Objective Bayesian Analysis

James O. Berger
Duke University and the
Statistical and Applied Mathematical Sciences Institute

In Honor of William H. Jefferys

Outline

- Introduction to objective Bayesian analysis
- A brief history of objective Bayesian, frequentist, and subjective Bayesian statistics
- Nice features of objective Bayesian analysis that might be of particular interest to astronomy.
  - Directly answering questions of interest, such as 'What is the probability that the theory is correct?'
  - Automatic Ockham's razor and multiplicity corrections
  - 'Correct' elimination of nuisance parameters

Introduction to Objective Bayesian Analysis

Bayesian analysis proceeds by

- modeling the data probabilistically;
- modeling unknown features of the data-model using *prior* probability distributions;
- using probability theory (often Bayes theorem) to find the *posterior* probability distribution of quantities of interest, given the data.

Example: A coin is (independently) spun $n = 10$ times, and $x = 3$ heads are observed.

Goal: Inference concerning θ, the probability of heads.

Likelihood function: $L(\theta) \propto \theta^3(1-\theta)^7$.

Objective Bayesian inference: Assign θ a prior density:

- Choice 1. The uniform density, $\pi(\theta) = 1$.
- Choice 2. The Jeffreys prior $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$

By Bayes theorem, the posterior density of θ, given the data $x = 3$, is (for the Jeffreys prior)

$$\pi^J(\theta \mid x = 3) \propto L(\theta)\,\theta^{-1/2}(1-\theta)^{-1/2} \propto \theta^{2.5}(1-\theta)^{6.5}\,,$$

which is the Beta$(2.5, 6.5)$ density.

History of Objective Bayesian, Frequentist, and Subjective Bayesian Statistics

The Reverend Thomas Bayes, began the objective Bayesian theory, by solving a particular problem

- Suppose X is Binomial (n,p); an 'objective' belief would be that each value of X occurs equally often.
- The only prior distribution on p consistent with this is the uniform distribution.
- Along the way, he codified Bayes theorem.
- Alas, he died before the work was finally published in 1763.

REV. T. BAYES

The real inventor of Objective Bayes was Simon Laplace (also a great mathematician, astronomer and civil servant) who wrote *Théorie Analytique des Probabilité* in 1812

- He virtually always utilized a ‘constant’ prior density (and clearly said why he did so).
- He established the ‘central limit theorem’ showing that, for large amounts of data, the posterior distribution is asymptotically normal (and the prior does not matter).
- He solved very many applications, especially in physical sciences.
- He had numerous methodological developments, e.g., a version of the Fisher exact test.

6. Laplace in his robes as Chancellor of the Senate.

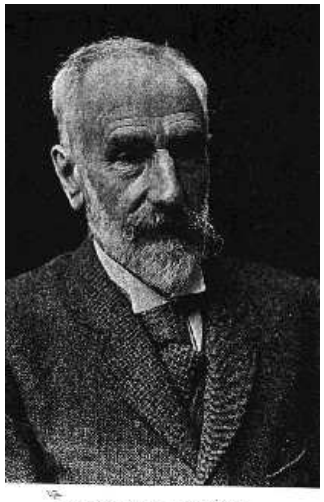
What's in a name, part I

- It was called *probability theory* until 1838.
- From 1838-1950, it was called *inverse probability,* apparently so named by Augustus de Morgan.
- From 1950 on it was called *Bayesian analysis* (as well as the other names).

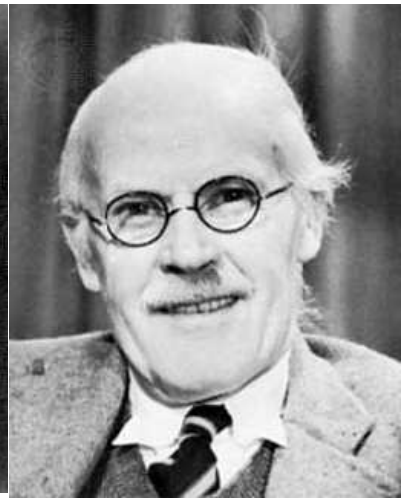
AUGUSTUS DE MORGAN

An example of the use of 'inverse probability' in the 19th century

- Luroth (in 1876) and Francis Edgeworth (in 1883) solved the problem of inference about a normal mean with unknown variance (using inverse probability with a constant prior on $h=1/\sigma$) , showing the inference should be based on the t-distribution with n degrees of freedom.
- But n-1 is the 'right' degrees of freedom, obtained by
  - R.A. Fisher first around 1920, using a frequentist argument;
  - Harold Jeffreys in the 1930's using a constant prior in $\log(\sigma)$.

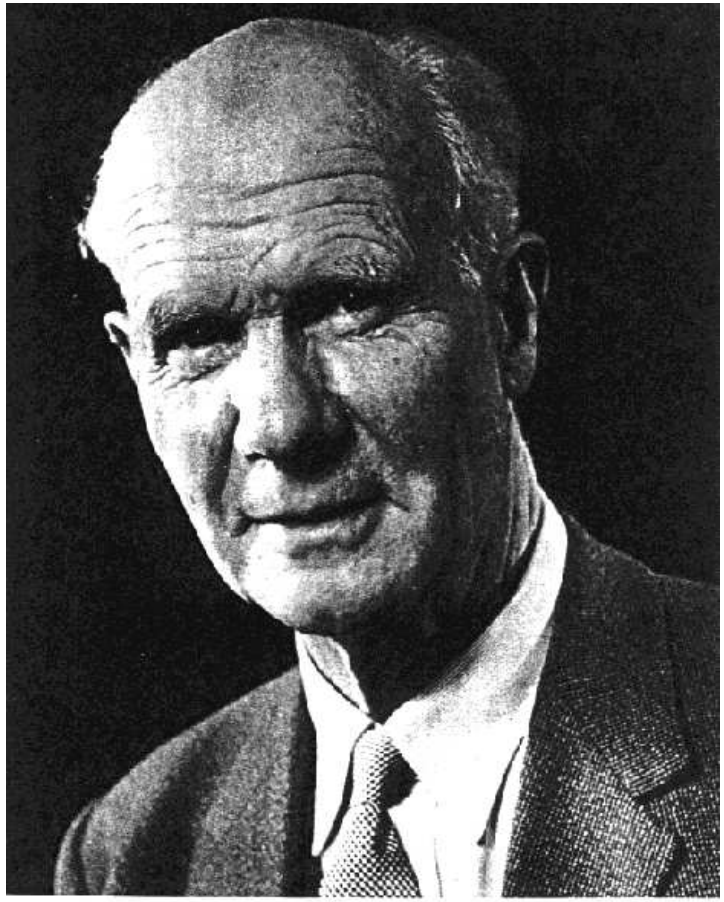


FRANCIS YSIDRO EDGEWORTH, F.B.A.



The importance of inverse probability b.f. (before Fisher): as an example, Egon Pearson in 1925 finding the ‘right’ objective prior for a binomial proportion

- Gathered a large number of estimates of proportions p_i from different binomial experiments
- Treated these as arising from the predictive distribution corresponding to a fixed prior.
- Estimated the underlying prior distribution (an early empirical Bayes analysis).
- Recommended something close to the currently recommended ‘Jeffreys prior’ $p^{-1/2}(1-p)^{-1/2}$.



EGON SHARPE PEARSON

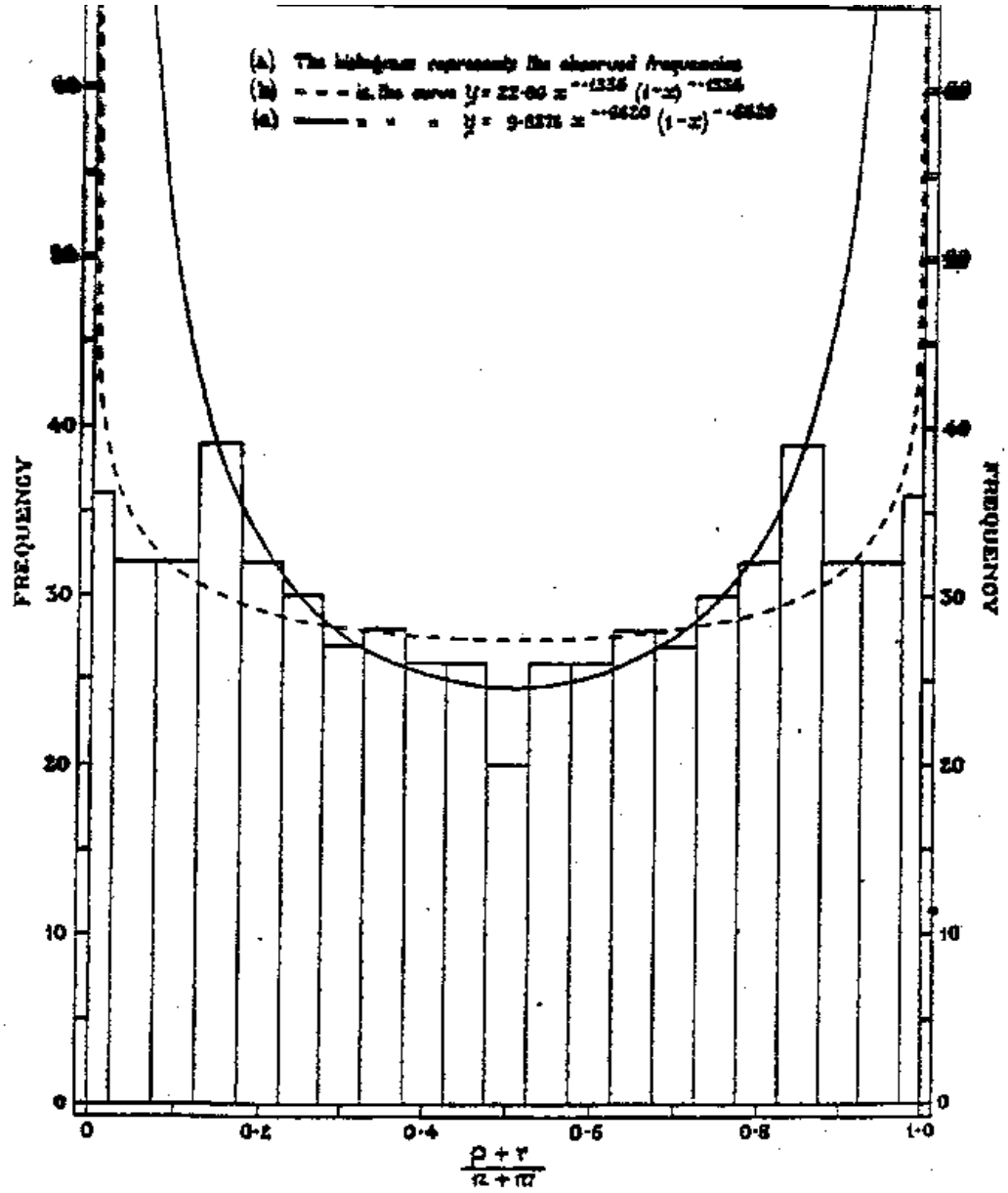
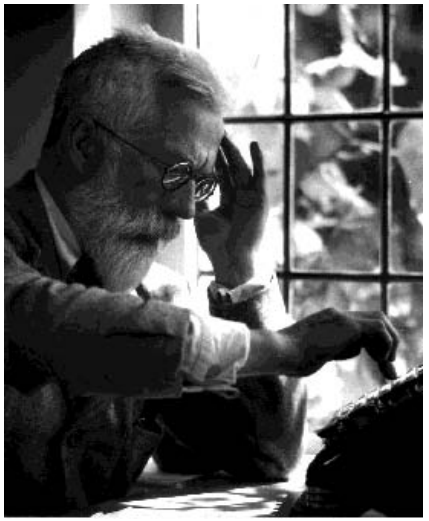


Fig. 3. Distribution of Frequencies of $\frac{p+r}{n+m}$ in 300 samples (made symmetrical).

1930's: 'inverse probability' gets 'replaced' in mainstream statistics by two alternatives

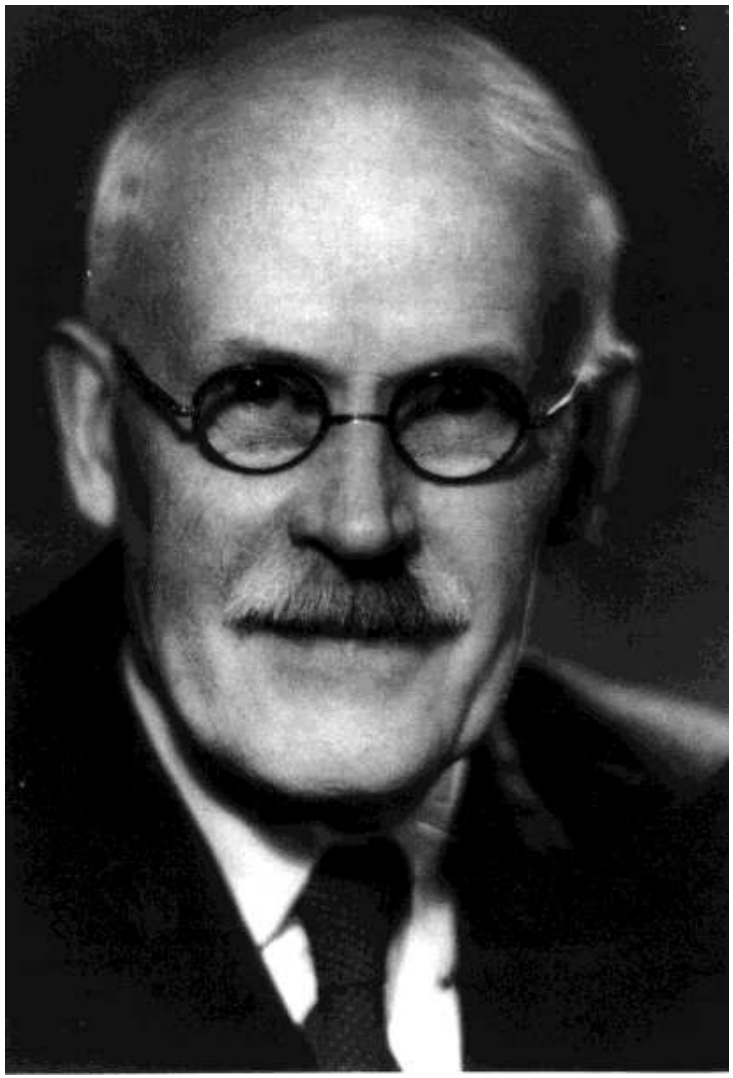
- For 50 years, Boole, Venn and others had been calling use of a constant prior logically unsound (since the answer depended on the choice of the parameter), so alternatives were desired.
- R.A. Fisher's developments of 'likelihood methods,' 'fiducial inference,' … appealed to many.
- Jerzy Neyman's development of the frequentist philosophy appealed to many others.



JERZY NEYMAN

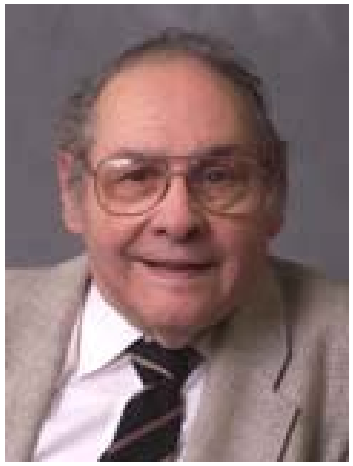
Harold Jeffreys (also a leading geophysicist) revived the Objective Bayesian viewpoint through his work, especially the *Theory of Probability* (1937, 1949, 1963)

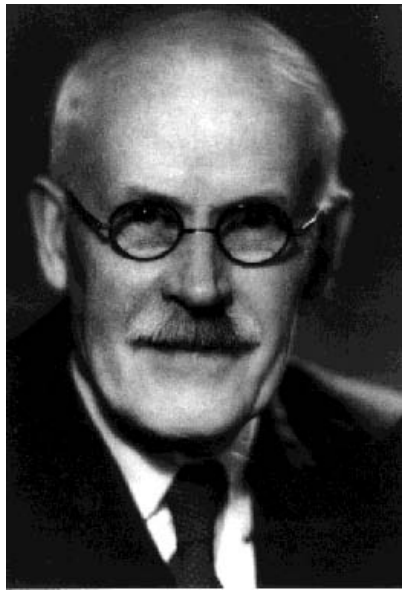
- The now famous *Jeffreys prior* yielded the same answer no matter what parameterization was used.
- His priors yielded the 'accepted' procedures in all of the standard statistical situations.
- He began to subject Fisherian and frequentist philosophies to critical examination, including his famous critique of p-values: "An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred."



What's in a name, part II

- In the 50's and 60's the *subjective* Bayesian approach was popularized (de Finetti, Rubin, Savage, Lindley, …)
- At the same time, the *objective* Bayesian approach was being revived by Jeffreys, but Bayesianism became incorrectly associated with the subjective viewpoint. Indeed,
  - only a small fraction of Bayesian analyses done today heavily utilize subjective priors;
  - objective Bayesian methodology dominates entire fields of application today.

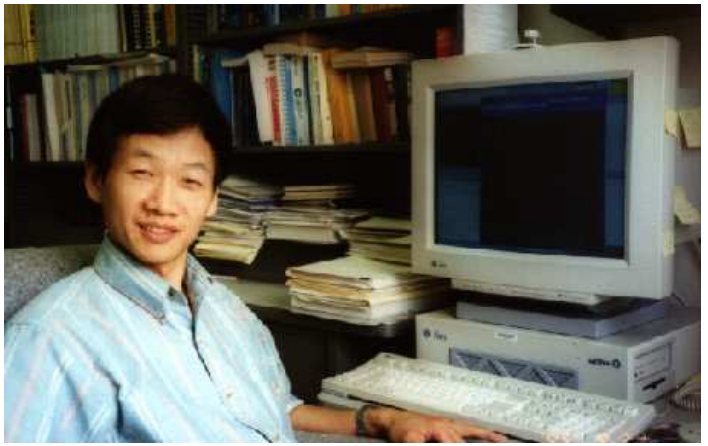




What's in a name, part III

- Some contenders for the name (other than Objective Bayes):
  - Probability
  - Inverse Probability
  - Noninformative Bayes
  - Default Bayes
  - Vague Bayes
  - Matching Bayes
  - Non-subjective Bayes
- But 'objective Bayes' has a website and soon will have *Objective Bayesian Inference*

(coming soon to a bookstore near you)



Nice Features of Objective Bayesian Analysis (for Astronomy)

1. Directly answering questions of interest, such as 'What is the probability that the theory is correct?'
2. Automatic Ockham's razor and multiplicity corrections
3. 'Correct' elimination of nuisance parameters

1. Directly Answering Questions of Interest

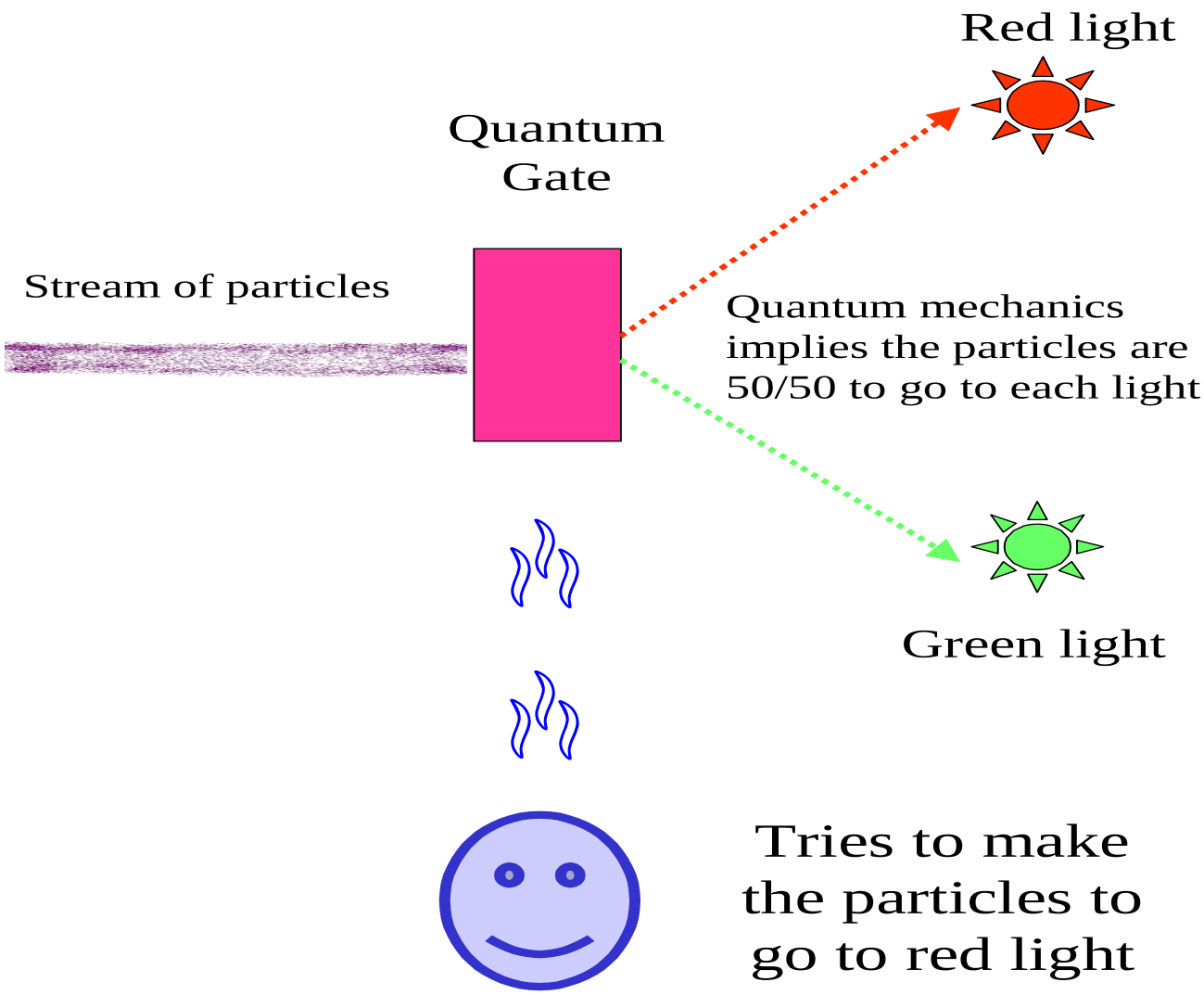
Objective Bayesian answers can be obtained for virtually all direct questions of interest, such as 'What is the probability that this hypothesis is correct?'

Psychokinesis Example:

Do subjects possess psychokinetic ability?

The experiment:

Schmidt, Jahn and Radin (1987) used electronic and quantum-mechanical random event generators with visual feedback; the subject with alleged psychokinetic ability tries to "influence" the generator.


Red light
Quantum
Gate
Stream of particles
Quantum mechanics
implies the particles are
50/50 to go to each light
Green light
Tries to make
the particles to
go to red light

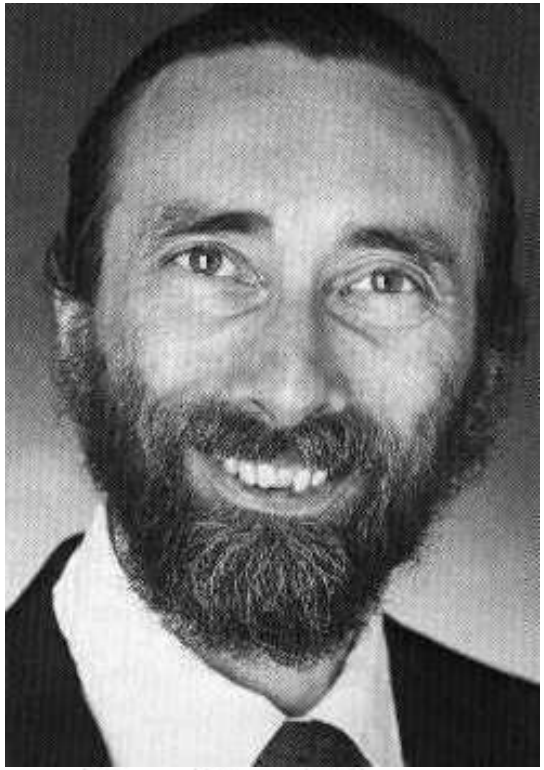
Data and model:

- Each particle is a Bernoulli trial (red = 1, green = 0)
  θ = probability of "1"
  $n = 104,490,000$ trials
  X = # "successes" (# of 1's), $\quad X \sim \text{Binomial}(n, \theta)$
  $x = 52,263,470$ is the actual observation

  $$\begin{aligned} \text{To test } H_0 : \theta &= \tfrac{1}{2} \text{ (subject has no influence)} \\ \text{versus } H_1 : \theta &\neq \tfrac{1}{2} \text{ (subject has influence)} \end{aligned}$$

- P-value $= P_{\theta=\frac{1}{2}}(|X - \frac{n}{2}| \geq |x - \frac{n}{2}|) \approx .0003.$
  Is there strong evidence against H_0 (i.e., strong evidence that the subject influences the particles) ?

Bayesian Analysis:
(Jefferys, 1990)



Prior distribution:

$Pr(H_i)$ = prior probability that
H_i is true, $i = 0, 1$;

On $H_1 : \theta \neq \frac{1}{2}$, $\pi(\theta)$ is
the prior density for θ.

Objective Bayes: choose
$Pr(H_0) = Pr(H_1) = \frac{1}{2}$, $\pi(\theta) = 1$ (on $0 < \theta < 1$).

Subjective Bayes: choose the $Pr(H_i)$ and $\pi(\theta)$ based on personal beliefs. For the latter, one might consider

$\pi_r(\theta)$ = uniform density on $(\frac{1}{2} - r, \frac{1}{2} + r)$;

r could be viewed as the largest expected success probability, given that psychokinesis exists.

Posterior probability of the null hypothesis:

$$\begin{aligned} Pr(H_0 \mid x) &= \text{probability } H_0 \text{ is true, given data } x \\ &= \frac{f(x \mid \theta{=}\frac{1}{2})\, Pr(H_0)}{Pr(H_0)\, f(x \mid \theta{=}\frac{1}{2}) + Pr(H_1) \int f(x \mid \theta)\pi(\theta)d\theta} \end{aligned}$$

For the objective prior,

$Pr(H_0 \mid x = 52,263,470) \approx 0.92$

(recall, p-value $\approx .0003$)

Posterior density on $H_1 : \theta \neq \frac{1}{2}$ is

$$\pi(\theta \mid x, H_1) \propto \pi(\theta) f(x \mid \theta) \propto 1 \times \theta^x (1-\theta)^{n-x},$$

the $Be(\theta \mid 52,263,470\ ,\ 52,226,530)$ density.

Issue 1. Approximating a believable null hypothesis by a precise null

A precise null, like $H_0 : \theta = \theta_0$, is typically never true *exactly*; rather, it is used as a surrogate for a 'real null'

$$H_0^\epsilon : |\theta - \theta_0| < \epsilon, \quad \epsilon \text{ small.}$$

Result (Berger and Delampady, 1989):
if $\epsilon < \frac{1}{4}\,\sigma_{\hat{\theta}}$, where $\sigma_{\hat{\theta}}$ is the standard error of the estimate of θ, then $Pr(H_0^\epsilon \mid \mathbf{x}) \approx Pr(H_0 \mid \mathbf{x})$.
(*Note:* this will typically be violated for very large n.)

Issue 2. Bayesian reporting in hypothesis testing

- The complete posterior distribution is given by
  - $Pr(H_0 \mid x)$, the posterior probability of null hypothesis
  - $\pi(\theta \mid x, H_1)$, the posterior distribution of θ under H_1
- A useful *summary* of the complete posterior is
  - $Pr(H_0 \mid x)$
  - C, a (say) 95% posterior credible set for θ under H_1
- In the psychokinesis example
  - $Pr(H_0 \mid x) = .92 \rightsquigarrow$ gives the probability of H_0
  - $C = (.50008, .50027) \rightsquigarrow$ shows where θ is if H_1 is true
- For testing precise hypotheses, confidence intervals alone are *not* a satisfactory inferential summary

Issue 3. Understanding the difference between p-values and Bayesian answers

In the psychokinesis example, p-value $\approx .0003$, but the objective posterior probability of the null ≈ 0.92.

- In the example, a factor of 30 is due to the difference between a tail area $\{X : \ |X - \frac{n}{2}| \geq |x - \frac{n}{2}|\}$ and the actual observation $x = 52,263,470$.
- The rest is due to the fact that the data is unusual under either hypothesis; but the degree of being 'unusual under H_1' depends on the prior $\pi(\theta)$. For the subjective $\pi_r(\theta)$ (uniform on $(0.5 - r, 0.5 + r)$), $P(H_0 \mid x)$ ranges between 0.009 (achieved at $r = 0.00022$) and 0.92.

- Only if the experimenter had a priori specified a value of r between 0.0001 and 0.0024, would the evidence for H_1 be at least 20 to 1.
- How can data arise that is unusual under either hypothesis?
  - Experimental bias in equipment? (But there were control runs.)
  - Incorrect model? (Indeed a binomial mixture model would have been better, but the p-value computation is not affected.)
  - Experimental bias from subjects or operators?
  - Optional stopping?

Calibration of p-values: (Sellke, Bayarri and Berger, 2001)

- A *proper* p-value satisfies $H_0 : p(X) \sim \text{Uniform}(0, 1)$.
- Consider testing this versus H_1, a reasonable nonparametric alternative for $p(X)$.
- Then it can be shown that, if $p < e^{-1}$, a *lower bound* on the objective posterior probability of H_0 (or the conditional Type I frequentist error probability) is

$$P(H_0 \mid p) \geq (1 + [-e\, p \log(p)]^{-1})^{-1}\,.$$

p	.2	.1	.05	.01	.005	.001
$P(H_0 \mid p)$	.465	.385	.289	.111	.067	.0184

Example: Are gamma ray bursts galactic or extra-galactic in origin?

- data in early 90's were 260 observed burst directions
- H_0: data are uniformly directionally distributed (implying extra-galactic origin)
- standard test for uniformity rejected at $p = 0.027$
- $P(H_0 \mid p) \geq (1 + [-e\,(.027)\log(.027)]^{-1})^{-1} = .21$, so the actual error rate in rejecting H_0 is *at least* .21

2. Automatic Ockham's Razor and Multiplicity Correction

- Bayesian analysis acts as an automatic Ockham's razor, greatly preferring simple models that reasonably explain the data to complex models (Jefferys and Berger, 1992)
- Bayesian analysis automatically corrects for multiple tests; no adhoc penalization is required.

Example of multiple comparisons (as would apply to microarray analysis) (Scott and Berger, 1993)

- Suppose $x_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, m$, are observed, with σ^2 known, and it is desired to determine which μ_i are nonzero.
- Most of the μ_i are thought to be zero; it is desired to find those that are nonzero. Let p denote the *unknown* common prior probability that μ_i is zero.
- Assume that the nonzero μ_i follow a $N(0, V)$ distribution, with V unknown.
- Assign p the uniform prior on $(0, 1)$ and V the prior density $\pi(V) = \sigma^2/(\sigma^2 + V)^2$.

- Then the posterior probability that $\mu_i \neq 0$ is

$$p_i = 1 - \frac{\int_0^1 \int_0^1 p \prod_{j \neq i} \left(p + (1-p)\sqrt{1-w}\, e^{w {x_j}^2/(2\sigma^2)} \right) dp dw}{\int_0^1 \int_0^1 \prod_{j=1}^m \left(p + (1-p)\sqrt{1-w}\, e^{w {x_j}^2/(2\sigma^2)} \right) dp dw}.$$

- $(p_1, p_2, \ldots, p_m)$ can be computed numerically if m is moderate. For large m, it is most efficient to do the computation via importance sampling, with a common importance sample for all p_i.

Example: Consider the following ten 'signal' observations:
-8.48, -5.43, -4.81, -2.64, -2.40, 3.32, 4.07, 4.81, 5.81, 6.24
Generate $n = 10, 50, 500,$ and 5000 $N(0,1)$ 'noise' observations.
Mix them together and try to identify the signals.

	Central seven 'signal' observations							#noise
n	-5.4	-4.8	-2.6	-2.4	3.3	4.1	4.81	$p_i > .6$
10	1	1	.94	.89	.99	1	1	1
50	1	1	.71	.59	.94	1	1	0
500	1	1	.26	.17	.67	.96	1	2
5000	1.0	.98	.03	.02	.16	.67	.98	1

Table 1: The posterior probabilities of being nonzero for the central 'signal' means (the others always had $p_i = 1$). *Note:* The penalty for multiple comparisons is automatic; one does not need any adjustments (e.g. Bonferoni).

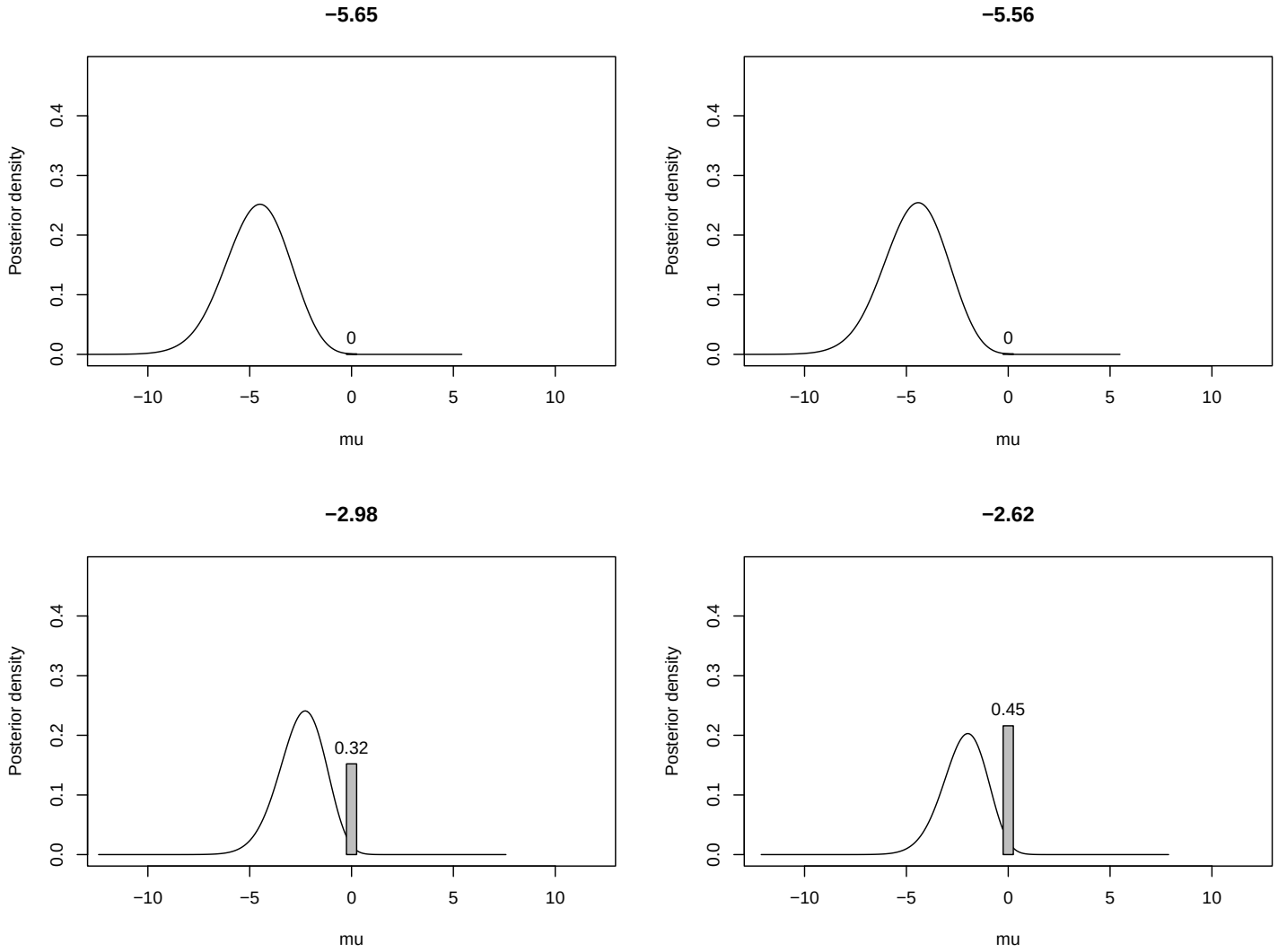


Figure 1: For four of the observations, $1 - p_i = \Pr(\mu_i = 0 \,|\, \boldsymbol{y})$ (the vertical bar), and the posterior densities for $\mu_i \neq 0$.

3. Eliminating Numerous Nuisance Parameters by Marginalization

Example: The Neyman-Scott problem: Suppose we observe

$$X_{ij} \sim \mathcal{N}(\mu_i,\ \sigma^2), \quad i = 1, \ldots, n;\ j = 1, 2.$$

Estimating σ^2 is of interest (or confidence sets for the μ_i). Defining $\bar{x}_i = (x_{i1} + x_{i2})/2$, $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_n)$, $S^2 = \sum_{i=1}^{n} \sum_{j=1}^{2} (x_{ij} - \bar{x}_i)^2$, and $\mu = (\mu_1, \ldots, \mu_n)$, the likelihood function (under M_2) can be written

$$L(\mu, \sigma) \propto \frac{1}{\sigma^{2n}} \exp\left[-\frac{1}{2\sigma^2}(2|\bar{\mathbf{x}} - \mu|^2 + S^2)\right].$$

The maximum likelihood estimates are $\hat{\mu}_i = \bar{x}_i$ and $\hat{\sigma}^2 = S^2/(2n)$. But $\hat{\sigma}^2 \to \sigma^2/2$ for large n, a bad estimate.

Objective Bayesian approach: The objective prior (reference or independence Jeffreys) for the problem is $\pi^N(\mu, \sigma) = 1/\sigma$, and the nuisance parameters are eliminated via marginalization, leading to the posterior distribution for σ^2

$$\begin{aligned} \pi(\sigma^2 \mid \mathbf{x}) &\propto \int \frac{1}{\sigma^{(2n+1)}} \exp\left[-\frac{1}{2\sigma^2}(2|\bar{\mathbf{x}} - \mu|^2 + S^2)\right] d\mu \\ &\propto \frac{1}{\sigma^{(n+1)}} \exp\left[-\frac{S^2}{2\sigma^2}\right]. \end{aligned}$$

with resulting estimates (posterior means) $\hat{\mu}_i = \bar{x}_i$ and $\hat{\sigma}^2 = S^2/n$.

Example: Trans-Neptunian Objects (Loredo, 1994).

The distribution of size D of TNOs follows a power law

$$f(D) \propto D^{-q} \, .$$

TNOs have a density distribution that varies with heliocentric radius, r, as

$$n(r) \propto r^{-\beta} \, .$$

The goal is to estimate q and β.

Key nuisance parameters are the magnitudes, m_i, of the optical flux for the observed TNOs, $i = 1, \ldots, N$.

- If estimates, $\hat{m}_i$, are simply plugged into the likelihood, bad estimates of q and β can result.
- Eliminating the m_i by marginalization works.

Bill will be riding away into the sunset, but

HE WILL BE BACK!

(among others, at the Spring 2006 Astrostatistics Program at SAMSI)