Homework 3

PSTAT 115, Winter 2022

Due on February 13, 2022 at 11:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

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 the distribution of possible unobserved values conditioned on the observed values. It depends on the observed values, NOT on unknown parameter θ. It is the distribution for future predicted data based on the data you have already seen. So basically the posterior predictive is used to predict new data values.
- The posterior distribution has a different meaning; it can be seen as an "adjustment" on prior probability and tells us what we know after the data has been observed. The posterior distribution is a way to summarize what we know about undetermined quantities in Bayesian analysis. It is a combination of the prior distribution and the likelihood function and depends on the unknown parameter θ.

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

- A posterior predictive model is conditional on just the data as it is the distribution of unobserved values conditioned on observed values.

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?

- If we are given the posterior distributions of the parameters of the model, then the posterior predictive distribution gives us some idea of what future data might look like.
- We can use posterior predictive models to perform posterior predictive checks in order to compare the predicted fitted model to the actual observed data. We want to see if the posterior predictive data (sampling data) looks more or less similar to the observed data or if it is an inadequate model to describe the data.
- When predict new values of Y we essentially are using observed values of the "older" random variable Y to predict what new values may look like for the same random variable. The posterior mean however is more commonly associated with finding the "best" value of model parameters by taking the expectation of the sampling model.

2. Posterior Credible Intervals.

One way for us to learn more about posterior distributions is to find some credible intervals. Suppose we have a posterior density of a parameter λ, defined by $\lambda|y \sim \text{Gamma}(4, 1)$.

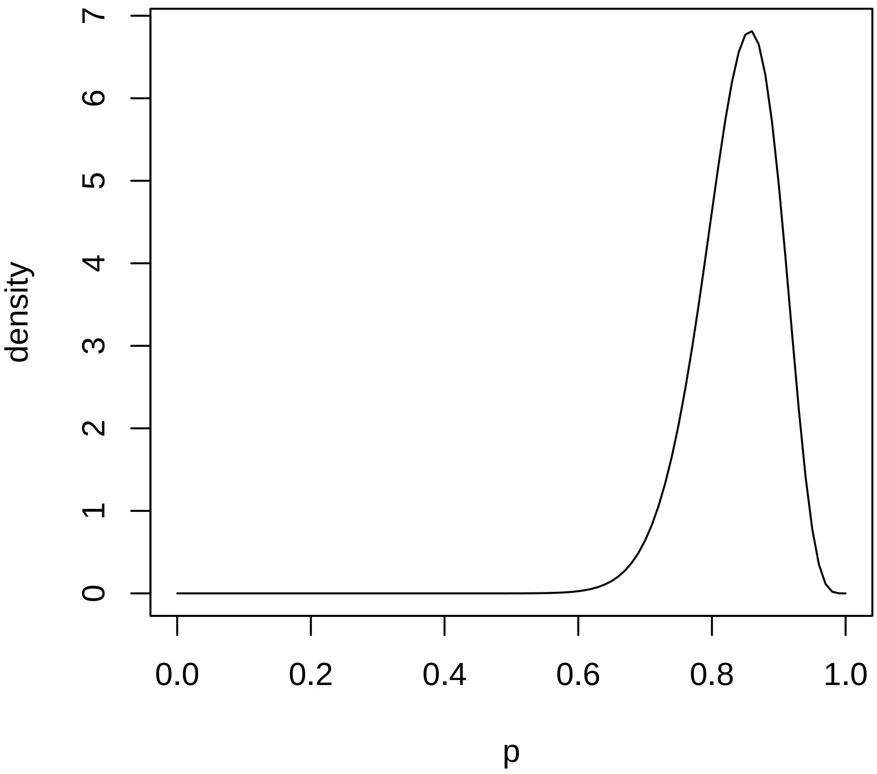
a. Plot this distribution. Construct the posterior middle 95% credible interval for λ. Save the lower and upper endpoints in a vector of length 2, called `middle_95`. Using either line segments or shading, display this interval on the plot.

```
# create interval - middle_95 should be a vector containing the two endpoints of the interval, in order
y <- c(4, 3, 6, 11, 3)
a <- 4
b <- 1
a_post = a + sum(y)
b_post = b + length(y)


alpha = 1 - 0.95
low = qgamma(alpha/2, a_post, b_post)
high = qgamma(1 - alpha/2, a_post, b_post)
middle_95 <- print(c(low,high))
```

```
## [1] 3.510499 7.137811
```

```
. = ottr::check("tests/q2a1.R")
```

```
# make the plot
curve(gamma(a_post + b_post)/gamma(a_post)/gamma(b_post) *
      p^(a_post - 1) * (1-p)^(b_post - 1), from = 0, to = 1, xname = "p",
      xlab = "p", ylab = "density")
abline(v = low, col = "red", lty = 2)
abline(v = high, col = "red", lty = 2)
```



b. Interpret the interval you found. How is it different from a frequentist confidence interval?

In the above plot,interval [0.6,1] is the central portion of the posterior distribution that contains 95% of the values.This is very different than frequentist statistics. In frequentists statistics a 95% Confident interval can be described as: if you were to repeat the experiment or sampling an infinite(or just many) times, 95% of the intervals constructed would contain the true value of the parameter.

c. Besides the middle 95% credible interval, we could also find the 95% highest posterior density (HPD) region. This region contains the 95% of posterior values with the highest posterior densities. The HPD region will always be the shortest credible interval for a given probability, since it by definition contains the values of λ with the highest probability of occurring. Use `HDInterval::hdi()` to construct the HPD region. Save the lower and upper endpoints of this region in a variable called `hdi_region`. Add this interval to the plot you made in part (a), making sure that both intervals are distinguishable on the plot.

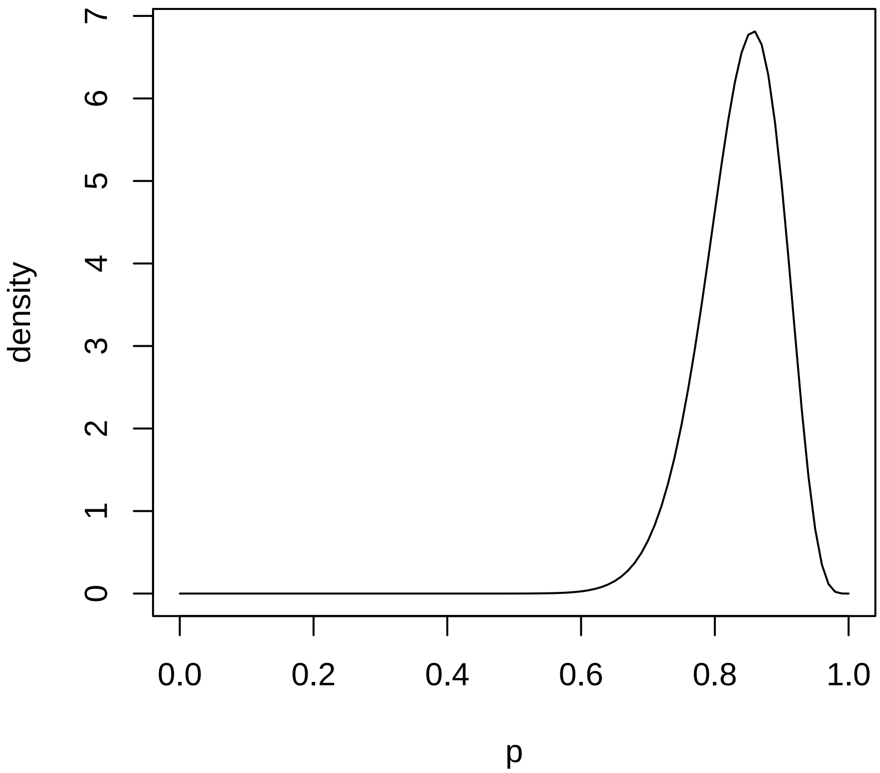
```
# create interval
tpd <- rgamma(100000, 4, 1)
hd_region <-  HDInterval::hdi(tpd, credMass=0.5)[1:2]
print(hd_region)
```

```
##    lower    upper
## 1.952612 4.349505
```

```
. = ottr::check("tests/q2c1.R")
```

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

```
# make the plot
curve(gamma(a_post + b_post)/gamma(a_post)/gamma(b_post) *
      p^(a_post - 1) * (1-p)^(b_post - 1), from = 0, to = 1, xname = "p",
      xlab = "p", ylab = "density")
abline(v = low, col = "red", lty = 2)
abline(v = high, col = "red", lty = 2)
abline(v = hd_region, col = "blue", lty = 2)
```



d. Based on your plot, how do the two kinds of 95% credible intervals differ? How long is the middle interval? The HDI interval?

The HDI interval is noticeably shorter than the middle interval. The Middle interval is [3.510499, 7.137811] and the HDI interval is [1.895055, 4.301931].

3. 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 are have higher rates of tumor formation than Type B mice.

a. For $n_0 \in \{1, 2, ..., 50\}$, obtain $Pr(\theta_B < \theta_A \mid y_A, y_B)$ via Monte Carlo sampling for $\theta_A \sim \text{gamma}(120, 10)$ and $\theta_B \sim \text{gamma}(12 \times n_0, n_0)$. Make a line plot of $Pr(\theta_B < \theta_A \mid y_A, y_B)$ vs n_0. Describe how sensitive the conclusions about the event $\{\theta_B < \theta_A\}$ are to the prior distribution on θ_B.

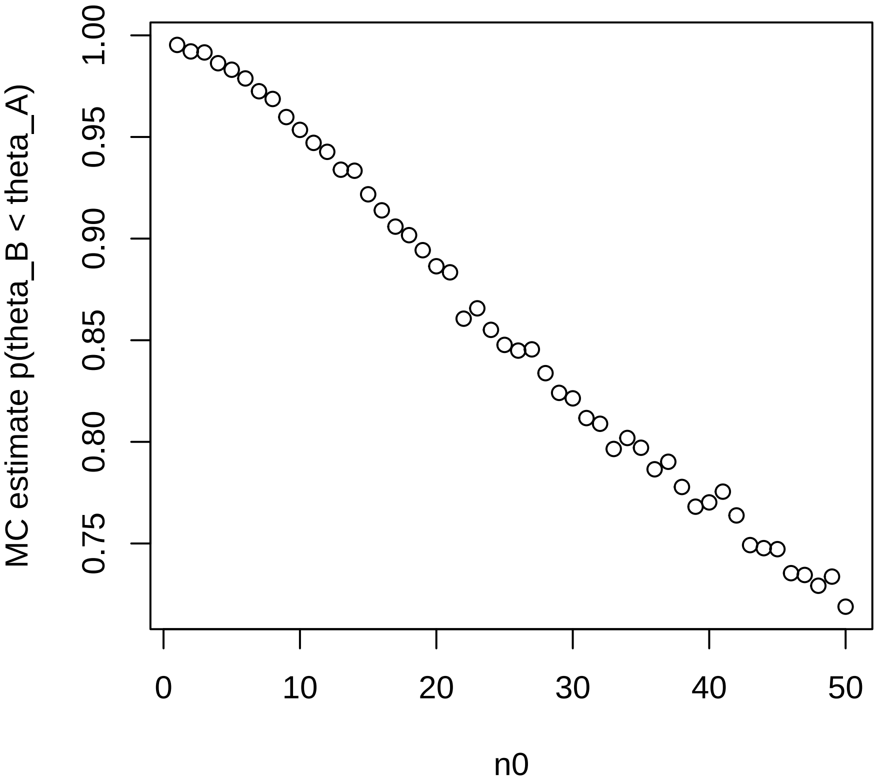
```
y_A <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
y_B <- c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)

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

```
pr = rep(0, 50)
n0s = 1:length(pr)
for(n0 in n0s) {
  theta_A = rgamma(10000, 120 + sum(y_A), 10 + length(y_A))
  theta_B = rgamma(10000, (12 * n0) + sum(y_B), n0 + length(y_B))

  pr[n0] = sum(theta_B < theta_A) / length(theta_B)
}
```

```
. = ottr::check("tests/q3a1.R")
```

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

```
# create the plot
plot(1:length(pr), pr, xlab = "n0", ylab = "MC estimate p(theta_B < theta_A)")
```



As the prior for B centered at 12 gets stronger, our posterior estimate that $p(\theta_B < \theta_A)$ gets lower.

b. Repeat the previous part replacing the event $\{\theta_B < \theta_A\}$ with the event $\{\tilde{Y}_B < \tilde{Y}_A\}$, where $\tilde{Y}_A$ and $\tilde{Y}_B$ are samples from the posterior predictive distribution.

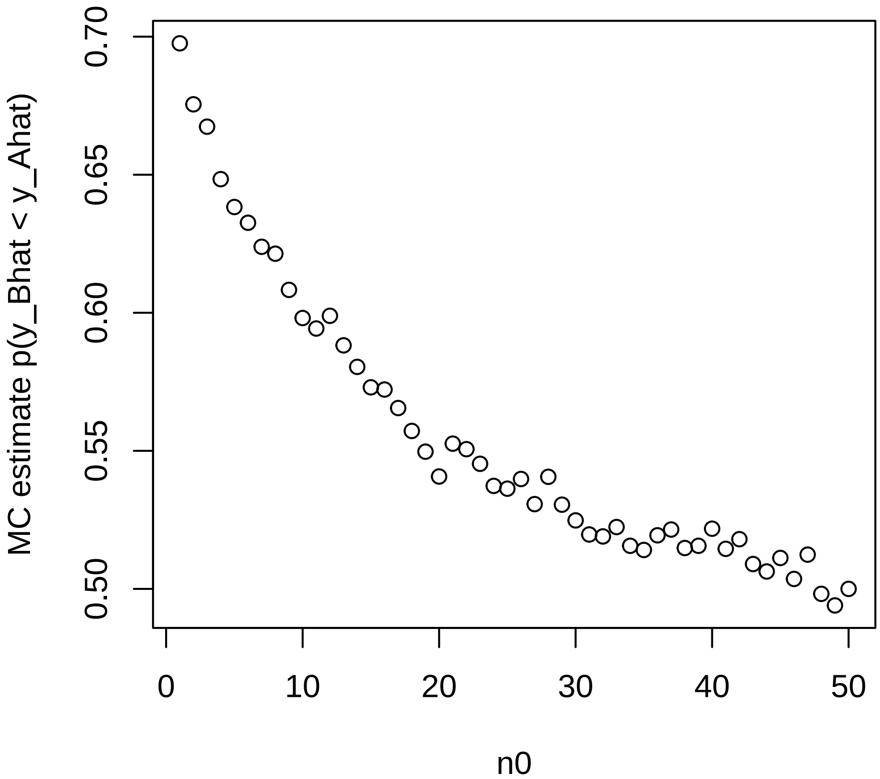
```
y_A = c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
y_B = c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)
```

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

pr = rep(0, 50)
n0s = 1:length(pr)
for(n0 in n0s) {
  theta_A = rgamma(10000, 120 + sum(y_A), 10 + length(y_A))
  theta_B = rgamma(10000, 12 * n0 + sum(y_B), n0 + length(y_B))

  y_Ahat = rpois(length(theta_A), theta_A)
  y_Bhat = rpois(length(theta_B), theta_B)

  pr[n0] = sum(y_Bhat < y_Ahat) / length(y_Bhat)
}
```

```
. = ottr::check("tests/q3b1.R")
```

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

```
# create the plot
plot(1:length(pr), pr, xlab = "n0", ylab = "MC estimate p(y_Bhat < y_Ahat)")
```



As the prior on θ_B around 12 gets stronger, the estimate of $p(\hat{y}_B < \hat{y}_A | y_B, y_A)$ gets lower.

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?

$\theta_B < \theta_A$ is the event that mice in group B acquire cancer at a lower rate than mice in group A.

$y_B < y_A$ is the event that the actual number of mice in group B with cancer is less than the actual number of mice in group A with cancer.

4. 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 $y_A^{(1)}, ..., y_A^{(1000)}$. Each $y_A^{(s)}$ is a sample of size $n_A = 10$ from the Poisson distribution with parameter $\theta_A^{(s)}$, $\theta_A^{(s)}$ is itself a sample from the posterior distribution $p(\theta_A \mid y_A)$ and y_A is the observed data. For each s, let $t^{(s)}$ be the sample average divided by the sample variance of $y_A^{(s)}$.

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

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

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

t <- c()
s = 1:1000
for (i in s){
  theta_A.posterior <- rgamma(1, 120 + sum(y_A), 10 + length(y_A))
  y_A.pred.datasets <- rpois(length(y_A), theta_A.posterior)
  t <- c(t, mean(y_A.pred.datasets)/var(y_A.pred.datasets))
}
```

```
. = ottr::check("tests/q4a1.R")
```

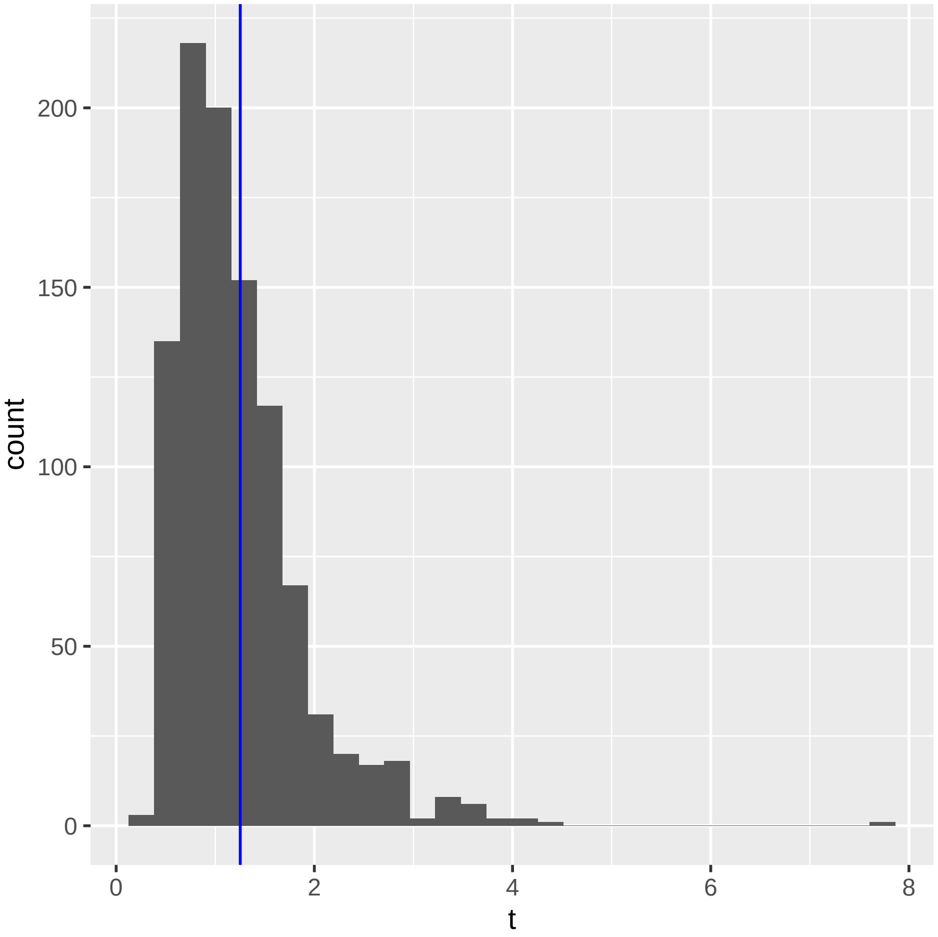
```
##
## All tests passed!
```

Given a possible Poisson model, A typical value for t^s is 1. The expected mean and variance for a poisson model is λ, thus $\lambda/\lambda = 1$.

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, $\frac{\text{mean}(y_A)}{\text{var}(y_A)}$. Based on this statistic, make a comment on if the Poisson model seems reasonable for these data (at least by this one metric).

```
# create the histogram, adding a vertical line at the observed value of the test statistic
t.df <- data.frame('S' = s, 't' = t)
ggplot(t.df, aes(t)) + geom_histogram()+ geom_vline(xintercept = mean(y_A)/var(y_A), col = "blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The Poisson model seems reasonable for these data considering the observed value is close to what we expect to be without the test statistic.

c. Repeat the part b) above for strain B mice, using Y_B and $n_B = 13$ to generate the samples. Assume the prior distribution $p(\theta_B) \sim \text{Gamma}(12, 1)$. Again make a comment on the Poisson model fit.

```
y_B = c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)
n_B = 13

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

tb <- c()
s = 1:1000
for (i in s){
  theta_B.posterior <- rgamma(1, (12*n0) + sum(y_B), n0 + n_B)
  y_B.pred.datasets <- rpois(n_B, theta_B.posterior)
  tb <- c(tb, mean(y_B.pred.datasets)/var(y_B.pred.datasets))
}

tb.df <-data.frame('S'= s, 't' = tb )
```

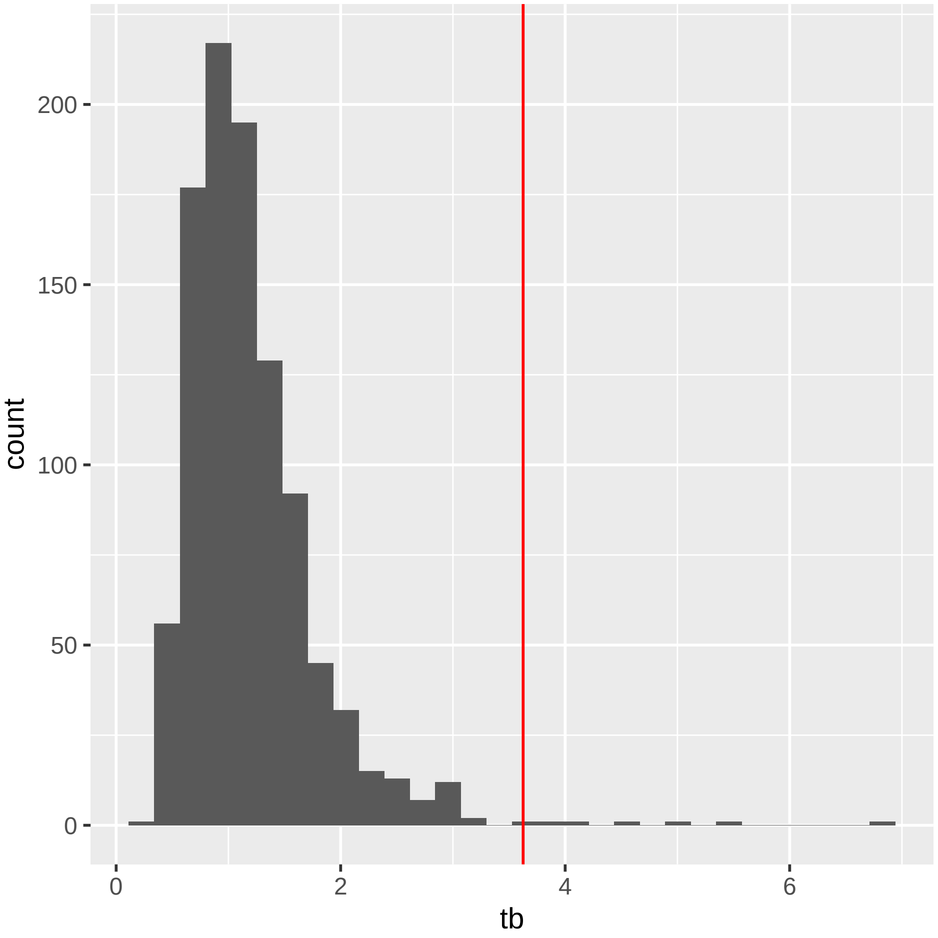
```
. = ottr::check("tests/q4c1.R")
```

```
##
```

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

```
# create the histogram, adding a vertical line at the observed value of the test statistic
ggplot(tb.df, aes(tb)) + geom_histogram() + geom_vline(xintercept = mean(y_B)/var(y_B), col = "red")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For strain B the poisson distribution does not seem reasonable as our observed value is very far from what our expected test statistic would be.