Lecture 13

X. Bayesian Statistics

A. Frequentist statistics was marked by “proof by contradiction”:

1. Most notably with the $p$-value:
   a. Answers question “How often would I see evidence this extreme or more extreme against $H_0$, if $H_0$ were true?”
   b. Recall example that likelihood can change very little even as the $p$-value changes dramatically.

2. No information is used about prior beliefs about hypotheses
   a. including information one might have from other studies.

B. Those not indoctrinated into frequentism want to ask

   “$P[H_0 \text{ true}] = P[\theta \in \Omega_0]$”, “$P[H_A \text{ true}] = P[\theta \in \Omega_A]$”.

1. Probabilities should be conditional on data (denoted by $X$).
2. We really want $P[H_0|X = x]$.
3. In order to do this for every null hypothesis, we need distribution of $\theta$ conditional on data
4. We also have model for $X$ dependent on $\theta$.
5. Now write $f(x|\theta)$ where as before I wrote $f(x; \theta)$.
   a. Think of it as distribution of data conditional on the parameter.
6. Math can show that the combination of the distribution in both

directions generally define $P[H_0]$

a. If the question has a solution.

7. Hence you cannot do this analysis without specification of $P[H_0]$.

a. If you want to ask about all possible null hypothesis, you need distribution of $\theta$.

b. Distribution is called a *prior*, since it represents ideas about the parameter before seeing data.

C. General conditions:

1. $\theta \in \Theta$, for $\Theta$ an interval.

2. Distribution is almost always continuous, and can be given by density $\varpi_\theta(\theta)$.

a. Sometimes this distribution is determined by parameters.

b. These should be known by the analyst.

c. These are termed *hyperparameters*.

D. Joint density for $\theta$ and $X$ is $f_{\theta,X}(\theta, x) = \varpi_\theta(\theta)f_{X|\theta}(x|\theta)$

1. $f_{X|\theta}(x|\theta)$ is likelihood $L(\theta)$. 
E. Density for $\theta \mid X$ is

$$f_{\theta \mid X}(\theta \mid x) = \frac{f_{\theta, X}(\theta \mid x)}{f_X(x)}$$

$$= \frac{f_{\theta, X}(\theta \mid x)}{\int_{\Theta} f_{\theta, X}(\theta, x) d\theta}$$

$$= \frac{\varpi_\theta(\theta) f_{X \mid \theta}(x, \theta)}{\int_{\Theta} \varpi_\theta(\theta) f_{X \mid \theta}(x \mid \theta) d\theta},$$

by Bayes theorem.

1. Same formula for discrete $X$, this time with mass function.
2. Result is called posterior distribution.

F. Difficulties with Bayesian analyses:

1. Computation of denominator:
   a. The two examples above illustrate conjugate priors:
      i. Families of distributions constructed so that the likelihood times the prior was in the same family as the prior.
      ii. Integral done by integrating the resulting other member of the family.
      iii. Not generally useful for survival models.
   b. Numerical Integration:
      i. Evaluate $L(\theta) \varpi_\theta(\theta)$ on a grid of points, separated by
         $$\Delta: \zeta_j = L(\theta_0 + (j - 1)\Delta) \varpi_\theta(\theta_0 + (j - 1)\Delta)$$
         for $j = \{1, \ldots, m\}$.  

ii. Approximation to integral is a linear combination of these evaluations

- Trapezoidal rule: \( \int_{\Theta} L(\theta) \omega_{i,\theta}(\theta) \, d\theta \approx \Delta (\omega_1 + 2 \sum_{j=1}^{m-1} \omega_j + \omega_m)/2 \).

- Simpson’s rule: if \( m \) odd, \( \int_{\Theta} L(\theta) \omega_{i,\theta}(\theta) \, d\theta \approx \Delta (\omega_1 + 4 \sum_{j=1, j \text{ odd}}^{m-1} \omega_j + 2 \sum_{j=1, j \text{ even}}^{m-1} \omega_j + \omega_m)/3 \).

c. Laplace’s method

i. We want \( \int_A \exp(\ell(\theta)) \omega(\theta) \, d\theta \)

ii. Let \( \hat{\theta} \) be the MLE.

iii. Do Taylor series approximation for log likelihood and prior separately.

iv. Extends to higher-order approximations.

2. Integration scales poorly as dimension of \( \theta \) increases

a. Deterministic integration is replaced by simulation.

G. Bayesian Inference

1. Estimation:

a. Generally use expectation of posterior distribution.

i. Minimizes expected squared error loss, analogously to material from lecture 1.
Lecture 13

b. Can also use posterior median or mode.

H. Recall that frequentist inference provides intervals of the form
\[ \theta \in (L, U) \], for \( L \) and \( U \), such that \( P_{\theta} [L \leq \theta \leq U] = 1 - \alpha \).

1. Probability uses \( f(data; \theta) \).

2. A similar interval using posterior probability is called a **credible interval**.
   a. Recall frequentist intervals were generally equal-tailed
   b. Bayesian intervals are often **highest posterior density**

I. Where do priors come from?

1. Represent subjective opinion about relative possibility of various values of parameter

2. What if you don’t have such a subjective opinion? Look for non-informative prior.
   a. Ex., \( \theta \) is a location parameter: non-informative prior should be uniform on \( (-\infty, \infty) \).
   b. Ex., \( \sigma \) is standard deviation: non-informative prior should be uniform on \( (0, \infty) \)?
      i. If \( \tau \) is variance: non-informative prior should be uniform on \( (0, \infty) \)?
ii. Problem: Density that is uniform for $\sigma$ is not uniform for $\sigma^2$, and vice versa.

iii. Solution: Log standard deviation (and hence log variance) uniform on $(-\infty, \infty)$.

c. Problem: All of these noninformative priors aren’t really distributions

i. since they integrate to infinity.

ii. Such priors are called “improper”.

- Conceptually justified as the limit of proper priors:
  \[ \theta \sim \mathcal{U}[-T, T], \ T \to \infty. \]

iii. In extreme cases, can make integral in denominator be infinite.

R Code

J. Bayesian hypothesis testing.

1. As before, decide between $H_0 : \theta \in \Omega_0$ vs. $H_A : \theta \in \Omega_a$.
   
   a. Here I used notation similar to that of frequentist analysis.
   
   b. At present, no “null” and “alternate” subtext.

2. Choose hypothesis with highest posterior probability.

3. Often report posterior odds $P[\Omega_0|\text{data}] / P[\Omega_a|\text{data}]$

4. Factor $B$ by which prior odds $P[\Omega_0] / P[\Omega_a]$ was changed is
Lecture 14

called Bayes factor.

a. \[ B = \frac{P[\Omega_0|\text{data}] \, P[\Omega_a]}{P[\Omega_a|\text{data}] \, P[\Omega_0]} \]

b. When hypothesis \(\Omega_0\) and \(\Omega_a\) are both simple, Bayes factor is the likelihood ratio.

c. Point hypotheses are only workable if there’s positive prior probability on them.