Lecture 5: Basics of Bayesian Hypothesis Testing

Merlise Clyde

September 9, 2021

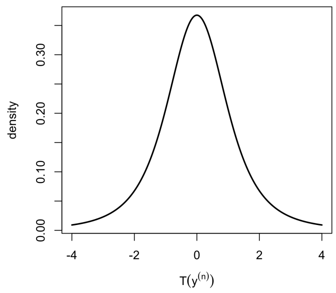
Hypothesis Testing

Suppose we have univariate data $y_i \overset{iid}{\sim} \mathcal{N}(\theta, 1)$

goal is to test $\mathcal{H}_0 : \theta = 0; \;\; \text{vs}\, \mathcal{H}_1 : \theta \neq 0$

Frequentist testing - likelihood ratio, Wald, score, UMP, confidence regions, etc

- Need a **test statistic** $T(y^{(n)})$ (and its sampling distribution)



- **p-value**: Calculate the probability of seeing a dataset/test statistics as extreme or more extreme than the oberved data with repeated sampling under the null hypothesis

Errors

if p-value is less than a pre-specified α then reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$

- Type I error: falsely concluding in favor of $\mathcal{H}_1$ when $\mathcal{H}_0$ is true
- To maintain a Type I error rate of α, then we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ when $p < \alpha$

For this to be a valid frequents test the p-value must have a uniform distribution under $\mathcal{H}_0$

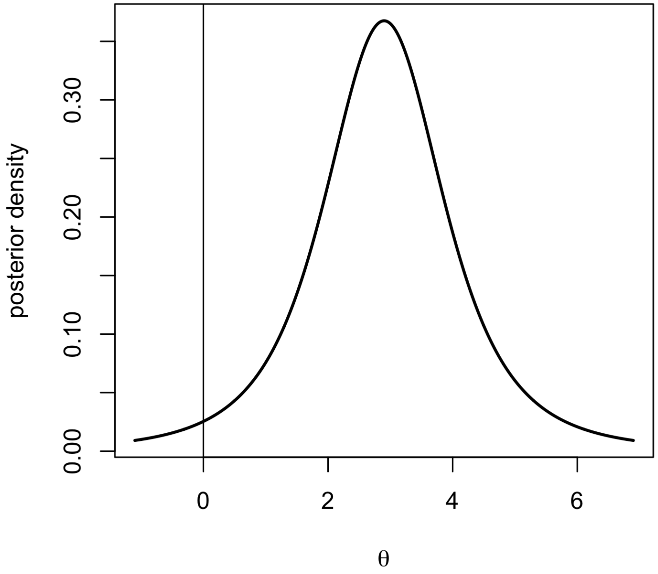
- Type II error: failing to conclude in favor of $\mathcal{H}_1$ when $\mathcal{H}_1$ is true
- 1 - P(Type II error) is the **power** of the test

Note: we *never* conclude in favor of $\mathcal{H}_0$. We are looking for enough evidence to reject $\mathcal{H}_0$. But if we fail to reject we do not conclude that it is true!

Bayesian Approach

1. Put a prior on θ, $\pi(\theta) = \mathcal{N}(\theta_0, 1/\tau_0^2)$.

2. Compute posterior $\theta \mid y^{(n)} \sim \mathcal{N}(\theta_n, 1/\tau_n^2)$ for updated parameters θ_n and τ_n^2.



Informal

Credible Intervals

1. Compute a 95% CI based on the posterior.

2. Reject $\mathcal{H}_0$ if interval does not contain zero.

Tail Areas:

1. Compute $\Pr(\theta > 0 \mid y^{(n)})$ and $\Pr(\theta < 0 \mid y^{(n)})$

2. Report minimum of these probabilities as a "Bayesian p-value"

Note: Tail probability is not the same as $\Pr(\mathcal{H}_0 \mid y^{(n)})$

Formal Bayesian Hypothesis Testing

Unknowns are $\mathcal{H}_0$ and $\mathcal{H}_1$

Put a prior on the actual hypotheses/models, that is, on $\pi(\mathcal{H}_0) = \Pr(\mathcal{H}_0 = \mathrm{True})$ and $\pi(\mathcal{H}_1) = \Pr(\mathcal{H}_1 = \mathrm{True})$.

- For example, set $\pi(\mathcal{H}_0) = 0.5$ and $\pi(\mathcal{H}_1) = 0.5$, if *a priori*, we believe the two hypotheses are equally likely. Likelihood of the hypotheses

$$\mathcal{L}(\mathcal{H}_i) \propto p(y^{(n)} \mid \mathcal{H}_i)$$

$$p(y^{(n)} \mid \mathcal{H}_0) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp -\frac{1}{2}(y_i - 0)^2$$

$$p(y^{(n)} \mid \mathcal{H}_1) = \int_{\Theta} p(y^{(n)} \mid \mathcal{H}_1, \theta) p(\theta \mid \mathcal{H}_1)\, d\theta$$

Bayesian Approach

Priors on parameters under each hypothesis

In our simple normal model, the only unknown parameter is θ

- under $\mathcal{H}_0$, $\theta = 0$ with probability 1
- under $\mathcal{H}_0$, $\theta \in \mathbb{R}$ Could take $\pi(\theta) = \mathcal{N}(\theta_0, 1/\tau_0^2)$.
- Compute marginal likelihoods for each hypothesis, that is, $\mathcal{L}(\mathcal{H}_0)$ and $\mathcal{L}(\mathcal{H}_1)$.
- Obtain posterior probabilities of $\mathcal{H}_0$ and $\mathcal{H}_1$ via Bayes Theorem.

Bayesian Approach - Decisions

Loss function for hypothesis testing

- $\hat{\mathcal{H}}$ is the chosen hypothesis
- $\mathcal{H}_{true}$ is the true hypothesis, $\mathcal{H}$ for short

Two types of errors:

- Type I error: $\hat{\mathcal{H}} = 1$ and $\mathcal{H} = 0$
- Type II error: $\hat{\mathcal{H}} = 0$ and $\mathcal{H} = 1$

Loss function:

$$L(\hat{\mathcal{H}}, \mathcal{H}) = w_1\, 1(\hat{\mathcal{H}} = 1, \mathcal{H} = 0) + w_2\, 1(\hat{\mathcal{H}} = 0, \mathcal{H} = 1)$$

- w_1 weights how bad making a Type I error
- w_2 weights how bad making a Type II error

Loss Function Functions and Decisions

- Relative weights

$$L(\hat{\mathcal{H}}, \mathcal{H}) = \mathbb{1}(\hat{\mathcal{H}} = 1, \mathcal{H} = 0) + w\, \mathbb{1}(\hat{\mathcal{H}} = 0, \mathcal{H} = 1)$$

- Special case $w = 1$

$$L(\hat{\mathcal{H}}, \mathcal{H}) = \mathbb{1}(\hat{\mathcal{H}} \neq \mathcal{H})$$

- known as 0-1 loss (most common)
- Bayes Risk (Posterior Expected Loss)

$$\mathsf{E}_{\mathcal{H} \mid y^{(n)}}[L(\hat{\mathcal{H}}, \mathcal{H})] = \mathbb{1}(\hat{\mathcal{H}} = 1)\pi(\mathcal{H}_0 \mid y^{(n)}) + \mathbb{1}(\hat{\mathcal{H}} = 0)\pi(\mathcal{H}_1 \mid y^{(n)})$$

- Minimize loss by picking hypothesis with the highest posterior probability

Bayesian hypothesis testing

- Using Bayes theorem,

$$\pi(\mathcal{H}_1 \mid Y) = \frac{p(y^{(n)} \mid \mathcal{H}_1)\pi(\mathcal{H}_1)}{p(y^{(n)} \mid \mathcal{H}_0)\pi(\mathcal{H}_0) + p(y^{(n)} \mid \mathcal{H}_1)\pi(\mathcal{H}_1)},$$

  where $p(y^{(n)} \mid \mathcal{H}_0)$ and $p(y^{(n)} \mid \mathcal{H}_1)$ are the marginal likelihoods hypotheses.

- If for example we set $\pi(\mathcal{H}_0) = 0.5$ and $\pi(\mathcal{H}_1) = 0.5$ *a priori*, then

$$\pi(\mathcal{H}_1 \mid Y) = \frac{0.5p(y^{(n)} \mid \mathcal{H}_1)}{0.5p(y^{(n)} \mid \mathcal{H}_0) + 0.5p(y^{(n)} \mid \mathcal{H}_1)}$$

$$= \frac{p(y^{(n)} \mid \mathcal{H}_1)}{p(y^{(n)} \mid \mathcal{H}_0) + p(y^{(n)} \mid \mathcal{H}_1)} = \frac{1}{\frac{p(y^{(n)}|\mathcal{H}_0)}{p(y^{(n)}|\mathcal{H}_1)} + 1}.$$

- The ratio $\frac{p(y^{(n)}|\mathcal{H}_0)}{p(y^{(n)}|\mathcal{H}_1)}$ is known as the **Bayes factor** in favor of $\mathcal{H}_0$, and often written as $\mathcal{BF}_{01}$. Similarly, we can compute $\mathcal{BF}_{10}$.

Bayes factors

- **Bayes factor**: is a ratio of marginal likelihoods and it provides a weight of evidence in the data in favor of one model over another.
- It is often used as an alternative to the frequentist p-value.
- **Rule of thumb**: $\mathcal{BF}_{01} > 10$ is strong evidence for $\mathcal{H}_0$; $\mathcal{BF}_{01} > 100$ is decisive evidence for $\mathcal{H}_0$.
- Notice that for our example,

$$\pi(\mathcal{H}_1 \mid Y) = \frac{1}{\frac{p(y^{(n)}|\mathcal{H}_0)}{p(y^{(n)}|\mathcal{H}_1)} + 1} = \frac{1}{\mathcal{BF}_{01} + 1}$$

  the higher the value of $\mathcal{BF}_{01}$, that is, the weight of evidence in the data in favor of $\mathcal{H}_0$, the lower the marginal posterior probability that $\mathcal{H}_1$ is true.
- That is, here, as $\mathcal{BF}_{01} \uparrow$, $\pi(\mathcal{H}_1 \mid Y) \downarrow$.

Bayes factors

- Let's look at another way to think of Bayes factors. First, recall that

$$\pi(\mathcal{H}_1 \mid Y) = \frac{p(y^{(n)} \mid \mathcal{H}_1)\pi(\mathcal{H}_1)}{p(y^{(n)} \mid \mathcal{H}_0)\pi(\mathcal{H}_0) + p(y^{(n)} \mid \mathcal{H}_1)\pi(\mathcal{H}_1)},$$

  so that

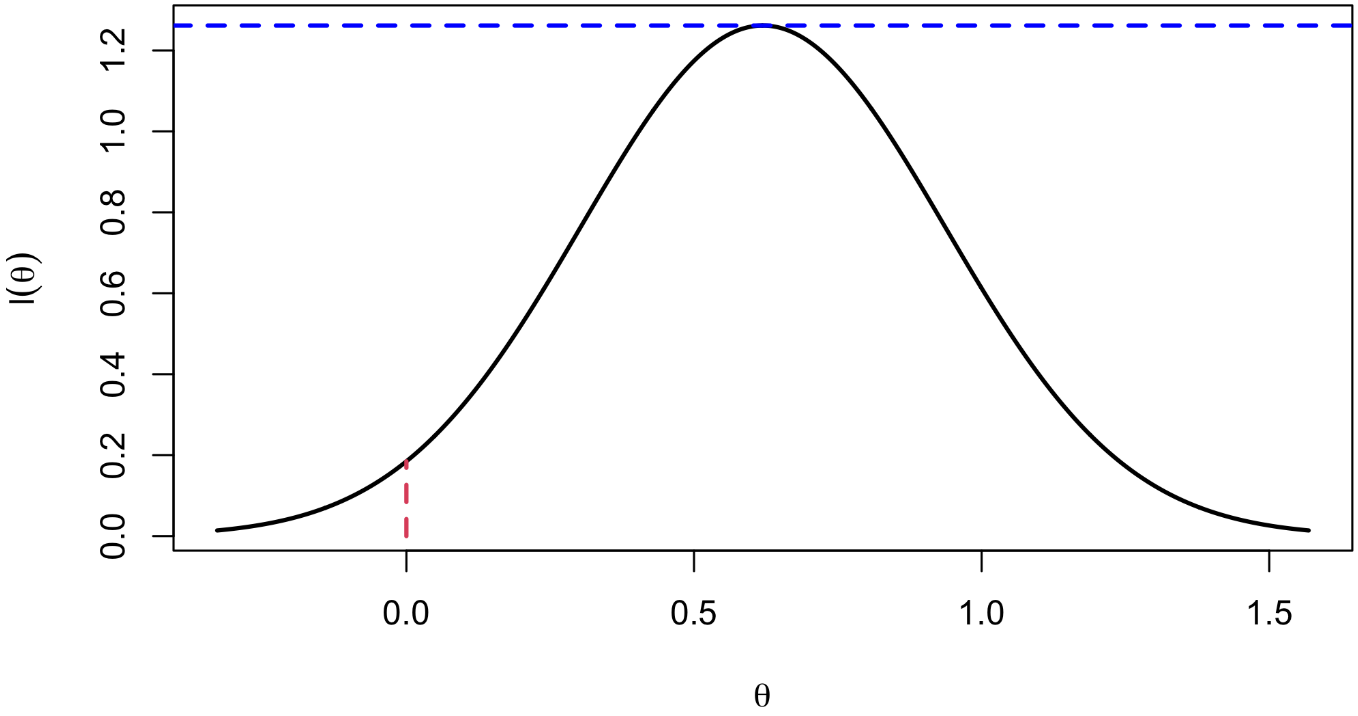
$$\begin{aligned}
\frac{\pi(\mathcal{H}_0|Y)}{\pi(\mathcal{H}_1|Y)} &= \frac{p(y^{(n)}|\mathcal{H}_0)\pi(\mathcal{H}_0)}{p(y^{(n)}|\mathcal{H}_0)\pi(\mathcal{H}_0) + p(y^{(n)}|\mathcal{H}_1)\pi(\mathcal{H}_1)} \div \frac{p(y^{(n)}|\mathcal{H}_1)\pi(\mathcal{H}_1)}{p(y^{(n)}\mathcal{H}_0)\pi(\mathcal{H}_0) + p(y^{(n)}|\mathcal{H}_1)\pi(\mathcal{H}_1)} \\
&= \frac{p(y^{(n)}|\mathcal{H}_0)\pi(\mathcal{H}_0)}{p(y^{(n)}|\mathcal{H}_0)\pi(\mathcal{H}_0) + p(y^{(n)}|\mathcal{H}_1)\pi(\mathcal{H}_1)} \times \frac{p(y^{(n)}|\mathcal{H}_0)\pi(\mathcal{H}_0) + p(y^{(n)}|\mathcal{H}_1)\pi(\mathcal{H}_1)}{p(y^{(n)}|\mathcal{H}_1)\pi(\mathcal{H}_1)}
\end{aligned}$$

$$\therefore \underbrace{\frac{\pi(\mathcal{H}_0 \mid Y)}{\pi(\mathcal{H}_1 \mid Y)}}_{\text{posterior odds}} = \underbrace{\frac{\pi(\mathcal{H}_0)}{\pi(\mathcal{H}_1)}}_{\text{prior odds}} \times \underbrace{\frac{p(y^{(n)} \mid \mathcal{H}_0)}{p(y^{(n)} \mid \mathcal{H}_1)}}_{\text{Bayes factor } \mathcal{BF}_{01}}$$

- Therefore, the Bayes factor can be thought of as the factor by which our prior odds change (towards the posterior odds) in the light of the data.

Likelihoods & Evidence

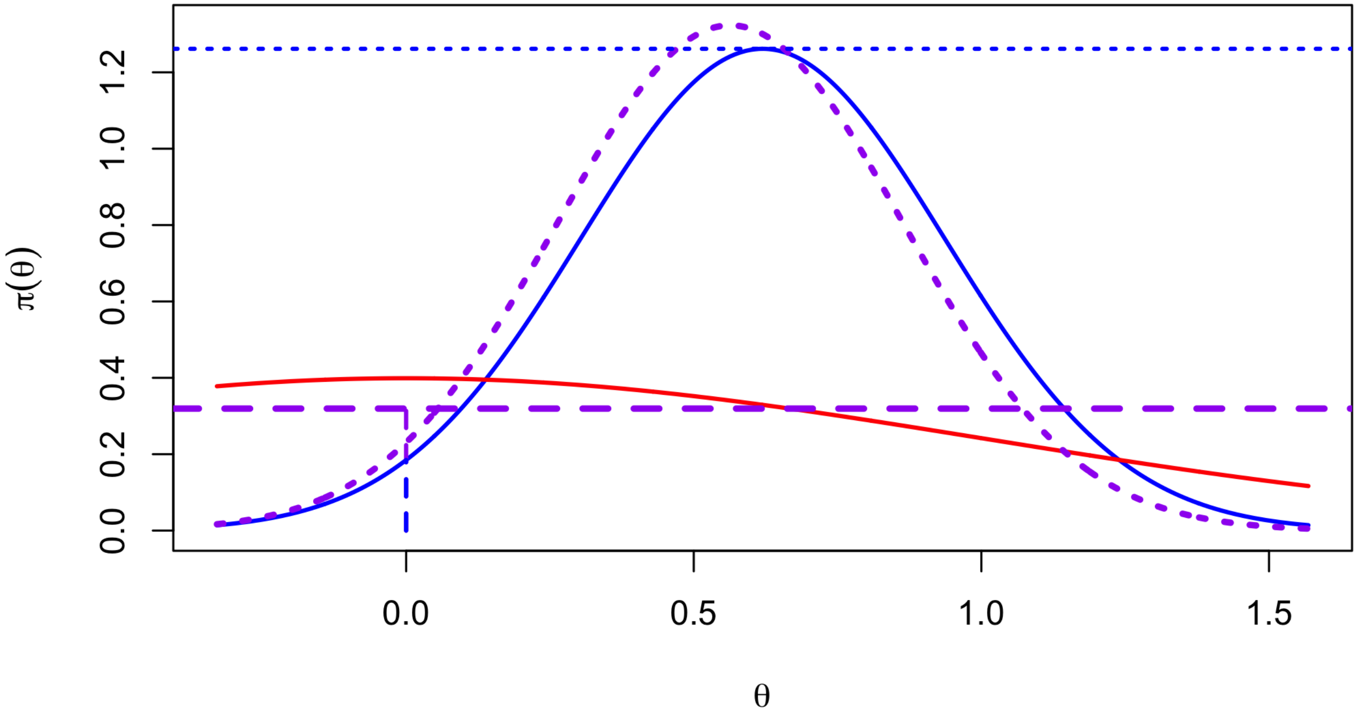
Maximized Likelihood



p-value = 0.05

Marginal Likelihoods & Evidence

Maximized Likelihood



$\mathcal{BF}_{10}$ = 1.73

Candidate's Formula (Besag 1989)

Alternative expression for Bayes Factor

$$\frac{p(y^{(n)} \mid \mathcal{H}_1)}{p(y^{(n)} \mid \mathcal{H}_0)} = \frac{\pi_\theta(0 \mid \mathcal{H}_1)}{\pi_\theta(0 \mid y^{(n)}, \mathcal{H}_1)}$$

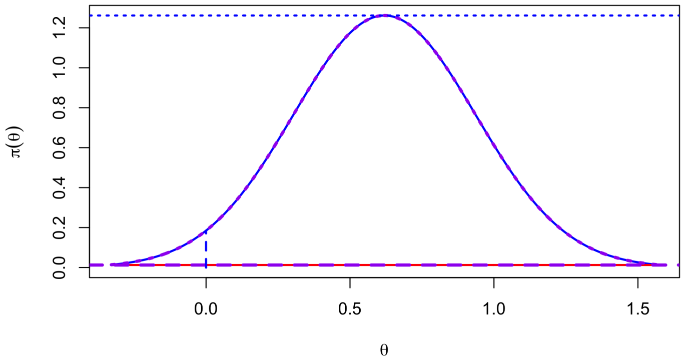
- ratio of the prior to posterior densities for θ evaluated at zero
- Savage-Dickey Ratio

Prior

Plots were based on a $\theta \mid \mathcal{H}_1 \sim \mathsf{N}(0, 1)$

- centered at value for θ under $\mathcal{H}_0$ (goes back to Jeffreys)
- "unit information prior" equivalent to a prior sample size is 1
- What happens if $n \to \infty$?
- What happens of $\tau_0 \to 0$?

Precision



- $\tau_0 = 1/1000$
- Posterior Probability of $\mathcal{H}_0$ = 0.9361
- As $\tau_0 \to 0$ the posterior probability of $\mathcal{H}_1$ goes to 0!

Bartlett's Paradox - the paradox is that a seemingly non-informative prior for θ is very informative about $\mathcal{H}$!