Practical Bayesian Data Analysis from a Former Frequentist

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 800717 Charlottesville VA 22908 USA
fharrell@virginia.edu
hesweb1.med.virginia.edu/biostat

MASTERING STATISTICAL ISSUES IN DRUG DEVELOPMENT
HENRY STEWART CONFERENCE STUDIES
15-16 MAY 2000

Abstract

Traditional statistical methods attempt to provide objective information about treatment effects through the use of easily computed P–values. However, much controversy surrounds the use of P–values, including statistical vs. clinical significance, artificiality of null hypotheses, 1–tailed vs. 2–tailed tests, difficulty in interpreting confidence intervals, falsely interpreting non–informative studies as ”negative”, arbitrariness in testing for equivalence, trading off type I and type II error, using P–values to quantify evidence, which statistical test should be used for 2×2 frequency tables, α–spending and adjusting for multiple comparisons, whether to adjust final P–values for the intention of terminating a trial early even though it completed as planned, complexity of group sequential monitoring procedures, and whether a promising but statistically insignificant trial can be extended. Bayesian methods allow calculation of probabilities that are usually of more interest to consumers, e.g. the probability that treatment A is similar to treatment B or the probability that treatment A is at least 5% better than treatment B, and these methods are simpler to use in monitoring ongoing trials. Bayesian methods are controversial in that they require the use of a prior distribution for the treatment effect, and calculations are more complex in spite of the concepts being simpler. This talk will discuss advantages of estimation over hypothesis testing, basics of the Bayesian approach, approaches to choosing the prior distribution and arguments for favoring non–informative priors in order to let the data speak for themselves, pros and cons of traditional and Bayesian inference, relating the bootstrap to the Bayesian approach, possible study design criteria, sample size and power issues, and implications for study design and review. The talk will

use several examples from clinical trials including GUSTO (t–PA vs. streptokinase for acute MI), a meta–analysis of possible harm from short–acting nifedipine, and interpreting results from an unplanned interim analysis. BUGS code will be given for these examples. The presentation will show how the Bayesian approach can solve many common problems such as not having to deal with how to “spend α” when considering multiple endpoints and sequential analyses. An example clinical trial design that allows for continuous monitoring for efficacy, safety, and similarity for two endpoints is given.

Major Topics and Suggested Schedule

- 9:00a – 10:30a
  - Overview of Methods for Quantifying and Acting on Evidence
  - Frequentist Statistical Inference
  - What’s Wrong with Hypothesis Testing?
  - Confidence Intervals
  - Overview of Bayesian Approach
  - The Standardized Likelihood Function
  - Bayesian Inferential Methods
- 10:30a – 11:00a: Break
- 11:00a – 12:30p
  - Three Types of Multiplicity
  - The Bootstrap
  - 2×2 Table Example
  - Software

  - Examples from Clinical Trials

- 12:30p – 1:30p: Lunch

- 1:30p – 2:15p
  - Meta–Analysis Example
  - Unplanned Interim Analysis Example
  - Example Study Designs
  - Power and Sample Size
  - Acceptance of Methods by Regulators & Industry
  - Summary

- 2:15p – 3:30p: General discussion

Outline

- Quantifying Evidence vs. Decision Making
- Frequentist Statistical Inference
  - Methods
  - Advantages
  - Disadvantages and Controversies
- What’s Wrong with Hypothesis Testing?
  - The Applied Statistician’s Creed
  - Has hypothesis testing hurt science?
- Confidence Intervals
- Bayesian Approach
  - Brief Overview of Methods
  - Advantages
  - Disadvantages and Controversies
- The Standardized Likelihood Function
- Bayesian Inferential Methods

  - Choosing the Prior Distribution
  - One–Sample Binomial
  - Two–Sample Binomial
  - Two–Sample Gaussian
  - One–Sample Gaussian
  - Deriving Posterior Distributions
  - Using Posterior Distributions
- Sequential Testing
- Subgroup Analysis
- Inference for Multiple Endpoints
- The Bootstrap
- 2×2 Table Example: Traditional, Bayes, Bootstrap
- Software: BUGS and S-Plus
- Examples from Clinical Trials
- Suggested Design Criteria
- Example Study Design

- Power and Sample Size
- Implications for Design & Evaluation
- Acceptance of Methods by Regulators & Industry
- Summary

Overview

- Point estimate for population treatment difference
- Probability of a *statistic* conditional on an assumption we hope to gather evidence against
- Binary decision based on this P–value
- Selection of the variable of interest by a stepwise variable selection algorithm
- Interval estimate: set of all parameter values that if hypothesized to hold would not be rejected at $1 - \alpha$ level or

  Interval that gives desired coverage probability for a parameter *estimate*
- Probability of a *parameter* (e.g., population treatment difference) conditional on *current data*
- Entire probability distribution for the *parameter*

- Optimal binary decision given model, prior beliefs, loss function (e.g., patient utilities), data

- Relative evidence: odds ratio, likelihood ratio, Bayes factor

  E.g.: Whatever my prior belief about the therapy, after receiving the current data the odds that the new therapy has positive efficacy is 18 times as high as it was before these data were available

Medical Diagnosis Framework

- Traditional (frequentist) approach analogous to consideration of probabilities of test outcomes $\mid$ disease status (sensitivity, specificity)
- Post–test probabilities of disease are much more useful
- Debate about use of direct probability models (e.g., logistic) vs. classification
  - Recursive partitioning (CART)
  - Discriminant analysis
  - Classify based on $\hat{P}$ from logistic model

Decisions vs. Simply Quantifying Evidence

- Decision tree to structure options and outcomes
- Uncertainty about each outcome quantified using probabilities
- Consequences valued on utility scale
- Derive thresholds corresponding to different actions
- Classic decision–making example: Berry et al.[12]
  - Vaccine trial in children in a Navajo reservation
  - Goal: minimize number of cases of *Haemophilus influenzae* b cases in the Navajo Nation

Problems with “Canned” Decisions

- See Spiegelhalter (1986): Probabilistic prediction in patient management and clinical trials[71]

  > However, such a complete specification and analysis of the problem, even when accompanied by elaborate sensitivity analyses, often does not appear convincing or transparent to the practising clinician. Indeed, Feinstein has stated that ‘quantitative decision analysis is unsatisfactory for the realities of clinical medicine’, primarily because of the problem in ascribing an agreed upon measure of ‘utility’ to a health outcome

- In medical diagnosis framework, utilities and patient preferences are not defined until the patient is in front of the doctor

- Example: decision re: cardiac cath is based on patient age, beyond how age enters into pre–test prob. of coronary disease

- It is presumptuous for the analyst to make classifications into “diseased” and “non–diseased”
- The preferred *published* output of diagnostic modeling is $\hat{P}(D|X)$
- In therapeutic studies, probabilities of efficacy and of cost are very useful; decisions can be made at the point of care when utilities are available (and relevant)

Methods

- Attempt to demonstrate S assuming $\bar{S}$ and showing it’s unlikely
- Treat unknowns as constants
- Choose a test statistic T
- Compute $\mathrm{Pr}[T$ as or more impressive as one observed$|H_0]$
- Probabilities “refer to the frequency with which different values of statistics (arising from sets of data other than those which have actually happened) could occur for some fixed but unknown values of the parameters”[15]

Advantages

- Simple to think of unknown parameter as a constant
- P–values relatively easy to compute
- Accepted by most of the world
- Prior beliefs not needed at computation time
- Robust nonparametric tests are available without modeling

Disadvantages and Controversies

- "Have to decide which 'reference set' of groups of data which have not actually occurred we are going to contemplate"[15]; what is "impressive"?
- Conditions on what is unknowable (parameters) and does not condition on what is *already* known (the data)
- H_0:no effect is a boring hypothesis that is often not really of interest. It is more of a mathematical convenience.
- Do we really think that most treatments have truly an effect of 0.0 in "negative trials"?
- Does not address clinical significance
- If real effect is mean decrease in BP by 0.2 mmHg, large enough n will yield $P < 0.05$
- By some mistake, $\alpha = 0.05$ is often used as magic cutoff

- Controversy surrounding 1–tailed vs. 2–tailed tests [68, Chapter 12]
- No method for trading off type I and type II error
- No uniquely accepted P–value for 2×2 table!
  What is “extreme”: of all possible tables or all tables with same total no. of deaths?
  No consensus on the optimum procedure for obtaining a P–value (e.g., Pearson χ^2 vs. Fisher’s so–called exact test, continuity correction, likelihood ratio test, new unconditional tests).
- For ECMO trial, 13 P–values have been computed for the same 2×2 table, ranging from 0.001 to 1.0
- P–values very often misinterpreted[a]
- Must interpret P–values in light of other evidence since it is a probability for a *statistic*, not for drug benefit

[a] Half of 24 cardiologists gave the correct response to a 4–choice question.[24]

- Berger and Berry: $n = 17$ matched pairs, $P = 0.049$, the **maximum** $\Pr[H_0] = 0.79$
- $P = 0.049$ deceptive because it involves probabilities of more extreme *unobserved* data[8]
- In testing a point $H_0, P = 0.05$ “essentially does not provide any evidence against the null hypothesis” (Berger et al.[9]) — $\Pr[H_1|P = 0.05]$ will be near 0.5 in many cases if prior probability of truth of H_0 is near 0.5
- Confidence intervals frequently misinterpreted — consumers act as if “degree of confidence” is uniform within the interval
- Very hard to directly answer interesting questions such as $\Pr[\text{similarity}]$
- Standard statistical methods use subjective input from “the producer rather than the consumer of the data”[8]
- P–values can only be used to provide evidence

against a hypothesis, not to give evidence in favor of a hypothesis. Schervish[67] gives examples where P–values are incoherent: if one uses a P–value to gauge the evidence in favor of an interval hypothesis for a certain dataset, the P–value based on the same dataset but for a *more restrictive* sub–hypothesis (i.e., one specifying a subset of the interval) actually gives more support (larger P).

- Equal P-values do not provide equal evidence about a hypothesis[63]
- If use $P < 0.05$ as a binary event, evidence is stronger in larger studies[63] [68, P. 179-183]
- If use actual P-value, evidence is stronger in smaller studies[63]
- Goodman[41] showed how P–values can provide misleading evidence by considering "replication probability" — prob. of getting a significant result in

a second study given P–value from first study and given true treatment effect = observed effect in first study

Initial P–value	Probability of Replication
.10	.37
.05	.50
.01	.73
.005	.80
.001	.91

- See also Berger & Sellke[7]
- See [65, 27] for interpretations of P–values under alternative hypotheses
- Why are P–values still used?

  Feinstein[33] believes their status "... is a lamentable demonstration of the credulity with which modern scientists will abandon biologic wisdom in favor of any quantitative ideology that offers the specious allure of a mathematical replacement for sensible thought."

The Multiplicity Mess

- Much controversy about need for/how to adjust for multiple comparisons
- Do you want Pr[Reject $\mid$ this H_0 true] = 0.05, or Pr[Reject $\mid$ this and other H_0s true] = 0.05?
- If the latter, C.L.s must use e.g. $1 - \frac{\alpha}{k}$ conf. level $\longrightarrow$precision of a parameter estimate depends on what other parameters were estimated
- Rothman[62]:“The theoretical basis for advocating a routine adjustment for multiple comparisons is the ‘universal null hypothesis’ that ‘chance’ serves as the first–order explanation for observed phenomena. This hypothesis undermines the basic premises of empirical research, which holds that nature follows regular laws that may be studied through observations. A policy of not making adjustments for multiple comparisons is preferable because it will lead to fewer errors of interpretation when the data under evaluation are not

random numbers but actual observations on nature. Furthermore, scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings."

- Cook and Farewell[21]: If results are intended to be interpreted marginally, there may be no need for controlling experimentwise error rate. See also [68, P. 142-143].
- Need to distinguish between H_0: at least one of five endpoints is improved by the drug and H_0: the fourth endpoint is improved by the drug
- Many conflicting alternative adjustment methods
- Bonferroni adjustment is consistent with a Bayesian prior distribution which specifies that the probability that all null hypotheses is true is a constant (say 0.5) no matter how many hypotheses are tested[80]
- Even with careful Bonferroni adjustment, a trial with

20 endpoints could be declared a success if only one endpoint was “significant” after adjustment; Bayesian approach allows more sensible specification of “success”

- Much controversy about need for adjusting for sequential testing. Frequentist approach is complicated.
  Example: 5 looks at data as trial proceeds
  Looks had no effect, trial proceeded to end
  Usual $P = 0.04$, need to adjust upwards for having looked

  Two studies with identical experiments and data but with investigators with different intentions $\longrightarrow$ one might claim “significance”, the other not (Berry[10])
  Example: one investigator may treat an interim analysis as a final analysis, another may intend to wait.

- It gets worse — need to adjust “final” point estimates for having done interim analyses

- Freedman et al.[36] give example where such adjustment yields 0.95 CI that includes 0.0 even for data indicating that study should be stopped at the first interim analysis
- As frequentist methods use intentions (e.g., stopping rule), they are not fully objective[8]

> If the investigator died after reporting the data but before reporting the design of the experiment, it would be impossible to calculate a P–value or other standard measures of evidence.

- Since P–values are probabilities of obtaining a result as or more extreme than the study's result under repeated experimentation, frequentists interpret results by inferring "what would have occurred following results that were not observed at analyses that were never performed"[29].

What’s Wrong with Hypothesis Testing?

- Hypotheses are often “straw men” that are imagined by the investigator just to fit into the traditional statistical framework
- Hypotheses are often inappropriately chosen (e.g., $H_0 : \rho = 0$)
- Most phenomena of interest are not all–or–nothing but represent a continuum
- See [50] for an interesting review

The Applied Statistician's Creed

- Nester[56]:

(a) TREATMENTS — all treatments differ;

(b) FACTORS — all factors interact;

(c) CORRELATIONS — all variables are correlated;

(d) POPULATIONS — no two populations are identical in any respect;

(e) NORMALITY — no data are normally distributed;

(f) VARIANCES — variances are never equal;

(g) MODELS — all models are wrong;

(h) EQUALITY — no two numbers are the same;

(i) SIZE — many numbers are very small.

- ⟶no two treatments actually yield identical patient outcomes
- ⟶Most hypotheses are irrelevant

Has Hypothesis Testing Hurt Science?

- Many studies are powered to be able to detect a huge treatment effect
- $\longrightarrow$sample size too small $\longrightarrow$confidence interval too wide to be able to reliably estimate treatment effects
- "Positive" study can have C.L. of $[.1, .99]$ for effect ratio
- "Negative" study can have C.L. of $[.1, 10]$
- Physicians, patients, payers need to know the magnitude of a therapeutic effect more than whether or not it is zero
- "It is incomparably more useful to have a plausible range for the value of a parameter than to know, with whatever degree of certitude, what single value is untenable." — Oakes[58]
- Study may yield precise enough estimates of

relative treatment effects but not of absolute effects

- C.L. for cost–effectiveness ratio may be extremely wide
- Hypothesis testing usually entails fixing n; many studies stop with $P = 0.06$ when adding 20 more patients could have resulted in a conclusive study
- Many “positive” studies are due to large n and not to clinically meaningful treatment effects
- Hypothesis testing usually implies inflexibility[69]

- Cornfield[23]:

  > “Of course a re–examination in the light of results of the assumptions on which the pre– observational partition of the sample space was based would be regarded in some circles as bad statistics. It would, however, be widely regarded as good science. I do not believe that anything that is good science can be bad statistics, and conclude my remarks with the hope that there are no statisticians so inflexible as to decline to analyze an honest body of scientific data simply because it fails to conform to some favored theoretical scheme. If there are such, however, clinical trials, in my opinion, are not for them.”

- If H_0 is rejected, practitioners often behave as if point estimate of treatment effect is population value

Confidence Intervals

- Misinterpreted twice as often as P–values
- Are one–dimensional: consumers interpret a confidence interval for OR of $[.35, 1.01]$ as saying that a 1% increase in mortality is as likely as a 10% decrease
- Confidence plots (with continuously varying $1-\alpha$) can help[13, 28], but their interpretation is complex

Methods

- Attempt to answer question by computing probability of the truth of a statement
- Let S denote a statement about the drug effect, e.g., patients on drug live longer than patients on placebo
- Want something like $\mathrm{Pr}[S|\text{ data}]$
- If θ is a parameter of interest (e.g., log odds ratio or difference in mean blood pressure), need a probability distribution of $\theta|$ data
- $\mathrm{Pr}[\theta|\text{data}] \propto \mathrm{Pr}[\text{data}|\theta]\,\mathrm{Pr}[\theta]$
- $\mathrm{Pr}[\theta]$ is the *prior* distribution for θ
- Assuming θ is an unknown random *variable*

Advantages

- "intended for measuring support for hypotheses when the data are fixed (the true state of affairs after the data are observed)"[67]

- "inferences are based on probabilities associated with *different* values of parameters which could have given rise to the *fixed* set of data which has actually occurred"[15]

- Results in a probability most clinicians think they're getting[a]

- Can compute (posterior) probability of interesting events, e.g.

  $\Pr[$drug is beneficial$]$

  $\Pr[$drug A clinically similar to drug B$]$

  $\Pr[$drug A is $> 5\%$ better than drug B$]$[19]

[a] Nineteen of 24 cardiologists rated the posterior probability as the quantity they would most like to know, from among three choices. [24]

$\Pr[$mortality reduction $\geq 0\cap$ cost reduction $> 0]$

$\Pr[$mortality reduction $\geq 0 \cup ($mortality reduction $> 0.02\cap$ cost reduction $> -\$5000)]$

$\Pr[$mortality reduction $\geq 0 \cup ($ cost reduction $> 0\cap$ morbidity reduction $\geq 0)]$

$\Pr[$ICER $\leq \$30,000/$ life year saved$]$

- Provides formal mechanism for using prior information/bias — $\Pr[\theta]$
- Places emphasis on estimation and graphical presentation rather than hypothesis testing
- Avoids 1–tailed/2–tailed issue
- Posterior (Berry prefers “current”) probabilities can be interpreted out of context better than P–values
- If $\Pr[$drug B is better than drug A$] = 0.92$, this is true whether drug C was compared to drug D or not
- Avoids many of complexities of sequential monitoring —

P–value adjustment is needed for frequentist methods because repeatedly computed test statistics no longer have a χ^2 or normal distribution;

A posterior probability is still a probability $\longrightarrow$ Can monitor continuously

- Allows accumulating information (from this as well as other trials) to be used as trial proceeds
- No need for sufficient statistics

Controversies

- Posterior probabilities may be hard to compute (often have to use numerical methods)
- How does one choose a prior distribution $\Pr[\theta]$?[49]
  - Biased prior – expert opinion difficult, can be manipulated, medical experts often wrong, whose opinion do you use?[34]
  - Skeptical prior (often useful in sequential monitoring)
  - Unbiased (flat, non–informative) prior
  - Truncated prior — allows one to pre–specify e.g. there is no chance the odds ratio could be outside $[\frac{1}{10}, 10]$
- For monitoring, Spiegelhalter et al.[74] suggest using "community of priors" (see [22] for pros and cons):
  - Skeptical prior with mean 0 against which judge early stopping for efficacy

  - Enthusiastic prior with mean δ_A (hypothesized effect) against which judge early stopping for no difference

- Rank–based analyses need to use models:
  Wilcoxon $\longrightarrow$proportional odds ordinal logistic model
  logrank $\longrightarrow$Cox PH model

Invalid Bayesian Analyses

- Choosing an improper model for the data (can be remedied by adding e.g. non–normality parameter with its own prior[15])

- Sampling to a foregone conclusion if a continuous prior is used but the investigators and the consumers were convinced that prob. of treatment effect is *exactly* zero > 0[a]

[a] This is easily solved by using a prior with a lump of probability at zero.

- Suppression of the latest data by an unscrupulous investigator:

  Current results using 200 patients nearly conclusive in favor of drug

  Decide to accrue 50 more patients to draw firm conclusion

  Results of 50 less favorable to drug

  Based final analysis on 200 patients[a]

[a] Note the martingale property of posterior probs.: $E[\Pr(\theta_1 > \theta_2 | \text{ data, data}')] = \Pr(\theta_1 > \theta_2 | \text{ data})$.

The Standardized Likelihood Function

- Unknown parameter θ, data vector y
- Let likelihood function be $l(\theta|y)$
- Standardized likelihood:

$$p(\theta|y) = \frac{l(\theta|y)}{\int l(\theta|y)d\theta} \tag{1}$$

- Don't need to choose a prior if willing to take the normalized likelihood as a basis for calculating probabilities of interest (Fisher's fiducial distributions)

One–Sample Binomial

- $Y_1, Y_2, \ldots, Y_n \sim$ Bernoulli(θ)
- $s =$ number of “successes”
- $l(\theta|y) = \theta^s(1-\theta)^{n-s}$
- $\int l(\theta|y)d\theta = \beta(s+1, n-s+1)$
- $p(\theta|y) = \frac{\theta^s(1-\theta)^{n-s}}{\beta(s+1,n-s+1)}$
- Solving for θ so that tail areas of $p(\theta|y) = \frac{\alpha}{2}$ gives exact $1-\alpha$ C.L. for 1–sample binomial

Basis

- $p(\theta|y) \propto l(\theta|y)p(\theta)$
- $l(\theta|y) =$ likelihood function
- Function through which data y modifies the prior knowledge of θ[15]
- Has the information about θ that comes from the data

Choosing the Prior Distribution

- Stylized or “automatic” priors[34, 49]
- Data quickly overwhelm all but the most skeptical priors, especially in clinical applications
- In scientific inference, let data speak for themselves
- →*A priori* relative ignorance, draw inference appropriate for an unpredudiced observer[15]
- Scientific studies usually not undertaken if precise estimates already known. Also, problems with informed consent.
- Even when researcher has strong prior beliefs, more convincing to analyze data using a *reference* prior dominated by likelihood[15]
- Box and Tiao[15] advocate *locally uniform priors* — considers local behavior of prior in region where the likelihood is appreciable, prior assumed not large outside that range

$\rightarrow$posterior $\approx$ standardized likelihood

- Choice of metric ϕ for uniformity of prior: Such that likelihood for $\phi(\theta)$ completely determined except for location ($\approx$ variance stabilizing transformation) — likelihood is *data translated*

> “Then to say we know little *a priori* relative to what the data is going to tell us, may be expressed by saying that we are almost equally willing to accept one value of $\phi(\theta)$ as another.”[15]

$\rightarrow$Highest likelihood intervals symmetric in $\phi(\theta)$

- Example: Gaussian dist.$\rightarrow\phi(\sigma) = \log(\sigma)$, or if use σ, prior $\propto \sigma^{-1}$

Consumer Specification of Prior

- Place statistics describing study results on web page

  Posterior computed and displayed using Java applet (Lehmann & Nguyen [53])

- Highly flexible approximate approach: store 1000 bootstrap $\hat{\theta}$, can quickly take a weighted sample from these to apply a non–uniform prior [57]

One–Sample Binomial, Continued

- $\hat{\theta} = \bar{y}$ (proportion)
- $\sin^{-1}\sqrt{\hat{\theta}}$ →nearly data–translated likelihood and locally uniform prior is nearly noninformative[15]
- Nearly noninformative prior on original scale $\propto [\theta(1-\theta)]^{-\frac{1}{2}}$
- Posterior using this prior is $p(\theta|y) = \frac{\theta^{s-\frac{1}{2}}(1-\theta)^{n-s-\frac{1}{2}}}{\beta(s+\frac{1}{2}, n-s+\frac{1}{2})}$

- Posterior using locally uniform priors on data–translated scale:
  $p(\theta_1, \theta_2|y) = \frac{\theta_1^{s_1-\frac{1}{2}}(1-\theta_1)^{n_1-s_1-\frac{1}{2}}\theta_2^{s_2-\frac{1}{2}}(1-\theta_2)^{n_2-s_2-\frac{1}{2}}}{\beta(s_1+\frac{1}{2}, n_1-s_1+\frac{1}{2})\beta(s_2+\frac{1}{2}, n_2-s_2+\frac{1}{2})}$
- Can integrate to get posterior distribution of any quantity of interest, e.g., $\frac{\theta_1}{1-\theta_1} \frac{1-\theta_2}{\theta_2}$
- See Hashemi et al.[44] for much more information about posterior distributions of ORs and other effect measures
- See Howard[45] for a discussion of the need to use priors that require θ_1 and θ_2 to be dependent.

Two–Sample Gaussian

- $Y_1 \sim N(\mu_1, \sigma^2)$ ind. of $Y_2 \sim N(\mu_2, \sigma^2)$
- $\mu_1, \mu_2, \log \sigma \sim$ constant independently[a]
- $\nu = n_1 + n_2 - 2$
- $\nu s^2 = \sum(y_{1i} - \bar{y}_1)^2 + \sum(y_{2i} - \bar{y}_2)^2$
- $\delta = \mu_2 - \mu_1, \hat{\delta} = \bar{y}_2 - \bar{y}_1$
- $p(\delta, \sigma^2 | y) = p(\sigma^2 | s^2) p(\delta | \sigma^2, \hat{\delta})$
- $\nu s^2 / \sigma^2 \sim \chi^2_\nu$
  $p(\delta | \sigma^2, \hat{\delta}) = N(\hat{\delta}, \sigma^2(1/n_1 + 1/n_2))$
- Integrate out σ^2 to get marginal posterior dist. of
  $\delta \sim t_\nu[\hat{\delta}, s^2(1/n_1 + 1/n_2)]$

[a] The prior for $\sigma \propto \sigma^{-1}$.

One–Sample Gaussian

- $Y \sim N(\mu, \sigma^2), \sigma$ known
- $\mu \sim N(\mu_0, \sigma_0^2)$
- $\mu|y \sim N(\mu', \sigma'^2)$
- $\mu' = \frac{w_0\mu_0 + wy}{w_0 + w}$
- $\sigma'^2 = \frac{1}{w_0 + w}$
- $w_0 = \sigma_0^{-2}, w = \sigma^{-2}$
- $\sigma_0 \to \infty : \mu \sim N(y, \sigma^2)$

Deriving Posterior Distribution

- Analytic integration sometimes possible
- Numerical integration/simulation methods, e.g., Gibbs sampler[20]
- Gibbs Markov Chain Monte Carlo method can handle huge number of parameters
- Simulated parameter values have correct marginal and joint distributions
- Uses a “burn–in” of say 1000 iterations which are discarded
- In some strange problems the realizations may not converge properly

Using Posterior Distribution

- Quantities such as $\Pr[OR < .9]$
- Credible (highest posterior density) intervals
- Posterior odds
- If posterior represented analytically (and especially if the CDF is), can compute any probability of interest quickly
- Simulation of realizations from the posterior makes for easy programming
- Example: Generate 5,000 ORs, compute fraction of ORs $< .9$, mean OR, median OR, credible interval (using sample quantiles)
- Kernel density estimate based on 5,000 realizations for graphical depiction
- Compute mode

Sequential Testing

- Frequentist approach to deciding when to quit watching a football game:
  Of all games which ended in a tie or with your team losing, what proportion had your team leading by 10 points with 12m to go in 4^{th} quarter?
  Must consider sample space
- Bayesian approach: at each moment can estimate the probability that your team will ultimately win based on the time left and the point spread
- No problem with estimating this probability every second
- Distribution of unknown parameters updated at any time[10]
- Evidence from experiment to date taken at face value[10]
- No need for independent increments

- No need for equal information time
- No scheduling
- No adjustment of point estimates, C.L. for monitoring strategy
- Determining number of “looks” (k) that minimizes expected sample size — frequentist: plot of avg. sample size vs. k is U–shaped[a]; Bayesian: the larger k the better[10]
- Example (Freireich et al. 1963): Patients treated in pairs to see which patient had better time to remission of leukemia[10]
- $\theta = \mathrm{Pr}[A \text{ better than } B], H_0 : \theta = \frac{1}{2}$
- Continuous monitoring $\rightarrow \alpha = 0.05$ corresponds with $P = 0.0075$
- Uniform prior for θ

[a] Because of α–adjustment

Patient	Preferred			Current
Pair	Treatment	$n_A - n_B$	$2P$	$\Pr[B > A]$
1	A	1	1.0	0.25
2	B	0	1.0	0.50
3	A	1	1.0	0.31
4	A	2	0.63	0.19
5	A	3	0.38	0.11
6	B	2	0.69	0.23
7	A	3	0.45	0.14
8	A	4	0.29	0.090
9	A	5	0.18	0.055
10	A	6	0.11	0.033
11	A	7	0.065	0.019
12	A	8	0.039	0.011
13	A	9	0.022	0.0065
14	B	8	0.057	0.018
15	A	9	0.035	0.011
16	A	10	0.021	0.0064
17	A	11	0.013	0.0038
18	A	12	0.0075	0.0022
19	A	13	0.0044	0.0013
21	A	14	0.0026	0.0008
21	A	15	0.0015	0.0005

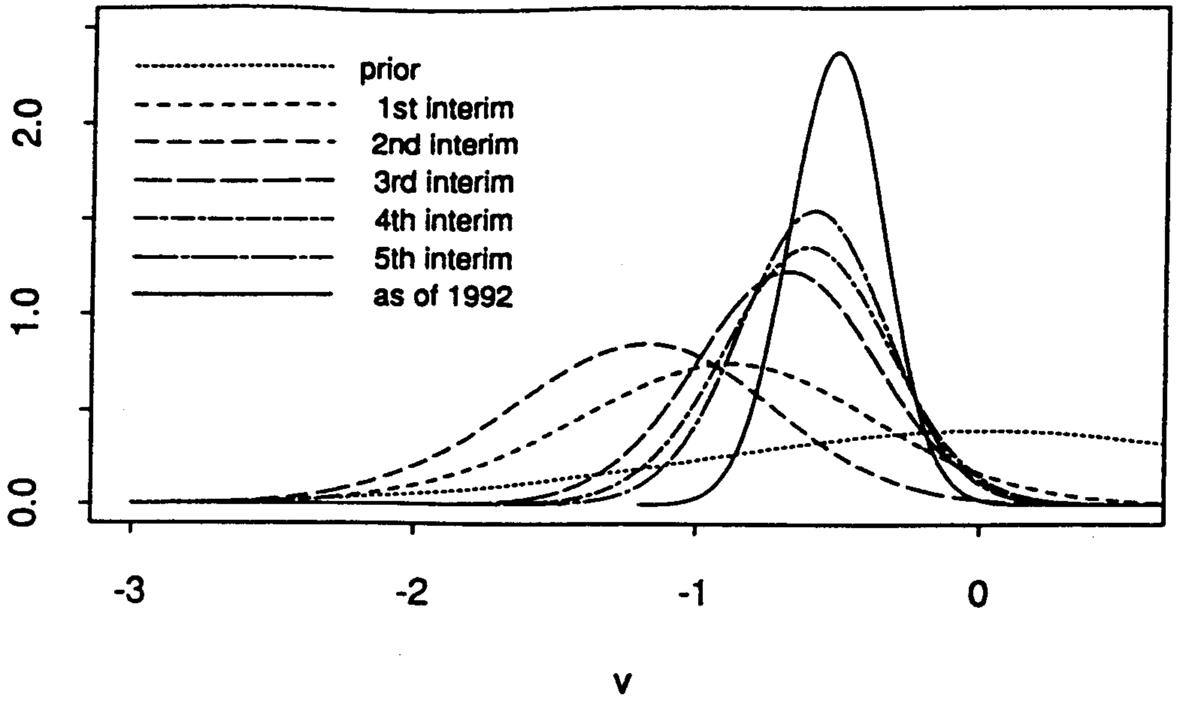


Figure 1: *Sequentially monitoring a clinical trial*[39]. v *is the log hazard ratio.*

Simulation Experiment

- Two–sample binomial
- Four replications of experiments with $\theta_1 = \theta_2 = 0.2$
- Four replications with $\theta_1 = 0.2, \theta_2 = 0.3$
- Non–informative prior on probs. using variance–stabilized scale
- Compute various posterior probs. by drawing 10,000 odds ratios from the posterior distribution
- Monitor results at $n = 20, 40, \ldots, 400,$ $500, 1000, 1500, \ldots, 8000, 16000$
- **S-PLUS Code** (File `sim.s`)

```
store()  ## in Hmisc library in Statlib ->
         ## divers objects to temporary storage

for(type in c(’null’,’non-null’)) {
  if(type==’null’) {
    p1 <- .2
    p2 <- .2
  ps.slide(’nullsim’,type=3,hor=F)
  ## ps.slide in Hmisc in Statlib
  ## (pretty defaults for postscript)
  set.seed(171)
} else {
  p1 <- .2
  p2 <- .3
  ps.slide(’nnullsim’,type=3,hor=F)
  set.seed(2193)
}

n.experiments <- 4
n.total       <- 16000
k             <- n.total/2

n.beta <- 10000

par(mfrow=c(2,2))
for(kx in 1:n.experiments) {

  ## Generate Bernoulli observations
  y1 <- sample(0:1,k,T,prob=c(1-p1,p1))
  y2 <- sample(0:1,k,T,prob=c(1-p2,p2))

  ## At any possible time of analysis,
  ## compute total # events
  s1 <- cumsum(y1)
  s2 <- cumsum(y2)
```

```
n1 <- n2 <- 1:k

ii <- c(seq(10,200,by=10),seq(250,4000,by=250),k
phi <- plow <- peq <- peff <- single(length(ii)

j <- 0

for(i in ii) {
  cat(i,’’)
  j <- j+1
  ss1 <- s1[i]
  ss2 <- s2[i]
  nn1 <- n1[i]
  nn2 <- n2[i]

  ## Get 10000 draws from posterior distribution
  ## event for each of the two groups, using pri
  ## noninformative on the variance-stabilized s
  ## (arcsin sqrt(p)).

  p1.u <- rbeta(n.beta,ss1+.5,nn1-ss1+.5)
  p2.u <- rbeta(n.beta,ss2+.5,nn2-ss2+.5)

  or <- p2.u/(1-p2.u)/ (p1.u/(1-p1.u))
  peff[j] <- mean(or < 1)
  plow[j] <- mean(or < .85)
  phi[j]  <- mean(or > 1/.85)
  peq[j]  <- mean(or >= .85 & or <= 1/.85)
```

```
    }

    x <- log(2*ii,2)
    labcurve(list(’OR < 1’         =list(x,peff),
                  ’OR < .85’       =list(x,plow),
                  ’OR > 1/.85’     =list(x,phi),
                  ’OR [.85,1/.85]’ =list(x,peq)),
             xlab=’log2(N)’, ylab=’Posterior Probabi
             ylim=c(0,1), keys=1:4, pl=T)
  }
  dev.off()
}
```

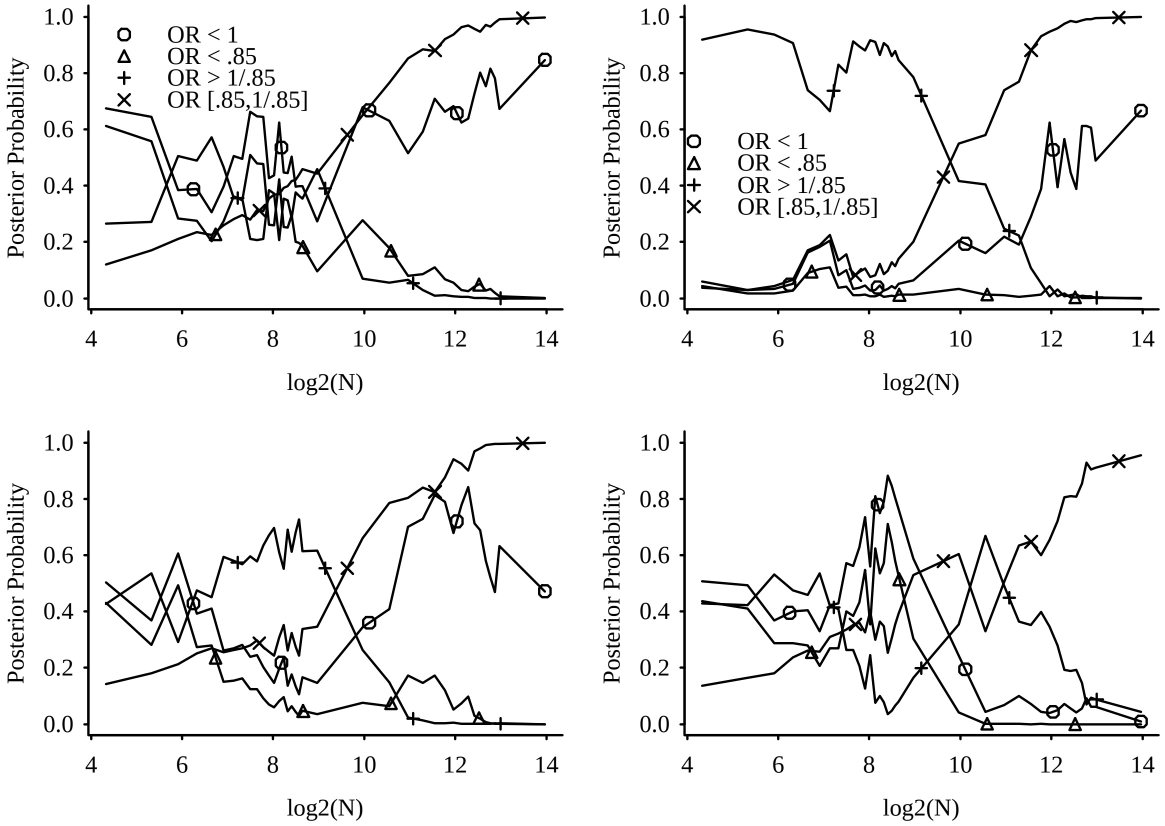


Figure 2: *Simulation of 4 sequentially monitored experiments each with* $n = 16000$*, for the null case where* $\theta_1 = \theta_2 = 0.2$*.*

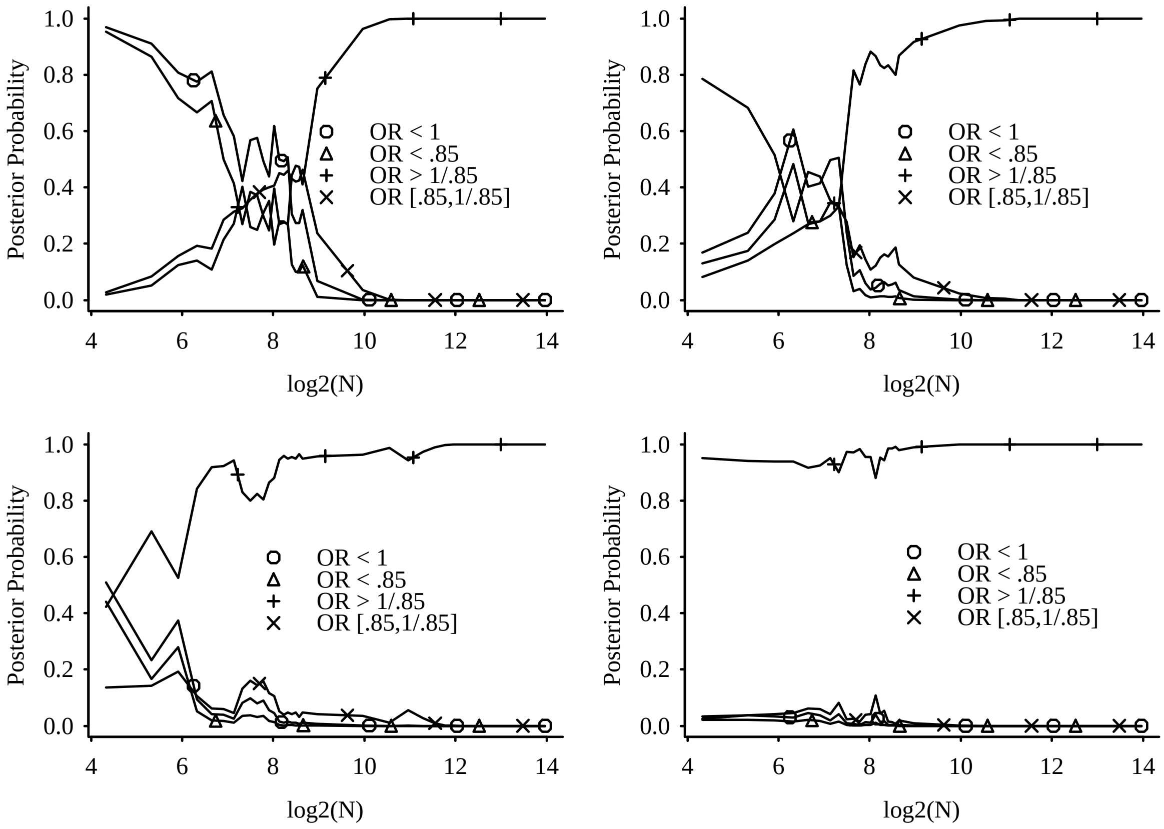


Figure 3: *Simulation of 4 sequentially monitored experiments, each with* $n = 16000$*, for the case where* $\theta_1 = 0.2, \theta_2 = 0.3$*.*

- Even with "significant" treatment effect in subgroup, point estimates of effects will be greatly exaggerated
- →Need to get away from hypothesis testing within subgroups
- Shrinkage methods needed
- Example: Represent differential treatment effects as random effects, shrinking them down to achieve optimal prediction[25, 26, 66, 3]
- If prior distribution for each parameter of interest is well- calibrated, posterior probabilities need no adjustment for the number of subgroups tested [80]

Inference for Multiple Endpoints

- Success criteria using the clinical and not the randomness scale
- Example: 3 endpoints
- Target Z_1: Population mean blood pressure $\downarrow \geq 5$ mmHg
- Target Z_2: Population exercise time $\uparrow \geq 1$ min.
- Target Z_3: Population mean angina score $\downarrow \geq 1$ point
- Posterior $\Pr[Z_1] = 0.97$
- Posterior $\Pr[Z_2] = 0.94$
- Posterior $\Pr[Z_3] = 0.6$
- $\Pr[Z_1 \cup Z_2 \cup Z_3] \geq 0.97$
- $\Pr[\bar{Z}_1 \cap \bar{Z}_2 \cap \bar{Z}_3] \leq 0.03$
- $\Pr[\# Z_i \geq 2] = 0.95$ for example

- To demonstrate that a drug improves at least one endpoint, study many endpoints!
- May want to show that at least $\frac{1}{2}$ of the endpoints are improved with high probability
- Alternative: Panel of experts rate importance of outcomes, e.g., $Z_1 = 1, Z_2 = 2, Z_3 = 3$
- Target could be ≥ 3 points
- Here $\Pr[Z_3 \cup (Z_1 \cap Z_2)] \geq 0.95$
- Simply count number of samples from posterior satisfying $Z_3 \cup (Z_1 \cap Z_2)$
- Another way to summarize results: Estimate $E[\#Z_i] = 0.97 + 0.94 + 0.6 = 2.51$ out of 3
- If all endpoints are binary, a kind of random effects model for the endpoints may be useful[52]
- If prior distribution for each parameter of interest is well- calibrated, posterior probabilities need no adjustment for the number of responses tested [80]

- See Thall and Sung[76] for formal Bayesian approaches to multiple endpoints in clinical trials
- See Berry [11] for a Bayesian perspective on data–generated hypothesis testing

The Bootstrap

- Distribution–free C.L.: Take e.g. 1000 samples with replacement from original sample $\rightarrow \hat{\theta}^1, \ldots, \hat{\theta}^{1000}$
- Sort, $[\hat{\theta}^{25}, \hat{\theta}^{975}]$
- Bootstrap can be used to form a posterior distribution when a somewhat odd reference prior putting mass only on observed values is used
  64, 57, 28, 70, 2
- Can use kernel density estimator based on $\hat{\theta}^1, \ldots, \hat{\theta}^{1000}$

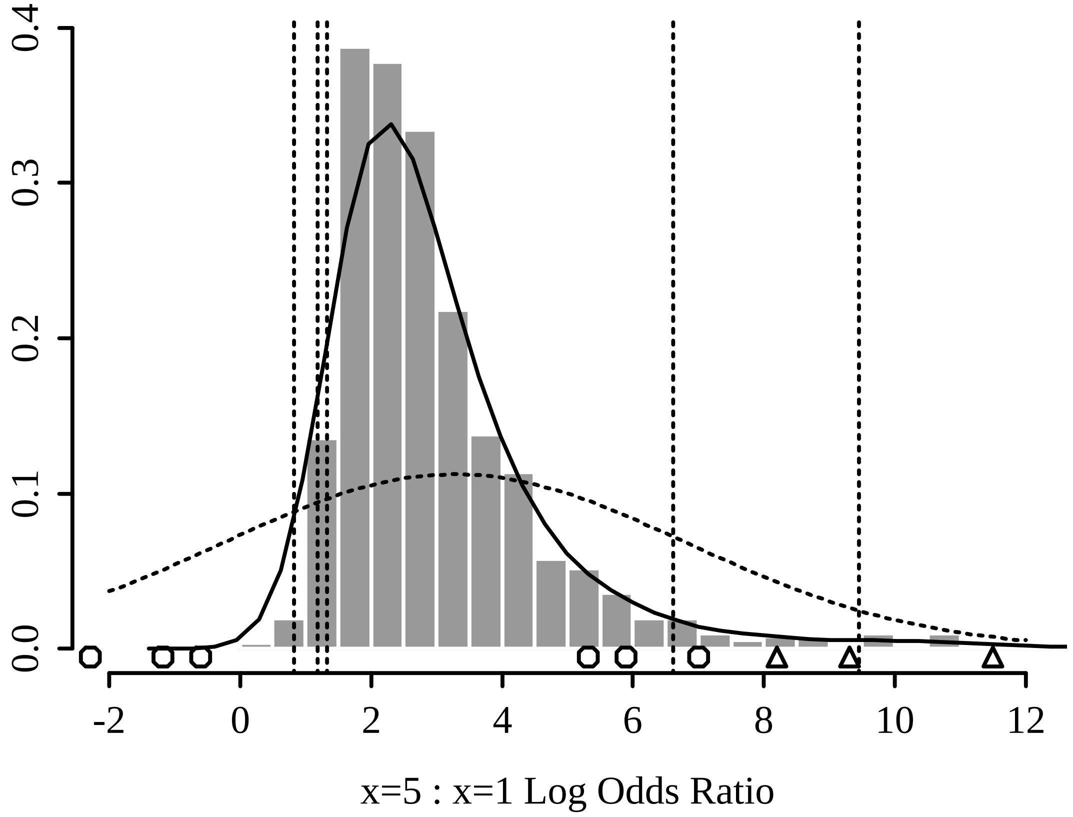


Figure 4: *Bootstrap distribution of the $x = 5 : x = 1$ log odds ratio from a quadratic logistic model with highly skewed x. The solid curve is a kernel density estimate, and the dashed curve is a normal density with the same mean and standard deviation as the bootstrapped values. Vertical lines indicate asymmetric 0.9, 0.95, and 0.99 two–sided confidence limits for the log odds ratio based on quantiles of the bootstrap values. The upper 0.99 confidence limit of 18.99 is not shown with a line. Triangles indicates corresponding symmetric confidence limits obtained assuming normality of estimates but using the bootstrap estimate of the standard error. The left limits for these are off the scale because of the high variance of the bootstrap estimates. Circles indicate confidence limits based on the usual normal theory–information matrix method.*

Two–Sample Binomial Example

- Advantageous to specify prior for OR instead of for the two probabilities of response θ_1, θ_2[74]
  Consider this later

- For now consider priors for θ_1, θ_2:
  - Flat
  - $\propto [\theta(1-\theta)]^{-\frac{1}{2}}$

- Data: Treatment A $\frac{30}{200}$
  Treatment B $\frac{18}{200}$

- OR = 0.56; $2P = 0.064$ (LR), 0.068 (Wald); $1P = 0.034$ (Wald)
  0.95 C.L. $[.304, 1.042]$ (Wald based on normality of log OR)

- **S-PLUS Code** (File betaboot.s)

```
store() ## store is in Hmisc library in statlib
        ## it causes objects to go into a temporary
library(Design,T)  ## Design library is in statlib

## Set number of events and number of trials for 2 g
s1 <- 30;  n1 <- 200
s2 <- 18;  n2 <- 200

or <- s2/(n2-s2) / (s1/(n1-s1))
or

## Get 50000 draws from posterior distribution of p
## event for each of the two groups, using flat pri
## (standardized likelihood)

nsim <- 50000
set.seed(179)  # not used in notes
p1   <- rbeta(nsim,s1+1,n1-s1+1) ## Generates 50000
p2   <- rbeta(nsim,s2+1,n2-s2+1)

## Use instead a prior that is noninformative on th
## variance-stabilized scale (arcsin sqrt(p)).
## The prior is  1/sqrt[p*(1-p)]

p1.u <- rbeta(nsim,s1+.5,n1-s1+.5)
p2.u <- rbeta(nsim,s2+.5,n2-s2+.5)
```

```
or.sim <- p2/(1-p2) / (p1/(1-p1)) ## 50000 simulated
or.sim.u <- p2.u/(1-p2.u)/ (p1.u/(1-p1.u))

## ---------------------------------------------------
## The following block of code is under development
## are (1) does this work with almost improper prio
## should probably be a correlation between the pri
## the log odds ratio (Note: cov(B-A) = -var(A), A=
## Use re-weighting to change prior distribution so
## normal distribution with mean 0 and variance 500
## logit(p1) is assumed to be normal with mean 0 an
## 1/[p(1-p)] terms are from the Jacobian

logit <- function(p) log(p/(1-p))

w <- dnorm(logit(p1.u),0,sqrt(10000))/(p1.u*(1-p1.u
     dnorm(logit(p2.u),0,sqrt(10000+500))/(p2.u*(1-p
## Make weight vector sum to 1
w <- w / sum(w)

## Estimate posterior Prob[or < 1] and Prob[or < .9
sum(w[or.sim.u < 1])
sum(w[or.sim.u < .9])

## Repeat this for a skeptical prior on the log or:
w2 <- dnorm(logit(p1.u),0,sqrt(10000))/(p1.u*(1-p1.u
     dnorm(logit(p2.u),0,sqrt(10000+1))/(p2.u*(1-p2
## Make weight vector sum to 1
w2 <- w2 / sum(w2)
sum(w2[or.sim.u < 1])
sum(w2[or.sim.u < .9])
## ------------------------------------------------



## Bootstrap distribution of OR
## First string out count data into vectors of bina

y1 <- c(rep(1,s1),rep(0,n1-s1))
y2 <- c(rep(1,s2),rep(0,n2-s2))

## Get L.R. chisq test and Wald C.L. from logistic r
y <- c(y1,y2)
group <- c(rep(1,n1),rep(2,n2))
f <- lrm(y ~ group)
## lrm, datadist, summary in Design library (in sta
## Gives chisq=3.44 2P=.064
dd <- datadist(group)
## stores distribution characteristics of group
options(datadist=’dd’)
summary(f, group=1:2) ## 0.95 C.L. for OR [.304,1.04

B <- 10000                   ## 10000 bootstrap samp
or.boot <- single(B)
for(j in 1:B) {
  if(j %% 100 ==0) cat(j,’’)
  i <- sample(n1,replace=T)  ## sample with replace
  y1.b <- y1[i]
  i <- sample(n2,replace=T)
  y2.b <- y2[i]
  odds.1 <- sum(y1.b)/(n1-sum(y1.b))
  odds.2 <- sum(y2.b)/(n2-sum(y2.b))
  or.boot[j] <- odds.2 / odds.1
}
store(or.boot)               ## store or.boot perma

ps.slide(’ordens’,type=3,hor=F,mar=c(4,3,2,1)+.1)
## in Hmisc library
labcurve(list(
  "Beta, Flat Prior"        =density(or.sim),
  "Beta, Prior=[p(1-p)]ˆ-.5"=c(density(or.sim.u),lty
  Bootstrap                 =c(density(or.boot),lwd=
  pl=T, xlab=’Odds Ratio’, ylab=’Density’, keys=’li
minor.tick(5,5)
## labcurve and minor.tick are in Hmisc library

usr <- par("usr")    ## x- and y-axis limits for pl
bot.arrow <- usr[3]  ## usr[3:4] = limits of y-axis
top.arrow <- bot.arrow + 0.05 * (usr[4] - usr[3])
quan <- quantile(or.sim,c(.025,.05,.95,.975))
for(i in 1:4)
  arrows(quan[i], top.arrow, quan[i], bot.arrow,
         rel = T, size = 0.5)
```

```
quan <- quantile(or.boot,c(.025,.975))
title(’Estimated Densities with 0.9 and 0.95\nProbab
title(sub=paste(
 ’Traditional 0.95 C.L. [.301,1.042], Bootstrap [’,
  round(quan[1],3),’,’,round(quan[2],3),’]’,sep=’’),
  adj=0,cex=.85)

text(1.03,.9,
     paste(’Prob[OR < 1  ]=’,round(mean(or.sim<1),3
           ’ (Beta)  ’, round(mean(or.boot<1),3),
           ’ (Bootstrap)\n’,
           ’Prob[OR < 0.9]=’,round(mean(or.sim<.9),3
           ’ (Beta)  ’, round(mean(or.boot<.9),3),
           ’ (Bootstrap)’,sep=’’),adj=0)
dev.off()
```

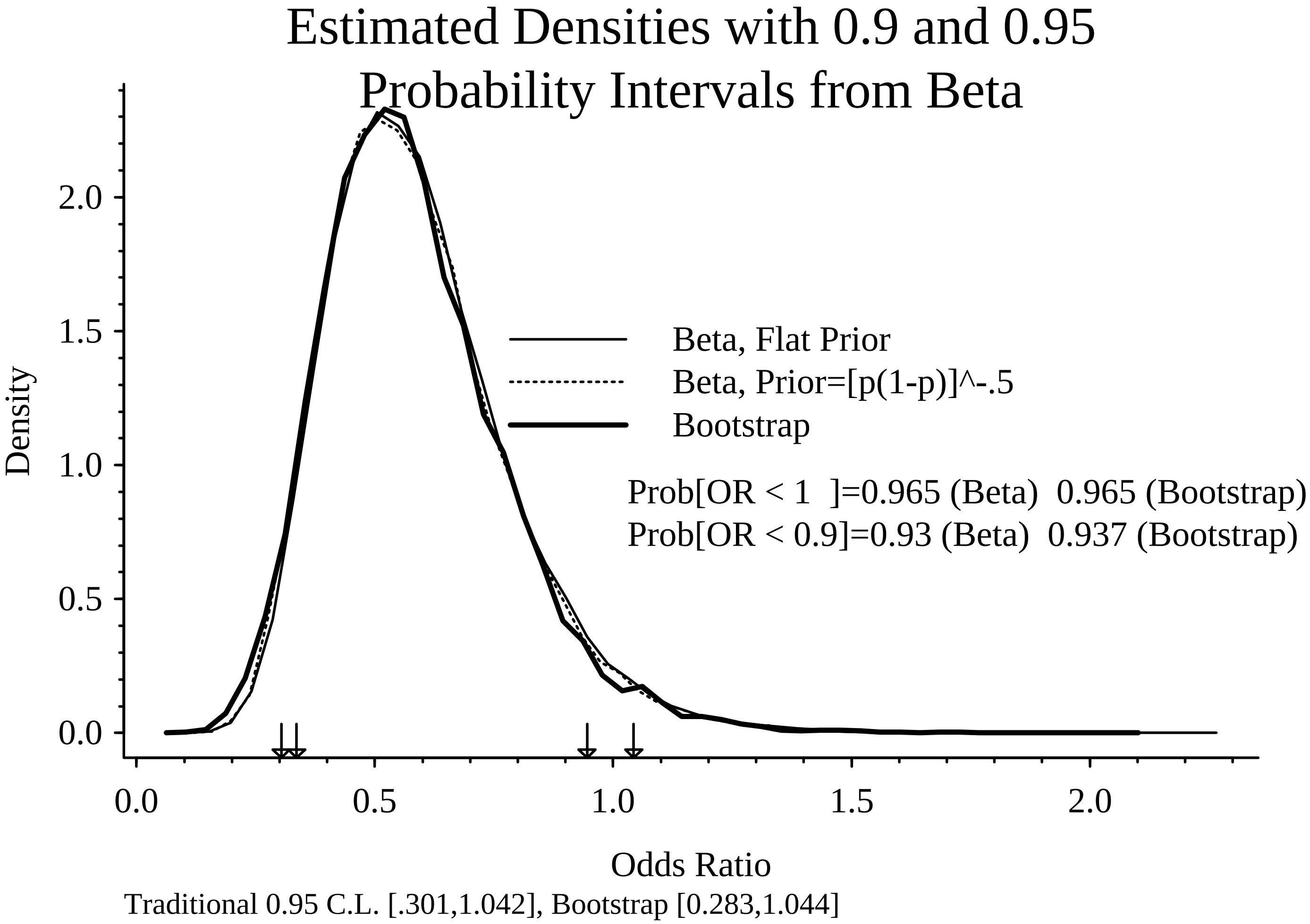


Figure 5: *Posterior density of OR from a kernel estimator. The posterior were derived using the bootstrap and using a Bayesian approach with 2 prior densities.*

Software

- BUGS (Bayesian Inference using Gibbs Sampling) package (public domain, Cambridge)[78, 40]
- Available for variety of computer systems
- `http://www.mrc-bsu.cam.ac.uk/bugs`
- `http://muskie.biostat.umn.edu/mirror/methodology/bugs/`.
- Works in conjunction with any version of S-PLUS using BUGS' `CODA` S-PLUS functions
- BUGS has a general modeling language
- WinBUGS allows for graphical specification of model, builtin interactive graphics for displaying results, report writing capability
- Two–volume Examples Guide is must reading!

GUSTO I

- Four thrombolytic strategies for acute MI, $n = 41,021$[77]
- SK=streptokinase, Combo=SK+t-PA
- Here consider only death ∪ disabling stroke

Treatment	N	Events	Fraction
t-PA	10393	712	0.068
Combo	10370	783	0.076
SK+IV	10409	811	0.078
SK+SQ	9837	752	0.076
SK	20246	1563	0.077

- t-PA:SK OR=0.879, $2P = 0.006$
- Bayesian analysis using 3 different priors
  - Flat (log OR Gaussian with variance 10^6)
  - log OR truncated Gaussian with $\Pr[OR > 4 \cup OR < \frac{1}{4}] = 0$

    - $\ast$ $\Pr[OR > 2 \cup OR < \frac{1}{2}] = 0.05$
    - $\ast$ $\Pr[OR > 1\frac{1}{3} \cup OR < \frac{3}{4}] = 0.05$

- Similarity region: $OR \in [0.9, \frac{1}{0.9}]$
- **BUGS Data File** (File `sk.tpa.dat`)

```
list(event=c(1563,712), treat=c(0,1), N=c(20246,1039
```

- **Initial Parameter Estimates** (File `bugs.in`)

```
list(int=0,b.treat=0)
```

- **Command File** (File `bugs.cmd`)

```
compile("bugs.bug")
update(1000)
monitor(or)
update(5000)
stats(or)
q()
```

- **Model Code** (File `bugs.bug`)

```
model logistic;

const
             M=2;
var
             event[M],
             treat[M],
             N[M],
             p[M],
             int,b.treat,or;

#data in "tpa.combo.dat";
#data in "sk.dat";
data in "sk.tpa.dat";
inits in "bugs.in";

{
             or <- exp(b.treat);

             for(i in 1:M) {
                           logit(p[i]) <- int+b.treat*treat[i];
                           event[i] ˜ dbin(p[i],N[i]);
             }

             #Prior distributions

             int ˜ dnorm(0.0, 1.0E-6);

#            b.treat ˜ dnorm(0.0, 7.989) I(-1.386,1.386);
#   trunc at or=4, .025 prob>2

             b.treat ˜ dnorm(0.0, 46.42723)I(-1.386,1.386);
#   .025 prob < .75

#            b.treat ˜ dnorm(0.0, 1.0E-6);   #flat prior
```

```
    }
```

- **S-Plus Code** (File `bugs.s`)

```
ind <- inddat()   ## inddat, readdat used here are from
out <- readdat()  ## an older version of BUGS.  These read BUGS outpu

which <- 5

ti <- c('Accelerated t-PA vs. Combination Therapy',
                 'SK+SQ Heparin vs. SK+IV Heparin',
                 'Combined SK vs. Accelerated t-PA')[min(which,3)]

##prior <- 'Noninformative prior'
##prior <- 'Skeptical prior'   ## (1/4,4) possible, .025 prob > 2, <
## sd=.3537729
prior   <- 'Very skeptical prior' ## (1/4,4) possible, .025 prob < .
## sd=.146762
fi <- c('or.tpa.combo','or.sk','or.sk.tpa','or.sk.tpa.skeptical',
        'or.sk.tpa.skeptical2')[which]

ps.slide('priors', type=3)   ## ps.slide in Hmisc library in Statlib
dtruncnorm <-
  function(x, mean = 0, sd = 1, lower = NA, upper = NA) {
             ##
             ## density of truncated normal  - taken from BART
             ##
             k.upper <- if(!is.na(upper)) pnorm((upper - mean)/sd) e
             k.lower <- if(!is.na(lower)) pnorm((lower - mean)/sd) e
             K <- 1/(k.upper - k.lower)
             y <- K * dnorm(x, mean, sd)
             y[x < lower] <- 0
             y[x > upper] <- 0
             y
  }

x <- seq(.1,3,length=200)
for(i in 1:2) {
  d <- dtruncnorm(log(x), mean=0, sd=c(.3537729,.146762)[i],
                  lower=-log(4), upper=log(4))
  if(i==1) plot(x, d, xlab='Odds Ratio', ylab='',
                ylim=c(0,3), type='l') else
  lines(x, d, lty=3)
}
abline(v=1, lty=2, lwd=1)
dev.off()
}

ps.slide(fi, type=3)

## The following uses drawdat2, a modified version of drawdat
## from a previous version of BUGS.
## drawdat2 has text() use cex=cex, remove cex= from plot(),
## comments out points(), par(), add xlab, posterior mode, remove ti
## get digits from options(), add xlim

cex <- .75  ## was 1.25 for large plot
options(digits=3)
drawdat2(v='or', trace=F, cex=cex, xlab='Odds Ratio', xlim=c(.5,1.5)

or <- out[,'or']
cl <- quantile(or, c(.025,.975))
options(digits=3)
fcl <- format(cl)
xpos <- c(1.01,1.09,.955,.963,.972)[which]
ypos <- c(5.4,5,6.2,6.4,6.4)[which]


text(xpos,ypos,
     paste('2.5% = ',fcl[1],
           '\n97.5% = ',fcl[2],
           '\n\nProb(OR < 1) = ',format(mean(or < 1)),
           '\nProb(OR < .95) = ',format(mean(or < .95)),
           '\nProb(OR < .90) = ',format(mean(or < .90)),
           '\nProb(.90 < OR < 1/.9) = ',format(mean(or > .9 & or < 1/
           sep=''),
     adj=0, cex=cex)
```

```
pstamp(paste(ti,prior,sep=’   ’))  ## pstamp is in Hmisc
```

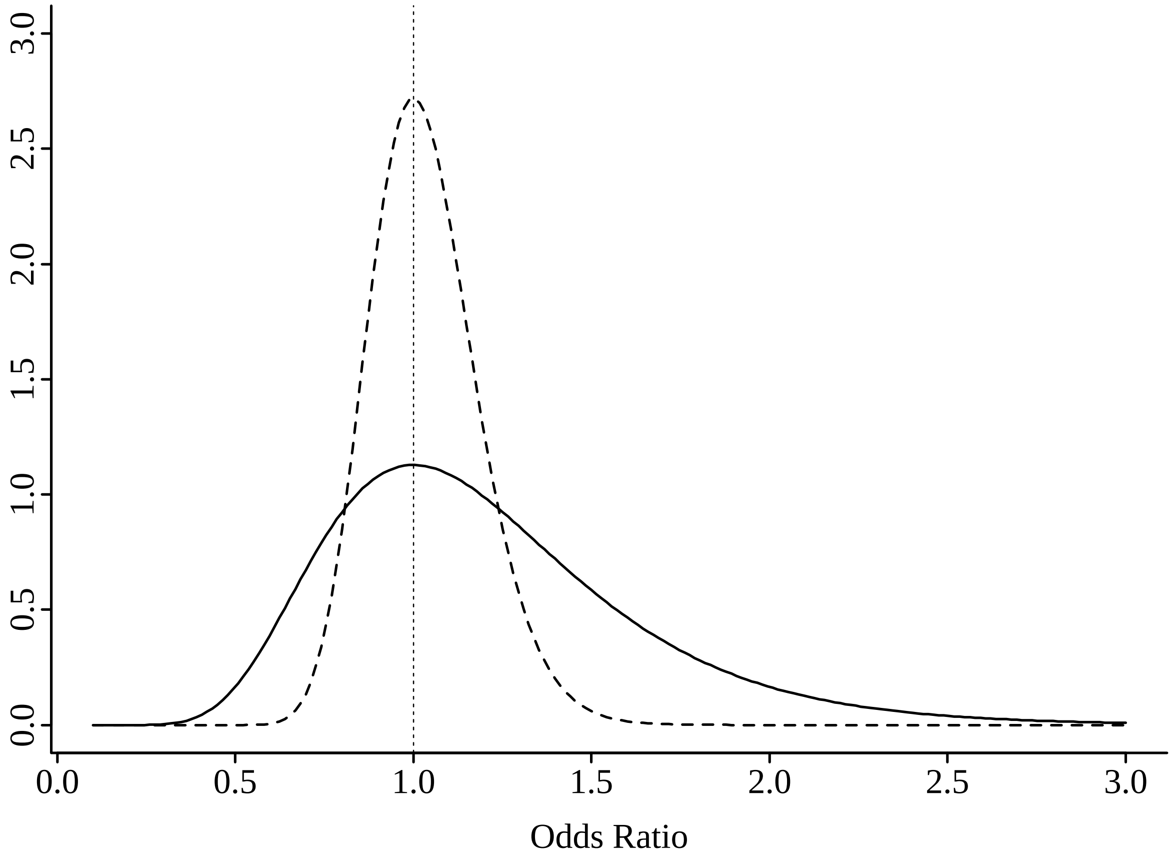
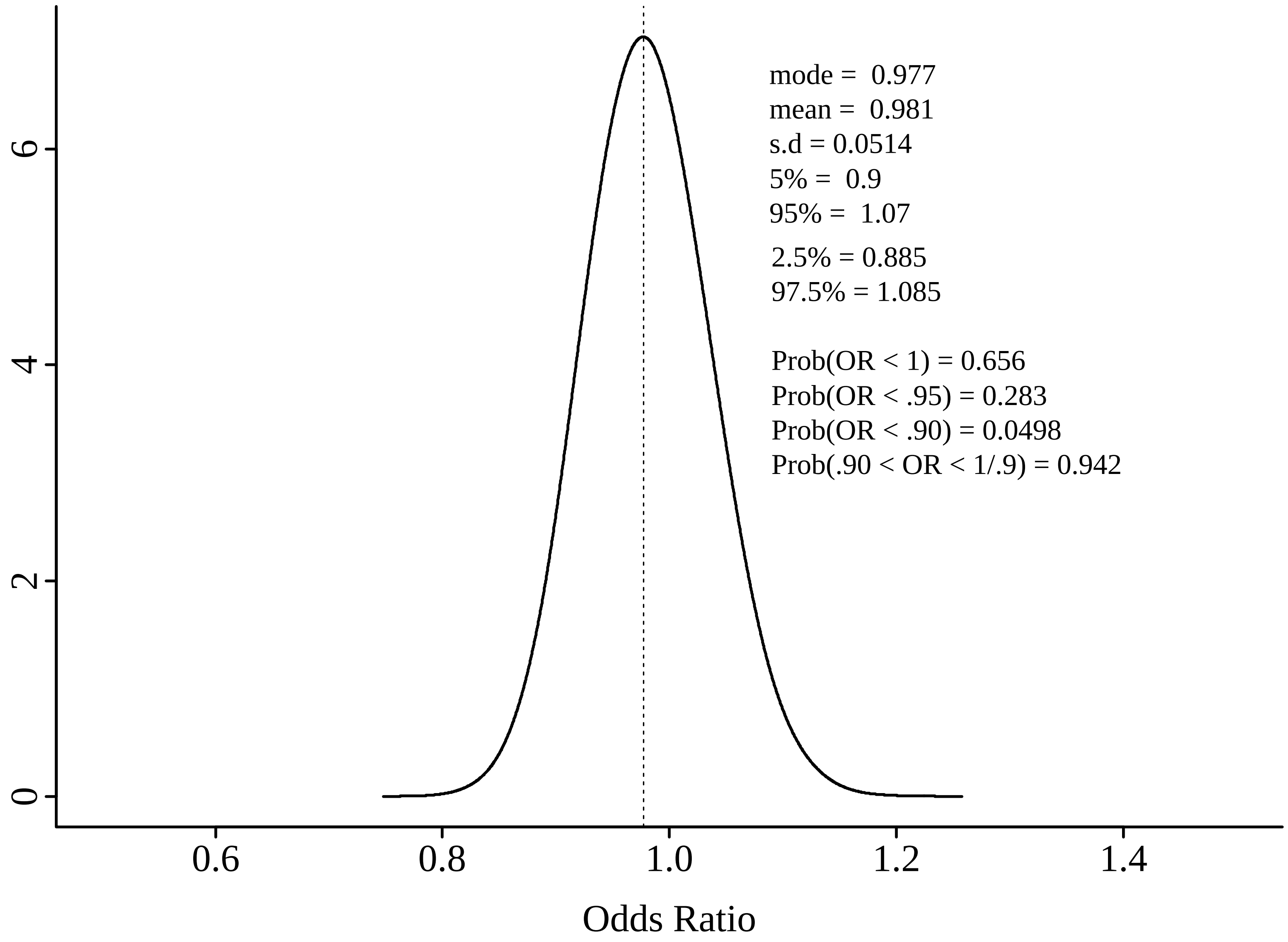
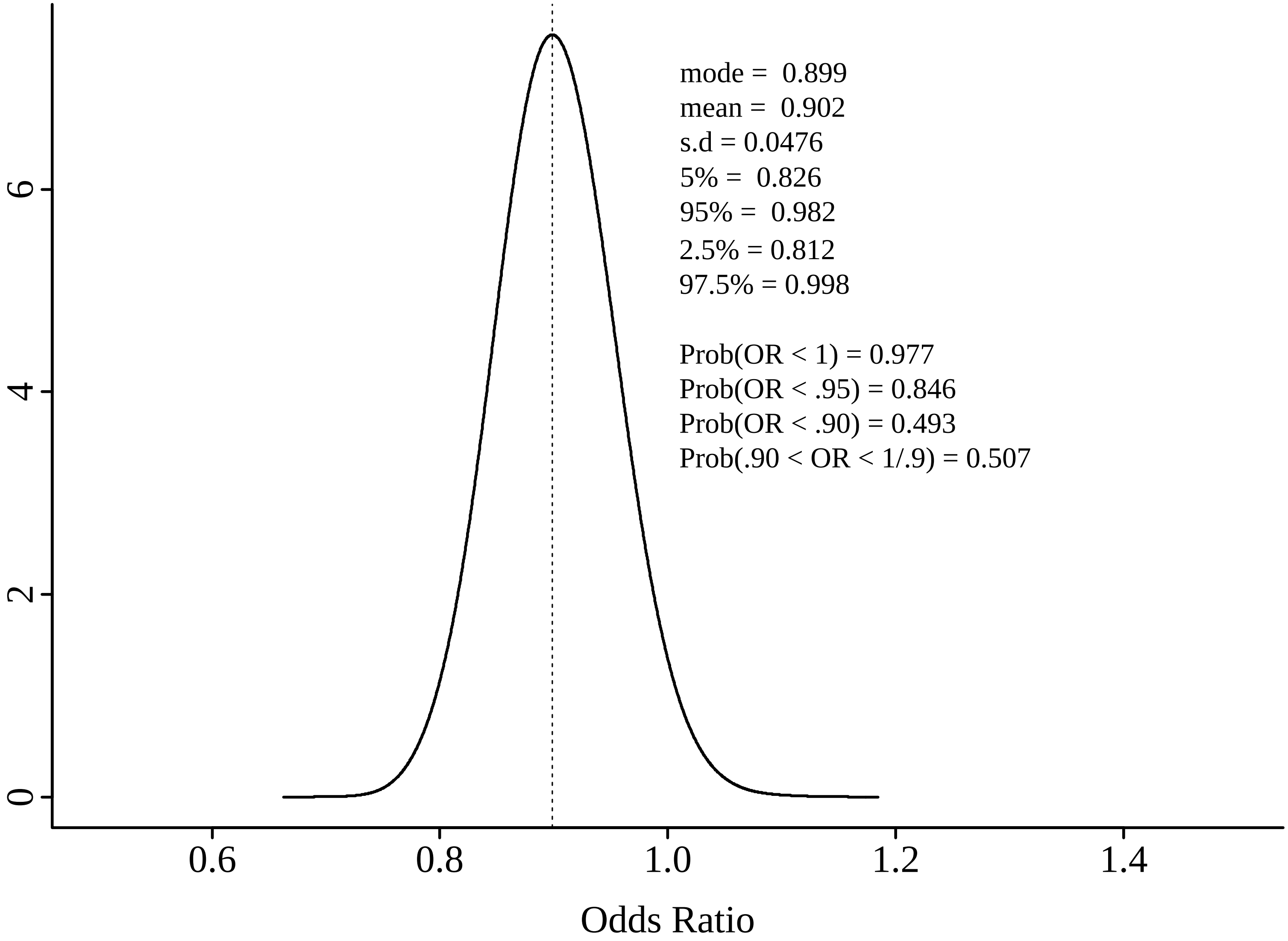


Figure 6: *Prior probability densities for $\mathrm{OR} = e^{\beta}$. Both distributions assume that OR = 1 (no effect) is the most likely value, and that ORs outside the interval $[\frac{1}{4}, 4]$ are impossible. The solid curve corresponds to a truncated normal distribution for $\log \mathrm{OR}$ having a standard deviation of 0.354. The dashed curve corresponds to a more skeptical prior distribution with a standard deviation of 0.147.*



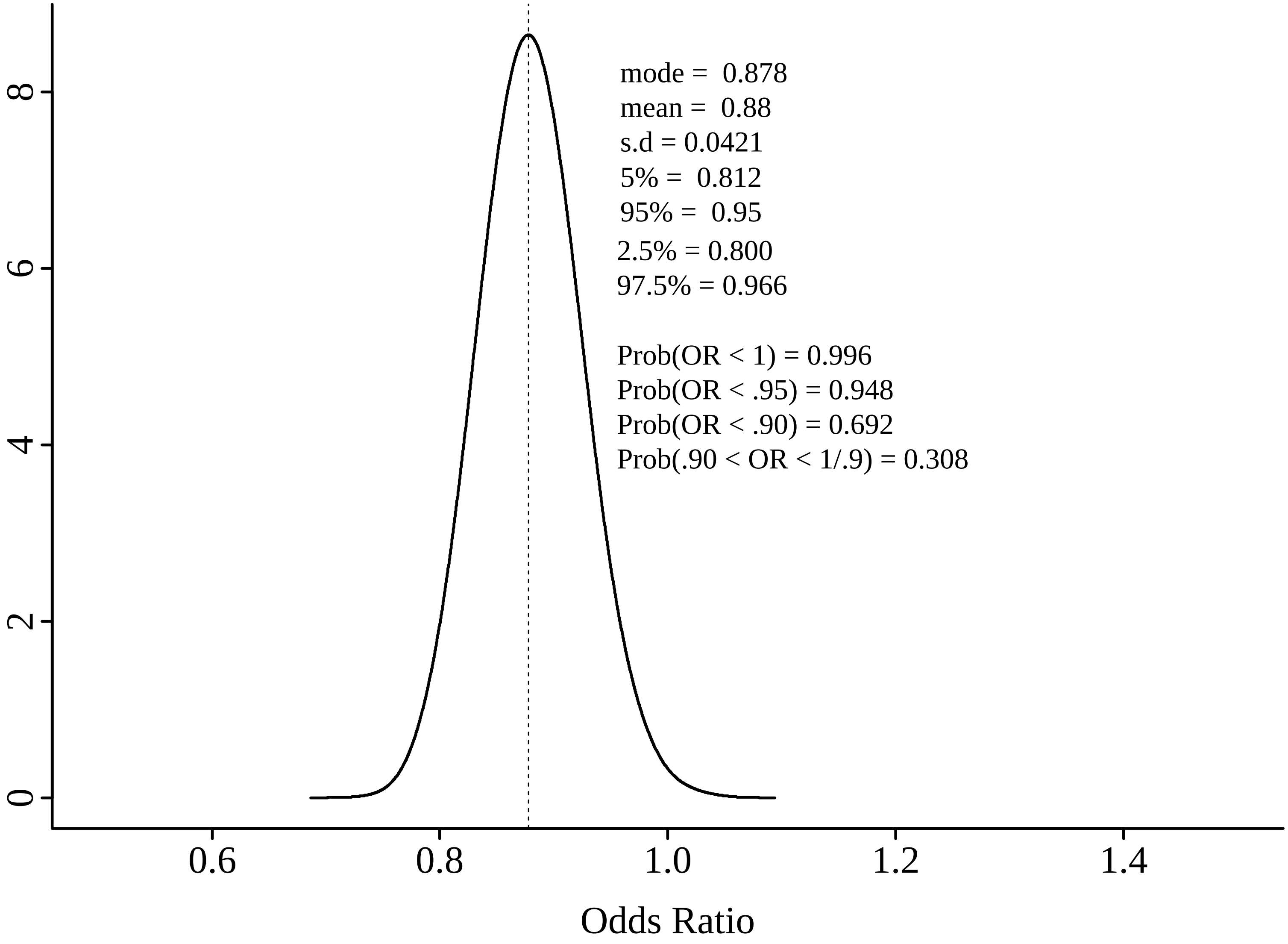
SK+SQ Heparin vs. SK+IV Heparin 11Jan96 10:52

Figure 7: *Posterior probability density for the ratio of the odds of a clinical endpoint for SK+SQ heparin divided by the odds for SK+IV heparin, using a flat prior distribution for log OR.*



Accelerated t-PA vs. Combination Therapy 11Jan96 11:04

Figure 8: *Posterior probability density for accelerated t-PA compared with non–accelerated t-PA with SK and heparin, using a flat prior distribution.*



Combined SK vs. Accelerated t-PA Noninformative prior 11Jan96 11:26

Figure 9: *Posterior probability density for accelerated t-PA compared with SK, using a flat prior for log OR.*

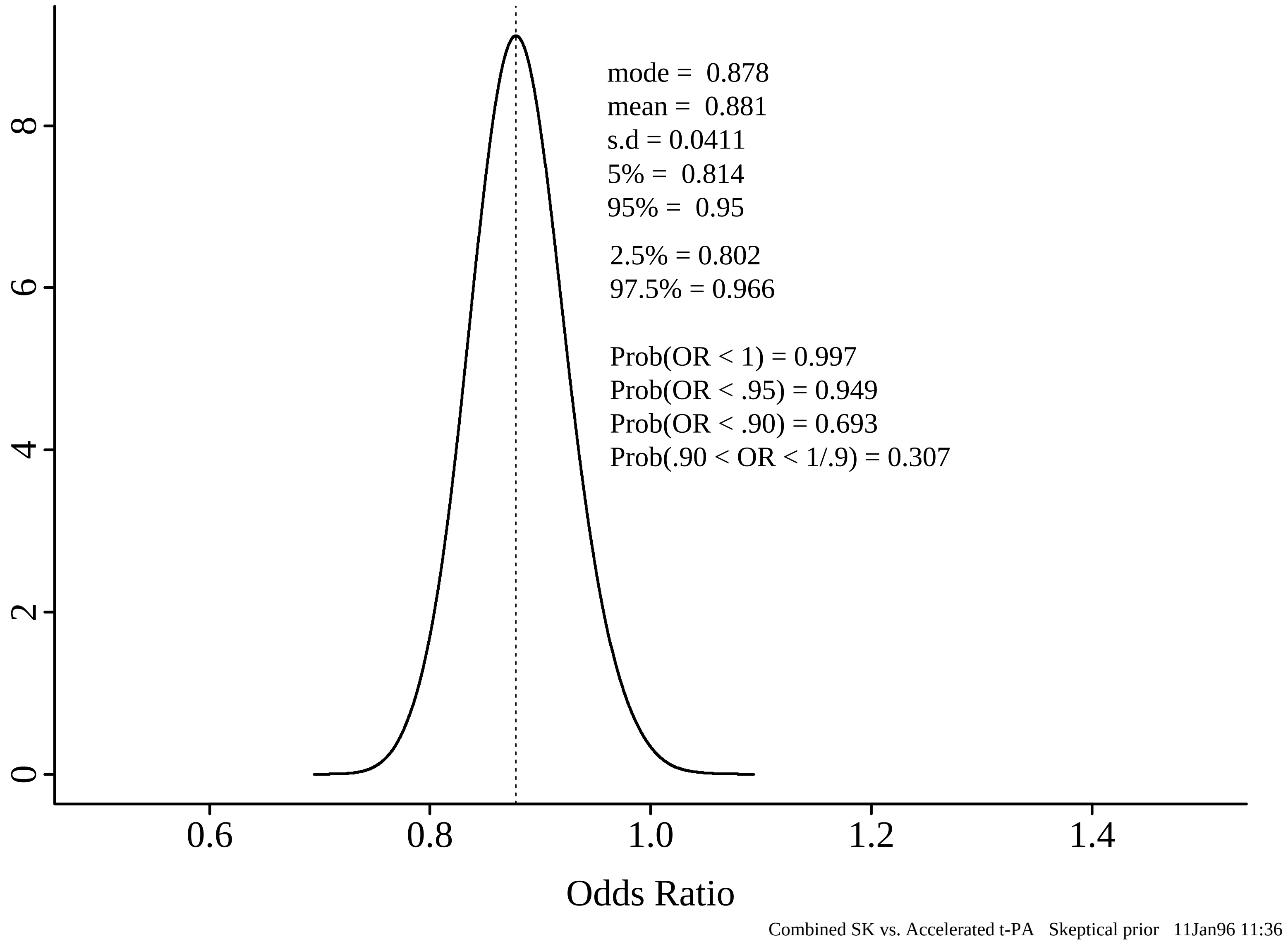
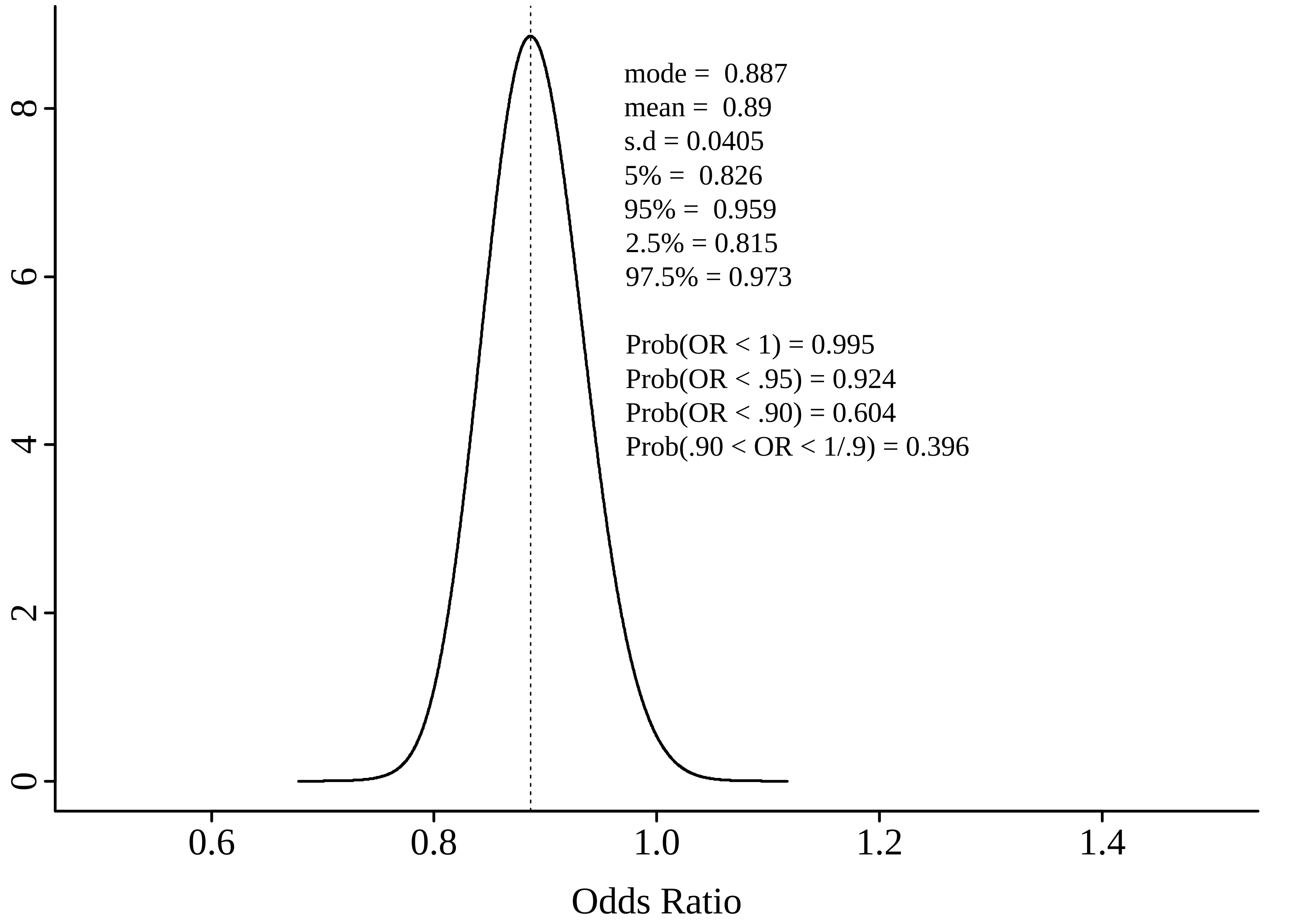


Figure 10: *Posterior probability density for accelerated t-PA compared with SK, using a prior distribution which assumed that* $\mathrm{Pr}(\mathrm{OR} > 2) = \mathrm{Pr}(\mathrm{OR} < \frac{1}{2}) = 0.025$.



Combined SK vs. Accelerated t-PA Very skeptical prior 11Jan96 13:59

Figure 11: *Posterior probability density for accelerated t-PA compared with SK, using a prior distribution which assumed that* $\Pr(\mathrm{OR} > 1\frac{1}{3}) = \Pr(\mathrm{OR} < \frac{3}{4}) = 0.025$.

Meta–analysis of Short–Acting Nifedipine

- From meta–analysis of 16 randomized trials by Furberg et al.[37][a]
- Individual subjects' data not available
- Used dead/alive; studies had varying follow–up and dose
- Model: logit $p_{ij} = \alpha +$study$_i$ + $\beta \times$ dose
- Fixed effects for β
- Random effects for studies[b]: Gaussian, σ^2 unknown but finite, has its own prior distribution ($\Gamma(10^{-4}, 10^{-4})$)[c]
- Quantity of interest: 100mg : placebo odds ratio for all–cause mortality

[a] With changes for the two Muller studies[59]

[b] For a single study, sites could be treated as random effects in the same way

[c] Use of a hyperprior to estimate σ^2 makes this similar to Empirical Bayes

- **BUGS Data File** (File `nifbugs.dat`)

```
list(dead = c(65, 65, 0, 0, 5, 7, 141, 150, 10, 10, 7, 7,
    6, 10, 90, 105, 2, 1, 10, 10, 0, 0, 5, 7, 2, 12, 4, 4,
    0, 1, 2, 5),
    dose = c(0, 30, 0, 40, 0,
             40, 0, 40, 0, 50, 0, 60, 0, 60, 0, 60, 0, 60, 0, 80, 0,
    80, 0, 80, 0, 80, 0, 100, 0, 100, 0, 100),
    study = c(12, 12, 5, 5, 1, 1, 16, 16, 14, 14,
    15, 15, 3, 3, 13, 13, 7, 7, 10, 10, 2, 2, 4, 4, 9, 9,
    6, 6, 8, 8, 11, 11),
    N = c(1146, 1130, 13, 13, 68, 60, 2251, 2240, 115, 112,
    120, 106, 75, 74, 678, 680, 327, 341, 88, 93, 25, 25,
    70, 68, 211, 214, 68, 64, 9, 13, 63, 63))
```

- **Initial Parameter Estimates** (File `bugs.in`)

```
list(int=0,b.dose=0,
     b.study=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),tau=
```

- **Command File** (File `bugs.cmd`)

```
compile("bugs.bug")
update(500)
monitor(int)
monitor(b.dose)
monitor(or)
monitor(c.study)
monitor(tau)
monitor(sigma)
update(2000)
stats(int)
stats(b.dose)
stats(or)
stats(c.study)
stats(sigma)
q()
```

- **Model Code** (File `bugs.bug`)

```
model logistic;

const
             S=16,                       # no. studies
             M=32;                       # no. records (2 * # studies
var
             dead[M],
             dose[M],
             study[M],
             N[M],
             p[M],
             int,b.dose,b.study[S],c.study[S],tau,sigma,or;
```

```
data in "nifbugs.dat";
inits in "bugs.in";

{
             for(k in 1:S) { # make random effects sum to zero
                          c.study[k] <- b.study[k] - mean(b.study[])
             }

             or <- exp(100*b.dose);

             for(i in 1:M) {
                          logit(p[i]) <- int+b.dose*dose[i]+ c.study
                          dead[i] ~ dbin(p[i],N[i]);
             }

             for(k in 1:S) {
                          b.study[k] ~ dnorm(0.0, tau);
             }

             #Prior distributions

             int ~ dnorm(0.0, 1.0E-6);
             b.dose ~ dnorm(0.0, 7.989E4) I(-0.01386,0.01386);
    # trunc at or=4, .025 prob>2

             tau ~ dgamma(0.0001, 0.0001);
             sigma <- 1/sqrt(tau);
    # s.d. of random effects

}
```

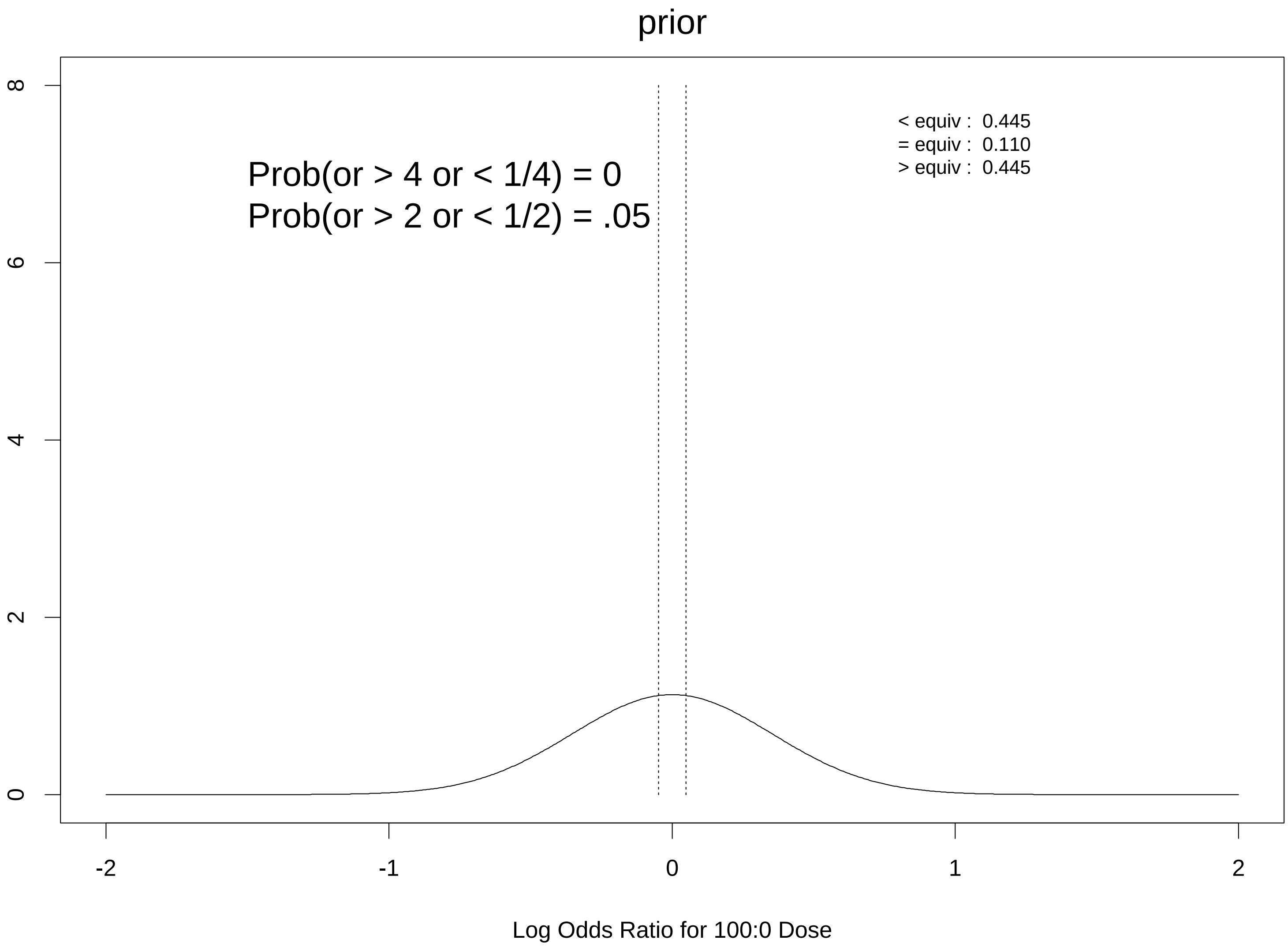


Figure 12: *Skeptical prior density for log OR; similary ("equivalence") zone is log odds* $\in [-0.05, 0.05]$

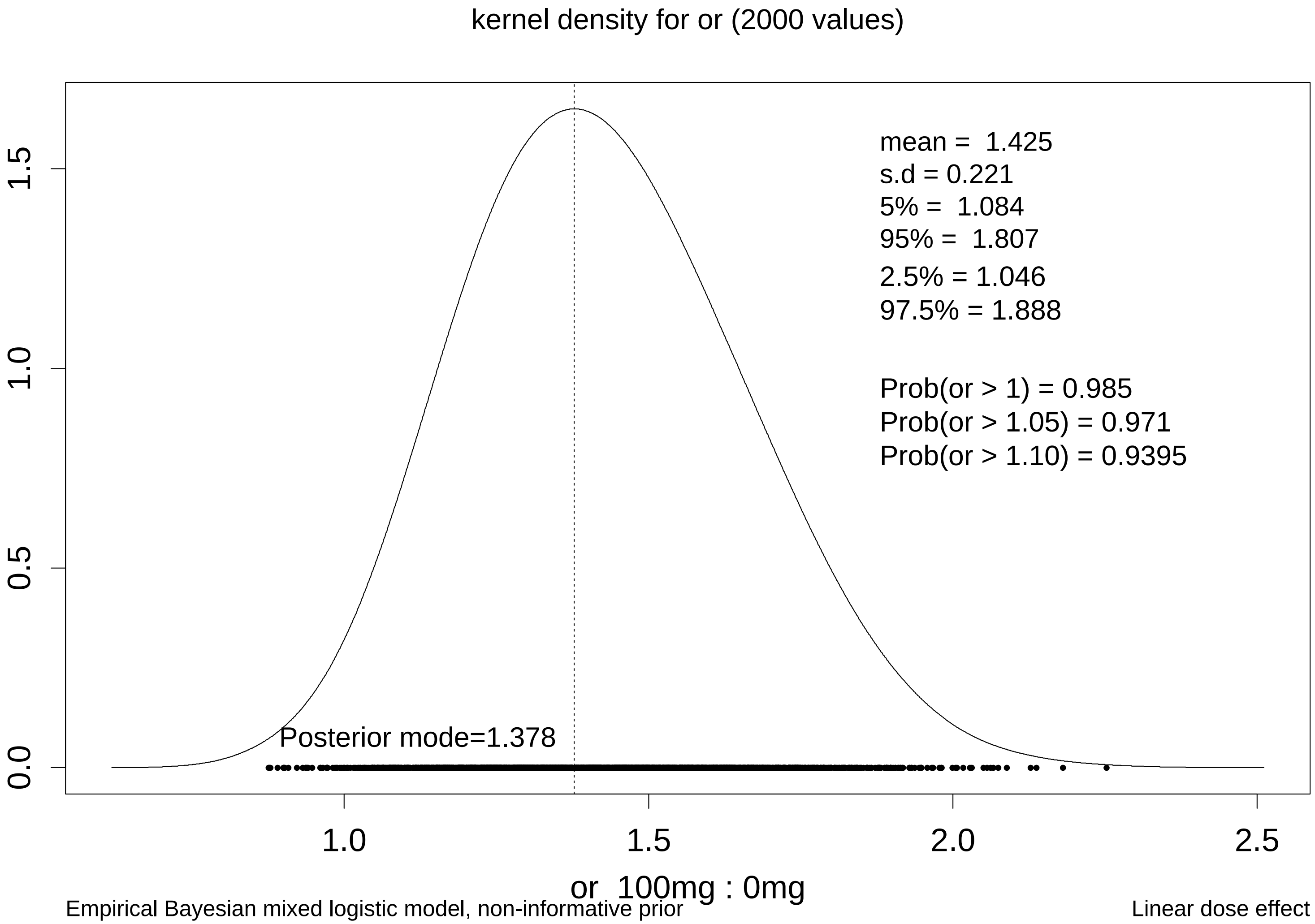


Figure 13: *Posterior density for pooled 100mg:0mg Nifedipine OR using a flat prior (Gaussian with variance 10^6) for β*

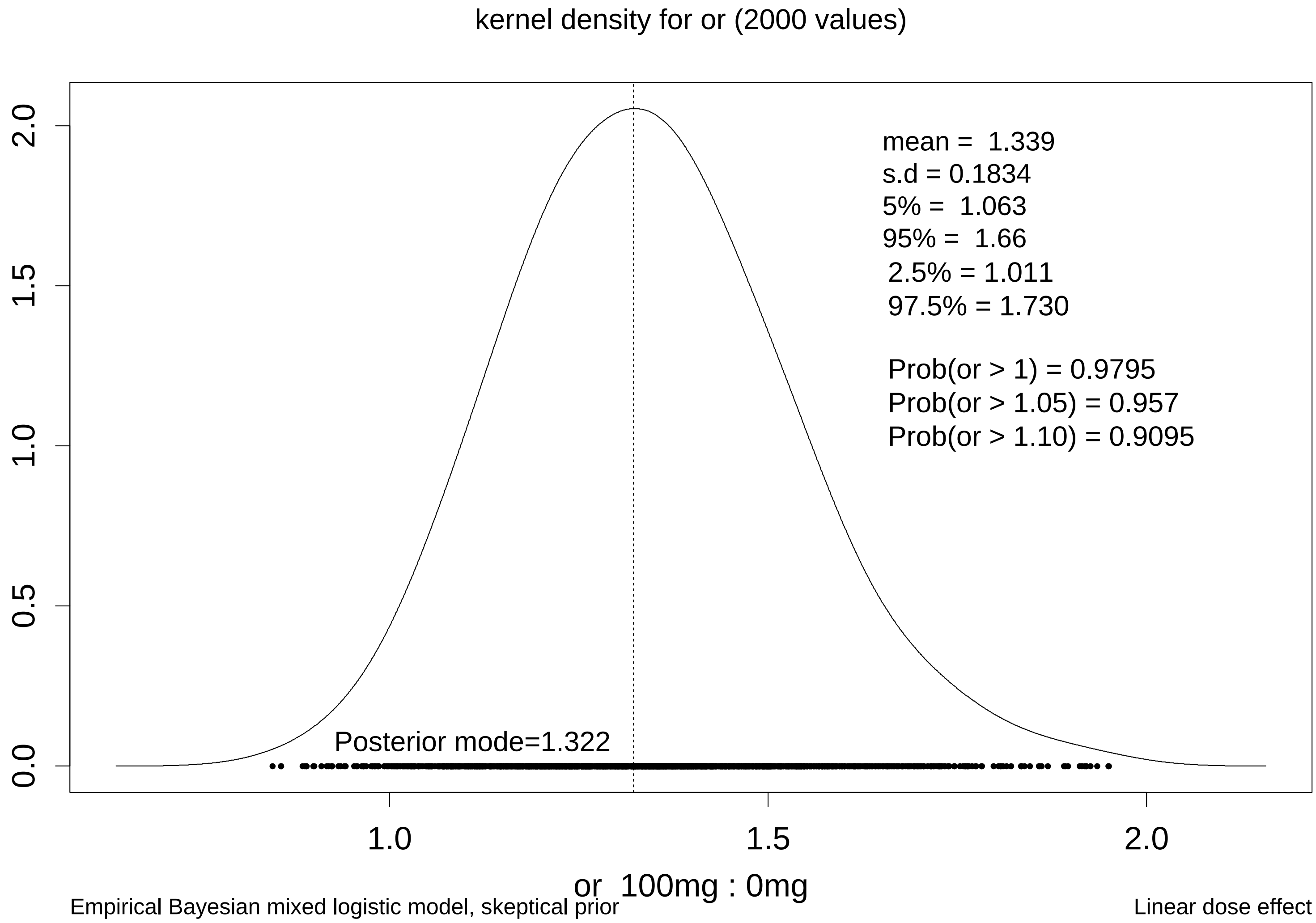


Figure 14: *Posterior density for pooled 100mg:0mg Nifedipine OR using a skeptical prior tilted toward no mortality effect*

Unplanned Interim Analysis

- Treatment A: $\frac{9}{44}$, Treatment B: $\frac{2}{43}$ events
- “Sensitivity analysis” using varying degrees of skepticism
- Larger prior variance $\longrightarrow\uparrow$ chance of large effect

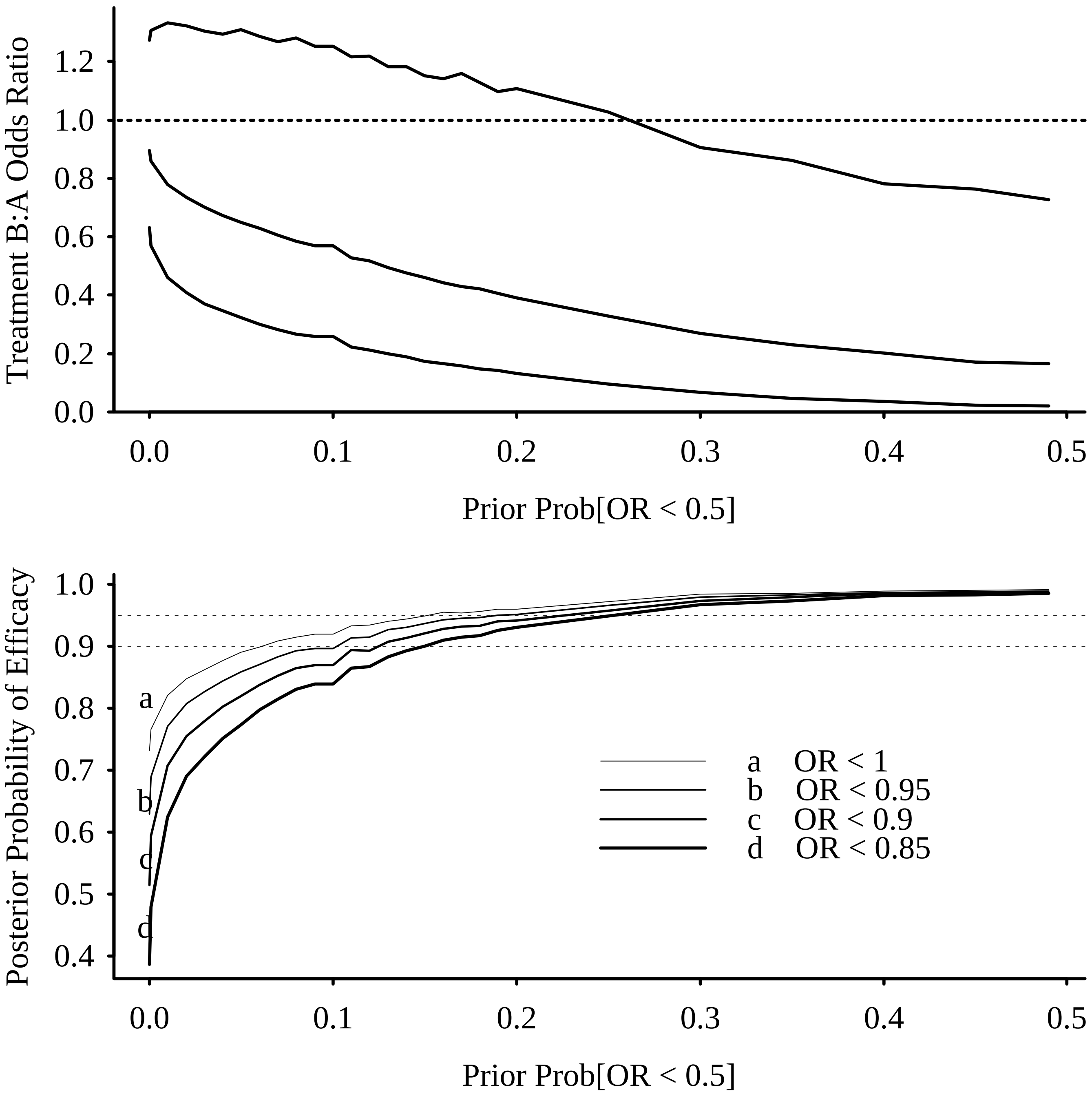


Figure 15: *Top panel: 0.95 credible interval and median B:A OR; bottom panel: posterior probabilities of any efficacy (OR < 1) or of clinically important efficacy*

Suggested Design Criteria

- Zone of clinical similarity is most important to pre–specify
- Mortality efficacy:
  $\Pr[OR < 1] \geq 0.95$
  $\Pr[OR < 0.9] \geq 0.9$
- Mortality safety: $\Pr[OR > 1] \geq 0.9$
- Similarity in an efficacy study:
  $\Pr[0.9 \leq OR \leq \frac{1}{0.9}] \geq 0.8$
- Similarity study: $\Pr[OR \leq \frac{1}{0.9}] \geq 0.9$
- Can accommodate relative and absolute effects simultaneously:
  $\Pr[OR < 0.9 \cup \theta_2 < \theta_1 - 0.05] > 0.9$

Example Study Design

- Study “powered” to detect a clinically relevant difference in a continuous measurement Y_1 measured at day 7 (in patients surviving ≥ 30 days); normality with equal variances is assumed
- Study not powered to detect a mortality difference but FDA wants to include mortality as a safety endpoint
- When mortality is relevant, should always union mortality reduction with improvement on Y_1, with no penalty, even though apriori power thought to be insufficient
- Second response variable is Y_2 = time to death, censored at 30 days; log–normal survival model expected to provide excellent fit
- Lachenbruch[51] derived a 2–stage 2 d.f. test for jointly testing for differences in Y_1 and Y_2 where Y_1 only applies depending on the value of Y_2

- Let μ_i = population mean Y_1 in treatment $i (i = 1, 2)$ given 30–day survival
- Let $\delta = \mu_2 - \mu_1$ and λ be the ratio (2:1) of population median survival time
- (Frequentist) sample estimates are differences in conditional means of Y_1 and anti–log of regression coefficient from log–normal model for Y_2
- Use Gibbs sampler to draw 30,000 realizations from the joint posterior density of δ and λ
- Prior for δ: normal with mean 0 and variance so that $\Pr[\delta < -20] = \frac{1}{3}$
- Prior for $\log \lambda$: normal with mean 0 and variance so that $\Pr[\lambda > 1.1] = \frac{1}{3}$
- Consider effect of skepticism:
  Variance of Y_1 estimated to be 100. If the variance was known and equal to 100 and if the difference in sample mean Y_1 was equal to -20, the following

probabilities of efficacy with respect to Y_1 obtain:

n Per Group	$\Pr[\delta < 0 \vert$ skeptical]	$\Pr[\delta < 0 \vert$ flat prior]
10	0.626	0.673
20	0.699	0.736
50	0.821	0.841
100	0.912	0.921
250	0.986	0.987
500	0.999	0.999
1000	1.000	1.000

- Main efficacy assessment: $\Pr[\delta < 0 \cup \lambda > 1]$

  Can also quote two separate probabilities

- Safety: $\Pr[\delta > 0]$, $\Pr[\lambda < 1]$

- Continuous monitoring:

  Stop when $\Pr[\delta < 0 \cup \lambda > 1] \geq 0.95$

  Stop when $\Pr[\delta > 0] \geq 0.99$ or

  $\Pr[\lambda < 1] \geq 0.90$ (mortality increase)

  Stop when

  $\Pr[-10 \leq \delta \leq 10 \cap \frac{1}{1.1} \leq \lambda \leq 1.1] \geq 0.8$

  (similarity)

- Frequentist design assumes a value for θ under H_a (θ_a)
- Many studies over-optimistically designed
  - Tried to detect a huge effect (one much larger than clinically useful) $\rightarrow n$ too small
  - Power calculation based on variances from small pilot studies[a]
- Does not formally recognize uncertainty about θ_a
- Pure Bayesian approach (no fixed sample size) is simple
- For fixed (maximum) sample size, standard C.L.–based formulas [17, 14], Bayesian confidence interval widths (see Thall & Simon for several examples) [48, 75]

[a] The power thus computed is actually a type of average power; one really needs to plot a power *distribution* and prehaps compute the 75^{th} percentile of power[72].

- Spiegelhalter and Freedman[72] show how predictive distributions can account for uncertainty about the treatment affect and σ (vs. trusting $\hat{\sigma}$ from pilot data)
- Can obtain an entire power distribution, not just the power under ideal parameter settings
- Bayesian “power” = posterior probability that left credible interval endpoint $>$ minimum worthwhile effect [72]
- Variety of approaches possible
  - Use no prior distribution for constructing C.I. or in specifying θ_a (θ_a = constant) (frequentist)
  - Use previous studies (informative prior) for C.I. but constant for θ_a
  - Allow uncertainty in θ_a but use no prior in constructing C.I. (frequentist test statistic) — Spiegelhalter & Freedman [72] main approach; can get *expected* power

- Different priors used in constructing C.I. and distribution for θ_a; former can come from regulators
- Harrell's S-PLUS `gbayes` function helps (hesweb1.med.virginia.edu/biostat/s/Hmisc.html)

Implications for Design/Evaluation

- Some studies can have lower sample sizes, e.g., more agressive monitoring/termination, one–tailed evaluation, no need to worry about spending α, use of data from similar studies
- Some studies will need to be larger because we are more interested in estimation than point–hypothesis testing or because we want to be able to conclude that a clinically significant difference exists
- Studies can be much more flexible
  - Formal incorporation of results from previous studies on the same treatment
  - Can use a prior which is a mixture of posterior from previous studies and non-informative prior; mixing probability = "applicability" of previous studies to current one, set by regulators
  - Adapt treatment during study

  - Unplanned analyses
  - With continuous monitoring, studies can be better designed — bailout still possible
  - Can extend a promising study
  - Reduce number of small, poorly designed studies by de-emphasizing power against a fixed θ
  - Reduce distinction between Phase II and III studies

- Most scientific approach is to experiment until you have the answer
- Allow for agressive, efficient designs
- Incentives for better design
- Let the data speak for themselves
- Trickery will still be apparent

Acceptance of Bayesian Methods by Regulatory Authorities and Industry

- FDA Center for Devices & Radiologic Health is actively courting Bayesian design & analytic plans
- Other FDA centers are not against any analytic philosophy; they are pro–science
- Biggest hurdle is industry, not regulators:
  Deadline mentality and risk aversion — "Let's do it the way we did it for our last successful application."
- Second biggest: It takes time to be a Bayesian
- Software will help
- Sound analytic plan before data analyzed, as always
- Consider letting reviewers specify priors

Summary

- Bayesian analysis actually reduces time spent in arguing about statistical tests/designs!
- Substitutes an argument about the choice of a prior for the following arguments:
  - Which treatment effect to use for sample size calculations
  - One–tailed vs. two–tailed test
  - “Exact” vs. approximate P–values (conditional vs. unconditional analyses)
  - How to test for similarity
  - Multiplicity adjustments for multiple endpoints
  - Scheduling, adjustments for sequential monitoring
  - How to penalize for extending a study

- How to translate results to clinical significance
- How to prevent the audience from misinterpreting a small or large P–value

- A little bit of skepticism goes a long way

References

[1] K. Abrams, D. Ashby, and D. Errington. Simple Bayesian analysis in clinical trials: A tutorial. *Controlled Clinical Trials*, 15:349–359, 1994.

[2] J. H. Albert. Teaching Bayesian statistics using sampling methods and MINITAB. *American Statistician*, 47:182–191, 1993.

[3] P. K. Andersen, J. P. Klein, and M. Zhang. Testing for centre effects in multi-centre survival studies: A monte carlo comparison of fixed and random effects tests. *Statistics in Medicine*, 18:1489–1500, 1999.

[4] V. Barnett. *Comparative Statistical Inference*. Wiley, second edition, 1982.

[5] E. J. Bedrick, R. Christensen, and W. Johnson. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, 91:1450–1460, 1996.

[6] E. J. Bedrick, R. Christensen, and W. Johnson. Bayesian binomial regression: Predicting survival at a trauma center. *American Statistician*, 51:211–218, 1997.

[7] J. Berger and T. Sellke. Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of the American Statistical Association*, 82:112–139, 1987.

[8] J. O. Berger and D. A. Berry. Statistical analysis and the illusion of objectivity (letters to editor p. 430-433). *American Scientist*, 76:159–165, 1988.

[9] J. O. Berger, B. Boukai, and Y. Wang. Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, 12:133–160, 1997.

[10] D. A. Berry. Interim analysis in clinical trials: The role of the likelihood principle. *American Statistician*, 41:117–122, 1987.

[11] D. A. Berry. Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 3, pages 79–94. Oxford University Press, 1988.

[12] D. A. Berry, M. C. Wolff, and D. Sack. Decision making during a phase III randomized controlled trial. *Controlled Clinical Trials*, 15:360–378, 1994.

[13] M. Borenstein. The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials*, 15:411–428, 1994.

[14] M. Borenstein. Planning for precision in survival studies. *Journal of Clinical Epidemiology*, 47:1277–1285, 1994.

[15] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA, 1973.

[16] N. Breslow. Biostatistics and Bayes (with discussion). *Statistical Science*, 5:269–298, 1990.

[17] D. R. Bristol. Sample sizes for constructing confidence intervals and testing hypotheses. *Statistics in Medicine*, 8:803–811, 1989.

[18] J. M. Brophy and L. Joseph. Placing trials in context using Bayesian analysis: GUSTO revisited by Reverend Bayes. *Journal of the American Medical Association*, 273:871–875, 1995.

[19] P. R. Burton. Helping doctors to draw appropriate inferences from the analysis of medical studies. *Statistics in Medicine*, 1994:1699–1713, 1994.

[20] G. Casella and E. I. George. Explaining the Gibbs sampler. *American Statistician*, 46:167–174, 1992.

[21] R. J. Cook and V. T. Farewell. Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society A*, 159:93–110, 1996.

[22] K. A. Cronin, L. S. Freedman, R. Lieberman, H. L. Weiss, S. W. Beenken, and G. J. Keloff. Bayesian monitoring of phase II trials in cancer cemo-prevention, with discussion by H. C. van Houwelingen. *Journal of Clinical Epidemiology*, 52:705–716, 1999.

[23] S. J. Cutler, S. W. Greenhouse, J. Cornfield, and M. A. Schneiderman. The role of hypothesis testing in clinical trials. *Journal of Chronic Diseases*, 19:857–882, 1966.

[24] G. A. Diamond and J. S. Forrester. Clinical trials and statistical verdicts: Probable grounds for appeal (*note: this article contains some serious statistical errors*). *Annals of Internal Medicine*, 98:385–394, 1983.

[25] D. O. Dixon and R. Simon. Bayesian subset analysis. *Biometrics*, 47:871–881, 1991.

[26] D. O. Dixon and R. Simon. Bayesian subset analysis in a colorecta cancer clinical trial. *Statistics in Medicine*, 11:13–22, 1992.

[27] R. M. J. Donahue. A note on information seldon reported via the p value. *American Statistician*, 53:303–306, 1999.

[28] B. Efron. Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80:3–26, 1993.

[29] S. S. Emerson. Stopping a clinical trial very early based on unplanned interim analysis: A group sequential approach. *Biometrics*, 51:1152–1162, 1995.

[30] R. D. Etzioni and J. B. Kadane. Bayesian statistical methods in public health and medicine. *Annual Review of Public Health*, 16:23–41, 1995.

[31] D. Faraggi and R. Simon. Large sample Bayesian inference on the parameters of the proportional hazard model. *Statistics in Medicine*, 16:2573–2585, 1997.

[32] P. M. Fayers, D. Ashby, and M. Parmar. Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in Medicine*, 16:1413–1430, 1997.

[33] A. R. Feinstein. *Clinical Biostatistics*. C. V. Mosby, St. Louis, 1977.

[34] L. D. Fisher. Comments on Bayesian and frequentist analysis and interpretation of clinical trials. *Controlled Clinical Trials*, 17:423–434, 1996.

[35] L. Freedman. Bayesian statistical methods. *British Medical Journal*, 313:569–570, 1996.

[36] L. S. Freedman, D. J. Spiegelhalter, and M. K. B. Parmar. The what, why and how of Bayesian clinical trials monitoring. *Statistics in Medicine*, 13:1371–1383, 1994.

[37] C. D. Furberg, B. M. Psaty, and J. V. Meyer. Nifedipine: Dose-related increase in mortality in patients with coronary heart disease. *Circulation*, 92:1326–1331, 1995.

[38] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.

[39] S. L. George, C. Li, D. A. Berry, and M. R. Green. Stopping a trial early: Frequentist and Bayesian approaches applied to a CALGB trial of non-small cell lung cancer. *Statistics in Medicine*, 13:1313–1328, 1994.

[40] W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. A language and program for complex Bayesian modeling. *The Statistician*, 43:169–177, 1994.

[41] S. N. Goodman. A comment on replication, p-values and evidence. *Statistics in Medicine*, 11:875–879, 1998.

[42] J. B. Greenhouse and L. Wasserman. Robust Bayesian methods for monitoring clinical trials. *Statistics in Medicine*, 14:1379–1391, 1995.

[43] J. Grossman, M. K. B. Parmar, and D. J. Spiegelhalter. A unified method for monitoring and analysing controlled trials. *Statistics in Medicine*, 13:1815–1826, 1994.

[44] L. Hashemi, B. Nandram, and R. Goldberg. Bayesian analysis for a single 2×2 table. *Statistics in Medicine*, 16:1311–1328, 1997.

[45] J. V. Howard. The 2×2 table: A discussion from a Bayesian viewpoint. *Statistical Science*, 13:351–367, 1998.

[46] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court, La Salle, IL, 1989.

[47] M. D. Hughes. Reporting Bayesian analyses of clinical trials. *Statistics in Medicine*, 12:1651–1663, 1993.

[48] L. Joseph, R. Du Berger, and P. Bélisle. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, 16:769–781, 1997.

[49] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370, 1996.

[50] D. H. Krantz. The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 44:1372–1381, 1999.

[51] P. A. Lachenbruch. Analysis of data with clumping at zero. *Biometrical Journal*, 18:351–356, 1976.

[52] J. M. Legler and L. M. Ryan. Latent variable models for teratogenesis using multiple binary outcomes. *Journal of the American Statistical Association*, 92:13–20, 1997.

[53] H. P. Lehmann and B. Nguyen. Bayesian communication of research results over the World Wide Web (see `http://omie.med.jhmi.edu/bayes`). *M.D. Computing*, 14(5):353–359, 1997.

[54] R. J. Lilford and D. Braunholtz. The statistical basis of public policy: A paradigm shift is overdue. *British Medical Journal*, 313:603–607, 1996.

[55] D. Malakoff. Bayes offers a 'new' way to make sense of numbers. *Science*, 286:1460–1464, 1999.

[56] M. R. Nester. An applied statistician's creed. *Applied Statistics*, 45:401–410, 1996.

[57] M. A. Newton and A. E. Rafter. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society B*, 56:3–48, 1994.

[58] M. Oakes. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. Wiley, New York, 1986.

[59] L. H. Opie and F. H. Messerli. Nifedipine and mortality: Grave defects in the dossier. *Circulation*, 92:1068–1073, 1995.

[60] G. L. Rosner and D. A. Berry. A Bayesian group sequential design for a multiple arm randomized clinical trial. *Statistics in Medicine*, 14:381–394, 1995.

[61] K. J. Rothman. A show of confidence (editorial). *New England Journal of Medicine*, 299:1362–3, 1978.

[62] K. J. Rothman. No adjustments are needed for multiple comparisons. *Epidemiology*, 1:43–46, 1990.

[63] R. M. Royall. The effect of sample size on the meaning of significance tests. *American Statistician*, 40:313–315, 1986.

[64] D. B. Rubin. The Bayesian bootstrap. *Applied Statistics*, 9:130–134, 1981.

[65] H. Sackrowitz and E. Samuel-Cahn. p values as random variables — Expected p values. *American Statistician*, 53:326–331, 1999.

[66] D. J. Sargent and J. S. Hodges. A hierarchical model method for subgroup analysis of time-to-event data in the Cox regression setting. Presented at the Joint Statistical Meetings, Chicago, 1996.

[67] M. J. Schervish. p values: What they are and what they are not. *American Statistician*, 50:203–206, 1996.

[68] S. Senn. *Statistical Issues in Drug Development*. Wiley, Chichester, England, 1997.

[69] L. B. Sheiner. The intellectual health of clinical drug evaluation. *Clinical Pharmacology and Therapeutics*, 50:4–9, 1991.

[70] A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: A sampling-resampling perspective. *American Statistician*, 46:84–88, 1992.

[71] D. J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5:421–433, 1986.

[72] D. J. Spiegelhalter and L. S. Freedman. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5:1–13, 1986.

[73] D. J. Spiegelhalter, L. S. Freedman, and M. K. B. Parmar. Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine*, 12:1501–1511, 1993.

[74] D. J. Spiegelhalter, L. S. Freedman, and M. K. B. Parmar. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society Series A*, 157:357–416, 1994.

[75] P. F. Thall and R. Simon. A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Controlled Clinical Trials*, 15:463–481, 1994.

[76] P. F. Thall and H. Sung. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine*, 17:1563–1580, 1998.

[77] The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine*, 329:673–682, 1993.

[78] A. Thomas, D. J. Spiegelhalter, and W. R. Gilks. Bugs: A program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 4, pages 837–842. Clarendon Press, Oxford, UK, 1992.

[79] R. Tibshirani. Noninformative priors for one parameter of many. *Biometrika*, 76:604–608, 1989.

[80] P. H. Westfall, W. O. Johnson, and J. M. Utts. A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, 84:419–427, 1997.

Information on the WWW

Simon Jackman’s web page: `http://tamarama.stanford.edu/mcmc/`