TO BAYES OR NOT TO BAYES?

Introductory Overview Lecture

Hal Stern
Department of Statistics
Iowa State University
hstern@iastate.edu

Hamlet, Act 3, Scene 1 - Shakespeare

http://server1.hypermart.net/hamlet/course.html (Ian Delaney)

To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and by a sleep to say we end
The heart-ache and **the thousand natural shocks**
That flesh is heir to, 'tis a consummation
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
For in that sleep of death **what dreams may come**
When we have shuffled off this mortal coil,
Must give us pause: there's the respect
That makes calamity of so long life;
For who would bear the whips and scorns of time,
The oppressor's wrong, the proud man's contumely,
The pangs of despised love, the law's delay,
The insolence of office and the spurns
That patient merit of the unworthy takes,
When he himself might his quietus make
With a bare bodkin? who would fardels bear,
To grunt and sweat under a weary life,
But that the dread of something after death,
The undiscover'd country from whose bourn
No traveller returns, puzzles the will
And makes us rather bear those ills we have
Than fly to others that we know not of?
Thus conscience does make cowards of us all;
And thus the native hue of resolution
Is sicklied o'er with the pale cast of thought,
And enterprises of great pith and moment
With this regard their currents turn awry,
And lose the name of action.–Soft you now!
The fair Ophelia! Nymph, in thy orisons
Be all my sins remember'd.

Another quote

"If they would all publish posthumously, as he did, we would all be better off"

- quote of unknown origin in reference to followers of Rev. T. Bayes

Next up: three introductory examples

- mapping of kidney cancer rates
- baseball batting average prediction
- SAT coaching study

Kidney cancer mortality rates
Manton et al. (JASA, 1989)

- Analyses of age-standardized death rates for cancer of kidney/ureter by U.S. county

- Two maps of estimated rates
  - Direct calculation: use observed rates in county/age-group cells to form estimates
  - Empirical Bayes: modeling to stabilize estimated rates

Kidney cancer mortality rates
Manton et al. (JASA, 1989)

- Maps here

Baseball batting averages
Efron & Morris (Sci. Amer., 1975)

Player	Data (success rate in first 45 attempts of 1970)	Inference (James-Stein) estimate)	Outcome (success rate in rest of 1970 attempts)
Clemente	.400	.290	.346
Robinson	.378	.286	.298
Howard	.356	.281	.276
Johnstone	.333	.277	.222
Berry	.311	.273	.273
Spencer	.311	.273	.270
Kessinger	.289	.268	.263
Alvarado	.267	.264	.210
Santo	.244	.259	.269
Swoboda	.244	.259	.230
Unser	.222	.254	.264
Williams	.222	.254	.256
Scott	.222	.254	.303
Petrocelli	.222	.254	.264
Rodriguez	.222	.254	.226
Campaneris	.200	.249	.285
Munson	.178	.244	.316
Alvis	.156	.239	.200

SAT coaching study
Rubin (J. Educ. Stat., 1981)

- Randomized experiments in 8 schools
- Separate analyses
- Outcome is SAT-Verbal score
- Effect of treatment (coaching) estimated using analysis of covariance

School	Estimated treatment effect	Standard error of effect estimate	Treatment effect
A	28	15	?
B	8	10	?
C	− 3	16	?
D	7	11	?
E	− 1	9	?
F	1	11	?
G	18	10	?
H	12	18	?

Bayesian inference: Two key ideas

- Explicit use of probability for quantifying uncertainty
  - probability models for data given parameters
  - probability distributions for parameters
- Inference for unknowns conditional on observed data
  - inverse probability
  - Bayes' theorem (hence the modern name)
  - formal decision-making

The Bayesian approach to inference

- A full probability model
  - likelihood $p(y|\theta) = p(\text{data} \mid \text{parameters})$
  - prior distribution $p(\theta)$
- Posterior inference
  - Bayes' thm to derive posterior distribution

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

  - probability statements about unknowns
  - formal decision-making
- Model checking/sensitivity analysis
  - does the model fit
  - are conclusions sensitive to choice of prior distn/likelihood

A simple (too simple?) example

- Normal distribution (known variance)
  - $Y_1, Y_2, \ldots, Y_n$ are n independent identically distributed $N(\mu, \sigma^2)$ random variables where σ^2 is known
  - Goal: inference for μ
  - Classical approach (frequentist)
    - maximum likelihood estimate is $\widehat{\mu} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
    - 95% confidence interval: $\bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$
    - test $H_o : \mu = 0$ with $Z = \sqrt{n}\bar{Y}/\sigma$
  - Interested in repeated sampling (with fixed μ) properties of procedures
    - $\bar{Y}$ is unbiased, minimum variance
    - CI contains true value 95% of the time
    - p-values used to interpret test results

A simple example (cont'd)

- Normal distribution (known variance)
  - Conditional on μ we have $y = (y_1, y_2, \ldots, y_n)$ are n independent identically distributed $N(\mu, \sigma^2)$ random variables where σ^2 is known
  - Goal: inference for μ
  - Bayesian approach
    - prior distribution for μ is $N(\mu_o, \tau^2)$ with μ_o, τ^2 specified by user (how? why?)
    - posterior distribution for μ :

$$\mu|y \ \sim \ N(\widehat{\mu}, V_\mu)$$

where

$$\widehat{\mu} = \frac{\bar{y}\left(\frac{n}{\sigma^2}\right) + \mu_o\left(\frac{1}{\tau^2}\right)}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

    - inference from posterior distribution (point estimates, interval estimates, tests)

A simple example (cont'd): Bayesian inference

- Point estimation
  - given loss function $L(\mu, t(y))$ we find optimal Bayes estimate $t(y)$ by minimizing posterior expected loss
  - e.g.: squared error loss, $L(\mu, t(y)) = (\mu - t(y))^2$, leads to $t(y) = E(\mu|y)$ (posterior mean)

- Interval estimation
  - 95% central posterior interval: $\widehat{\mu} \pm 1.96\sqrt{V_\mu}$
  - interpretation – what people want to say
  - alternative: highest posterior density interval

- Hypothesis testing (e.g., $H_o : \mu = 0$ vs $H_a : \mu > 0$)
  - can compute $\Pr(\mu > 0|y)$
  - formal approach is Bayes factors

- A key point: the Bayesian approach is a way of generating procedures
  - can then ask about properties, frequentist or otherwise, of the procedures

A simple example (cont'd): Interpreting the posterior distribution

- Posterior distribution for μ :

$$\mu|y_1, \ldots, y_n \sim N(\widehat{\mu}, V_\mu)$$

where

$$\widehat{\mu} = \frac{\bar{y}\left(\frac{n}{\sigma^2}\right) + \mu_o\left(\frac{1}{\tau^2}\right)}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

- Interpretation
  - prior distn and posterior distn are both normal (conjugate prior family)
  - posterior mean is weighted average of prior mean and sample mean (weights = precisions = 1/variance)
  - posterior precision is sum of prior precision and data precision
  - for any μ_o, τ^2 let n get very large:

$$\widehat{\mu} \approx \bar{y} \quad \text{and} \quad V_\mu \approx \sigma^2/n$$

(the posterior distn looks like the traditional sampling distn)

Bayesian calculation

- Difficulty in calculating posterior distributions made Bayesian analysis impractical (except in simple problems) for a long time
- Recent advances in computational algorithms have essentially eliminated this problem
- Today, brief remarks on calculation as time permits
- For our normal example:

$$
\begin{aligned}
p(\mu|y) &= \frac{p(y|\mu)p(\mu)}{p(y)} \\
&= \frac{\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2} \left(\frac{1}{2\pi\tau^2}\right)^{1/2} e^{-\frac{1}{2\tau^2}(\mu-\mu_o)^2}}{\int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2} \left(\frac{1}{2\pi\tau^2}\right)^{1/2} e^{-\frac{1}{2\tau^2}(\mu-\mu_o)^2} d\mu} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{1}{2\tau^2}(\mu-\mu_o)^2\right)
\end{aligned}
$$

  - can do the integral in denominator or use following shortcut
    - consider only terms involving μ (last line)
    - note last line is quadratic in μ in exponent
    - conclude posterior distribution is normal
    - identify mean and variance

Prior distributions

- Where do prior distributions come from?

- Subjective prior distributions
  - "honest" prior opinion
  - elicit prior distn from experts

- Conjugate prior distributions
  - prior distribution/likelihood pairs such that prior distn and posterior distn are from same family (e.g., normal/normal in the example)
  - mathematically convenient
  - may be a bit limiting (but can use mixtures)

- Noninformative prior distns (more to come)

- Large samples: likelihood dominates the prior distn

Noninformative prior distns: An interesting result for our example

- Posterior distribution for μ :

$$\mu|y_1,\ldots,y_n \sim N(\widehat{\mu}, V_\mu)$$

where

$$\widehat{\mu} = \frac{\bar{y}\left(\frac{n}{\sigma^2}\right) + \mu_o\left(\frac{1}{\tau^2}\right)}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} \quad \text{and} \quad V_\mu = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$$

- For given μ_o, n let τ^2 get very large:

$$\widehat{\mu} \to \bar{y} \quad \text{and} \quad V_\mu \to \sigma^2/n$$

(the posterior distn looks like the traditional sampling distn)

- Normal prior distn with "infinite" τ^2 appears to add no information to the likelihood

- Remarks
  - obtain same posterior mean ($\bar{y}$) and variance (σ^2/n) with $p(\mu) = \text{Unif}(-\infty, -\infty)$ as prior distn (but this is not a proper distn)
  - an improper prior distribution can lead to a proper posterior distribution but won't always
  - the flat (uniform) prior distn is known as a non-informative prior distn for this example

Noninformative prior distributions

- Often attempt to portray prior ignorance
- Sometimes referred to as "objective" Bayes
- Frequently flat prior distributions
  (but on what scale?)
- Resulting distns may be improper
- Difficulty with an improper prior distn:
  need to check that the posterior distn is proper
- Alternative is to use a proper prior distribution
  with large variance
- One difficulty with noninformative prior distns:
  the concept of noninformative is not well defined

Lessons from the simple example

- Bayesian analysis combines information from data and prior distn
- For large samples: Bayesian analysis is approximately the same as the classical analysis
- For flat prior distn: Bayesian analysis is essentially the same as the classical analysis
- Why be Bayesian?
- Let's make things a bit more sophisticated

SAT coaching study: A normal-normal hierarchical model (a.k.a. a random effects model)

- Separate randomized experiments in 8 high schools
- Treatment is local SAT coaching program
- Outcome is SAT-Verbal score (200 to 800)
- Treatment effect estimated using analysis of covariance to adjust for PSAT (preliminary SAT)

$$SAT = \beta_0 + \beta_1 \ PSAT + \beta_2 \ Trt$$

- Separate regression done for each school
- Notation:
  - quantities of interest θ_j: average effects of coaching programs
  - data y_j: separate estimated treatment effects ($\widehat{\beta}_2$)
  - standard errors σ_j

SAT coaching study: Data and model

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j	Average treatment effect, θ_j
A	28	15	?
B	8	10	?
C	− 3	16	?
D	7	11	?
E	− 1	9	?
F	1	11	?
G	18	10	?
H	12	18	?

- Model overview
  - data model / likelihood:
    $$y_j|\theta_j \overset{\text{ind}}{\sim} \text{N}(\theta_j, \sigma_j^2), \text{ for } j = 1, \ldots, 8$$
    with σ_j^2's assumed known
  - prior distn: $\theta_j|\mu, \tau^2 \overset{\text{iid}}{\sim} \text{N}(\mu, \tau^2)$ for $j = 1, \ldots, 8$
  - hyperprior distribution: $p(\mu, \tau^2) \propto 1/\tau$

SAT coaching study: Model

- Data model / likelihood:

  $$y_j|\theta_j \overset{\text{ind}}{\sim} \text{N}(\theta_j, \sigma_j^2), \text{ for } j = 1, \ldots, 8$$
  with σ_j^2's assumed known

  – normality and known variance justified by large sample size in each school

- Prior distribution: $\theta_j|\mu, \tau^2 \overset{\text{iid}}{\sim} \text{N}(\mu, \tau^2)$ for $j = 1, \ldots, 8$

  – exchangeable prior distn for θ_j's

  – traditional random effects model

  * note $\tau \to 0$ reduces to complete pooling
  * note $\tau \to \infty$ reduces to separate estimates

  – frequentists also use this model

  * don't call this a prior distn
  * don't usually consider θ_j's as being of interest

- Hyperprior distribution: $p(\mu, \tau^2) \propto 1/\tau$

  – an improper, "noninformative" prior distn

  – equivalent to $p(\mu, \tau) \propto 1$

SAT coaching study: Preliminary data analysis

- Consider the 8 programs separately
  - two programs appear to work (18-28 points)
  - four programs appear to have a small effect
  - two programs appear to have negative effects
  - large standard errors imply overlapping CIs

- A pooled estimate
  - classical test fails to reject hypothesis that all θ_j's are equal
  - pooled estimate $= \sum_j (y_j/\sigma_j^2) / \sum_j (1/\sigma_j^2) = 7.9$ (standard error is 4.2)
  - pooled estimate applies to each school

- Neither separate nor pooled estimates seem right

- We will see that the hierarchical model provides a compromise between the two estimates

SAT coaching study: Bayesian computation

- Joint posterior distribution:

$$
\begin{aligned}
& p(\theta, \mu, \tau | y) \\
& \quad \propto p(y|\theta) p(\theta|\mu, \tau) p(\mu, \tau) \\
& \quad \propto \prod_{j=1}^{8} N(y_j|\theta_j, \sigma_j^2) \prod_{j=1}^{8} N(\theta_j|\mu, \tau^2) \\
& \quad \propto \tau^{-8} \exp\left[-\frac{1}{2}\sum_j \frac{1}{\tau^2}(\theta_j - \mu)^2\right] \exp\left[-\frac{1}{2}\sum_j \frac{1}{\sigma_j^2}(y_j - \theta_j)^2\right]
\end{aligned}
$$

- Factors that depend only on y and $\{\sigma_j\}$ are treated as constants because they are known

- Joint posterior distn has 10 parameters ... don't recognize it as any known 10-dim distn

SAT coaching study: Bayesian computation (cont'd)

- Computational approaches include:
  - approximation
    - normal approximation to posterior distn
    - "empirical" Bayes methods: estimate μ and τ^2 and then proceed as in simple example
  - simulation: approximate the posterior distn with a random sample from the distn
    - Gibbs sampling (a Markov chain Monte Carlo (MCMC) approach)
      - use sequence of full conditional posterior distns (μ given others, τ^2 given others, etc.)
      - in this case each full conditional distn is easy to recognize
      - software available (BUGS Project at www.mrc-bsu.cam.ac.uk/bugs/welcome.html)
      - more in next hour's tutorial
    - hierarchical computation provides an alternative here (see next slide)

SAT coaching study: Bayesian computation (cont'd)

- Hierarchical computation - note that we can write
$$p(\mu, \theta, \tau^2|y) = p(\tau^2|y)p(\mu|\tau^2, y)p(\theta|\mu, \tau^2, y)$$
where
  - first term is messy but one-dimensional
  - second term turns out to be normal
  - third term is just like our simple example (normal observation y_j with unknown mean θ_j and normal prior distn for θ_j)

- To simulate from joint posterior distribution $p(\theta, \mu, \tau|y)$:
  1. draw τ from $p(\tau|y)$ (grid approximation)
  2. draw μ from $p(\mu|\tau, y)$ (normal distribution)
  3. draw $\theta = (\theta_1, \ldots, \theta_8)$ from $p(\theta|\mu, \tau, y)$ (independent normal distribution for each θ_j)

- Repeat 1000 times to obtain 1000 simulations

Simulation draw	Posterior simulations of model parameters				
	τ	μ	θ_1	...	θ_8
1	x	x	x	x	x
2	x	x	x	x	x
⋮	x	x	x	x	x
1000	x	x	x	x	x

SAT coaching study: Sample program

- S-plus code listing here

SAT coaching study: Results

- graph of $p(\tau|y)$

- histogram of posterior draws of μ

SAT coaching study: Results

- histogram of posterior draws of $\theta's$ (A, C, E, max)

SAT coaching study: Results

School	Posterior quantiles 2.5%	25%	50%	75%	97.5%	Estimates pooled	separate
A	− 2	6	10	16	32	8	28
B	− 5	4	8	12	20	8	8
C	−12	3	7	11	22	8	− 3
D	− 6	4	8	12	21	8	7
E	−10	2	6	10	17	8	− 1
F	− 9	2	6	10	19	8	1
G	− 1	6	10	15	27	8	18
H	− 7	4	8	13	23	8	12
μ	− 2	5	8	11	18		
τ	0.3	2.3	5.1	8.8	21.0		

Discussion:

compare Bayes results to those from complete pooling and separate analyses

SAT coaching study: Results

We can easily address more complicated inferential questions:

Pr(school A's effect is the max $\mid y$) = 0.25

Pr(school B's effect is the max $\mid y$) = 0.10

Pr(school C's effect is the max $\mid y$) = 0.10

Pr(school A's effect is the min $\mid y$) = 0.07

Pr(school B's effect is the min $\mid y$) = 0.09

Pr(school C's effect is the min $\mid y$) = 0.17

Pr(sch. A's effect $>$ sch. C's effect $\mid y$) = 0.67

SAT coaching study: Model checking and sensitivity analysis

- Model checking
  - are results plausible?
  - diagnostics: e.g., is max observed y_j what we'd expect under the model?

- Sensitivity analysis
  - often done by reanalysis with different prior distn or likelihood
  - e.g., t-distn in place of normal distn for θ_j's
  - e.g., different prior distn on τ
  - in this case we can easily study sensitivity to the prior distn on τ (because μ and θ_j's have normal distn given τ and y)

SAT coaching study: Sensitivity analysis

- Results for θ conditional on τ, y here

SAT coaching example: Summary

- Classical estimates (no pooling, complete pooling) provide starting point for analysis
- Data-determined degree of pooling across studies
- Inference about the individual schools (θ_j's)
- Inference about the population of schools (μ, τ^2)
- Importance of model checking/sensitivity analysis

Other common hierarchical models

- Poisson-Gamma model
  - data are modeled as Poisson counts (e.g., disease counts in different counties)
  - Poisson mean (or rate) parameters vary across observations
  - parameters modeled as a sample from a gamma population distn

- Binomial-Beta model
  - data are modeled as binomial random variables (e.g., hits in first 45 at-bats)
  - individual probabilities of success vary
  - success parameters modeled as a sample from a beta population distn

To Bayes or not to Bayes

- To Bayes:
  - use of probability to describe uncertainty
  - there is often a natural prescription for accommodating modifications/complications (e.g., missing data, handling outliers via t-distn)
  - flexible inference (e.g., distn on ranks of schools)
  - good frequentist properties

- Not to Bayes:
  - prior distributions are an additional layer of formalism
  - interpretation of "non-frequentist" probabilities
  - communication with subject area scientists

Additional reading

- Texts:
  - Bayesian Data Analysis - A. Gelman et al.
  - Bayes and Empirical Bayes Methods for Data Analysis - Carlin and Louis

    (both at Chapman and Hall/CRC Press)

- Articles:
  - Efron (1986, American Statistician) "Why isn't everyone a Bayesian?" (with discussion)
  - Lindley (1990, Statistical Science) "The present position of Bayesian statistics" (with discussion)
  - Stern (1998, Stats) "A primer on the Bayesian approach to statistical inference"

- Contact me: hstern@iastate.edu