Bayesian hypothesis testing

Dr. Jarad Niemi

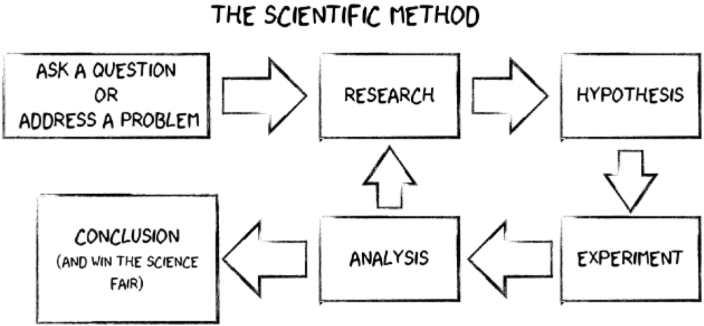
STAT 544 - Iowa State University

March 5, 2019

Outline

- Scientific method
  - Statistical hypothesis testing
  - Simple vs composite hypotheses
- Simple Bayesian hypothesis testing
  - All simple hypotheses
  - All composite hypotheses
- Propriety
  - Posterior
  - Prior predictive distribution
- Bayesian hypothesis testing with mixed hypotheses (models)
  - Prior model probability
  - Prior for parameters in composite hypotheses
    - WARNING: do not use non-informative priors
  - Posterior model probability

Scientific method



http://www.wired.com/wiredscience/2013/04/whats-wrong-with-the-scientific-method/

Statistical hypothesis testing

Definition

A simple hypothesis specifies the value for all parameters while a composite hypothesis does not.

Let $Y_i \stackrel{ind}{\sim} Ber(\theta)$ and

- $H_0 : \theta = 0.5$ (simple)
- $H_1 : \theta \neq 0.5$ (composite)

Prior probabilities on simple hypotheses

What is your prior probability for the following hypotheses:

- a coin flip has exactly 0.5 probability of landing heads
- a fertilizer treatment has zero effect on plant growth
- inactivation of a mouse growth gene has zero effect on mouse hair color
- a butterfly flapping its wings in Australia has no effect on temperature in Ames
- guessing the color of a card drawn from a deck has probability 0.5

Many null hypotheses have zero probability *a priori*, so why bother performing the hypothesis test?

Bayesian hypothesis testing with all simple hypotheses

Let $Y \sim p(y|\theta)$ and $H_j : \theta = d_j$ for $j = 1, \ldots, J$. Treat this as a discrete prior on the d_j, i.e.

$$P(\theta = d_j) = p_j.$$

The posterior is then

$$P(\theta = d_j|y) = \frac{p_j p(y|d_j)}{\sum_{k=1}^J p_k p(y|\theta_k)} \propto p_j p(y|d_j).$$

For example, suppose $Y_i \stackrel{ind}{\sim} Ber(\theta)$ and $P(\theta = d_j) = 1/11$ where $d_j = j/10$ for $j = 0, \ldots, 10$. The posterior is

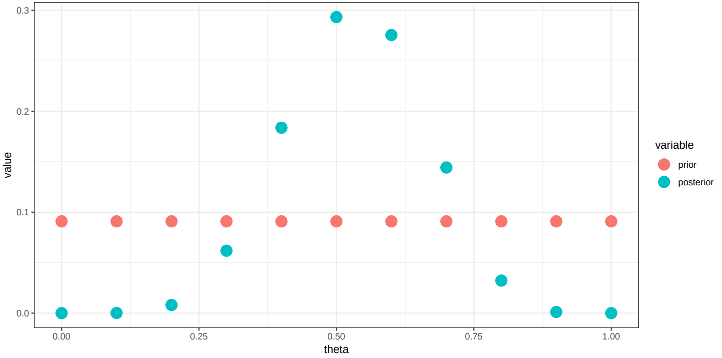
$$P(\theta = d_j|y) \propto \frac{1}{11} \prod_{i=1}^n (d_j)^{y_i} (1 - d_j)^{1-y_i} = \frac{1}{11} (d_j)^{n\overline{y}} (1 - d_j)^{n(1-\overline{y})}$$

If $j = 0$ ($j = 10$), any $y_i = 1$ ($y_i = 0$) will make the posterior probability of H_0 ($H_1 1$) zero.

Discrete prior example

```r
n = 13; y = rbinom(n,1,.45); sum(y)

[1] 7
```



Bayesian hypothesis testing with all composite hypotheses

Let $Y \sim p(y|\theta)$ and $H_j : \theta \in (E_{j-1}, E_j]$ for $j = 1, \ldots, J$. Just calculate the area under the curve, i.e. prior probabilities are

$$P(H_j) = P(E_{j-1} < \theta < E_j) = \int_{E_{j-1}}^{E_j} p(\theta) d\theta.$$

and posterior probabilities are

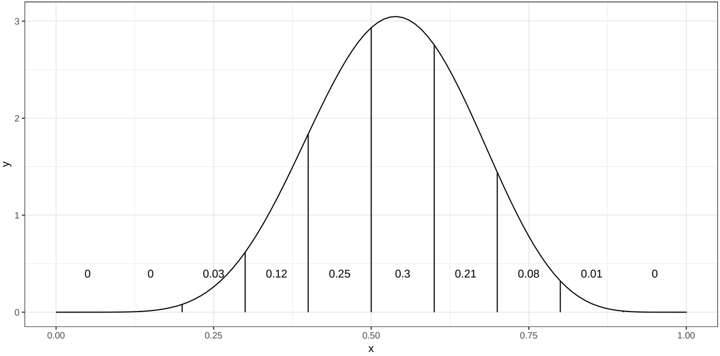
$$P(H_j|y) = P(E_{j-1} < \theta < E_j|y) = \int_{E_{j-1}}^{E_j} p(\theta|y) d\theta$$

For example, suppose $Y_i \stackrel{ind}{\sim} Ber(\theta)$ and $E_j = j/10$ for $j = 0, \ldots, 10$. Now, assume

$$\theta \sim Be(1,1) \quad \text{and thus} \quad \theta|y \sim Be(1 + n\overline{y}, 1 + n[1 - \overline{y}]).$$

Beta example

The posterior probabilities are



Tonelli's Theorem (successor to Fubini's Theorem)

Theorem

Tonelli's Theorem states that if $\mathcal{X}$ and $\mathcal{Y}$ are σ-finite measure spaces and f is non-negative and measureable, then

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)dydx = \int_{\mathcal{Y}}\int_{\mathcal{X}} f(x,y)dxdy$$

i.e. you can interchange the integrals (or sums).

On the following slides, the use of this theorem will be indicated by TT.

Proper priors with discrete data

Theorem

If the prior is proper and the data are discrete, then the posterior is always proper.

Proof.

Let $p(\theta)$ be the prior and $p(y|\theta)$ be the statistical model. Thus, we need to show that

$$p(y) = \int_\Theta p(y|\theta)p(\theta)d\theta < \infty \quad \forall y.$$

For discrete y, we have

$$\begin{aligned} p(y) &\le \sum_{z\in\mathcal{Y}} p(z) = \sum_{z\in\mathcal{Y}} \int_\Theta p(z|\theta)p(\theta)d\theta \stackrel{TT}{=} \int_\Theta \sum_{z\in\mathcal{Y}} p(z|\theta)p(\theta)d\theta \\ &= \int_\Theta p(\theta)d\theta = 1. \end{aligned}$$

Thus the posterior is always proper if y is discrete and the prior is proper. □

Proper priors with continuous data

Theorem

If the prior is proper and the data are continuous, then the posterior is almost always proper.

Proof.

Let $p(\theta)$ be the prior and $p(y|\theta)$ be the statistical model. Thus, we need to show that

$$p(y) = \int_\Theta p(y|\theta)p(\theta)d\theta < \infty \quad \text{for almost all } y.$$

For continuous y, we have

$$\int_\mathcal{Y} p(z)dz = \int_\mathcal{Y} \int_\Theta p(z|\theta)p(\theta)d\theta dz \overset{TT}{=} \int_\Theta \int_\mathcal{Y} p(z|\theta)dz\, p(\theta)d\theta = \int_\Theta p(\theta)d\theta = 1$$

thus $p(y)$ is finite except on a set of measure zero, i.e. $p(\theta|y)$ is almost always proper. □

Proper prior predictive distributions

In the previous derivations when the prior is proper, we showed that

$$\sum_{z \in \mathcal{Y}} p(z) = 1 \qquad \text{and} \qquad \int_{\mathcal{Y}} p(z) dz = 1$$

for discrete and continuous data, respectively.

Corollary

When the prior is proper, the prior predictive distribution is also proper.

Improper prior predictive distributions

Theorem

If $p(\theta)$ is improper, then $p(y) = \int p(y|\theta)p(\theta)d\theta$ is improper.

Proof.

$$\begin{aligned} \int p(y)dy &= \int \int p(y|\theta)p(\theta)d\theta dy \stackrel{TT}{=} \int p(\theta) \int p(y|\theta)dyd\theta \\ &= \int p(\theta)d\theta \end{aligned}$$

since $p(\theta)$ is improper, so is $p(y)$. A similar result holds for discrete y replacing the integral with a sum. □

Bayesian hypothesis testing

To evaluate the relative plausibility of a hypothesis (model), we use the posterior model probability:

$$p(H_j|y) = \frac{p(y|H_j)p(H_j)}{p(y)} = \frac{p(y|H_j)p(H_j)}{\sum_{k=1}^{J} p(y|H_k)p(H_k)} \propto p(y|H_j)p(H_j).$$

where $p(H_j)$ is the prior model probability and

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

is the marginal likelihood under model H_j and $p(\theta|H_j)$ is the prior for parameters θ when model H_j is true.

Marginal likelihood

The marginal likelihood calculation differs for simple vs composite hypotheses:

- Simple hypotheses can be considered to have a Dirac delta function for a prior, e.g. if $H_0 : \theta = \theta_0$ then $\theta|H_0 \sim \delta_{\theta_0}$. Then the marginal likelihood is

$$p(y|H_0) = \int p(y|\theta)p(\theta|H_0)d\theta = p(y|\theta_0).$$

- Composite hypotheses have a continuous prior and thus

$$p(y|H_j) = \int p(y|\theta)p(\theta|H_j)d\theta.$$

Two models

If we only have two models: H_0 and H_1, then

$$p(H_0|y) = \frac{p(y|H_0)p(H_0)}{p(y|H_0)p(H_0) + p(y|H_1)p(H_1)} = \frac{1}{1 + \frac{p(y|H_1)}{p(y|H_0)}\frac{p(H_1)}{p(H_0)}}$$

where

$$\frac{p(H_1)}{p(H_0)} = \frac{p(H_1)}{1 - p(H_1)}$$

is the prior odds in favor of H_1 and

$$BF(H_1 : H_0) = \frac{p(y|H_1)}{p(y|H_0)} = \frac{1}{BF(H_0 : H_1)}$$

is the Bayes Factor for model H_1 relative to H_0.

Binomial model

Consider a coin flipping experiment so that $Y_i \stackrel{ind}{\sim} Ber(\theta)$ and the null hypothesis $H_0 : \theta = 0.5$ versus the alternative $H_1 : \theta \neq 0.5$ and $\theta | H_1 \sim Be(a,b)$.

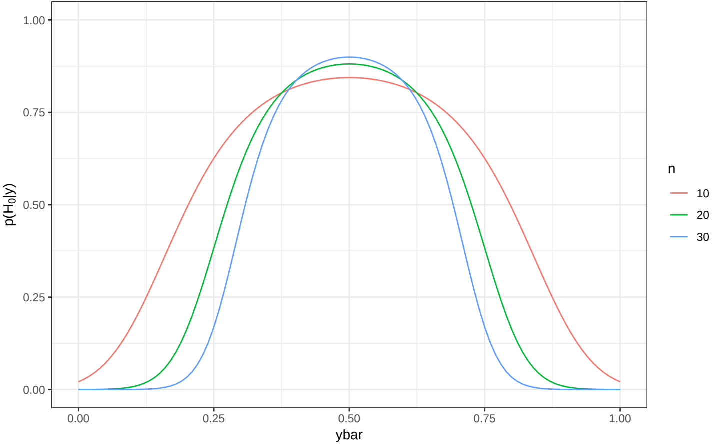
$$\begin{aligned} BF(H_0 : H_1) &= \frac{0.5^n}{\int_0^1 \theta^{n\overline{y}}(1-\theta)^{n(1-\overline{y})} \frac{\theta^{a-1}(1-\theta)^{b-1}}{Beta(a,b)} d\theta} \\ &= \frac{0.5^n}{\frac{1}{Beta(a,b)} \int_0^1 \theta^{a+n\overline{y}-1}(1-\theta)^{b+n-n\overline{y}-1}\theta} \\ &= \frac{0.5^n}{\frac{Beta(a+n\overline{y},b+n-n\overline{y})}{Beta(a,b)}} \\ &= \frac{0.5^n Beta(a,b)}{Beta(a+n\overline{y},b+n-n\overline{y})} \end{aligned}$$

and with $p(H_0) = p(H_1)$ the posterior model probability is

$$P(H_0|y) = \frac{1}{1 + \frac{1}{BF(H_0:H_1)}}.$$

Sample size and sample average

$P(H_0) = P(H_1) = 0.5$ and $\theta | H_1 \sim Be(1,1)$:



"Non-informative" prior

Recall that $\theta \sim Be(a,b)$ has

- a prior successes and
- b prior failures.

Thus, in some sense $a,b \to 0$ puts minimal prior data into the analysis.

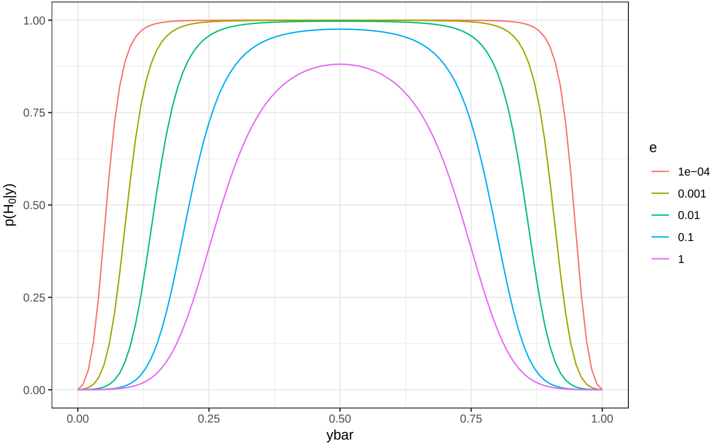
If $\theta|H_1 \sim Be(e,e)$, then

$$BF(H_0 : H_1) = \frac{0.5^n Be(e,e)}{Be(e+n\overline{y}, e+n-n\overline{y})} \xrightarrow{e\to 0} \infty \quad \text{for any } \overline{y} \in (0,1)$$

since $Be(e,e) \xrightarrow{e\to 0} \infty$.

Limit of proper prior

$P(H_0) = P(H_1) = 0.5$ and $\theta|H_1 \sim Be(e,e)$:



Normal example

Consider the model $Y \sim N(\theta, 1)$ and the hypothesis test

- $H_0 : \theta = 0$ versus
- $H_1 : \theta \neq 0$ with prior $\theta|H_1 \sim N(0, C)$.

The predictive distribution under H_1 is

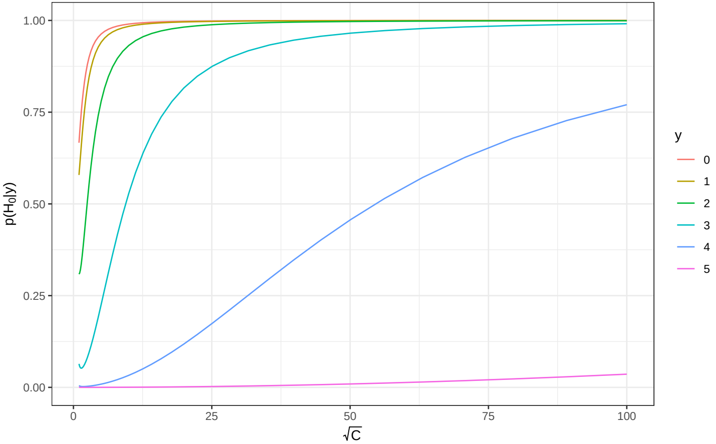
$$p(y|H_1) = \int p(y|\theta)p(\theta|H_1)d\theta = N(y; 0, 1 + C)$$

and the Bayes factor is

$$BF(H_0 : H_1) = \frac{N(y; 0, 1)}{N(y; 0, 1 + C)}.$$

The Bayes factor will increase as $C \to \infty$ for any y and this only gets worse if you use an improper prior.

Normal example



Summary

- Treat hypothesis testing as parameter estimation
  - All simple hypotheses: discrete prior
  - All composite hypotheses: continuous prior
- Formal Bayesian hypothesis testing
  (simple and composite hypotheses)
  - Specify prior model probabilities
  - Specify parameter priors for composite hypotheses
    WARNING: Do not use non-informative priors!
  - Calculate Bayes Factors or posterior model probabilities

Scientific method updated

All models are wrong, but some are useful.

George Box 1987