Objective Bayesian Analysis: its uses in practice and its role in the unification of statistics

James O. Berger

Duke University and the

Statistical and Applied Mathematical Sciences Institute

Allen T. Craig Lectures, University of Iowa,
April 8-9, 2004

Outline

- Introduction to objective Bayesian analysis through a medical diagnosis example.
- Objective and subjective Bayesian analysis - history.
- Nice features of objective Bayesian analysis, through examples.
- Unification of frequentist and objective Bayesian statistics
  - in estimation (correlation coefficient);
  - the conditioning hurdle;
  - hypothesis testing.

I. Introduction to 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.

A Medical Diagnosis Example (with Mossman, 2001)

The Medical Problem:

- Within a population, $p_0 = Pr(\text{Disease } D)$.
- A diagnosic test results in either a Positive (P) or Negative (N) reading.
- $p_1 = Pr(P \,|\, \text{patient has } D)$.
- $p_2 = Pr(P \,|\, \text{patient does not have } D)$.

It follows from Bayes theorem that

$$\theta = Pr(D|P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0) p_2}.$$

The Statistical Problem: The p_i are unknown. Based on (independent) data $X_i \sim \text{Binomial}(n_i, p_i)$ (arising from medical studies), find a $100(1-\alpha)\%$ confidence set for θ.

Suggested Solution: Assign p_i the Jeffreys-rule prior

$$\pi(p_i) \propto p_i^{-1/2}(1-p_i)^{-1/2}.$$

By Bayes theorem, the posterior distribution of p_i given the data, x_i, is

$$\pi(p_i \mid x_i) = \frac{p_i^{-1/2}(1-p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}}{\int p_i^{-1/2}(1-p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n_i - x_i}\, dp_i},$$

which is the $\text{Beta}(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ distribution.

Finally, compute the desired confidence set (formally, the $100(1-\alpha)\%$ equal-tailed posterior credible set) by

- drawing random p_i from the Beta$(x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2})$ distributions, $i = 0, 1, 2$;
- computing the associated $\theta = p_0 p_1 / [p_0 p_1 + (1 - p_0) p_2]$;
- repeating this process $10,000$ times;
- using the $\frac{\alpha}{2}\%$ upper and lower percentiles of these generated θ to form the desired confidence limits.

$n_0 = n_1 = n_2$	(x_0, x_1, x_2)	95% confidence interval
20	(2,18,2)	(0.107, 0.872)
20	(10,18,0)	(0.857, 1.000)
80	(20,60,20)	(0.346, 0.658)
80	(40,72,8)	(0.808, 0.952)

Table 1: The 95% equal-tailed posterior credible interval for $\theta = p_0 p_1 / [p_0 p_1 + (1 - p_0) p_2]$, for various values of the n_i and x_i.

II. Subjective vs Objective Bayes Analysis

- In *subjective Bayesian* analysis
  - prior distributions represent beliefs or knowledge;
  - Bayes theorem shows how prior beliefs become posterior beliefs, through learning from data.
  - The subjective approach is crucial in many areas of design, metaanalysis, and decision making.
- In *objective Bayesian* analysis
  - prior distributions are chosen to represent 'neutral' knowledge about unknowns;
  - the posterior distribution is claimed to give the probability of unknowns arising from just data.

History of Objective and Subjective Bayes

- *Objective Bayesian* inference, using constant prior densities for unknowns, was prominent from 1775–1925, under the name *inverse probability*.
- By 1940, the prevailing statistical philosophies were either *Fisherian* (associated with Ronald Fisher) or *frequentist* (associated with Jerzy Neyman).
- *Subjective Bayesian* analysis became prominent by 1960; it is often mistakenly equated with Bayesianism
- Harold Jeffreys revived the *objective Bayesian* school from 1940-1970.

III. Some Nice Features of Objective Bayesian Analysis

- It is as objective as anything else in statistics.
- It allows direct answer of questions of interest (examples).
- It is often the easiest way to obtain good frequentist procedures (examples).
- It is great for difficult problems, such as dealing with many nuisance parameters and multiplicity of hypotheses (examples).

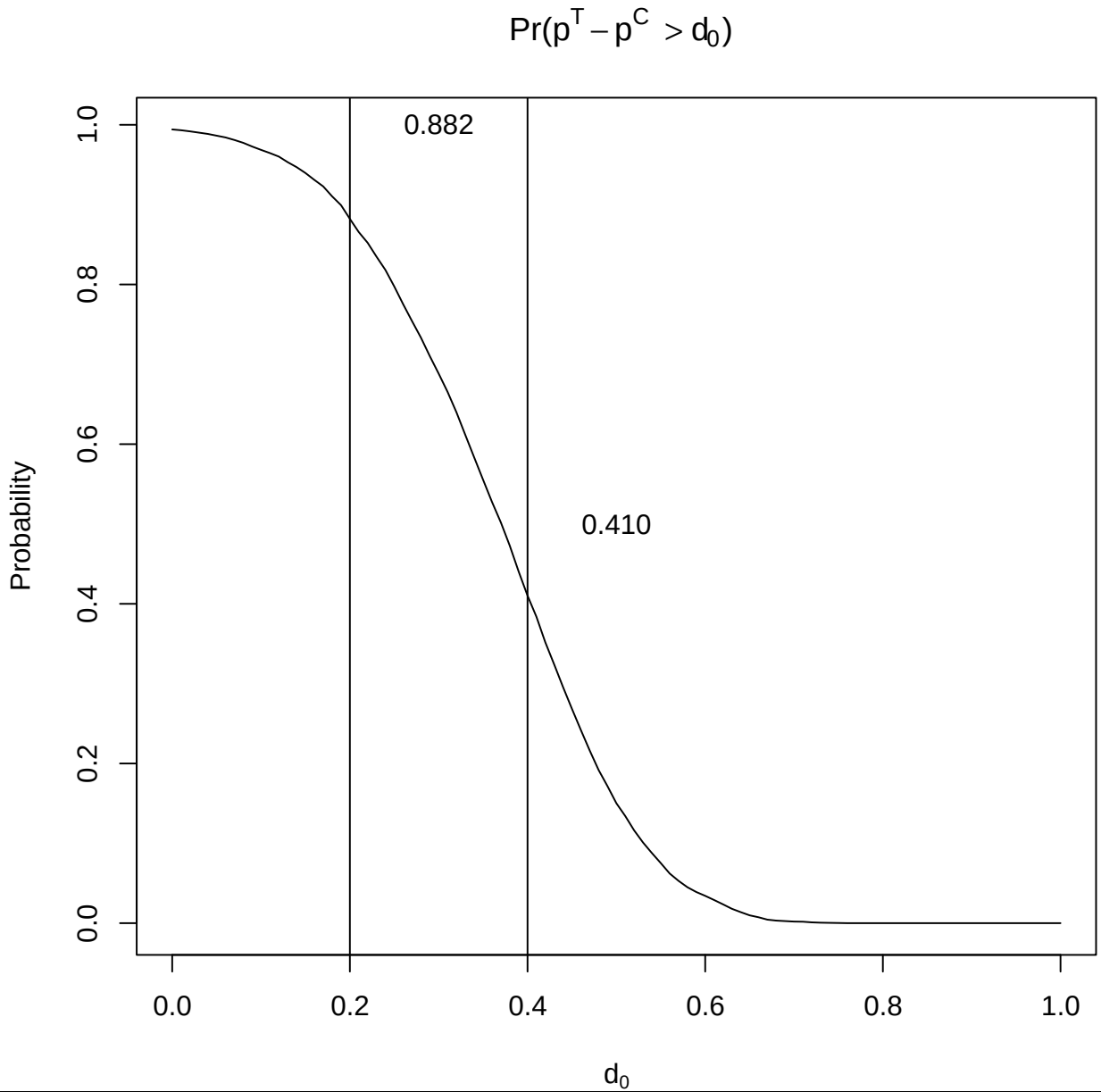
1. Directly Answering Questions of Interest

Bayesian answers can be obtained for virtually all direct questions of interest, such as

- What is the probability that drug A improves average survival by at least 3 months over drug B?
- What is the probability that a new drug is as effective as the standard drug?

Questions like this are answered only indirectly using classical methods.

Indeed, one can give an entire probability curve for the difference in performance of a treatment and control.


$Pr(p^T - p^C > d_0)$
0.882
0.410
Probability
0.0
0.2
0.4
0.6
0.8
1.0
d_0

2. Obtaining Good Frequentist Procedures

Objective Bayesian procedures are usually great frequentist procedures. In the medical diagnosis example, consider the frequentist percentage of the time that the 95% Bayesian sets for $\theta = Pr(D|P) = \frac{p_0 p_1}{p_0 p_1 + (1-p_0)p_2}$ miss on the left and on the right (ideal would be 2.5% each) for the indicated parameter values when $n_0 = n_1 = n_2 = 20$.

(p_0, p_1, p_2)	O-Bayes	Log Odds	Gart-Nam	Delta
$(\frac{1}{4}, \frac{3}{4}, \frac{1}{4})$	2.86,2.71	1.53,1.55	2.77,2.57	2.68,2.45
$(\frac{1}{10}, \frac{9}{10}, \frac{1}{10})$	2.23,2.47	0.17,0.03	1.58,2.14	0.83,0.41
$(\frac{1}{2}, \frac{9}{10}, \frac{1}{10})$	2.81,2.40	0.04,4.40	2.40,2.12	1.25,1.91

3. Dealing with Nuisance Parameters and Multiplicity in Hypothesis Testing

A. Nuisance Parameters: Complex models usually have many nuisance parameters, and it is difficult to account for the uncertainty in their values when eliminated in non-Bayesian ways (e.g., by using plug-in estimates).

By eliminating nuisance parameters through integration, objective Bayesian analysis automatically accounts for the uncertainty in the nuisance parameters.

Example: Hierarchical or random effects or mixed models or multilevel models or ...

A very simple version: For $i = 1, \ldots, p$,

$$X_i \sim \text{Normal}(\mu_i, 1) \quad \text{and} \quad \mu_i \sim \text{Normal}(0, \tau^2).$$

If $S^2 = \sum X_i^2 < p$, the mle for τ^2 is $\hat{\tau}^2 = 0$ and the unbiased estimate is negative. (With numerous variance components, this is a common occurrence. Even here, for $p = 4$ and $\tau^2 = 1$, $\Pr(S^2 < p) = 0.264$.)

The likelihood, $L(\tau^2)$, decreases away from 0 very slowly, indicating considerable uncertainty about τ^2, even though the mle is 0.

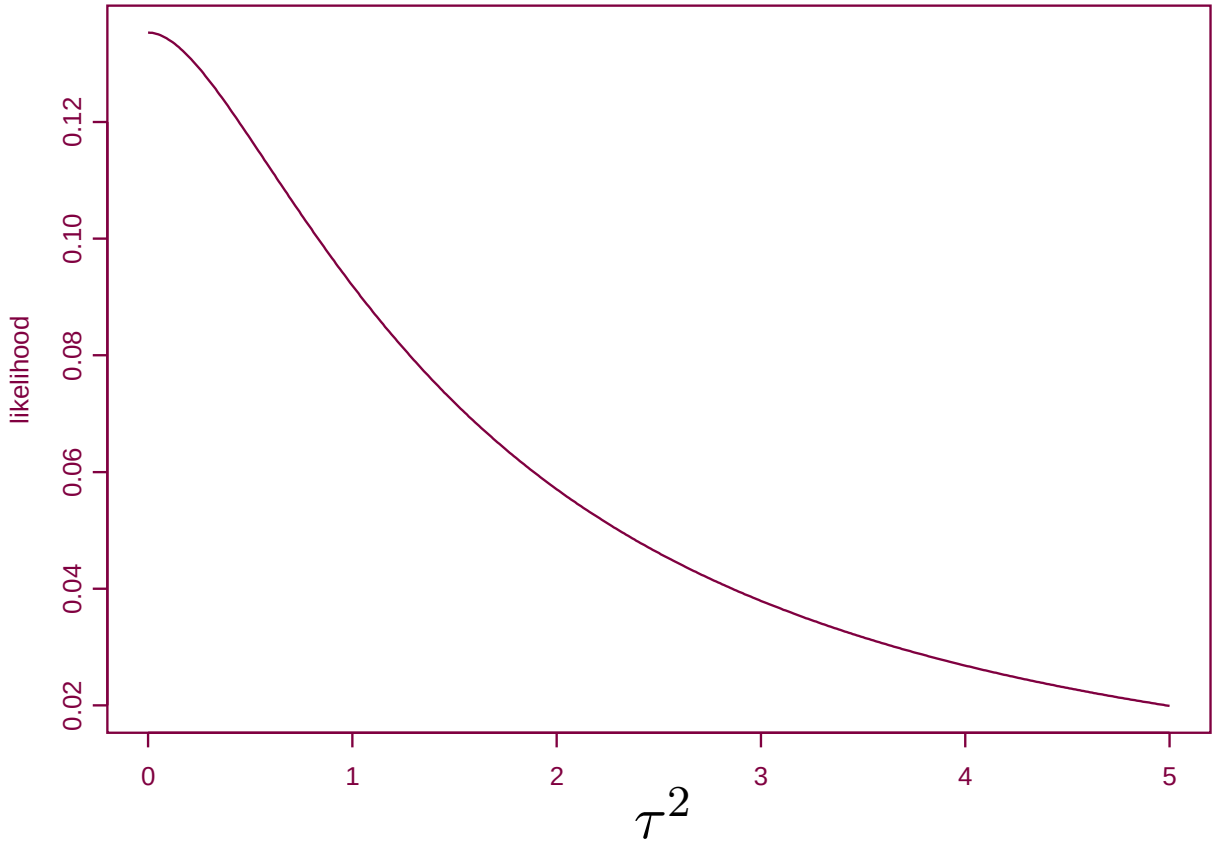


Figure 1: Likelihood function of τ^2 when $p = 4$ and $S^2 = 4$ is observed.

Neglecting uncertainty in τ^2 affects the analysis in an incorrectly aggressive fashion.

Setting τ^2 to 0 is equivalent to setting $\mu_1 = \ldots = \mu_p = 0$.

Classical methods have difficulty incorporating the uncertainty in τ^2, because the maximum is achieved at a boundary.

Objective Bayes analysis

- leads to a posterior for τ^2 that reflects the uncertainty in the likelihood;
- can be easily implemented computationally for very complex hierarchical models using BUGS.

B. Dealing with Multiplicity

Objective Bayesian analysis does not require Bonferroni type adjustments for multiple testing; adjustments are automatic.

Example: Microarray Analysis

- The probability model and analysis goal for the data:
  - Suppose, for $i = 1, \ldots, m$, that observation X_i is normally distributed with mean μ_i and variance 1, to be denoted $X_i \sim N(\mu_i, 1)$.
  - Most of the μ_i are thought likely to be zero, and it is desired to detect those that are nonzero.

- The probability model (prior distribution) for the unknown μ_i:
  - Let p denote the *unknown* prior probability that a given μ_i is zero; thus $1 - p$ is the prior probability that it is nonzero:
    - assign p the uniform density on the interval $(0, 1)$.
  - Assume that the nonzero μ_i follow a $N(0, V)$ distribution, with V unknown:
    - assign V the prior density $\pi(V) = 1/(1 + V)^2$.

- Of primary interest are the posterior probabilities that the μ_j are nonzero, given by

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

- $(p_1, p_2, \ldots, p_m)$ 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_j.

- Note that, in the Bayesian approach, the 'penalty' for adding additional (mean zero) comparisons is that p will concentrate on smaller values. This is automatic, and one need not do any adjustments (e.g. Bonferoni).

Example: Draw ten $N(0, 4^2)$ 'signal' observations:
-8.48, -5.43, -4.81, -2.64, -2.40, 3.32, 4.07, 4.81, 5.81, 6.24
Generate $n = 10, 50, 500, \text{and } 5000$ $N(0, 1)$ 'noise' observations.

Mix them together and try to identify the signals.

n	Central seven 'signal' observations							#noise
	-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 2: The posterior probabilities that the central 'signal' means are nonzero (the others always had $p_i = 1$). *Note:* The penalty for multiple comparisons is automatic.

IV. Objective Bayes and the Unification of Statistics

Estimation Example - the Correlation Coefficient:

The bivariate normal distribution of (x_1, x_2) has mean (μ_1, μ_2) and covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, where ρ is the correlation between x_1 and x_2.

For a sample $(x_{11}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{1n}, x_{2n})$, define

$$\boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{x}})(\boldsymbol{X}_i - \overline{\boldsymbol{x}})' = \begin{pmatrix} s_{11} & r\sqrt{s_{11}s_{22}} \\ r\sqrt{s_{11}s_{22}} & s_{22} \end{pmatrix},$$

where $s_{ij} = \sum_{k=1}^{n} (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)$, $r = s_{12}/\sqrt{s_{11}s_{22}}$.

Credible Intervals for ρ, under the right-Haar prior

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) \quad = \quad \frac{1}{\sigma_1^2(1-\rho^2)}\,,$$

can be found by

- drawing independent $Z \sim \mathrm{N}(0,1)$, χ^2_{n-1} and χ^2_{n-2};
- setting
  $\rho = \frac{Y}{\sqrt{1+Y^2}}$, where $Y = -\frac{Z}{\sqrt{\chi^2_{n-1}}} + \frac{\sqrt{\chi^2_{n-2}}}{\sqrt{\chi^2_{n-1}}}\frac{r}{\sqrt{1-r^2}}$;
- repeating this process $10,000$ times;
- using the $\frac{\alpha}{2}\%$ upper and lower percentiles of these generated ρ to form the desired confidence limits.

Lemma 1

1. This Bayesian credible set, $C(r)$, when considered as a frequentist confidence interval, has exact coverage $1-\alpha$.

2. This credible set can be shown to be the same as the fiducial confidence interval obtained by Fisher in 1930.

Two Historical Curiosities:

1. Was it known that Fisher's fiducial interval has exact frequentist coverage of $1-\alpha$?

2. In the early 60's, results of Lindley and Brillinger showed that, if one starts with the density $f(r \mid \rho)$ of r, there is no prior distribution for ρ which has a posterior equal to the fiducial distribution.

The Main Hurdle to Unification: The Conditioning Problem

Basic question for a frequentist: What is the sequence of possible data for which to consider frequentist evaluations?

(Fisher: "relevant subset;" Lehmann: "frame of reference.")

Artificial example: Observe X_1 and X_2, where

$$X_i = \begin{cases} \theta + 1 & \text{with probability } 1/2 \\ \theta - 1 & \text{with probability } 1/2. \end{cases}$$

Consider the confidence set for θ

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{if } X_1 \neq X_2 \\ X_1 - 1 & \text{if } X_1 = X_2. \end{cases}$$

Unconditional coverage:

$$P_\theta(C(X_1, X_2) \text{ contains } \theta) = 0.75.$$

This is silly: if $x_1 \neq x_2$, we know $C(x_1, x_2) = \theta$;
if $x_1 = x_2$, $C(X_1, X_2)$ equals θ only with probability 1/2.

One must use the *conditional frequentist* **approach:**

- Define the conditioning statistic $S = |X_1 - X_2|$, measuring the "strength of evidence" in the data (here ranging from $s = 0$ to $s = 2$);
- Compute frequentist coverage conditional on the strength of evidence S.

$$\begin{aligned} P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, s = 2) &= 1 \\ P_\theta(C(X_1, X_2) \text{ contains } \theta \,|\, s = 0) &= \tfrac{1}{2}. \end{aligned}$$

Note: The correct answer is obtained trivially by objective Bayesian analysis, using a constant prior density for θ.

Genetics Example (A. Kong): Mapping genes for complex traits based on affected half-sib data X.

Goal: A 95% confidence set for

$\theta =$ true location of a susceptibility gene.

Unconditional method: Kruglyak and Lander (1995) found $C^u(X)$ such that $P_\theta(C^u(X) \text{ contains } \theta) = 0.95$.

Conditional method: Conditioning on the location ancillary statistic S results in a confidence set $C^c(X)$ such that $P_\theta(C^c(X) \text{ contains } \theta|S) = 0.95$.

Differences can be significant:

$1 - \alpha^u(s) = P_\theta(C^u(X) \text{ contains } \theta|S)$ can be near zero for some s. Furthermore, $Pr(1 - \alpha_u(S) < 0.90) = 0.11$.

Note: Objective Bayes easily gives the correct answer.

The Example of Simple versus Simple Testing:

For testing simple H_0 vs. simple H_1,

- classical N-P tests yield unconditional frequentist error rates (α, β), but these rates do not reflect the 'conditional' evidence as the data varies within the rejection or acceptance regions;
- p-values are data-dependent but do not have a strict frequentist interpretation.

One indication of the non-frequentist nature of p-values can be seen from the following *applet* (available on my web page www.stat.duke.edu/∼berger). The situation considered has:

- normal data with unknown mean θ and known variance;
- tests of the form $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

The applet simulates a long series of such tests, and records how often H_0 is true for p-values in given ranges.

The *Conditional Frequentist Approach* provides true conditional frequentist error rates, and also unifies frequentist and objective Bayesian testing.

- Let p_i be the p-value from testing H_i against the other hypothesis.
- While not frequentist error probabilities, Fisher argued that p-values are good 'measures of evidence,' so define the conditioning statistic $S = \max\{p_0, p_1\}$; its use is equivalent to deciding that data (in either the rejection or acceptance regions) with the same p-value has the same 'strength of evidence.'
- Accept H_0 when $p_0 > p_1$, and reject otherwise.

- Compute Type I and Type II conditional error probabilities (CEP) as

$$\begin{aligned}\alpha(s) &= P_0(\text{rejecting } H_0|S=s) \equiv P_0(p_0 \le p_1|S(X)=s)\\ \beta(s) &= P_1(\text{accepting } H_0|S=s) \equiv P_1(p_0 > p_1|S(X)=s).\end{aligned}$$

- The potential *unification:*
  - The evidentiary content of p-values is acknowledged, but 'converted' to error probabilities by conditioning.
  - The conditional error probabilities $\alpha(s)$ and $\beta(s)$ are fully data-dependent, yet fully frequentist.
  - $\alpha(s)$ and $\beta(s)$ are exactly equal to the (objective) posterior probabilities of H_0 and H_1, respectively.

History of conditional frequentist testing

- Many Fisherian precursors (the Fisher exact test).
- General theory in Kiefer (1977).
- Brown (1978) found optimal conditional frequentist tests for testing 'symmetric' simple hypotheses.
- Berger, Brown and Wolpert (1994) developed the theory we discuss for testing simple hypotheses.
- Berger, Boukai and Wang (1997a, 1997b) generalized to simple versus composite hypothesis testing.
- Dass (2001) generalized to discrete settings;
- Dass and Berger (2003) to composite hypotheses.

Final Comments

- Foundational Comments
  - We do not pay enough attention to good conditional performance:
    - Bayesians, because it is automatic;
    - frequentists, because a practical theory is elusive.
  - Focusing on conditional performance brings frequentists and Bayesians (and Fisherians) closer together foundationally and methodologically.

- Practical Comments on Objective Bayesian Analysis
  - It allows direct answer of natural questions.
  - It is often the easiest way to obtain good frequentist procedures.
  - It is great for difficult problems, including problems with many nuisance parameters or hypotheses.
  - It has other methodological benefits, such as
    - not having a 'spending α' penalty in sequential trials;
    - being able to deal with multiple possible models through model averaging;
    - ...