Why Try Bayesian Methods?
(Lecture 5)

Tom Loredo

Dept. of Astronomy, Cornell University

http://www.astro.cornell.edu/staff/loredo/bayes/

Today's Lecture

- Problems you avoid
  - Ambiguity in what is "random"
  - Recognizable subsets
  - The "nuisance" of nuisance parameters
  - Misleading measures of "significance"

  $\Rightarrow$ *Irrelevance of long run behavior/sample averages*
- Advantages you gain
- Foundations

The Randomness of Randomness

Theory (H_0):

The number of “A” stars in a cluster should be 0.1 of the total.

Observations:

5 A stars found out of 96 total stars observed.

Theorist’s analysis:

Calculate χ^2 using $\bar{n}_A = 9.6$ and $\bar{n}_X = 86.4$.

Significance level is $p(> \chi^2 | H_0) = 0.12$ (or 0.07 using more rigorous binomial tail area). Theory is accepted.

Observer's analysis:

Actual observing plan was to keep observing until 5 A stars seen!

"Random" quantity is $N_{\rm tot}$, not n_A; it should follow the negative binomial dist'n. Expect $N_{\rm tot} = 50 \pm 21$.

$p(> \chi^2 | H_0) = 0.03$. Theory is rejected.

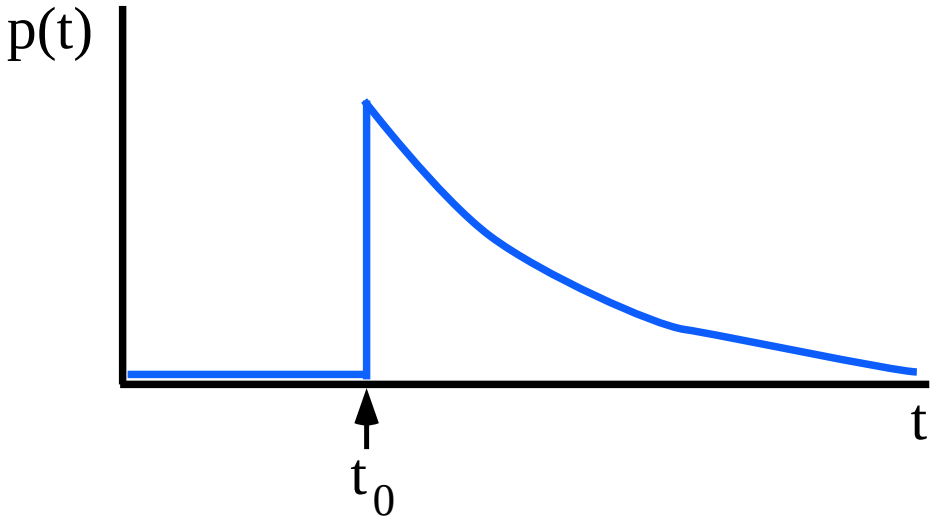
Telescope technician's analysis:

A storm was coming in, so the observations would have ended whether 5 A stars had been seen or not. The proper ensemble should take into account $p(\text{storm}) \ldots$

Bayesian analysis:

The Bayes factor is the same for binomial or negative binomial likelihoods, and slightly favors H_0. Include $p(\text{storm})$ if you want—it will drop out!

Recognizable Subsets



Goal:

Locate the start time, t_0, of a burst of events with an abrupt start and a decay time $\tau = 1$ s.

$$p(t|t_0) = \frac{1}{\tau}\exp(-\frac{t-t_0}{\tau}) \quad \text{for} \quad t \geq t_0$$

Frequentist method of moments:

Note $\langle t \rangle \equiv \int t\, p(t) dt = t_0 + \tau$ so unbiased estimator for t_0 is

$$\hat{t}_0 \equiv \frac{1}{N} \sum_{i=1}^{N} (t_i - \tau)$$

Also, $p(\hat{t}_0 | t_0)$ is analytic $\rightarrow$ find confidence intervals.

Suppose $\{t_i\} = \{12, 14, 16\}$. Find

$$\hat{t}_0 = 13; \qquad\qquad 12.15 < t_0 < 13.83\ (90\%)$$

But $\hat{t}$ and entire interval lie *after* first event!

Can show that the confidence region will not include the true value *100% of the time* in the subset of samples that have $\hat{t} > t_1 + 0.85$, and *we can tell from the data* whether or not any particular sample lies in this subset.

Bayesian solution:

Product of event time probabilities and flat prior $\rightarrow$

$$\begin{aligned} p(t_0 \mid \{t_i\}, I) &= N \exp\left[N(t_0 - t_1)\right] \quad \text{for} \quad t_0 \leq t_1 \\ &= 0 \quad \text{for} \quad t_0 > t_1 \end{aligned}$$

Summaries:

- $\hat{t}_0 = 12$; $\langle t_0 \rangle = 11.66$
- 90% HPD region is $11.23 < t_0 < 12.0$

Marginals *vs.* Profiles

General problem:

Measure several quantities with an "uncalibrated" instrument that adds Gaussian noise with *unknown* but constant σ.

Learn about σ by pooling the data.

Example—Pairs of measurements:

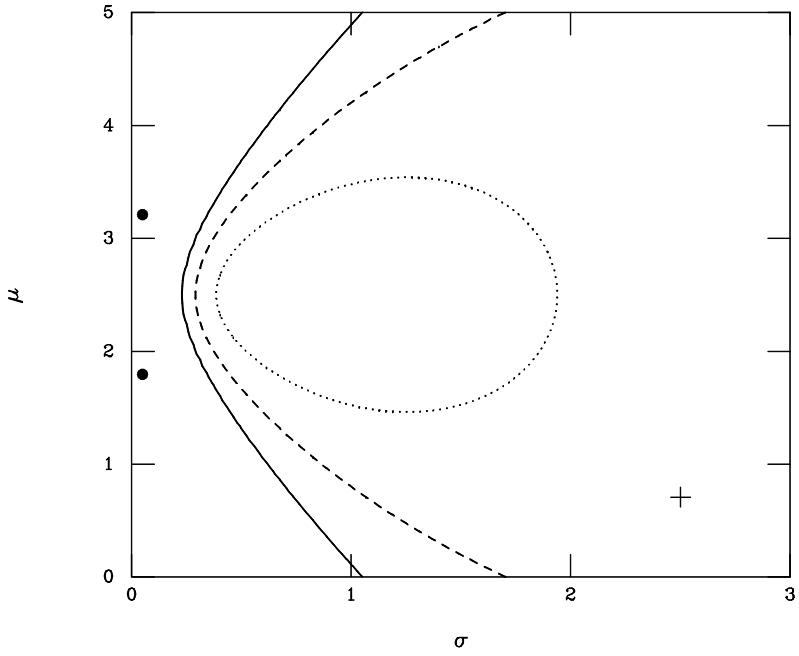
Make 2 measurements (x_i, y_i) for each of N quantities μ_i.

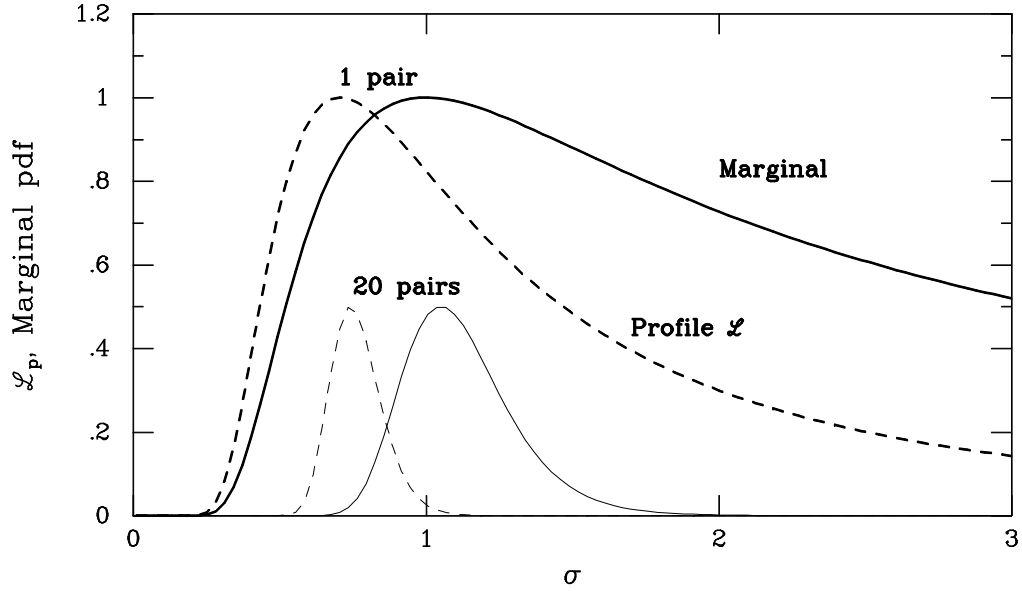
$$\mathcal{L}(\{\mu_i\}, \sigma) = \prod_i \frac{\exp\left[-\frac{(x_i-\mu_i)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}} \times \frac{\exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right]}{\sigma\sqrt{2\pi}}$$

Frequentist approach uses $\mathcal{L}_p(\sigma) = \max_{\{\mu_i\}} \mathcal{L}(\{\mu_i\}, \sigma)$

But $p(\sigma|D)$ and $\mathcal{L}_p(\sigma)$ differ dramatically!

Joint & Marginal Results for $\sigma = 1$





A Simple Significance Test

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

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

Test statistic:

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

The Significance of Significance

Collect the α 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 α levels.

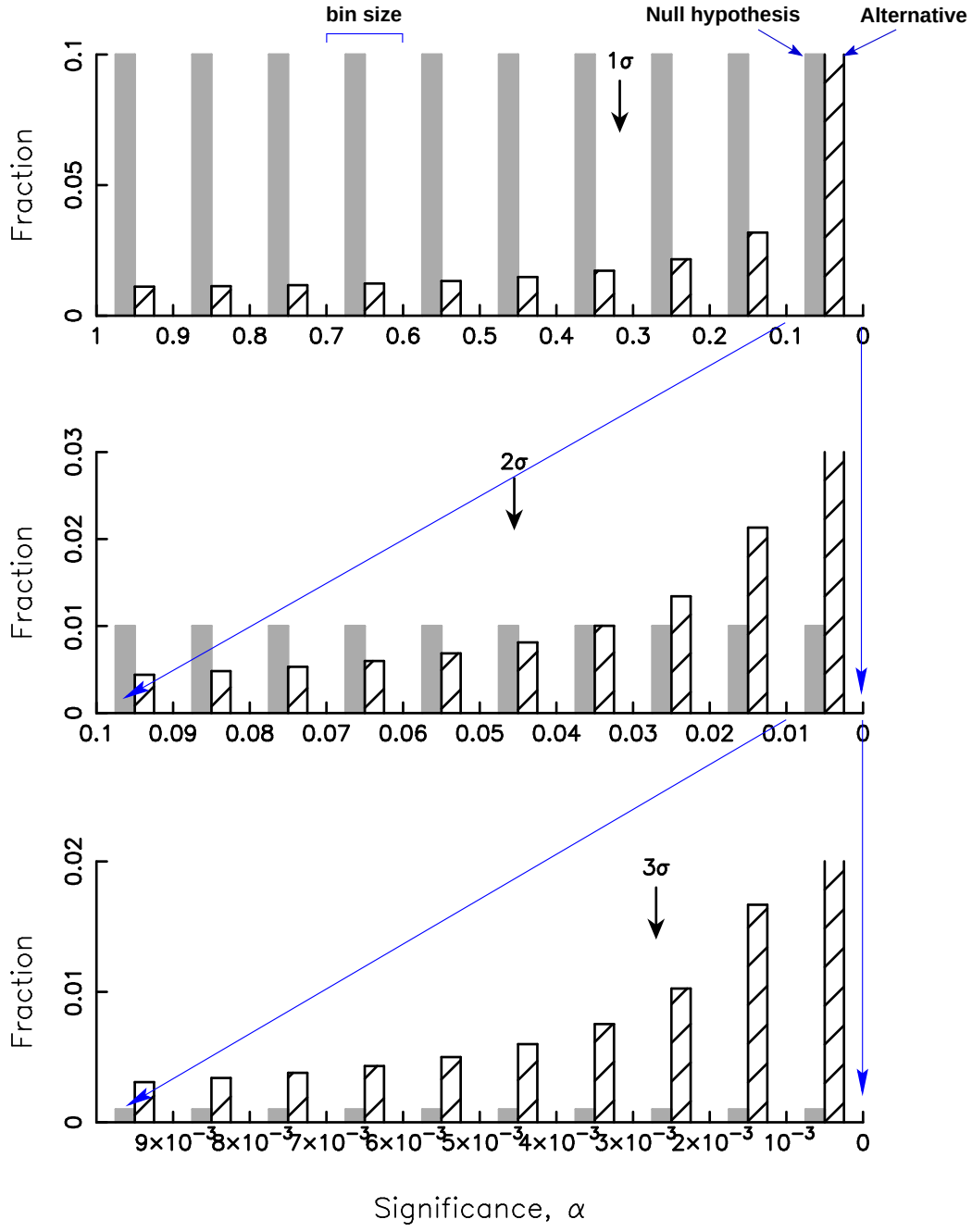
- Suppose that, overall, H_0 was true about half of the time.
- Focus on the subset with $t \approx 2$ (say, $[1.95, 2.05]$ so $\alpha \in [.04, .05]$, so that H_0 was rejected at the 0.05 level.
- Find out how many times in that subset H_0 turned out to be true.
- Do the same for other significance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

A Monte Carlo experiment:

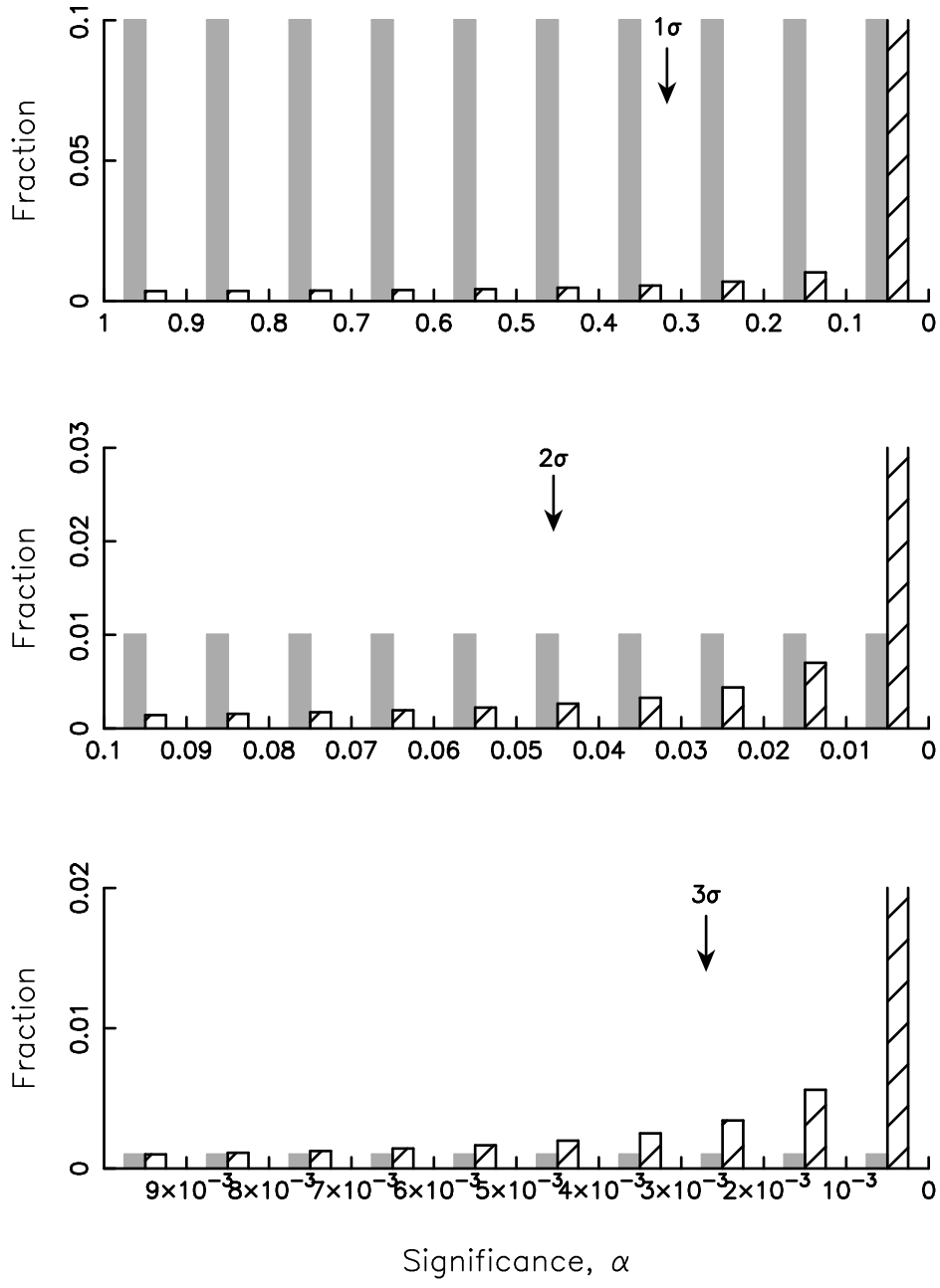
- Choose $\mu = 0$ OR $\mu \sim N(0, 4\sigma^2)$ with a fair coin flip
- Simulate $n = 20$ data, $x_i \sim N(\mu, \sigma^2)$
- Calculate $t_{\rm obs} = \frac{|\bar{x}|}{\sigma/\sqrt{n}}$ and $\alpha(t_{\rm obs}) = P(t > t_{\rm obs}|\mu = 0)$
- Bin $\alpha(t)$ separately for each hypothesis; repeat

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

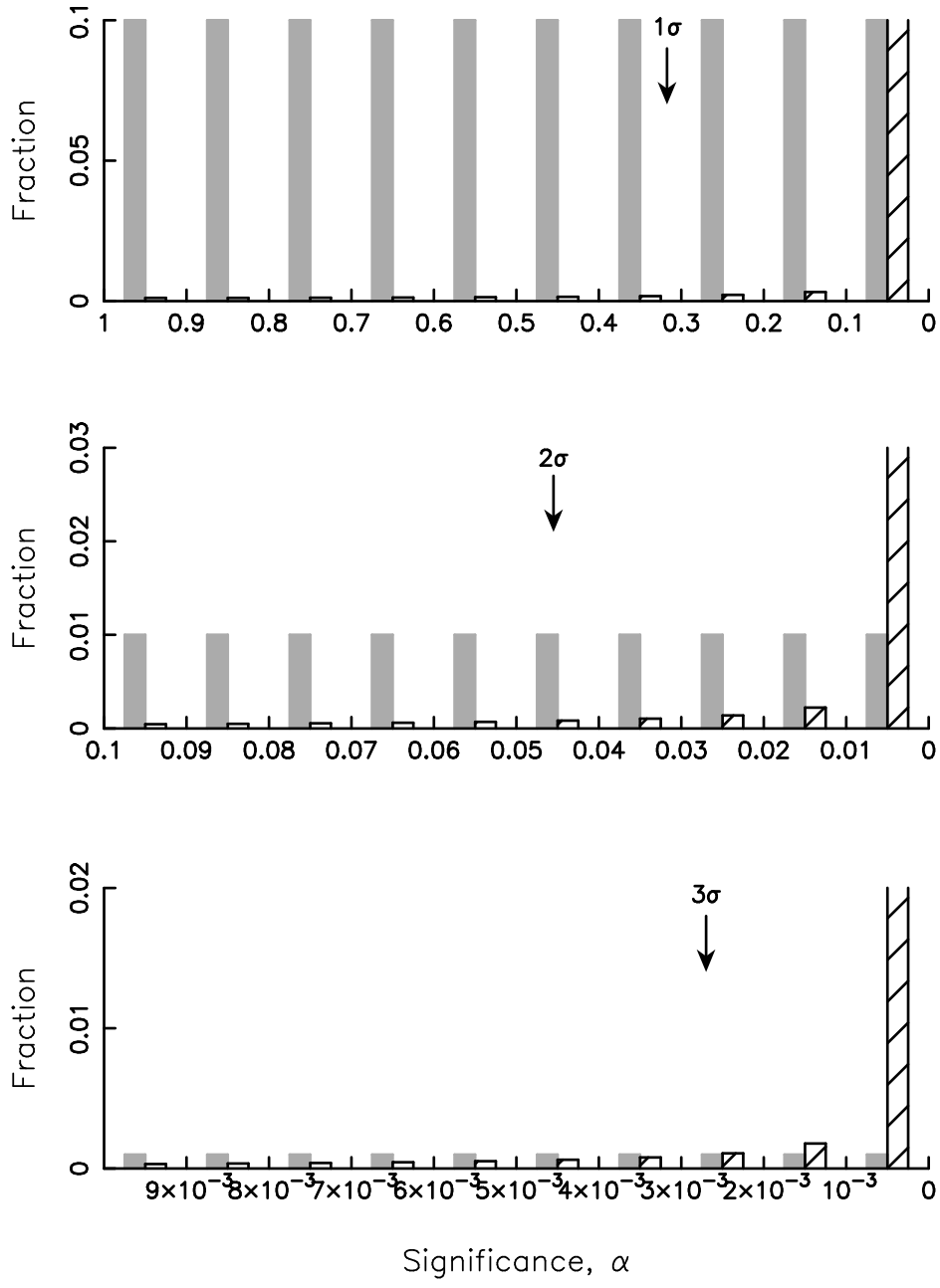
Significance Level Frequencies, $n = 20$



Significance Level Frequencies, $n = 200$



Significance Level Frequencies, $n = 2000$



What about another μ prior?

- For data sets with H_0 rejected at $\alpha \approx 0.05$, 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 $\alpha \approx 0.01$, 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 $\alpha \approx 0.05$, the null is true at least 72% of the time, and typically over 90%.

In addition . . .

- At a fixed α, 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)

Significance is not an easily interpretable measure of the weight of evidence against the null.

- Significance does not accurately measure how often the null will be wrongly rejected among similar data sets.
- The “obvious” (and recommended!) interpretation overestimates the evidence.
- For fixed significance, the weight of the evidence decreases with increasing sample size.

A History of Criticism

Significance tests have been under suspicion since their creation

- "Some difficulties of interpretation encountered in the application of the chi-square test" (Berkson 1938)
- Jeffreys 1939ff: the $\sqrt{n}$ effect (Jeffreys-Lindley paradox)

W. L. Thompson's on-line bibliography of works criticizing significance tests contains > 300 entries spanning a dozen disciplines!

- "Significance tests die hard: the amazing persistence of a probabilistic misconception"
- "Needed: A ban on the significance test"
- "The insignificance of statistical significance"

H. Jeffreys, addressing an audience of statisticians:

> For n from about 10 to 500 the usual result is that $K = 1$ when $(a - \alpha_0)/s_\alpha$ is about 2... not far from the rough rule long known to astronomers, i.e. that differences up to twice the standard error usually disappear when more or better observations become available... I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the [observed] trial was improbable; that is, that it has not predicted something that has not happened. As an argument astronomer's experience is far better. (Jeffreys 1980)

A Bayesian Look at Significance

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

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

Why is significance a poor measure of the weight of evidence?

- We should be *comparing hypotheses*, not trying to identify rare events.
- Comparison should use the *actual data*, not merely membership of the data in some larger set. Significance level conditions on incomplete information.

The Weatherman

Joint Frequencies of

Actual & Predicted Weather

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

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 he get the job?

Weatherman: You'll never be in an unpredicted rain.

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

The value of an inference lies in its usefulness in the individual case.

Long run performance is not an adequate criterion for assessing the usefulness of inferences.

What you avoid

- Hidden subjectivity/arbitrariness
- Dependence on “stopping rules”
- Recognizable subsets
- Defining number of “independent” trials in searches
- Inconsistency & incoherence (e.g., inadmissable estimators)
- Inconsistency with prior information
- Complexity of interpretation (e.g., significance vs. sample size)

What you get

- Probabilities *for hypotheses*
  - Straightforward interpretation
  - Identify weak experiments
  - Crucial for global (hierarchical) analyses (e.g., pop'n studies)
  - Forces analyst to be explicit about assumptions
- Handle nuisance parameters
- Valid for all sample sizes
- Handles multimodality
- Quantitative Occam's razor
- Model comparison for > 2 alternatives; needn't be nested

And there's more . . .

- Use prior info/combine experiments
- Systematic error treatable
- Straightforward experimental design
- Good frequentist properties:
  - Consistent
  - Calibrated—E.g., if you choose a model only if odds > 100, you will be right $\approx 99\%$ of the time
  - Coverage as good or better than common methods
- Unity/simplicity

Foundations

"Many Ways To Bayes"

- Consistency with logic $+$ internal consistency $\rightarrow$ BI
  (Cox; Jaynes; Garrett)
- "Coherence"/Optimal betting $\rightarrow$ BI (Ramsey; DeFinetti; Wald)
- Avoiding recognizable subsets $\rightarrow$ BI (Cornfield)
- Avoiding stopping rule problems $\rightarrow$ $\mathcal{L}$-principle
  (Birnbaum; Berger & Wolpert)
- Algorithmic information theory $\rightarrow$ BI
  (Rissanen; Wallace & Freeman)
- Optimal information processing $\rightarrow$ BI (Good; Zellner)

There is probably something to all of this!

What the theorems mean

When reporting numbers ordering hypotheses, values must be consistent with calculus of probabilities for hypotheses.

Many frequentist methods satisfy this requirement.

Role of priors

Priors are not fundamental!

Priors are analogous to initial conditions for ODEs.

- Sometimes crucial
- Sometimes a nuisance

Key Ideas

- To make inferences, calculate probabilities *for hypotheses*
- What's different about this approach:
  - Must be explicit about alternatives (think about models instead of statistics)
  - Sum/integrate in parameter space rather than sample space
- This avoids a variety of frequentist difficulties
- This provides many practical benefits
- It's what one *should* do