An Introduction to Bayesian Methods with Clinical Applications

Frank E Harrell Jr and Mario Peruggia
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
School of Medicine, University of Virginia
Box 600 Charlottesville VA 22908

fharrell@virginia.edu

July 8, 1998

Outline

- The Scientific Method and the Problem of Induction (General discussion pp. 1–10 based on Howson and Urbach[22])
- Traditional Approaches
  - Philosophy of Science (Popper)
  - Statistics (Fisher, Neyman-Pearson)
- Bayesian Approach
  - Brief Overview
  - Simple 2×2 Table Example
  - Methods
  - Advantages
  - Disadvantages and Controversies

- Non–informative Prior Distributions
- Sequential Testing
- Two-Sample Binomial Example
- Has hypothesis testing hurt science?
- Implications for Designing and Evaluating Clinical Studies
- In addition to Howson and Urback[22], good general references (roughly in order of difficulty) are: Berry[5], DeGroot[13], Barnett[2] and Berger[3]

The problem of Induction

- General character of scientific hypotheses
- Information derives from empirical observations
- How can we ensure that any given theory is right?
- Fairly general agreement:
  No positive solution is available.
- One possibility: Appraise scientific theories in terms of their probabilities. This leads to
  - Probabilistic Induction
  - Bayesianism

Philosophy of Science

- Probabilistic induction has historically been resisted by both philosophers of science and statisticians, who think that "objective" conclusions can be reached.
- **Karl Popper**
- Some theories can be established to be objectively superior to competing ones.
  - Theories can be refuted by empirical observations.
  - Deductive consequences can be observationally verified.
- Example: All swans are white:
  - Refuted by sighting of black swan.
  - corroborated by sighting of white swan.

- Problems:
  - How to choose *objectively* among the theories that have not been refuted?
  - Focuses only on logical consequences of a theory, but:
    1. Many deterministic theories cannot be checked directly.
    2. Many theories are explicitly probabilistic (e.g. Mendel’s theory of inheritance).
  - Evidence is often measured with error and is therefore not certain.

Statistics

- **Sir R.A. Fisher**
- Evidence can have a negative impact on a statistical hypothesis (null hypothesis)
- An experiment gives “the facts a chance of disproving the null hypothesis.”
- Statistical refutation of the null hypothesis is different from its logical refutation.
- Hypothesis is contradicted by the data (c.f. small P–value) means that either an improbable event has occurred, or the null hypothesis is false, or both.
- What to do it the evidence does not contradict the null hypothesis?
- According to Fisher a null hypothesis is never accepted.

- What test statistic to use?
  - No general prescription on what statistic to use.
  - Different test statistics can lead to different conclusions from the same analysis.
  - May get *logically* inconsistent conclusions (c.f. collapsing of contingency tables and χ^2 test).
- Fisher's other important contributions
  - Testing of causal hypothesis (agricultural and clinical trials).
  - Controlled experiment.
  - Principle of randomization.

Neyman and Pearson

- Test a statistical hypothesis (H_0) not in isolation but *against* competing theories (H_1).
- This is different from Fisher's idea that one should be able to reject a theory regardless of how other theories perform.
- Two possible errors:
  - H_0 true $\rightarrow$ accept H_1 (Type I)
  - H_0 false $\rightarrow$ accept H_0 (Type II)
- Goal: Try to keep the probabilities of both types of error small.

N.P. Theory Helps But

- Basic interpretation problem remains:
  What do acceptance and rejection mean?

- Both Fisher and N.P. agree on one point:
  There are no *probabilities* of theories being correct.

- One problem with P–values: $P = 0.05$ "essentially does not provide any evidence against the null hypothesis" (Berger et al.[4]) — $\Pr[H_1|P = 0.05]$ will be near 0.5 in many cases if prior probability of truth of H_0 is near 0.5

Bayesian Approach

- This approach formally recognizes the inherent uncertainty about scientific theories.
- Degrees of certainty are translated into probabilities.
- These probabilities are subjective:
  They reflect an investigator's personal views.
- Probabilities are revised each time that new evidence becomes available using Bayes theorem.
- The more new data are collected, the less the impact of the original subjective assessment becomes.

Simple Example

- Automated Test (+ or -) for Hypertension (Y or N)

	Y	N	
+	15	25	40
-	5	55	60
	20	80	100

- Unconditional probability of disease (prevalence)

$$P(Y) = 20/100 =$$

- Conditional probability of positive test given diseased (sensitivity)

$$\begin{aligned} P(+|Y) &= 15/20 \\ &= \frac{15/100}{20/100} = \frac{P(+\,\mathrm{and}\,Y)}{P(Y)} \end{aligned}$$

- Conditional probability of negative test given healthy (specificity)

$$\begin{aligned} P(-|N) &= 55/80 \\ &= \frac{55/100}{80/100} = \frac{P(-\,\text{and}\,N)}{P(N)} \end{aligned}$$

- Suppose table is representative of the whole population.

  What is the probability that an individual who tests positive is hypertensive?

- Want conditional probability of diseased given positive test (predictive value positive)

$$\begin{aligned} P(Y|+) &= 15/40 \\ &= \frac{15/100}{40/100} = \frac{15/100}{15/100 + 25/100} \\ &= \frac{(15/20)(20/100)}{(15/20)(20/100) + (25/80)(80/100)} \\ &= \frac{P(+|Y)P(Y)}{P(+|Y)P(Y) + P(+|N)P(N)} \end{aligned}$$

- Last formula is called Bayes rule or Bayes theorem.
- We have:
  - “Prior” knowledge of the proportion of diseased people in the population (prevalence)
  - A statistical model for how the test performs (sensitivity and specificity)
- Mr. Smith comes to the clinic.
  - Before administering the test, our prior beliefs of his being hypertensive coincide with the prevalence.
  - After administering the test, we use Bayes theorem to update our prior beliefs of his being hypertensive and obtain the “posterior” probability that he has high blood pressure given the outcome of the test.

Methods

- The basic approach to scientific learning described in the previous example characterizes the Bayesian method.
- Use Bayes' rule to update degree of evidence given observed data.
- 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 probability that S is true given the data.
- If θ is a parameter of interest (e.g., log odds ratio or difference in mean blood pressure), need a probability distribution of θ given the data.
- Evidence for effect θ $=$ evidence from data $\times$ prior knowledge about θ.

- Evidence quantified by probability distribution (prior & posterior distributions).
- 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)."[30]
- Results in a probability most clinicians think they're getting while p-values are often misinterpreted [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].[11]
- Provides formal mechanism for using prior information/bias — prior distribution for θ.

[a] Nineteen of 24 cardiologists rated the posterior probability as the quantity they would most like to know, from among three choices. Half of 24 cardiologists gave the correct response to a 4–choice question concerning p-values.[14]

- Places emphasis on estimation and graphical presentation rather than hypothesis testing.
- Avoids 1–tailed/2–tailed test controversy.
- If $\Pr[\text{drug B is better than drug A}] = 0.92$, this is true whether drug C was compared to drug D or not. No need for specialized multiple comparisons techniques.
- Avoids many of the complexities of sequential monitoring —
  P–value adjustment is needed in frequentist methods for the type I error to have its intended meaning
  A posterior probability is still a probability → Can monitor continuously.
- Allows accumulating information (from this as well as other trials) to be used as trial proceeds.

Controversies

- Posterior probabilities may be hard to compute (often have to use numerical methods).
- How does one choose a prior distribution?[24]
  - Biased prior – expert opinion
    difficult, can be manipulated, medical experts often wrong, whose opinion do you use?[16]
  - 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. $[\frac{1}{10}, 10]$

Non–Informative Prior Distributions

- 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 unprejudiced observer.[8]
- 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 assuming no prior beliefs either way.

Consumer or Reviewer Specification of Prior

- Place statistics describing study results on web page
- Posterior computed and displayed using Java applet (Lehmann & Nguyen [25])

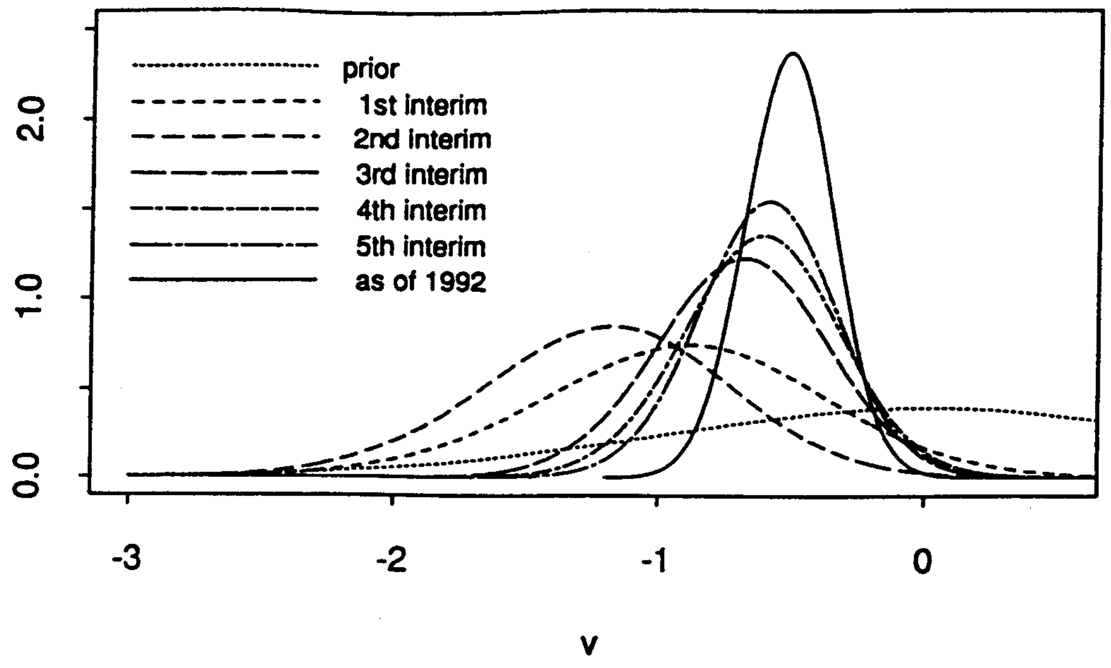

Figure 1: Sequentially monitoring a clinical trial[20]. v is the log hazard ratio.

Two–Sample Binomial Example

- Considered two prior distributions (one flat)
- Data: Treatment A $\frac{30}{200}$
  Treatment B $\frac{18}{200}$
- OR = 0.56; $2P = 0.064$
  0.95 C.L. $[.304, 1.042]$

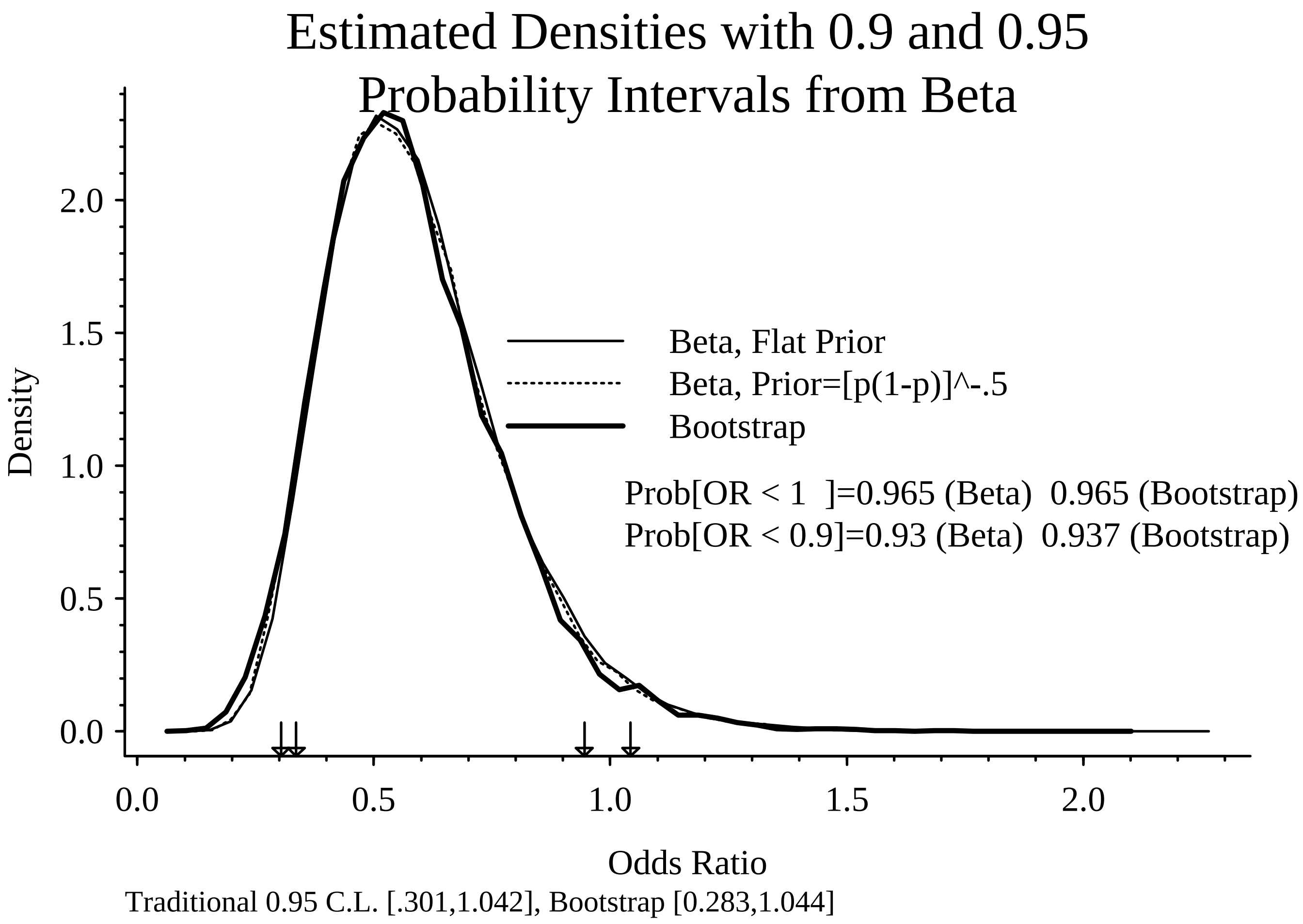

Figure 2: Posterior distribution of the odds ratio. The posterior were derived using the bootstrap and using a Bayesian approach with 2 prior densities.

Has Hypothesis Testing Hurt Science?

- Many studies are powered to be able to detect a huge treatment effect
- → sample size too small → 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[27]

- 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[31]

Implications for Design/Evaluation

- Many studies overoptimistically designed
  - Tried to detect a huge effect (one much larger than clinically useful) → n too small
  - Power calculation based on variances from small pilot studies[a]
- Some studies can have lower sample sizes, e.g., more aggressive monitoring/termination, one–tailed evaluation, no need to worry about spending α
- 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

[a] The power thus computed is actually a type of average power; one really needs to plot a power *distribution* and perhaps compute the 75^{th} percentile of power[32].

  - 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
  - Reduce distinction between Phase II and III studies
- Most scientific approach is to experiment until you have the answer
- Allow for aggressive, efficient, better designs
- Let the data speak for themselves

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

[3] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer–Verlag, New York, 1985.

[4] 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.

[5] D. A. Berry. *Statistics: A Bayesian Perspective*. Duxbury Press, Belmont, CA, 1996.

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

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

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

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

[10] 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.

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

[12] 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.

[13] M. H. DeGroot. *Probability and Statistics*. Addison Wesley, Reading, MA, 1986.

[14] 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.

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

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

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

[18] 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.

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

[20] 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.

[21] S. N. Goodman and J. A. Berlin. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, 121:200–206, 1994.

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

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

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

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

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

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

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

[29] K. J. Rothman. Significance questing. *Annals of Internal Medicine*, 105:445–447, 1986.

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

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

[32] 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.

[33] 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.

[34] 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.