STA 360/601: Bayesian and Modern Statistics

Bayesian hypothesis testing & Bayes factors

Jeff Miller

Department of Statistical Science, Duke University

Bayesian hypothesis testing

- Problem: You have two or more competing hypotheses $H_0, H_1, \ldots,$ and want to consider the evidence in favor of each, based on some data.
- Examples:
  1. Does drug X reduce the risk of stroke (H_1) or not (H_0)?
  2. Does Patient X have disease Y (H_1) or not (H_0)?
  3. Does the Higgs boson exist (H_1) or not (H_0)?
  4. You are Gregor Mendel. Which of several models of trait inheritance $H_0, H_1, \ldots, H_m$ is correct?
  5. Data on 5000 subjects was collected over 60 years. Which variables are predictive of heart disease risk? (Each subset of variables is a competing hypothesis.)

A simple example

- Data: $X_1, \ldots, X_n \overset{iid}{\sim} \text{N}(\mu, \sigma^2)$, where σ is known.
- Hypotheses: $\text{H}_0 : \mu = 0$ versus $\text{H}_1 : \mu \neq 0$
- Same setup as a classical frequentist hypothesis test.
- Let's say the data is

  $$x = (x_1, \ldots, x_8) = (0.8, -0.4, 0.1, 0.0, 1.2, 0.8, 1.0, 0.9).$$

  What is your intuitive judgment of the plausibility of H_0 and H_1?
- What would be a natural Bayesian approach? Any ideas?

A Bayesian approach

- Put a prior on the hypotheses, say, $p(\mathrm{H}_0) = \pi$ and $p(\mathrm{H}_1) = 1 - \pi$.
- Under $\mathrm{H}_0 : \mu = 0$, the data is simply $\mathrm{N}(0, \sigma^2)$.
- Under $\mathrm{H}_1 : \mu \neq 0$, we don't know μ, so let's put a prior on it: $\mu \sim \mathrm{N}(0, \sigma_1^2)$. (Technically, perhaps we should exclude the point $\mu = 0$ from the prior, but it makes no difference since this has probability zero anyways.)
- Now, we want to know the posterior probabilities $p(\mathrm{H}_0|x)$ and $p(\mathrm{H}_1|x)$ where $x = (x_1, \ldots, x_n)$.
- By Bayes' rule, $p(\mathrm{H}_k|x) \propto p(x|\mathrm{H}_k)p(\mathrm{H}_k)$. So, we need $p(x|\mathrm{H}_0)$ and $p(x|\mathrm{H}_1)$ (the *marginal likelihoods*).

Computing the marginal likelihoods

- H_0 is easy: $p(x|\mathrm{H}_0) = \prod_{i=1}^{n} \mathrm{N}(x_i \mid 0, \sigma^2)$
- ...and H_1 is not too hard:

$$
\begin{aligned}
p(x|\mathrm{H}_1) &= \int p(x|\mu, \mathrm{H}_1) p(\mu|\mathrm{H}_1) d\mu \\
&= \int \Big(\prod_{i=1}^{n} \mathrm{N}(x_i \mid \mu, \sigma^2) \Big) \mathrm{N}(\mu \mid 0, \sigma_1^2) d\mu \\
&= \text{(typical Gaussian integral... complete the square, etc.)} \\
&= \frac{s}{\sigma_1} \exp\left(\tfrac{1}{2} m^2 / s^2\right) \prod_{i=1}^{n} \mathrm{N}(x_i \mid 0, \sigma^2),
\end{aligned}
$$

where $1/s^2 = n/\sigma^2 + 1/\sigma_1^2$ and $m = (s^2/\sigma^2) \sum_i x_i$.

Outcome for our simple example

- Our data is

$$x = (x_1, \ldots, x_8) = (0.8, -0.4, 0.1, 0.0, 1.2, 0.8, 1.0, 0.9).$$

- Let's suppose $p(\mathrm{H}_0) = p(\mathrm{H}_1) = 1/2$, $\sigma = 1$, and $\sigma_1 = 1$.
- Plugging the marginal likelihood and prior into $p(\mathrm{H}_k|x) \propto p(x|\mathrm{H}_k)p(\mathrm{H}_k)$ we get

$$p(\mathrm{H}_0|x) = 0.506 \text{ and } p(\mathrm{H}_1|x) = 0.494.$$

- So, basically, we have no idea.

Decisions, decisions, . . .

- Suppose we have to choose one of the hypotheses.
- Suppose that when we choose d and the truth is h, we incur a loss $L(h, d)$.
- Since we have put a prior on h, we may as well consider it as a random variable, H.
- The *posterior expected loss* associated with choosing d given data x is

$$\mathrm{E}\big(L(H, d) \mid x\big) = \sum_{h} L(h, d) p(H = h \mid x)$$

where the sum is over all hypotheses $h = \mathrm{H}_0, \mathrm{H}_1, \ldots$.

Example: 0 - 1 loss

- *0 - 1 loss* is the loss function $L(h, d) = \mathbb{1}(h \neq d)$, i.e., you lose 1 if wrong, 0 if right.
- The posterior expected loss in this case is

$$\begin{aligned} \mathrm{E}\big(L(H, d) \mid x\big) &= \sum_h L(h, d) p(H = h \mid x) \\ &= \sum_h \mathbb{1}(h \neq d) p(H = h \mid x) \\ &= 1 - p(H = d \mid x). \end{aligned}$$

- So, to minimize our posterior expected loss, the optimal decision d^* (under 0 - 1 loss) is the hypothesis with highest posterior probability $p(H = d|x)$.
- In the case of two hypotheses, H_0 and H_1,

$$d^* = \begin{cases} \mathrm{H}_0 & \text{if } p(\mathrm{H}_0|x) > 1/2 \\ \mathrm{H}_1 & \text{if } p(\mathrm{H}_1|x) > 1/2 \\ \text{either} & \text{otherwise.} \end{cases}$$

A few remarks

- If $L(h, d)$ is not 0 - 1 loss, the optimal decision will not necessarily be the hypothesis with highest posterior probability.
- The Bayesian hypothesis testing approach described above is very different than frequentist hypothesis testing.
- For frequentist hypothesis testing of H_0 versus H_1:
  - The usual approach is to minimize Type II errors (choosing H_0 when H_1 is true) subject to an upper bound on the probability of Type I error (choosing H_1 when H_0 is true).
  - There is an asymmetry in the frequentist approach: H_0 is a *null hypothesis*, i.e., a default position (the reigning champion), and H_1 is an *alternative hypothesis* (the challenger).
  - Metaphor: It is like a criminal trial, in which the defendant is presumed innocent (H_0) unless proven guilty beyond all reasonable doubt (H_1).
- The Bayesian approach does not have this asymmetry, allowing for a more balanced approach to minimize overall loss. However, as always, the outcome depends on the prior.

Bayes factors

- Bayes factors provide a way to be a little less dependent on the prior.
- The *Bayes factor* in favor of H_1 over H_0, for data $x = (x_1, \ldots, x_n)$, is

$$B_{10} = \frac{p(x|\mathrm{H}_1)}{p(x|\mathrm{H}_0)}.$$

- Note that this doesn't depend on $p(\mathrm{H}_0)$ or $p(\mathrm{H}_1)$ ...
- ...but it does still depend on the priors we choose for parameters required to define the distribution of x given H_0 or H_1 (e.g., μ in our simple example).
- When $B_{10} > 1$, this is evidence in favor of H_1, when $B_{10} < 1$, it is evidence in favor of H_0.
- Some have suggested scales for interpreting Bayes factors, e.g., 10 – 30 is "strong evidence", but this is purely heuristic and not universally accepted.

Some properties of Bayes factors

- In the case of two competing hypotheses, the Bayes factor is related to the posterior probability as follows:

$$p(\mathrm{H}_0|x) = \frac{p(x|\mathrm{H}_0)p(\mathrm{H}_0)}{p(x|\mathrm{H}_0)p(\mathrm{H}_0) + p(x|\mathrm{H}_1)p(\mathrm{H}_1)}$$
$$= \frac{1}{1 + \frac{p(x|\mathrm{H}_1)p(\mathrm{H}_1)}{p(x|\mathrm{H}_0)p(\mathrm{H}_0)}}$$
$$= \frac{1}{1 + \text{Bayes factor} \times \text{Prior odds}}$$

- Also, "Posterior odds = Bayes factor $\times$ Prior odds", i.e.,

$$\frac{p(\mathrm{H}_1|x)}{p(\mathrm{H}_0|x)} = B_{10}\frac{p(\mathrm{H}_1)}{p(\mathrm{H}_0)}.$$

Back to our example

- Data: $x = (0.8, -0.4, 0.1, 0.0, 1.2, 0.8, 1.0, 0.9)$.
- $p(\mathrm{H}_0) = p(\mathrm{H}_1) = 1/2$, $\sigma = 1$, and $\sigma_1 = 1$.
- Posterior probabilities:

$$p(\mathrm{H}_0|x) = 0.506 \text{ and } p(\mathrm{H}_1|x) = 0.494.$$

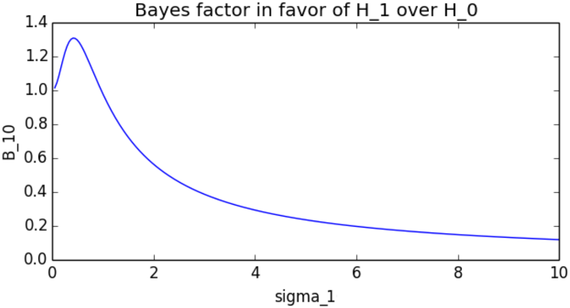
- Bayes factors:

$$B_{10} = \frac{p(x|\mathrm{H}_1)}{p(x|\mathrm{H}_0)} = 0.98$$

$$B_{01} = \frac{p(x|\mathrm{H}_0)}{p(x|\mathrm{H}_1)} = 1.02$$

Sensitivity to the prior

- Bayes factors can depend strongly on the prior on parameters (e.g., μ in our example).
- In our example, the prior standard deviation σ_1 of μ given H_1 has a significant effect on the Bayes factor:



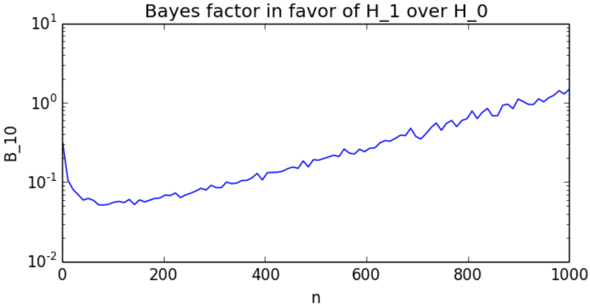
- In particular, $B_{10} \to 0$ as $\sigma_1 \to \infty$.
- Improper priors CANNOT be used here.

Lindley's "paradox"

- This sensitivity is the issue underlying Lindley's "paradox" (which is, as usual, not actually a paradox).
- The original "paradox" is that it is possible for very reasonable frequentist and Bayesian approaches to give contradictory answers about which hypothesis is favored by the evidence.
- e.g., frequentist rejects H_0 while Bayesian finds strong evidence for H_0.
- This underlying issue also shows up in Bayesian models over variable-dimension parameter spaces, e.g., mixture models.

Non-monotonicity wrt sample size

- Another thing to be careful of is that Bayes factors can be non-monotone in the sample size n.
- Example: Same as before, but with $\sigma_1 = 5$ and $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{N}(0.1, 1)$. Plot is averaged over many samples:



- H_1 is true, but if we only had 100 samples, we would only see B_{10} decreasing down to ≈ 0.05, seeming to suggest that it is converging to 0, and we might mistakenly be convinced of H_0.

Remarks

- The Bayesian approach allows for principled (but subjective) decision-theoretic hypothesis testing.
- Also, the Bayesian approach extends naturally to more complicated models.
- The prior really matters here — only trust the results to the extent that you trust the prior.
- It's a good idea to do a sensitivity analysis: vary the prior and see how the result changes.
- Careful: Bayes factors can be non-monotone in n.

Homework exercise

- You have data from an experiment collecting cell counts for a control group and treatment group.
- Control group:

$$x_{1:n} = (204, 215, 182, 225, 207, 188, 205, 227, 190, 211, 196, 203)$$

- Treatment group:

$$y_{1:m} = (211, 233, 244, 241, 195, 252, 238, 249, 220, 213)$$

- The counts are assumed to be Poisson distributed.
- There are two hypotheses, H_0: Poisson with same mean, vs. H_1: Poisson with different means.

Homework exercise (continued)

- Model this as follows.
- $p(\mathrm{H}_0) = 3/4$, $p(\mathrm{H}_1) = 1/4$.
- Under H_0: $X_1, \ldots, X_n, Y_1, \ldots, Y_m \sim \text{Poisson}(\lambda)$ i.i.d. given λ, and $\lambda \sim \text{Gamma}(a, b)$ where $a = 4 =$ shape and $b = 0.02 =$ rate (i.e., λ has pdf $b^a \lambda^{a-1} \exp(-b\lambda)/\Gamma(a)$).
- Under H_1: $X_1, \ldots, X_n \sim \text{Poisson}(\lambda_c)$ i.i.d. given λ_c, and $Y_1, \ldots, Y_m \sim \text{Poisson}(\lambda_t)$ i.i.d. given λ_t, and $\lambda_c, \lambda_t \sim \text{Gamma}(a, b)$ independently, with the same a, b as above.
- Compute $p(\mathrm{H}_k | x, y)$ for $k = 0, 1$. Compute B_{10}.
- Compute the prior odds and posterior odds. Interpret your results.
- Does the prior on the λ's appear to be reasonable (judging by the data)? Why or why not? Try different values of a and b and interpret what you see.

Further reading

- Kass & Raftery, *Bayes factors*, JASA, 1995.