simulated data

Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$.

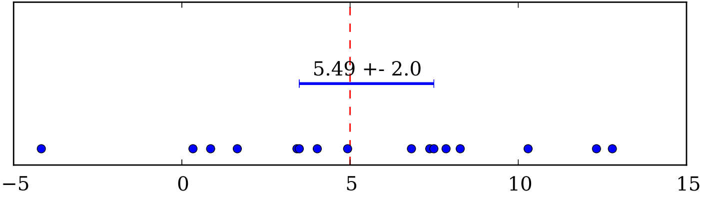
What is the CL associated with this interval?



Some simulated data

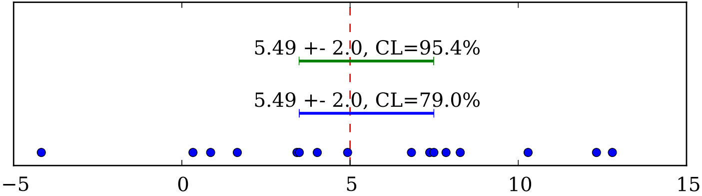
Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$.

What is the CL associated with this interval?



The confidence level for this interval is 79.0%.

Two intervals

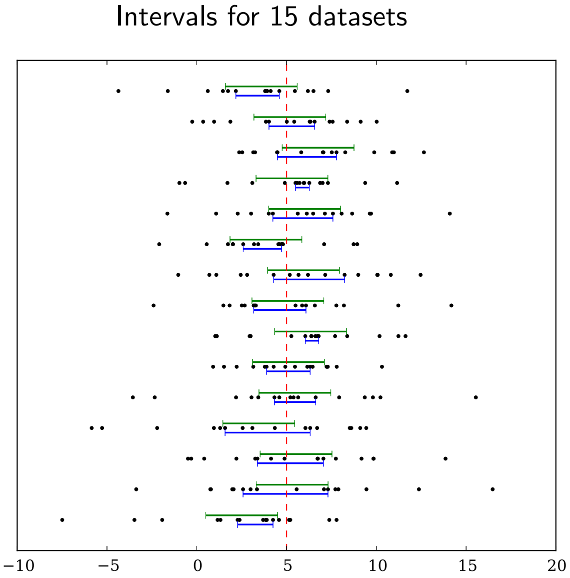


- Green interval: $\bar{x} \pm 2\sigma/\sqrt{N}$
- Blue interval: Let $x_{(k)} \equiv k$'th order statistic
  Report $[x_{(6)}, x_{(11)}]$ (i.e., leave out 5 outermost each side)

The point

> *The (frequentist) confidence level is a property of the* **procedure**, *not of the particular interval reported for a given dataset*

Performance of intervals



Gaussian problem posterior distribution

For the Gaussian example, a bit of algebra ("complete the square") gives:

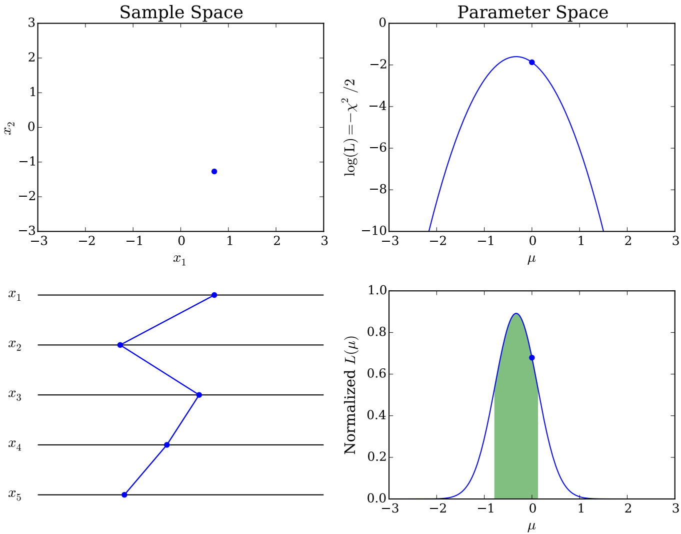
$$
\begin{aligned}
\mathcal{L}(\mu) &\propto \prod_i \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{1}{2}\sum_i \frac{(x_i-\mu)^2}{\sigma^2}\right] \\
&\propto \exp\left[-\frac{(\mu-\bar{x})^2}{2(\sigma/\sqrt{N})^2}\right]
\end{aligned}
$$

The likelihood is Gaussian in μ
Flat prior $\rightarrow$ posterior density for μ is $\texttt{Norm}(\bar{x}, \sigma^2/N)$

Highest posterior density (HPD) credible region by integrating:
$\bar{x} \pm \sigma/\sqrt{N}$ with $P = 68.3\%$

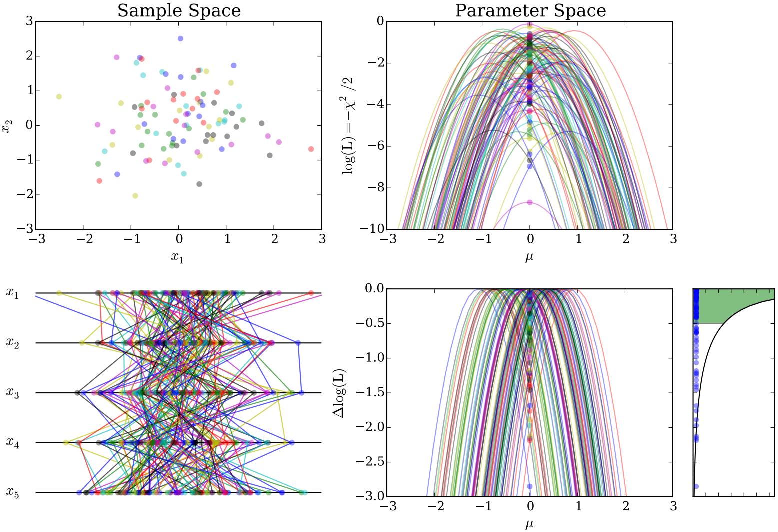
Bayesian credible region

Normalize the likelihood for the observed sample; report the region that includes 68.3% of the normalized likelihood:



"Root-N" confidence region calibration

Likelihoods or χ^2 curves for 100 *simulated* data sets, $\mu = 0$



Parameter estimation take-aways

- $\mathcal{F}$ and $\mathcal{B}$ approaches do very different kinds of summing/averaging
  - $\mathcal{F}$: Sum/average over sample space
  - $\mathcal{B}$: Sum/average over parameter space
- The observed data play very different roles
  - $\mathcal{F}$ probabilities *do not* (must not!) use the observed data
  - $\mathcal{B}$ probabilities *only* use the observed data
- They both produce the same reported estimates in the normal mean setting, but this is a *coincidence* that won't hold in general

Plan

1. Frequentist and Bayesian parameter estimation

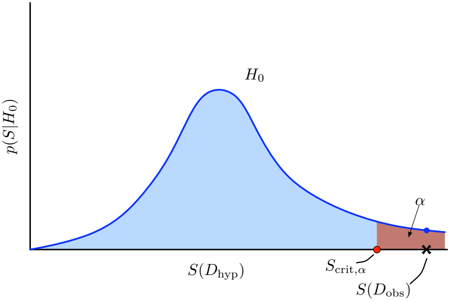
2. **Frequentist and Bayesian model assessment**

3. Demographics and detection

Null hypothesis significance testing (NHST)

Neyman-Pearson testing

- Specify simple null hypothesis H_0 such that rejecting it implies an interesting effect is present
- Devise statistic $S(D)$ measuring departure from null predictions
- Divide sample space into probable and improbable parts (for H_0); $p(\text{improbable}|H_0) = \alpha$ (Type I error rate), with α specified a priori
- If $S(D_{\rm obs})$ lies in improbable region, reject H_0; otherwise accept it
- Report: "H_0 was rejected (or not) with a procedure with false-alarm frequency α"



Neyman and Pearson devised this approach guided by Neyman's *frequentist principle*:

> *In repeated practical use of a statistical procedure, the long-run average actual error should not be greater than (and ideally should equal) the long-run average reported error. (Berger 2003)*

A *confidence region* is an example of a familiar procedure satisfying the frequentist principle

They insisted that one also specify an alternative, and find the error rate for falsely rejecting it (Type II error)

For *simple* null and alternative hypotheses, the optimal $S(D)$ is the (log) *likelihood ratio*. For composite hypotheses, the *maximum* likelihood ratio is popular.

Fisher's p-value testing

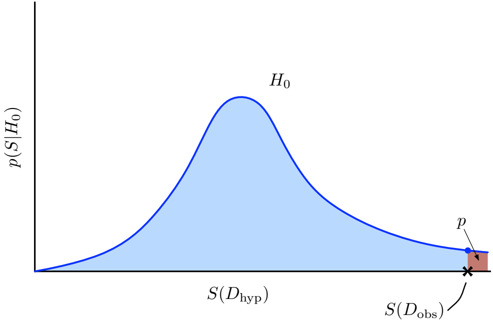
Fisher (and others) felt reporting a rejection frequency of α no matter where $S(D_{\rm obs})$ lies in the rejection region does not accurately communicate the strength of evidence against H_0

He advocated reporting the *p-value*:

$$p = P(S(D) > S(D_{\rm obs})|H_0)$$

Smaller p-values indicate stronger evidence against H_0

Astronomers call this the *significance level* or the *false-alarm probability* (FAP). Statisticians don't—for good reason!



ASA 2016 statement on statistical significance and p-values

- **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**
- **Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**
- **By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.**
- ...

p-values and the FAP fallacy

From the exoplanets literature:

> "...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 2.2 $M_{\odot}$..."

> "This detection has a signal-to-noise ratio of 4.1 with an empirically estimated upper limit on false alarm probability of 1.0%."

> "We find a false-alarm probability $< 10^{-4}$ that the RV oscillations attributed to CoRoT-7b and CoRoT-7c are spurious effects of noise and activity."

All of these statements incorrectly describe the weight of evidence for a planet, and almost certainly greatly exaggerate the weight of the evidence

What's wrong?

"This *signal, with* $S(D_{\rm obs}) = X$*, has a* FAP *of* $p\ldots$"

$$p = P(\{D_{\rm hyp} : S(D_{\rm hyp}) \ge S(D_{\rm obs})\}|H_0)$$

Probability ... **given** $\boldsymbol{H_0}$

p is computed assuming that H_0 *always operates*

Every alarm is false (i.e., with FAP$=1$) in this "world"

Probability... including **worse** *departures from null predictions*

p refers, not specifically to $D_{\rm obs}$, but to a *set* including *more extreme data*

$D_{\rm obs}$ bounds this set on the *weakest* side

What a p-value really means

In the voice of Don LaFontaine or Lake Bell:

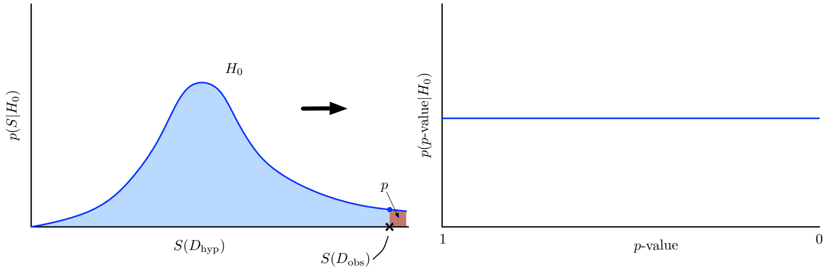
In a world. . . *with absolutely no exoplanets, with a threshold set so we wrongly claim to detect planets* $100 \times p\%$ *of the time, this data would wrongly be considered a detection—and it would be the data providing the* **weakest** *evidence for a planet in that world.*

Who wants to say *that*?! Whence "p-value."

p's one intuitive property

Under the null, the fraction of time $p > X$ is. . . X.

Think of p as an alternative test statistic—a nonlinear mapping of $S(D)$ that has a *uniform distribution under the null*



p is a surprise-ordered relabeling of the data, with a $U(0,1)$ PDF, and a linearly rising CDF

Surprise isn't enough

The rarity of data "like" $D_{\rm obs}$ under H_0 is evidence against H_0 only if *plausible alternatives* make $D_{\rm obs}$ *less* surprising

Expand the "world" of the *p*-value calculation:

- Let an alternative, H_1, sometimes operate, with probability π_1 (with null prevalence $\pi_0 = 1 - \pi_1$)
- Compare the rates for getting the observed *p*-value under H_0 and H_1 (*not* "observed or smaller *p*-value")
- Equivalently: Compare the rates for getting $S(D_{\rm obs})$ under H_0 and H_1

A Monte Carlo experiment (Berger 2003)

Consider measurements of μ with Gaussian noise, σ known:

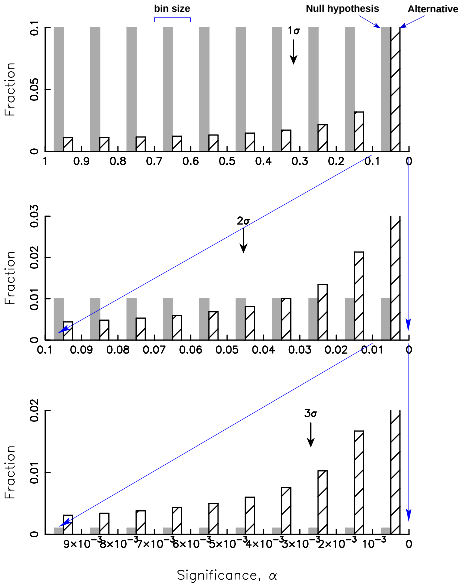
- Choose $\mu = 0$ OR $\mu \sim N(0, 4\sigma^2)$ with a fair coin flip*
- Simulate n data, $x_i \sim N(\mu, \sigma^2)$ (use $n = 20$, 200, 2000)
- Calculate $z_{\rm obs} = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$ and $p(z_{\rm obs}) = P(z > z_{\rm obs}|\mu = 0)$
- Bin $p(z)$ separately for each hypothesis;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

Compare how often the two hypotheses produce data with a 1–, 2–, or 3–σ effect $\rightarrow$ *conditional error probabilities* (real FAPs!)

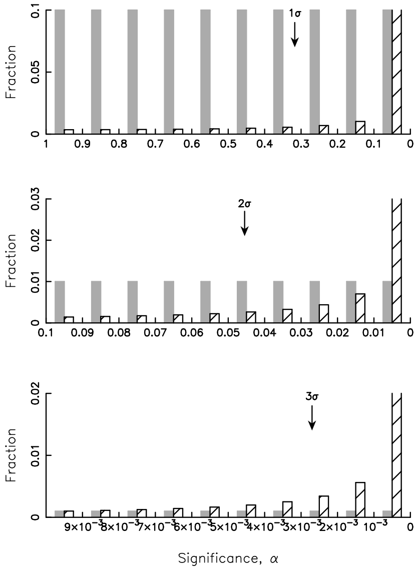
z	p-value
1	0.317
2	0.046
3	0.003

*An assumption that gives alternatives a "fair" chance and would *over*estimate the evidence against H_0 in settings where the null is more prevalent

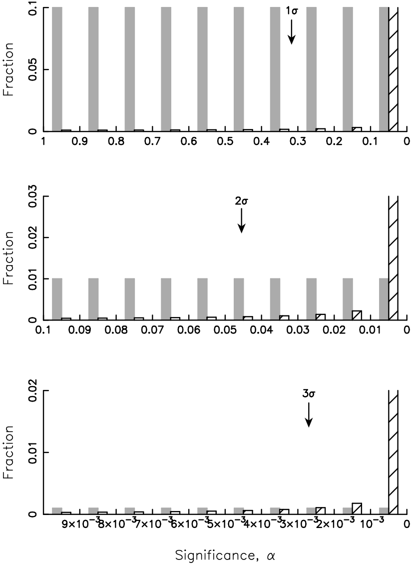
Significance Level Frequencies, $n = 20$



Significance Level Frequencies, $n = 200$



Significance Level Frequencies, $n = 2000$



Conditional error rates and posterior odds

Bayes's theorem comparing two hypotheses $\rightarrow$ posterior odds:

$$
\begin{aligned}
O_{10} &\equiv \frac{p(H_1|D)}{p(H_0|D)} \\
&= \frac{p(H_1)}{p(H_0)} \times \frac{p(D|H_1)}{p(D|H_0)}
\end{aligned}
$$

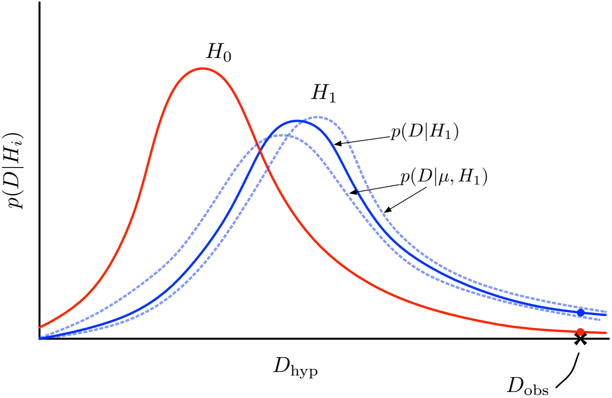
Here $D = \{x_i\}$, and the *Bayes factor* is:

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

$\rightarrow B$ here is just the ratio calculated in the Monte Carlo!

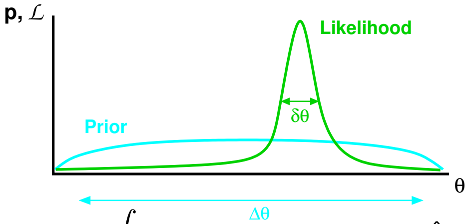
For *compound* hypotheses (H_1 here), the *marginal likelihood* accounts for parameter uncertainty that is ignored by p-values (which typically set parameters equal to best-fit values):

$$p(D|H_i) = \int d\theta_i \; p(\theta_i) \; p(D|\theta_i, H_i)$$



Also, the marginal likelihood uses *all* of the data, not just the value of a test statistic: in general $p(D|H_i) \neq p(S(D)|H_i)$

Marginal vs. maximum likelihood & Ockham's razor



$$
\begin{aligned}
p(D|M_i) &= \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i \\
&\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\
&= \text{Maximum Likelihood} \times \text{Ockham Factor}
\end{aligned}
$$

Models with more parameters often make the data more probable — *for the best fit*

Ockham factor penalizes models for "wasted" volume of parameter space

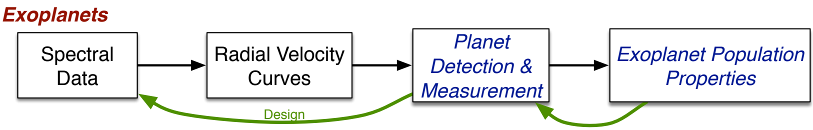
Quantifies intuition that models shouldn't require fine-tuning

Plan

1. Frequentist and Bayesian parameter estimation
2. Frequentist and Bayesian model assessment
3. **Demographics and detection**

A circularity problem

- We need to know population properties—prevalences, parameter distributions—to quantify detection uncertainty for a particular member using conditional error rates/posterior odds.
- We try to detect individual members in order to learn about the population.



Bayesian discovery chains have feedback loops

Hierarchical Bayes (HB) for detection

In a population context, we can learn features of priors by pooling the data—including learning prevalences/occurrence rates

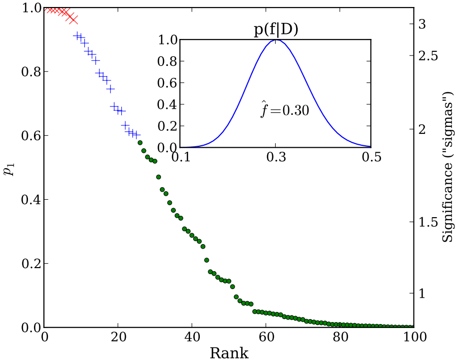
Measure $N = 100$ targets with additive Gaussian noise, $\sigma = 1$

- 70 have $A = 0$ (M_0)
- 30 have $A = 2.2$ (M_1)

Let f = fraction of objects with $A = 2.2$.

If f were known, it would be the prior probability for a Bayesian odds calculation.

Treat f as unknown (flat prior); infer it from the data



One can say there are about 30 sources present, without being able to say for sure whether many of the candidates are sources or not.

What to report for individual discoveries?

This is an open issue!

Do report the p-value (perhaps a posterior predictive p-value) — but just call it... a p-value! View it mainly as a *model checking* tool, a rough measure of misfit of the null.

Supplement it with a result that speaks more directly to the false/true alarm probability—a Bayes factor or conditional $\mathcal{F}$ error rate

The FAP depends on the prior odds, $\Pi_{01} = \pi_0/\pi_1$ and priors for any uncertain parameters (pop'n dist'ns), motivating suggestions:

- Establish default/consensus *interim priors* for analyzing data from individual systems; report interim posteriors and detection odds; eventually update using HB results
- Report p-value and the prior odds that would be needed to produce a specified FAP (such as 5%, or 1%)

Recap: Exoplanet discovery answers

- **Single-host planet detection:**
  - $\mathcal{F}$: Null hypothesis ("no planet") significance testing via *maximum* likelihood ratio:
    - Fixed-threshold Type I (false alarm) & II (false no-alarm) error probabilities
    - p-values [Note: *p-value $\neq$ FAP!*]
  - $\mathcal{B}$: Posterior probability (or odds) for a planet via *marginal* likelihood
- **Planet demographics:**
  - $\mathcal{B}$: Hierarchical Bayesian modeling—learning priors from populations
  - $\mathcal{F}$: Adaptive thresholding via p-value distribution (e.g., controlling false discovery *rate* — FDR; out-of-scope!)
- **Feedback**: Single-host inference $\leftrightarrow$ demographic inference

Entries to the p-value literature

- Bibliographies: "402 Citations..." (Thompson 2001) [web site]; "Papers Discussing Significance Testing" (2001–2011) [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 [URL]
- "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]

Jetsam!

Generalizing Berger's Monte Carlo expt

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)

Feedback Example: Adaptive Threshold vs. Hier. Bayes

Setting: Counting sources (real vs. spurious)

Measure $N = 100$ objects with additive Gaussian noise, $\sigma = 1$:

- 30 have $A = 2.2$
- 70 have $A = 0$

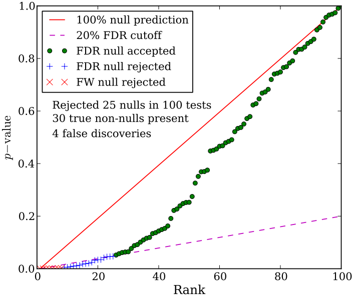
Detect via 100 tests of $H_0 : A = 0$

Source Present	Detection Result: *Negative*	*Positive*	Total
H_0: *No*	T_-	F_+	ν_0
H_1: *Yes*	F_-	T_+	ν_1
Total	N_-	N_+	N

Thresholding controlling FWER and FDR

Threshold criteria:

- Fixed: Control *family-wise error rate* at level α: accept objects with p-values$p = \alpha/N$, aiming to not make a single false discovery $\rightarrow$ 9 (accurate) discoveries for FWER = 20%
- Adaptive Control *false discovery rate*, $\langle F_+/N_+ \rangle = 20\%$ via Benjamini-Hochberg $\rightarrow$ 25 discoveries (4 false)
- Other choices possible



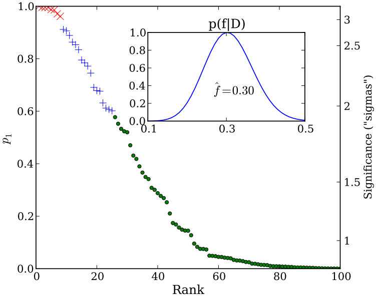
Issue with FDR control: Astronomers will use detections to infer distributions; will be biased for dim sources

Hierarchical Bayes approach

Let f = fraction of objects with $A = 2.2$.

If f were known, it would be the prior probability for a Bayesian odds calculation.

Treat f as *unknown* (flat prior); infer it from the data:



One can say there are about 30 sources present, without being able to say for sure whether many of the candidates are sources or not.
Caution: The "upper level" prior needs some care in more complex settings (Scott & Berger 2008; MLM literature)

Confidence regions

"Confidence region"

- Frequentist quantification of uncertainty in a parameter estimate
- A *procedure* that takes data as input, and gives a region as output
- The *specific region* found by applying a CR procedure to an observed dataset

"Confidence level"

- Lower bound on coverage $C(\theta)$: how often $\mathrm{CR}(D_{\mathrm{hyp}})$ contains the parameter value θ used to generate D_{hyp} (conservative guarantee of coverage)

"Calibration" of credible regions

How often may we expect an HPD region with probability P to include the true value if we analyze many datasets? I.e., what's the frequentist coverage of an interval rule $\Delta(D)$ defined by calculating the Bayesian HPD region each time?

Suppose we generate datasets by picking a parameter value from $p(\theta)$ and simulating data from $p(D|\theta)$

The fraction of time θ will be in the HPD region is:

$$Q = \int d\theta \; p(\theta) \int dD \; p(D|\theta) \; [\![\theta \in \Delta(D)]\!]$$

Note $p(\theta)p(D|\theta) = p(\theta, D) = p(D)p(\theta|D)$, so

$$Q \;\; = \;\; \int dD \int d\theta \; p(\theta|D) \, p(D) \; [\![\theta \in \Delta(D)]\!]$$

$$
\begin{aligned}
Q &= \int dD \int d\theta \; p(\theta|D)\, p(D) \; [\![\theta \in \Delta(D)]\!] \\
&= \int dD \; p(D) \int d\theta \; p(\theta|D) \; [\![\theta \in \Delta(D)]\!] \\
&= \int dD \; p(D) \; \int_{\Delta(D)} d\theta \; p(\theta|D) \\
&= \int dD \; p(D) P \\
&= P
\end{aligned}
$$

The HPD region includes the true parameters $100P\%$ of the time

This is exactly true for any problem, even for small datasets

Keep in mind it involves drawing θ from the prior; credible regions are "calibrated with respect to the prior"

Credible regions guarantee average coverage

Recall the original Q integral:

$$\begin{aligned} Q &= \int d\theta\, p(\theta) \int dD\, p(D|\theta)\, [\![\theta \in \Delta(D)]\!] \\ &= \int d\theta\, p(\theta) C(\theta) \end{aligned}$$

where $C(\theta)$ is the (frequentist) coverage of the HPD region when the data are generated using θ

This indicates Bayesian regions have guaranteed *average coverage*

The prior can be interpreted as quantifying how much we care about coverage in different parts of the parameter space