Homework 3

PSTAT 115, Winter 2023

Due on February 20, 2023 at 11:59 pm

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?

Type your answer here, replacing this text.

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

Type your answer here, replacing this text.

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?

Type your answer here, replacing this text.

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.

# YOUR CODE HERE
pr <- NULL # YOUR CODE HERE
print(pr)
```

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.

# YOUR CODE HERE
```

```
ppr <- NULL # YOUR CODE HERE
print(ppr)
```

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?

Type your answer here, replacing this text.

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?

Type your answer here, replacing this text.

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

# YOUR CODE HERE
```

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

```
# create the histogram, adding a vertical line at the observed value of the test statistic
# YOUR CODE HERE
```

Type your answer here, replacing this text.

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