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?

The posterior predictive distribution is a distribution which uses old data to predict new data points. The posterior predictive distribution is the distribution of the random variable based on the assumed distribution, like poisson, with a parameter which is updated with data. On the other hand, the posterior distribution of the parameter is the distribution of the random parameter of the assumed distribution of the variable.

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 conditional just on the data, no parameters are involved. Of course the model itself depends on the parameter as well, which is evidenced by the integration over all possible values of the parameter with the posterior distribution of the parameter.

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 a posterior predictive distribution because it is a distribution, which means it incorporates variability. If we were to just take the posterior mean of the parameter we would fail to represent the fact that the parameter is random. That is because we take a single fixed value as a parameter instead of representing other possible values in the parameter distribution. It would not make sense for all new data to be based on a single parameter value such as the parameter mean. Thus a posterior predictive distribution which has sampling variability and posterior variability will be much more useful for predicting new data than a posterior mean.

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.

```
# Calculating Posterior parameters (done in hw2)
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)

# Prior parameters here
alpha_A = 120
```

```r
beta_A = 10

alpha_B = 12
beta_B = 1

# Posterior parameters here
alpha_A_posterior = sum(y_A) + alpha_A
beta_A_posterior = length(y_A) + beta_A

alpha_B_posterior = sum(y_B) + alpha_B
beta_B_posterior = length(y_B) + beta_B
```

```r
# Vector to store results
results <- rep(NA, 10000)

# Monte Carlo Sampling
for (i in 1:10000){
  theta_A <- rgamma(1, alpha_A_posterior, beta_A_posterior)
  theta_B <- rgamma(1, alpha_B_posterior, beta_B_posterior)
  if (theta_B < theta_A){
    results[i] <- 1
  } else{
    results[i] <- 0
  }
}

# store your probability in a vector called "pr" for testing.
# Calculating P(theta_B < theta_A)
pr <- sum(results) / length(results)
print(pr)
```

```
## [1] 0.9957
```

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.

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

# Vector to store results
results2 <- rep(NA, 10000)

# Posterior Predictive Sampling
for (i in 1:10000){
  theta_A <- rgamma(1, alpha_A_posterior, beta_A_posterior)
  theta_B <- rgamma(1, alpha_B_posterior, beta_B_posterior)
  Ytilde_A <- rpois(1, theta_A)
  Ytilde_B <- rpois(1, theta_B)
  if (Ytilde_B < Ytilde_A){
    results2[i] <- 1
  } else{
    results2[i] <- 0
  }
}

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

```
ppr <- sum(results2) / length(results2)
print(ppr)
```

```
## [1] 0.6979
```

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?

The events $\{\theta_B < \theta_A\}$ represents when the rate of tumors in Type B mice is less than the rate of tumors in Type A mice.
The events $\{\tilde{Y}_B < \tilde{Y}_A\}$ represent when the actual number of tumors in a simulated sample from Type B mice is less than the number of tumors in a simulated sample from Type A mice.
The two events are different because one event represents the rate at which tumors will appear, while the other represents the actual simulated tumor counts. The relative values of the answers in parts a and b make sense, because if Type B mice have a smaller rate of tumors than type A, it would make sense that in simulation, Type B mice will indeed have less tumors than Type A mice, which is the case.

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 B 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?

If a poisson model was a reasonable one, the "typical" value of $t^{(s)}$ should be 1, because for a poisson distribution, the mean and variance are the same.

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 <- rep(NA, 1000)
for (i in 1:1000){
  theta_B <- rgamma(1, alpha_B_posterior, beta_B_posterior)
  Ytilde_B <- rpois(13, theta_B)
  mu <- mean(Ytilde_B)
  var <- var(Ytilde_B)
  tb[i] <- mu / var
}
# tb
```

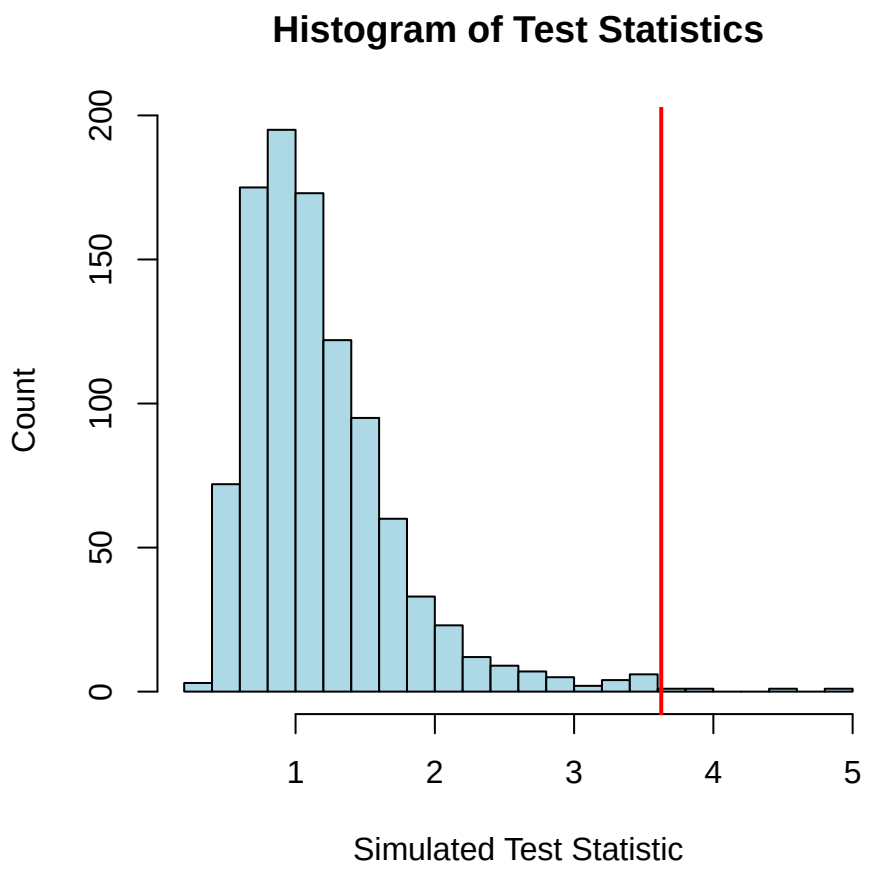
```
. = ottr::check("tests/q4c1.R")
```

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

```
test_stat <- mean(y_B) / var(y_B)

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

```
hist(tb, xlab = 'Simulated Test Statistic', ylab = 'Count', main = 'Histogram of Test Statistics', col =
abline(v = test_stat, col='red', lwd = 2)
```



A poisson model does not seem reasonable for this data because the test statistic of the original data is 3.625668 which is much more than 1. This test statistic value is telling us that the mean is larger than the variance, which does not support a poisson model where mean and variance are the same.

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?

Since the test statistic for the observed data is 3.625668 we know the mean is much larger than the variance, which would give us evidence that the data is underdispersed.