Lecture 19: Bayesian Hypothesis Testing (Model Comparison)
Lecture Outline

• Bayesian Hypothesis Testing
• Bayes Factor
• BIC
Bayesian Approach to Testing

• What is cumbersome with the Frequentist approach to hypothesis testing?
• $p$-value: the probability of rejecting the null hypothesis given the null is actually true. Written as a probability statement:

$$p\text{-value} = P(\text{our results} \mid H_0 \text{ is true})$$

• This is often incorrectly interpreted to be:

$$P(H_0 \text{ is true} \mid \text{our results}) \neq p\text{-value}$$

• How could we calculate the second probability statement?
• If we knew the overall/prior probability: $P(H_0 \text{ is true})$.
• In real life this is impossible to know. But in some situations, we could make an assumed prior probability on the model and take a Bayes approach 😊
Comparing two Models

• The probability of a model, $M$, given the data, $D$, based on Bayes’ formula is:

$$P(M \mid D) = \frac{P(D \mid M) \cdot P(M)}{P(D)}$$

• Then to compare two models:

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1) \cdot P(M_1) / P(D)}{P(D \mid M_2) \cdot P(M_2) / P(D)} = \left( \frac{P(M_1)}{P(M_2)} \right) \left( \frac{P(D \mid M_1)}{P(D \mid M_2)} \right)$$

Posterior model odds = prior model odds \times Bayes factor
Lecture Outline

• Bayesian Hypothesis Testing
• Bayes Factor
• BIC
Bayes Factor

• The information in the data to compare the models is then measured in the *Bayes factor*:

\[
B = \frac{P(D | M_1)}{P(D | M_2)}
\]

• To interpret the Bayes factor, use the following heuristic scale:

<table>
<thead>
<tr>
<th>Bayes Factor</th>
<th>Strength of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B \leq (1/10) )</td>
<td>Strong against ( M_1 )</td>
</tr>
<tr>
<td>((1/10) &lt; B \leq (1/3))</td>
<td>Substantial against ( M_1 )</td>
</tr>
<tr>
<td>((1/3) &lt; B &lt; 1)</td>
<td>Barely against ( M_1 )</td>
</tr>
<tr>
<td>(1 \leq B &lt; 3)</td>
<td>Barely for ( M_1 )</td>
</tr>
<tr>
<td>(3 \leq B &lt; 10)</td>
<td>Substantial for ( M_1 )</td>
</tr>
<tr>
<td>(B \leq 10)</td>
<td>Strong for ( M_1 )</td>
</tr>
</tbody>
</table>
Bayes Factor

• So what the heck is $P(D \mid M) = P(\text{Data} \mid \text{Model})$?
• In the Bayesian framework, it is:

$$P(D \mid M) = \int P(\theta \mid M) P(D \mid \theta, M) d\theta$$

• The model is a fixed value, so you can re-write the above as:

$$P(D) = \int P(\theta) P(D \mid \theta) d\theta$$
Bayes Factor Example #1

• Let’s use the simplest case: when the models we are comparing are specific values of a parameter.
• Let’s say we are observe 4 heads and 1 tail from a coin-flipping game.
• Let $p$ = probability of landing heads.
• We know that they were either flipped using a fair coin with $p = 0.5$ (Model 1), or a biased coin with $p = 0.9$ (Model 2).
• Calculate the *Bayes factor* for these two models.
• Interpret the results.
Bayes Factor Example #1

\[ P(D \mid M_1) = \]

\[ P(D \mid M_2) = \]

\[ B = \frac{P(D \mid M_1)}{P(D \mid M_2)} = \]

Interpretation:
Bayes Factor Example #2

• More realistically, just like in hypothesis testing, we may want to test a hypothesis of the form:
  
  \[ H_0: \theta = \theta_0 \text{ vs. } H_A: \theta \neq \theta_0 \text{ (really, } \theta \text{ is unrestricted)} \]

• To calculate the Bayes factor, we have to be more careful. The larger model (one without the restriction on \( \theta \)) is usually put in the numerator. And we need to put a prior on \( \theta \) for this model.

\[
B = \frac{f(D \mid M_1: \theta \text{ unconstrained})}{f(D \mid M_2: \theta = \theta_0)} = \frac{\int_{\theta \in \Omega|M_1} f(\theta)f(D \mid \theta)\,d\theta}{f(D \mid \theta_0)}
\]
Bayes Factor Example #2

• Let iid $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\sigma^2$ known.

$$H_0: \mu = \mu_0 \text{ vs. } H_A: \mu \neq \mu_0 \text{ (really, } \mu \text{ is unrestricted)}$$

• We could then set up two models to compare: one where we force $\mu = \mu_0$, and one where we allow $\mu$ to vary (but we will need to put a prior on it). We could put a minimally informative prior on $\mu$ for model #2.

• For our sleep dataset (for which we know $\sigma^2 = 1$), we wanted to test:

$$H_0: \mu = 8 \text{ vs. } H_A: \mu \neq 8$$

• What would be a reasonable pair of models to compare for these two different hypotheses?

• Model #1: $X_i \sim N(\mu = 8, \sigma^2 = 1)$

• Model #2: $X_i | \mu \sim N(\mu, \sigma^2)$ where $\mu \sim N(\mu_{prior} = 8, \sigma^2_{prior} = 4^2)$. 
Bayes Factor Example #2

• We observed $\bar{x} = 7.319$ for $n = 91$ observations. What are the probabilities of seeing the data for each model?
• Model #1 \{ $X_i \sim N(\mu = 8, \sigma^2 = 1)$ \}:

$$f(D \mid M_1 : \mu = 8) = f(\bar{x} = 7.319 \mid \mu = 8) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left( -\frac{\sum (x_i - \mu_0)^2}{2\sigma^2} \right)$$

$$= \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left( -\frac{n(\bar{x} - \mu_0)^2}{2\sigma^2} \right) = \left( \frac{1}{\sqrt{2\pi}} \right)^{91} \exp \left( -\frac{91(7.319 - 8)^2}{2} \right)$$

$$= 3.3015 \times 10^{-46}$$

• Wow, that’s a small number! Why is that not surprising?
• What issues could this lead to?
Bayes Factor Example #2

• Model #2 \(\{ X_i | \mu \sim N(\mu, \sigma^2) \text{ where } \mu \sim N(\mu = 8, \sigma^2 = 4^2) \}\).

\[ f(D | M_2 : \mu \text{ unconstrained}) = \int_{\mu|M_2} f(\mu) f(\bar{x} | \mu) d\mu \]

\[ = \int_{-\infty}^{\infty} \left[ \left( \frac{1}{\sigma_{\text{prior}} \sqrt{2\pi}} \right) \exp\left( - \frac{(\mu - \mu_{\text{prior}})^2}{2\sigma_{\text{prior}}^2} \right) \right] \left[ \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp\left( - \frac{n(\bar{x} - \mu)^2}{2\sigma^2} \right) \right] d\mu \]

\[ = \int_{-\infty}^{\infty} \left[ \left( \frac{1}{4\sqrt{2\pi}} \right) \exp\left( - \frac{(\mu - 8)^2}{2(4)^2} \right) \right] \left[ \left( \frac{1}{\sqrt{2\pi}} \right)^{91} \exp\left( - \frac{91(7.319 - \mu)^2}{2} \right) \right] d\mu \]

\[ \approx 1.15 \times 10^{-37} \]

• Thus the Bayes factor is \( B \approx (3.3 \times 10^{-46}) / (1.15 \times 10^{-37}) = 2.9 \times 10^{-9} \)

• So which model is preferred?
Bayes Factor Example #2: R-code

# Read the data
f=file.choose()
data=read.csv(f,header=T)
attach(data)

joint.dist=function(mu,x,hyper){
xbar=mean(x)
n=length(x)
mu0=hyper[1]
sigma0=sqrt(hyper[2])
prior=(1/(sigma0*sqrt(2*pi)))*exp(-(mu-mu0)^2/(2*sigma0^2))
likelihood=(1/sqrt(2*pi))^n * exp(-(xbar-mu)^2/2)
joint=prior*likelihood
return(joint)
}
integrate(joint.dist,lower=-Inf,upper=Inf,x=sleep,hyper=c(8,4^2))
integrate(joint.dist,lower=5,upper=10,x=sleep,hyper=c(8,4^2))
Lecture Outline

• Bayesian Hypothesis Testing
• Bayes Factor
• Bayesian Information Criterion (BIC)
Methods for Model Comparison

• Both in the Frequentist and Bayesian paradigms, there are times when you would like to compare models, and choose a best model (usually for future prediction).

There are a lot of ways to do this:

• If the models were built off of likelihood methods, then you could perform a likelihood ratio test (models need to be nested: one model’s parameters are a subset of another)

• If you cannot perform a likelihood ratio test, then you need another measure/statistic to compare models

• Statistics for model comparison: adjusted $R^2$ (bad), Akaike’s Information Criterion (AIC: very good), Bayesian Information Criterion (BIC: very good)
BIC for Model Comparison

• The *Bayesian Information Criterion (BIC)* can be used to compare two models.

• For a specific model, the BIC can be calculated to be:

\[
BIC = -2 \left[ l(\hat{\theta}) \right] + k \left[ \log(n) + \log(2\pi) \right]
\]

where \( l(\hat{\theta}) \) is the log-likelihood at the MLEs under for this model, \( k \) is the number of free parameters in the model, and \( n \) is the sample size.

• Then the **best model** will be the one with the **smallest** BIC. Note: the choice of prior has no effect on this measure.

• Key: your measured observations (response variable) must not change.

• Where does this formula come from?
BIC Example

- So let’s compare two models for iid $X_1, \ldots, X_n = \#$ siblings. $n = 92$, $\bar{X} = 1.326$, $\min(X) = 0$, $\max(X) = 9$.

- Model #1: $X_i \sim \text{Pois}(\lambda)$ \( \hat{\lambda}_{MLE} = \bar{X} \)

- Model #2: $X_i \sim \text{Unif}(\theta)$ [discrete uniform]

- Model #1:
  \[
  n = 92, \ k = 1 \\
  l(\hat{\lambda}_{MLE}) = -n\hat{\lambda} + (\log(\hat{\lambda}))\sum X_i - \sum \log(X_i!) = -136.47
  \]

- Model #2:
  \[
  n = 92, \ k = 1 \\
  l(\hat{\theta}_{MLE}) = -n\log(\hat{\theta}_{MLE}) = -91 \cdot \log(10) = -211.84
  \]
BIC Example (cont.)

• For this dataset and these models:
  • Model #1’s BIC:
    \[
    BIC = -2\left[l(\hat{\theta})\right] + k\left[\log(n) + \log(2\pi)\right]
    = -2[-136.47] + 1[\log(92) + \log(2\pi)] = 279.30
    \]
  • Model #2’s BIC:
    \[
    BIC = -2\left[l(\hat{\theta})\right] + k\left[\log(n) + \log(2\pi)\right]
    = -2[-211.84] + 1[\log(92) + \log(2\pi)] = 430.04
    \]
  • Model #1 is preferred (by a lot!)
Model Selection Details (BIC)

• In this example, the BIC was calculated for the two models, and they both had the same number of parameters \((k = 1)\).
• In most applications, the BIC will be used to compare two models in which the number of parameters vary.
• For example, you may want to decide between a variety of models to predict an outcome variable \((Y)\), each model with a different set of predictors \((X’s)\) or possibly based on different underlying distributions.
• The BIC is good for choosing the best out of these potential models as long as the response variable is not changing: (1) not being transformed and (2) there is no missing data in the predictors: number of observations is not changing.
Take Home Message

• Hypothesis testing in the Bayesian setting is not nearly as important. It’s really a matter of model selection.

• Instead of calculating a p-value, we can now calculate the actual probability that practitioners want: $P(H_0 \text{ is true } | \text{ data})$. This of course will assume a prior distribution on the parameters.

• The Bayes factor ($B$) can be used to test hypotheses and compare models. In the two-sided hypothesis framework, it can be calculated as:

$$B = \frac{f(D | M_1 : \theta \text{ unconstrained})}{f(D | M_2 : \theta = \theta_0)} = \frac{\int f(\theta) f(D | \theta) d\theta}{f(D | \theta_0)}$$

• The Bayesian Information Criterion (BIC) can also be used to compare models.

$$BIC = -2[l(\hat{\theta})] + k[\log(n) + \log(2\pi)]$$