Bayesian hypothesis testing (cont.)

Dr. Jarad Niemi

STAT 544 - Iowa State University

March 7, 2019

Outline

- Review of formal Bayesian hypothesis testing
- Likelihood ratio tests
- Jeffrey-Lindley paradox
- p-value interpretation

Bayes tests = evaluate predictive models

Consider a standard hypothesis test scenario:

$$H_0 : \theta = \theta_0, \qquad H_1 : \theta \neq \theta_0$$

A Bayesian measure of the support for the null hypothesis is the Bayes Factor:

$$BF(H_0 : H_1) = \frac{p(y|H_0)}{p(y|H_1)} = \frac{p(y|\theta_0)}{\int p(y|\theta)p(\theta|H_1)d\theta}$$

where $p(\theta|H_1)$ is the prior distribution for θ under the alternative hypothesis. Thus the Bayes Factor measures the predictive ability of the two Bayesian models. Both models say $p(y|\theta)$ are the data model if we know θ, but

1. Model 0 says $\theta = \theta_0$ and thus $p(y|\theta_0)$ is our predictive distribution for y under model H_0 while
2. Model 1 says $p(\theta|H_1)$ is our uncertainty about θ and thus

$$p(y|H_1) = \int p(y|\theta)p(\theta|H_1)d\theta$$

is our predictive distribution for y under model H_1.

Normal example

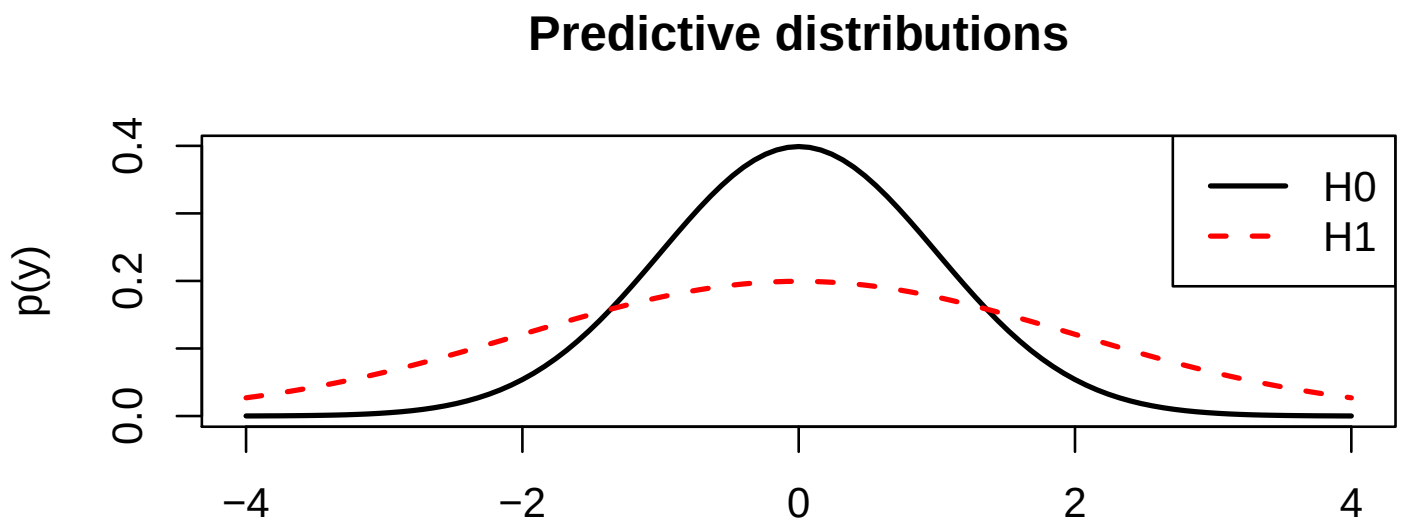
Consider $y \sim N(\theta, 1)$ and

$$H_0 : \theta = 0, \qquad H_1 : \theta \neq 0$$

and we assume $\theta|H_1 \sim N(0, C)$. Thus,

$$BF(H_0 : H_1) = \frac{p(y|H_0)}{p(y|H_1)} = \frac{p(y|\theta_0)}{\int p(y|\theta)p(\theta|H_1)d\theta} = \frac{N(y; 0, 1)}{N(y; 0, 1 + C)}.$$

Now, as $C \to \infty$, our predictions about y become less sharp.



Likelihood Ratio Tests

Consider a likelihood $L(\theta) = p(y|\theta)$, then the likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$ with $\Theta = \Theta_0 \cup \Theta_0^c$ is

$$\lambda(y) = \frac{\sup_{\Theta_0} L(\theta)}{\sup_{\Theta} L(\theta)} = \frac{L(\hat{\theta}_{0,MLE})}{L(\hat{\theta}_{MLE})}$$

where $\hat{\theta}_{MLE}$ and $\hat{\theta}_{0,MLE}$ are the (restricted) MLEs. The likelihood ratio test (LRT) is any test that has a rejection region of the form $\{y : \lambda(y) \leq c\}$. (Casella & Berger Def 8.2.1)

Under certain conditions (see Casella & Berger 10.3.3), as $n \to \infty$

$$-2 \log \lambda(y) \to \chi^2_\nu$$

where ν us the difference between the number of free parameters specified by $\theta \in \theta_0$ and the number of free parameters specified by $\theta \in \Theta$.

Binomial example

Consider a coin flipping experiment so that $Y_i \overset{iid}{\sim} Ber(\theta)$ and the null hypothesis $H_0 : \theta = 0.5$ versus the alternative $H_1 : \theta \neq 0.5$. Then

$$\lambda(y) = \frac{\sup_{\Theta_0} L(\theta)}{\sup_{\Theta} L(\theta)} = \frac{0.5^n}{\hat{\theta}_{MLE}^{n\bar{y}}(1-\hat{\theta}_{MLE})^{n-n\bar{y}}} = \frac{0.5^n}{\bar{y}^{n\bar{y}}(1-\bar{y})^{n-n\bar{y}}}$$

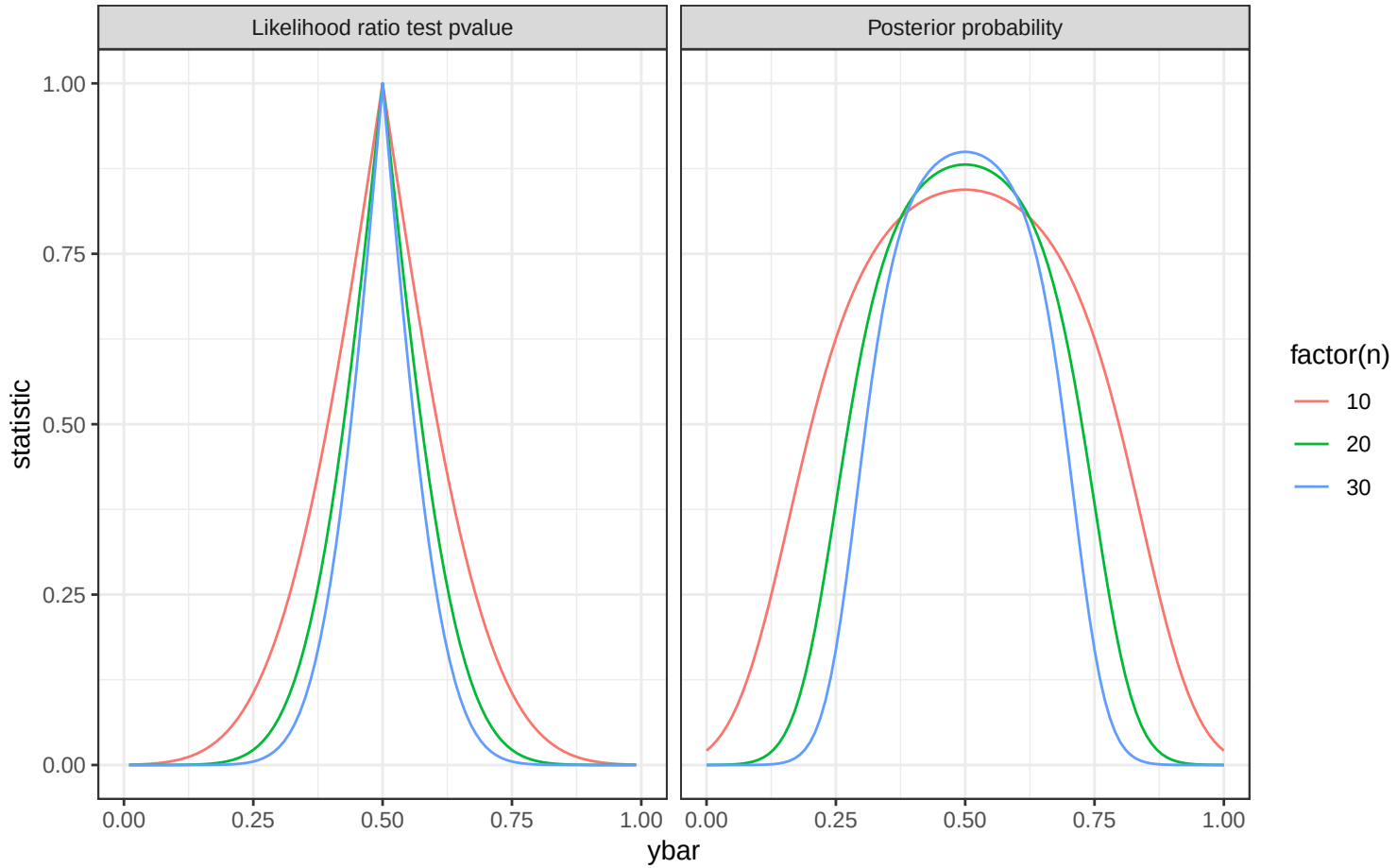
and $-2\log\lambda(y) \to \chi_1^2$ as $n \to \infty$ so

$$p\text{-value} \approx P(\chi_1^2 > -2\log\lambda(y)).$$

If p-value$< a$, then we reject H_0 at significance level a. Typically $a = 0.05$.

Binomial example

$Y \sim Bin(n, \theta)$ and, for the Bayesian analysis, $\theta|H_1 \sim Be(1, 1)$ and $p(H_0) = p(H_1) = 0.5$:



Do p-values and posterior probabilities agree?

Suppose $n = 10,000$ and $y = 4,900$, then the p-value is

$$p\text{-value} \approx P(\chi^2_1 > -2\log(0.135)) = 0.045$$

so we would reject H_0 at the 0.05 level.

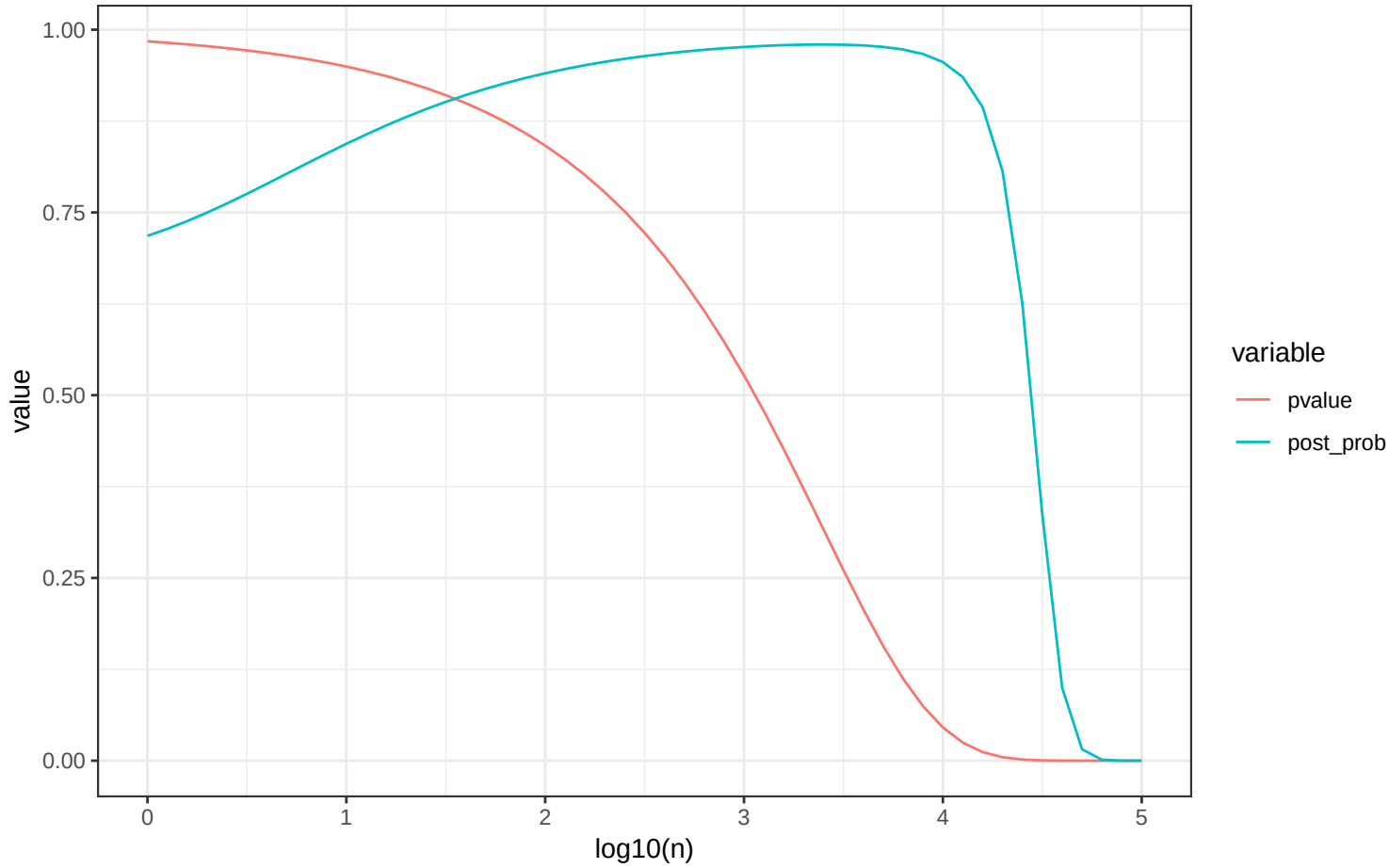
The posterior probability of H_0 is

$$p(H_0|y) \approx \frac{1}{1 + 1/10.8} = 0.96,$$

so the probability of H_0 being true is 96%.

It appears the Bayesian and LRT p-value completely disagree!

Binomial $\overline{y} = 0.49$ with $n \to \infty$



Jeffrey-Lindley Paradox

Definition

The Jeffrey-Lindley Paradox concerns a situation when comparing two hypotheses H_0 and H_1 given data y and find

- a frequentist test result is significant leading to rejection of H_0, but
- our posterior belief in H_0 being true is high.

This can happen when

- the effect size is small,
- n is large,
- H_0 is relatively precise,
- H_1 is relatively diffuse, and
- the prior model odds is ≈ 1.

Comparison

The test statistic with point null hypotheses:

$$\lambda(y) = \frac{p(y|\theta_0)}{p\left(y|\hat{\theta}_{MLE}\right)}$$

$$BF(H_0 : H_1) = \frac{p(y|\theta_0)}{\int p(y|\theta)p(\theta|H_1)d\theta} = \frac{p(y|H_0)}{p(y|H_1)}$$

A few comments:

- The LRT chooses the best possible alternative value.
- The Bayesian test penalizes for vagueness in the prior.
- The LRT can be interpreted as a Bayesian point mass prior exactly at the MLE.
- Generally, p-values provide a measure of lack-of-fit of the data to the null model.
- Bayesian tests compare predictive performance of two Bayesian models (model+prior).

Normal mean testing

Let $y \sim N(\theta, 1)$ and we are testing

$$H_0 : \theta = 0 \qquad \text{vs} \qquad H_1 : \theta \neq 0$$

We can compute a two-sided p-value via

$$p\text{-value} = 2\Phi(-|y|)$$

where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal.

Typically, we set our Type I error rate at level a, i.e.

$$P(\text{reject } H_0 | H_0 \text{ true}) = a.$$

But, if the p-value is less than a, we should be interested in

$$P(H_0 \text{ true} | \text{reject } H_0).$$

p-value interpretation

Let $y \sim N(\theta, 1)$ and we are testing

$$H_0 : \theta = 0 \qquad \text{vs} \qquad H_1 : \theta \neq 0$$

For the following activity, you need to tell me

1. the observed p-value,
2. the relative frequencies of null and alternative hypotheses, and
3. the distribution for θ under the alternative.

Then this p-value app below will calculate (via simulation) the probability the null hypothesis is true.

```
shiny::runGitHub('jarad/pvalue')
```

p-value app approach

The idea is that a scientist performs a series of experiments. For each experiment,

- whether H_0 or H_1 is true is randomly determined,
- θ is sampled according to which hypothesis is true, and
- the p-value is calculated.

This process is repeated until a p-value of the desired value is achieved, e.g. p-value=0.05, and the true hypothesis is recorded. Thus,

$$P(H_0 \text{ true} \mid p\text{-value } = 0.05) \approx \frac{1}{K}\sum_{k=1}^{K} \mathrm{I}(H_0 \text{ true} \mid p\text{-value } \approx 0.05).$$

Thus, there is nothing Bayesian happening here except that the probability being calculated has the unknown quantity on the left and the known quantity on the right.

Prosecutor's Fallacy

It is common for those using statistics to equate the following

$$p\text{-value} \stackrel{?}{\approx} P(\text{data}|H_0 \text{ true}) \neq P(H_0 \text{ true}|\text{data}).$$

but we can use Bayes rule to show us that these probabilities cannot be equated

$$p(H_0|y) = \frac{p(y|H_0)p(H_0)}{p(y)} = \frac{p(y|H_0)p(H_0)}{p(y|H_1)p(H_1) + p(y|H_1)p(H_1)}$$

This situation is common enough that it is called The Prosecutor's Fallacy.

ASA Statement on p-values

https://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

Principles:

1. P-values can indicate how incompatible the data are with a specified statistical model[, the model associated with the null hypothesis.]
2. P-values do not measure the probability the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based solely on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of the result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.