Comments on Jim Berger's Lectures, Part 1

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
www.ams.ucsc.edu/∼draper

CBMS-MUM
Santa Cruz CA

24 July 2012

Themes So Far

- Lecture 1: Reproducibility of science: misuse and technical failures of frequentist hypothesis-testing
  - Publication bias (only studies with $p < 0.05$ get into print)
  - Experimental biases not properly assessed statistically
    - Jim's plot (Lecture 1, page 8) of estimated neutron lifetimes against time
    - In the language of meta-analysis, physicists have been using fixed-effects models when they should have been using random-effects models
    - Let y_t be the estimate of the physical constant θ at time $t = 1, \ldots, T$; the physicists' fixed-effects model is

$$y_t = \theta + b + e_t\,, \quad b = 0 \tag{1}$$

(this assumes no bias at time t), which can be rewritten more fully as

$$\begin{aligned} (\theta, \sigma^2) &\sim \text{diffuse} \\ (y_t|\theta, \sigma^2) &\overset{\text{IID}}{\sim} N\left(\theta, \sigma^2\right)\,; \end{aligned} \tag{2}$$

Experimental Biases

- Reproducibility of science: misuse and technical failures of frequentist hypothesis-testing (continued)
  - Experimental biases not properly assessed statistically (continued)
    - instead they should have been working with the random-effects model

$$\begin{aligned} y_t &= \theta + b_t + e_t^* \\ b_t &= b + \epsilon_t^* , \quad b = 0 \end{aligned} \tag{3}$$

(this allows for an unknown bias at time t), which can be rewritten more fully as

$$\begin{aligned} (\theta, \sigma_y^2, \sigma_\theta^2) &\sim \text{diffuse} \\ (\theta_t | \theta, \sigma_\theta^2) &\overset{\text{IID}}{\sim} N\left(\theta, \sigma_\theta^2\right) \\ (y_t | \theta_t, \sigma_y^2) &\overset{\text{indep}}{\sim} N\left(\theta_t, \sigma_y^2\right) \end{aligned} \tag{4}$$

    - As soon as new data come in that sharply contradict the previous estimate (e.g., in the late 1960s for the neutron-lifetime problem), the "random bias term" in model (4) will sharply widen the uncertainty bands appropriately

Bayesian Hypothesis Testing

- Reproducibility of science: misuse and technical failures of frequentist hypothesis-testing (continued)
  - Misinterpretation of p values
  - Failure to adjust for multiplicities
    - "The tradition in epidemiology is to ignore multiple testing, usually arguing that the purpose is to find anomalies for further study"; this would be OK if, e.g., an observational nutritional study examining many possible effects and showing a negative association between consumption of beta carotene and incidence of colon cancer were labeled as exploratory, but instead the headline is "beta carotene prevents colon cancer"
    - "The tradition in psychology is to ignore optional stopping; if you're close to $p = 0.05$, go get more data to try to get to $p = 0.05$ (with no adjustment)"
    - Multiple statistical analyses, until you get the answer you want
- How Bayesian hypothesis-testing can fix some of these problems
  - Example: Alvac (no effect as an HIV vaccine), Aidsvax (no effect as an HIV vaccine); could Alvac as a primer and Aidsvax as a booster work?

Bayesian Hypothesis Testing (continued)

- How Bayesian hypothesis-testing can fix some of these problems (continued)
  - Big trial conducted in Thailand with 16,395 (general-population, non-high-risk) individuals randomized to treatment (first Alvac, then Aidsvax two weeks later) or control (two saline shots two weeks apart)
  - HIV rate 74 out of 8198 ($\bar{y}_C = 9.0$ per 1,000 people) in control group, 51 out of 8197 ($\bar{y}_T = 6.2$ per 1,000) in treatment group; treatment rate is 30% less than control rate (highly practically significant)
  - With the usual sampling model $(\bar{y}_C - \bar{y}_T | \Delta, \sigma^2) \sim N(\Delta, \sigma^2)$, the estimated improvement (frequentist or diffuse-prior-Bayesian) under the treatment is $\hat{\Delta} = \bar{y}_C - \bar{y}_T = (9.0 - 6.2) = 2.8$ per 1,000, with a frequentist standard error/diffuse-prior-posterior standard deviation of 1.35 per 1,000
  - Frequentist hypothesis-testing people with (H_0: $\Delta = 0$ versus H_1: $\Delta \neq 0$) get a p value of 0.039; with (H_0: $\Delta \leq 0$ versus H_1: $\Delta > 0$) or (H_0: $\Delta = 0$ versus H_1: $\Delta > 0$) such people

Bayesian Hypothesis Testing (continued)

- How Bayesian hypothesis-testing can fix some of these problems (continued)
  - get a p value of 0.019; in both cases this is "statistically significant at the usual 0.05 false-positive level"; on a standard frequentist analysis the treatment therefore looks effective
  - 95% interval (frequentist or diffuse-prior-Bayesian) for the population difference Δ in rates (control – treatment) runs from 0.1 to 5.5 to per 1,000 people ("statistically significant at the usual 0.05 false-positive level"); diffuse-prior $P(\Delta > 0|\mathrm{data}) = 0.98 = (1-$ the one-tailed p value); on a naive Bayesian analysis the treatment looks effective
  - However, HIV experts (other than those who ran the study) nearly unanimously agree that both Alvac and Aidsvax by themselves show no biological activity, so they are highly skeptical of this finding
  - Jim proposes to resolve this conflict by performing a robust-Bayesian sharp-null test of (H_0: $\Delta = 0$ versus H_1: $\Delta > 0$); he gets a robust Bayes factor of H_1 to H_0 of $B_{10} \doteq 5.6$, and $-e\,p\log p$ for the one-tailed p value of 0.019 yields an approximate upper bound on B_{10} of about 4.8;

Bayesian Hypothesis Testing (continued)

- How Bayesian hypothesis-testing can fix some of these problems (continued)
  - he interprets this as much weaker evidence in favor of the treatment than the p values would imply, which makes the HIV experts (other than those who ran the study) happy
  - I have two problems with this analysis:
    - (1) To motivate his choice of (H_0: $\Delta = 0$ versus H_1: $\Delta > 0$) Jim invokes the idea of plausible precise hypotheses
    - "A precise hypothesis is an hypothesis of lower dimension than the alternative (e.g., H_0: $\mu = 0$ versus H_1: $\mu \neq 0$)"
    - "A precise hypothesis is plausible if it has a reasonable prior probability of being true"
    - He now argues that, in the HIV example, because both Alvac and Aidsvax by themselves show no biological activity, the hypothesis that the population difference Δ in HIV rates under treatment versus placebo is EXACTLY 0 is plausible, so you should test (H_0: $\Delta = 0$ versus H_1: $\Delta > 0$)
    - However, there are many instances in chemistry in which compounds A and B by themselves have no effect on a chemical process but the introduction of both A and B

Bayesian Hypothesis Testing (continued)

- How Bayesian hypothesis-testing can fix some of these problems (continued)
  - has a profound effect; when you absorb this fact, the possibility that Alvac and Aidsvax together might make things WORSE (e.g., by jointly weakening the immune system when the effect of this weakening singly is too small to see) also has to be regarded as plausible, in which case the right null hypothesis on Jim's reasoning should be H_0: $\Delta \leq 0$
  - With this null hypothesis the posterior probability of H_1: $\Delta > 0$ given the data, using a naive diffuse prior, becomes 0.98 and the scientific conclusion is completely different
  - So under Jim's approach, people have to get into a big argument about whether H_0: $\Delta = 0$ or H_0: $\Delta \leq 0$ is the "right" (more plausible?) null hypothesis to test; the people who go with H_0: $\Delta = 0$ conclude "insufficient evidence to say that the treatment is effective" and the people who go with H_0: $\Delta \leq 0$ conclude that the treatment is highly likely to be effective by an amount that's large in practical terms
  - This is not a satisfactory state of affairs

Bayesian Hypothesis Testing (continued)

- How Bayesian hypothesis-testing can fix some of these problems (continued)
  - The fact that we're prepared to write down a sampling model for the data values $\bar{y}_C$ and $\bar{y}_T$ of the form $(\bar{y}_C - \bar{y}_T | \Delta, \sigma^2) \sim N(\Delta, \sigma^2)$ means that our uncertainty about Δ is inherently continuous, so we know scientifically before the experiment is performed that Δ is not EXACTLY 0
  - For me the relevant hypotheses are therefore (H_0: $\Delta \leq c$ versus H_1: $\Delta > c$), where c is a practical significance threshold specified by the subject-matter experts (e.g., if the non-intervention HIV rate is 9 per 1,000, then there's no point in further investigating any vaccine that doesn't lower this rate by at least (e.g.) 10%, so $c = 0.9$ per 1,000)
  - (2) My preferred solution to the discrepancy between {the frequentist analysis of this data set} and {the views of the HIV experts (other than those who ran the study)} is to compare the posterior plausibility of H_0: $\Delta \leq c$ and H_1: $\Delta > c$ in an analysis in which the views of the experts are brought in via a non-diffuse prior on Δ that's quite skeptical about

Bayesian Hypothesis Testing (continued)

- How Bayesian hypothesis-testing can fix some of these problems (continued)
  - the possibility that Δ is large and positive (Spiegelhalter has shown how to do this well, using sensitivity analysis)
  - I'm not claiming that there is NO role for Bayesian testing of sharp-null hypotheses; Jim and I just differ substantially on how often we think it's appropriate to conduct such a test
  - Lecture 2, pages 18 and 19:
    - "Let θ denote the [population] difference in mean treatment effects for cancer treatments A and B";
    - Scenario 1: Treatment A = standard chemotherapy versus Treatment B: standard chemotherapy + steroids
    - Berger: $\theta = 0$ is plausible, so use H_0: $\theta = 0$; Draper: the basic scientific question is "Does the average effect of treatment B differ from the average effect of treatment A by an amount that's large enough to be worth pursuing with further studies?", so use H_0: $\theta < c$ as above
    - Berger: $\{H_0$: Males and females of a species have the same characteristic $A\}$ is plausible, so do a sharp-null test;

Bayesian Hypothesis Testing

- How Bayesian hypothesis-testing can fix some of these problems (continued)
  - Draper: the rates at which characteristic A arises in males and females of this species differ by an amount that's not large enough to be biologically relevant, so use H_0: $|\theta| < c$, where c is specified scientifically and will often be large relative to $\frac{c_1}{\sqrt{n}}$ (so that Jim's result on page 22 doesn't apply)
  - Berger: $\{H_0$: There is no psychokinetic effect$\}$ is plausible, so do a sharp-null test; Draper: there either is or is not a biological mechanism in the human brain that makes psychokinesis possible, so do a sharp-null test
- Other topics: "Hypothesis testing is drastically overused" (amen to that)
- Referring to particular Bayesian choices (e.g., prior distributions) as "objective": more on this later
- "Approximating a believable precise hypothesis by an exact precise null hypothesis": more on this later, when we talk about Lindley's "Paradox"

Additional Topics

- The Bayesian approach to multiple testing (more on this later)
- "Summary 1: There is a lack of recognition that better statistics is the solution to much of the reproducibility problem" (YES)
- "Summary 2: How Bayesian analysis can help"
- "There is no optional stopping issue; formal Bayesian answers do not depend on the stopping rule"

  This is only partially true: if the underlying problem were inferential, it's true that Bayesians can look at data arriving sequentially as often as they want with no "penalty"; the posterior from the last batch of data just becomes the prior for the next batch

  However, the underlying problem often actually has a decision-making (not inferential) character (action 1: take the drug to phase III; action 2: don't), and in such cases there are very real downsides from both false positives and false negatives at each interim analysis; when this is correctly quantified in the utility function, in effect Bayesians have to worry about "spending α" just like frequentists

Additional Topics (continued)

- Conditional frequentist inference as a bridge between frequency and Bayes (yes, but there are such benefits for frequentists from partial conditioning that they have to ask themselves: why not condition on the entire data set and be Bayesian, as long as we still pay attention to how often we get the right answer? this should arguably move ALL of us toward WELL-CALIBRATED (not "objective") Bayesian analyses; more on this later)
- End of comments on Lecture 1 and pages 1–22 of Lecture 2