Homework 3

Carly Greutert

February 20, 2023

1. Warmup: Posterior Predictive Distributions

a. What is a posterior predictive distribution (i.e., what does it give probabilities for)? How is this different from the posterior distribution of a parameter?

A Posterior predictive distribution observes the distribution of a possible value given previous observations. It incorporates sampling and posterior uncertainty. It gives probabilities for observing a new variable. This is different from the posterior distribution of a parameter because it considers uncertainty about our parameter.

b. Is a posterior predictive model conditional on just the data, just the parameter, or on both the data and the parameter?

The posterior predictive model is just conditional on the data.

c. Why do we need posterior predictive distributions? For example, if we wanted to predict new values of Y, why couldn't we just use the posterior mean of the parameter?

We need posterior predictive distributions because our parameters derived from the prior are not certain, it is good to predict based on previous observations, which is what PPDs can do since they do not depend on the parameters. In other words, the posterior mean is not necessarily the most accurate approach for predicting new values since it just captures the data that data that has been observed.

2. Cancer Research in Laboratory Mice

As a reminder from homework 2, a laboratory is estimating the rate of tumorigenesis (the formation of tumors) in two strains of mice, A and B. They have tumor count data for 10 mice in strain A and 13 mice in strain B. Type A mice have been well studied, and information from other laboratories suggests that type A mice have tumor counts that are approximately Poisson-distributed. Tumor count rates for type B mice are unknown, but type B mice are related to type A mice. Assuming a Poisson sampling distribution for each group with rates θ_A and θ_B. We assume $\theta_A \sim \text{gamma}(120, 10)$ and $\theta_B \sim \text{gamma}(12, 1)$. We observe $y_A = (12, 9, 12, 14, 13, 13, 15, 8, 15, 6)$ and $y_B = (11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)$. Now we will actually investigate evidence that Type A mice have higher rates of tumor formation than Type B mice.

a. Obtain $Pr(\theta_B < \theta_A \mid y_A, y_B)$ via Monte Carlo sampling. Report the value.

```
y_A <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
y_B <- c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

# store your probabilities in a vector called "pr" for testing.
S <- 1000
pr <- c()
for(s in 1:S){
 theta_A <- rgamma(1, 120 + sum(y_A), 10 + length(y_A))
 theta_B <- rgamma(1, 12 + sum(y_B), 1 + length(y_B))
 pr <- append(pr, theta_B < theta_A)}
pr <- sum(pr)/S
print(pr)
```

```
## [1] 0.998
```

b. Now compute $P\tilde{Y}_B < \tilde{Y}_A \mid Y_B, Y_A)$, where $\tilde{Y}_A$ and $\tilde{Y}_B$ are samples from the posterior predictive distribution.

```
y_A <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)

# store your probabilities in a vector called "ppr" for testing.

ppr <- c()
for (s in 1:S){
  theta_A <- rgamma(1, 120 + sum(y_A), 10 + length(y_A))
  theta_B <- rgamma(1, 12 + sum(y_B), 1 + length(y_B))
  tilde_Y_A <- rpois(1, theta_A)
  tilde_Y_B <- rpois(1, theta_B)
  ppr <- append(ppr, tilde_Y_B < tilde_Y_A)
}
ppr <- sum(ppr)/S
print(ppr)
```

```
## [1] 0.698
```

c. In the context of this problem, describe the meaning of the events $\{\theta_B < \theta_A\}$ and $\{\tilde{Y}_B < \tilde{Y}_A\}$. How are they different? Why do the relative values of the answers in parts a and b make sense?

When we're analyzing the probabilities that $\{\theta_B < \theta_A\}$, we are observing whether group A is more likely to develop tumors at a higher rate than group B. Our analysis of $\{\tilde{Y}_B < \tilde{Y}_A\}$ tells us how likely it is that group A will have higher tumor counts than group B. The relative values of the answers make sense because we are seeing the proportion of overall times that it is likely for one group to have a higher probability over the other and it helps with replicability.

3. Posterior Predictive Model Checking

Model checking and refinement is an essential part of Bayesian data analysis. Let's investigate the adequacy of the Poisson model for the tumor count data. Consider strain A mice only for now, and generate posterior predictive datasets $y_B^{(1)}, ..., y_A^{(1000)}$. Each $y_B^{(s)}$ is a sample of size $n_B = 13$ from the Poisson distribution with parameter $\theta_B^{(s)}$, $\theta_B^{(s)}$ is itself a sample from the posterior distribution $p(\theta_B \mid y_B)$ and y_B is the observed data. For each s, let $t^{(s)}$ be the sample average divided by the sample variance of $y_B^{(s)}$.

a. If the Poisson model was a reasonable one, what would a "typical" value $t^{(s)}$ be? Why?

A "typical" value of $t^{(s)}$, which is calculated by taking the sample average and dividing by the sample variance, should be roughly 1 since for a poisson distribution the mean and variance are both just the parameter θ_B, it is essentially dividing by itself.

b. In any given experiment, the realized value of t^s will not be exactly the "typical value" due to sampling variability. Make a histogram of $t^{(s)}$ and compare to the observed value of this statistic, $\frac{\text{mean}(y_A)}{\text{var}(y_B)}$. Can sampling variability alone explain the observed test statistic? It may help to compute the fraction of posterior predictive draws which are larger than the observed draws. Make a comment on if the Poisson model seems reasonable for these data (at least by this one metric).

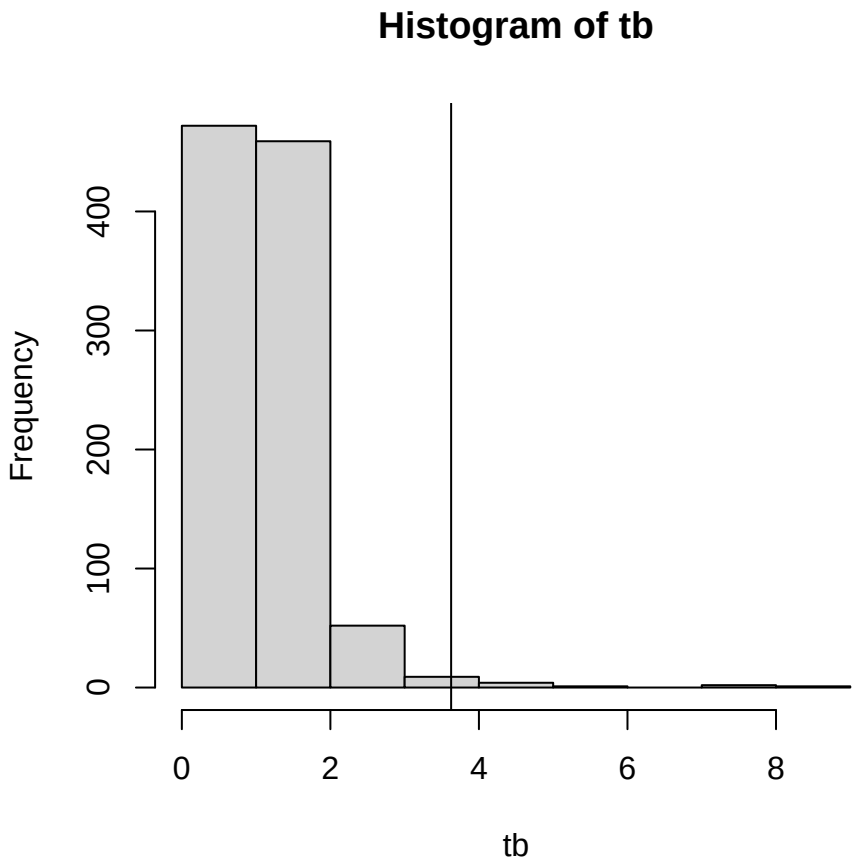
```
y_B = c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

# generate posterior predictive datasets and find test statistic for each one
# store your test statistics in a vector called "tb" for testing

tb <- c()
for (s in 1:S){
```

```
  theta_B <- rgamma(1, 12 + sum(y_B), 1 + length(y_B))
  tilde_YB <- rpois(13, theta_B)
  tb <- append(tb, mean(tilde_YB)/var(tilde_YB))
}
```

```
. = ottr::check("tests/q4c1.R")
```

```
##
## All tests passed!
```

```
# create the histogram, adding a vertical line at the observed value of the test statistic
hist(tb)
abline(v = mean(y_B)/var(y_B))
```



It seems like the test statistics under-calculated the actual observed value of this statistic. Thus, we would need to adjust our model so that this is reflected in our sample test statistics, as it is not completely reasonable.

c. When the mean is less than the variance we say that the data is *overdispersed*. When the mean is more than the variance we say that the data is *underdispersed*. Do you have any evidence that the data is underdispered? Overdispered?

This data appears to be underdispersed since the mean, 8.692308, of the data (y_B) is larger than the variance, 2.397436. Especially considering our sample test statistics were predicting a much lower value, we need to adjust our model.