PSTAT 115 Homework 3

Aaron Barel

Due October 28, 2018

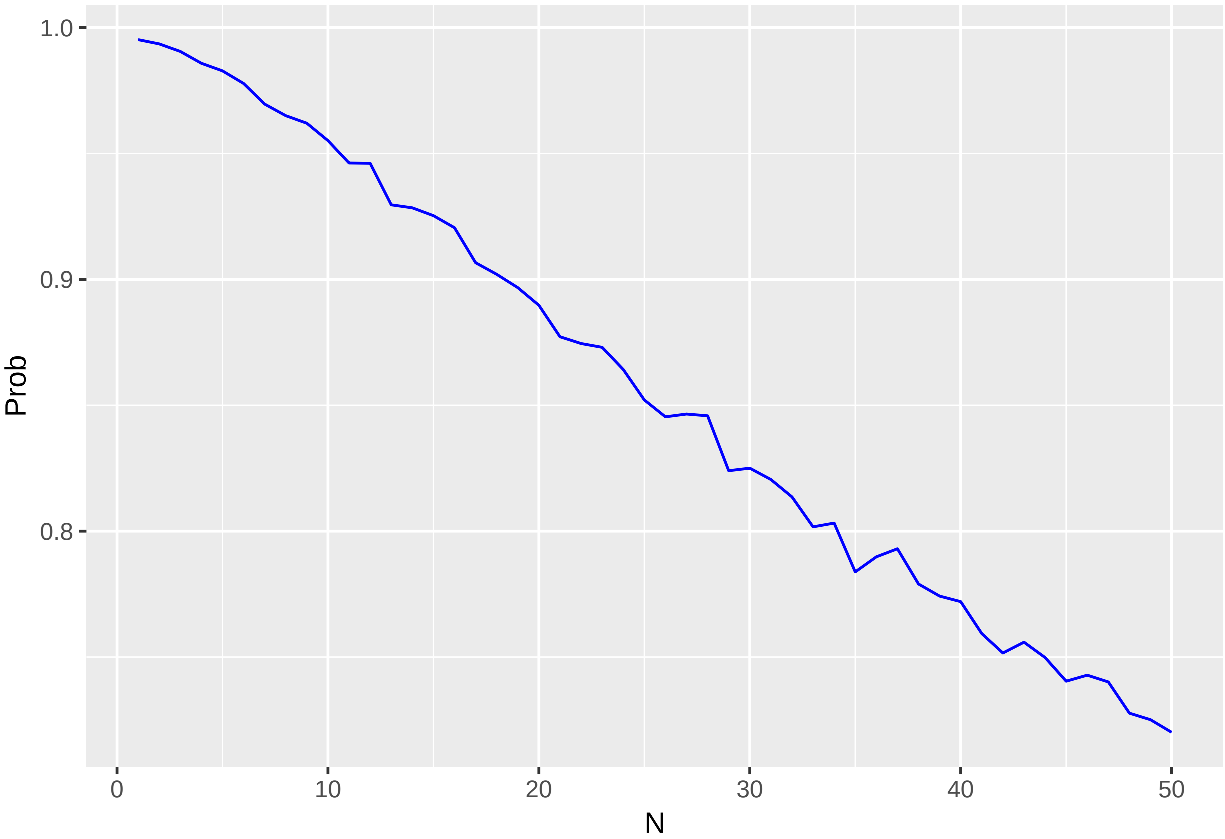
1

(a)

```
library(tidyverse)
```

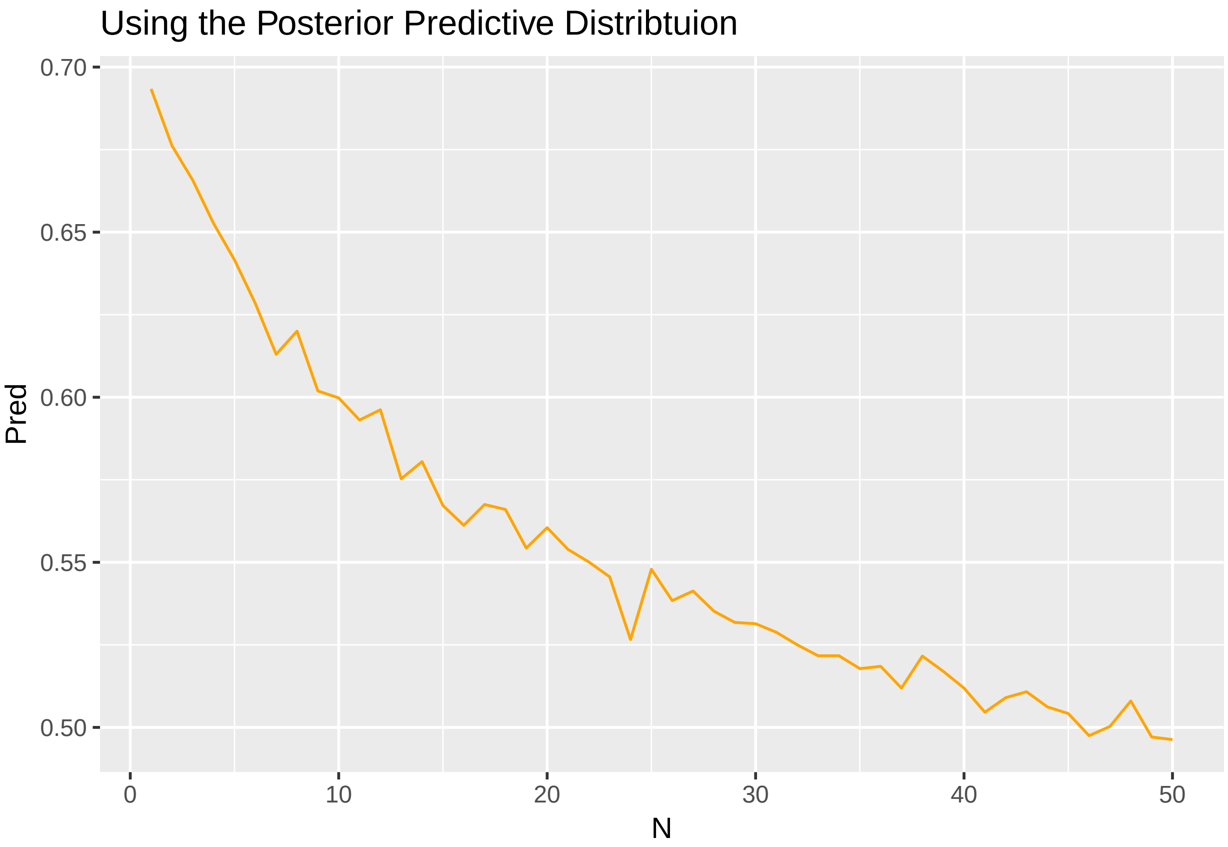
```
## -- Attaching packages ---------------------------------------------------------------------------------------------

## √ ggplot2 3.0.0     √ purrr   0.2.5
## √ tibble  1.4.2     √ dplyr   0.7.6
## √ tidyr   0.8.1     √ stringr 1.3.1
## √ readr   1.1.1     √ forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
n_0 <- c(1:50)
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)

sum.ya <- sum(y_a); sum.yb <- sum(y_b)

n_a <- length(y_a); n_b <- length(y_b)

set.seed(11111)

#proportion of theta_b < theta_a will be stored in this vector
prop <- c()

for(n in n_0){
  theta.a.posterior <- rgamma(10000, 120 + sum.ya, 10 + n_a)
  theta.b.posterior <- rgamma(10000, (12*n) + sum.yb, n + n_b)
  prop <- c(prop, mean(theta.b.posterior < theta.a.posterior))
}

prob <- data.frame("N" = n_0, "Prob" = prop)

ggplot(prob, aes(x = N, y = Prob)) + geom_line(color = 'blue')
```

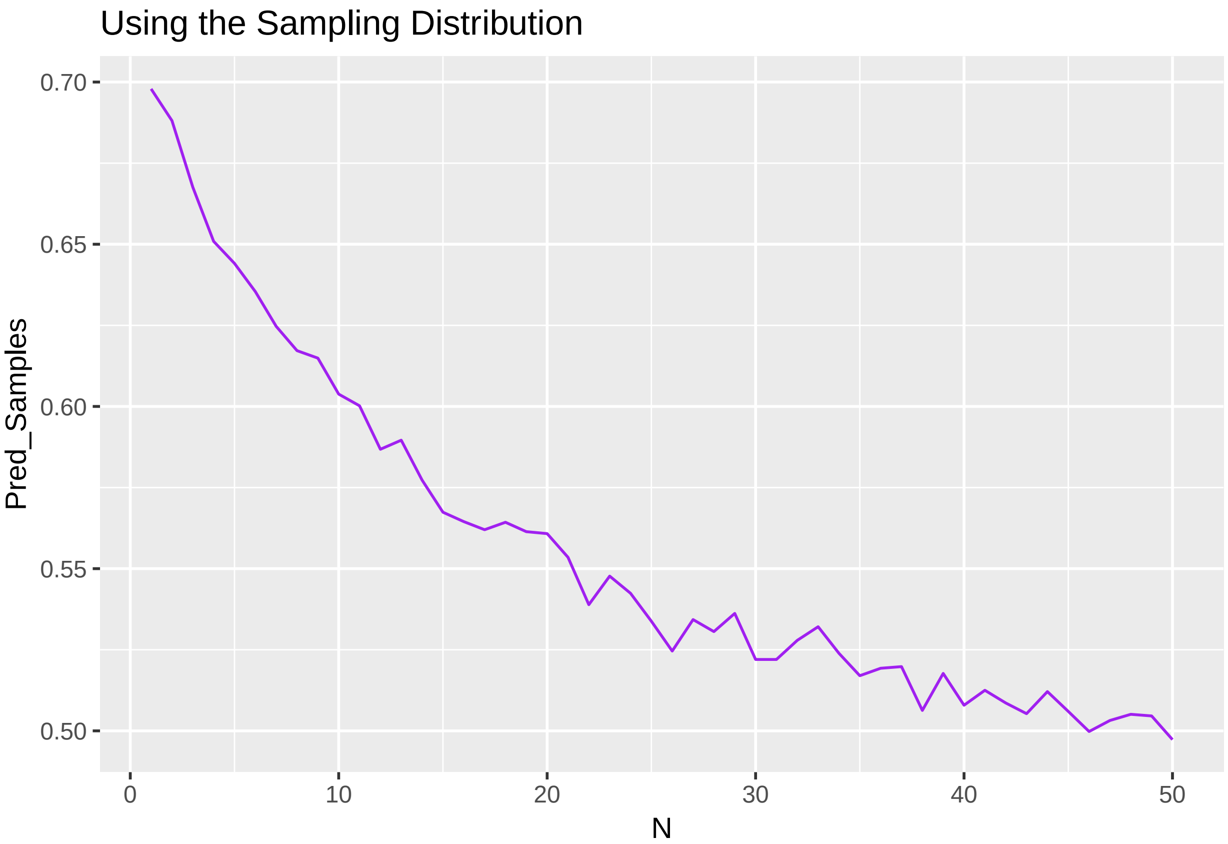


(b)

```
#proportion of the prediction with prior b is less than prediciton with prior a will be stored in this
pred.prop <- c()


#USING THE PREDICTIVE POSTERIOR DISTRIBUTION
for(n in n_0){
  y_a.pred <- rnbinom(10000, size = 120 + sum.ya, mu = (120 + sum.ya)/(10 + n_a))
  y_b.pred <- rnbinom(10000, size = (12*n) + sum.yb, mu = ((12*n) + sum.yb)/(n + n_b))
  pred.prop <- c(pred.prop, mean(y_b.pred < y_a.pred))
}

pred <- data.frame("N" = n_0, "Pred" = pred.prop)


ggplot(pred, aes(x = N, y = Pred)) + geom_line(color = 'orange') + labs(title = 'Using the Posterior Pr
```



```
#USING THE SAMPLING DISTRIBUTION

sampling.pred.vector <- c()

for(n in n_0){
  theta.a.posterior <- rgamma(10000, 120 + sum.ya, 10 + n_a)
  theta.b.posterior <- rgamma(10000, (12*n) + sum.yb, n + n_b)
  #USE EXPECTED VALUE OF THETA WHEN SAMPLING
  sample.pred.ya <- rpois(10000, mean(theta.a.posterior))
  sample.pred.yb <- rpois(10000, mean(theta.b.posterior))
  sampling.pred.vector <- c(sampling.pred.vector, mean(sample.pred.yb < sample.pred.ya))
}

pred.samples.pois <- data.frame("N" = n_0, 'Pred_Samples' = sampling.pred.vector)


ggplot(pred.samples.pois, aes(x = N, y = Pred_Samples)) + geom_line(color = 'purple') + labs(title = 'U
```



```
#NOTE THE PLOTS ARE VERY SIMILAR AS EXPECTED
```

(c) In the context of this problem, the event $\{\theta_B < \theta_A\}$ is the event that the rate of tumorigenesis in group B is smaller than the rate of tumorigenesis in group A given the data and prior information about the rates. The event $\{\widetilde{Y_B} < \widetilde{Y_A}\}$ is the event that we predict number of mice with tumors from group B is less than group A given the data and prior information of the rates.

2

```
#Model checking process
t <- c()
s = 1:1000
for(i in s){
  theta.a.posterior <- rgamma(1, 120 + sum.ya, 10 + n_a)
  y_a.pred.datasets <- rpois(n_a, theta.a.posterior)
  t <- c(t, mean(y_a.pred.datasets)/var(y_a.pred.datasets))
}
```

(a)

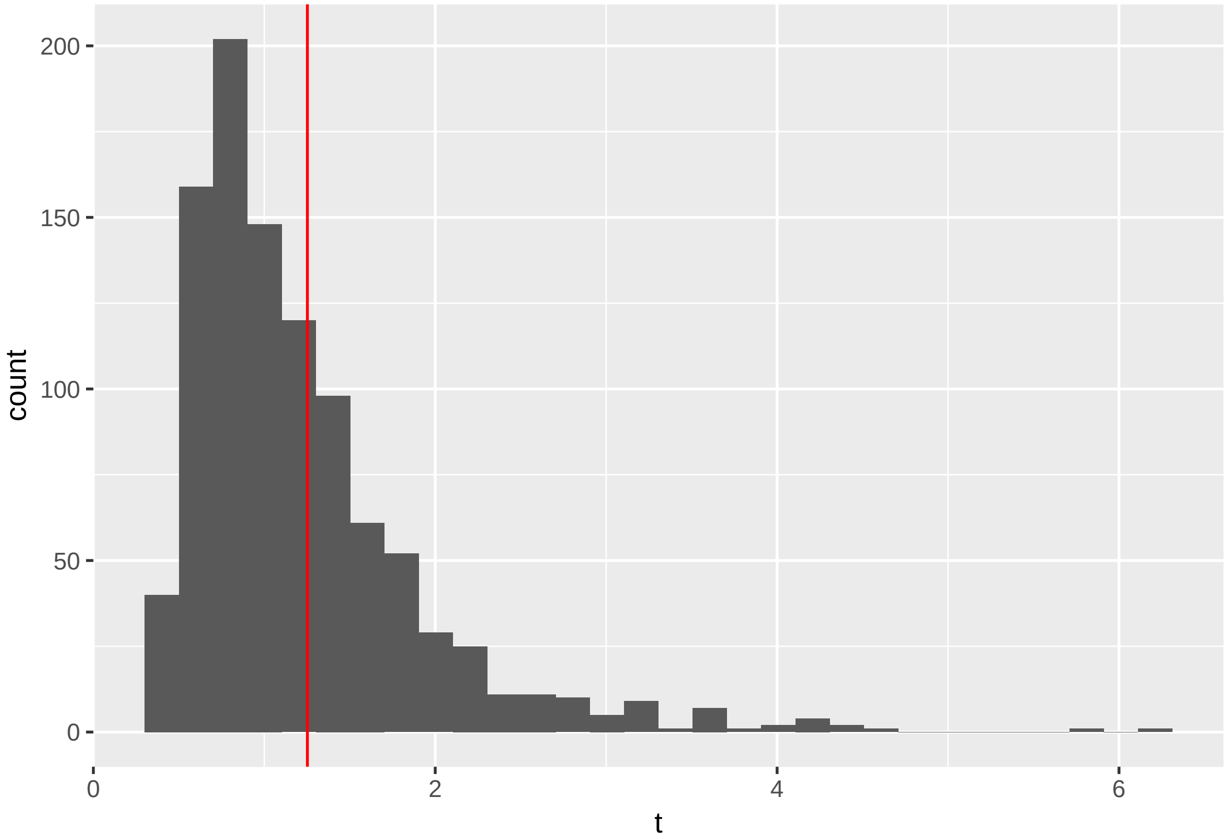
If the Poisson model is appropriate, a typical value of $t^{(s)}$ is 1 because the expected value and variance of $Pois(\lambda)$ are both λ thus $\frac{\lambda}{\lambda} = 1$.

(b)

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

```
ggplot(t.df, aes(t)) + geom_histogram() + geom_vline(xintercept = mean(y_a)/var(y_a), color = 'red')
```

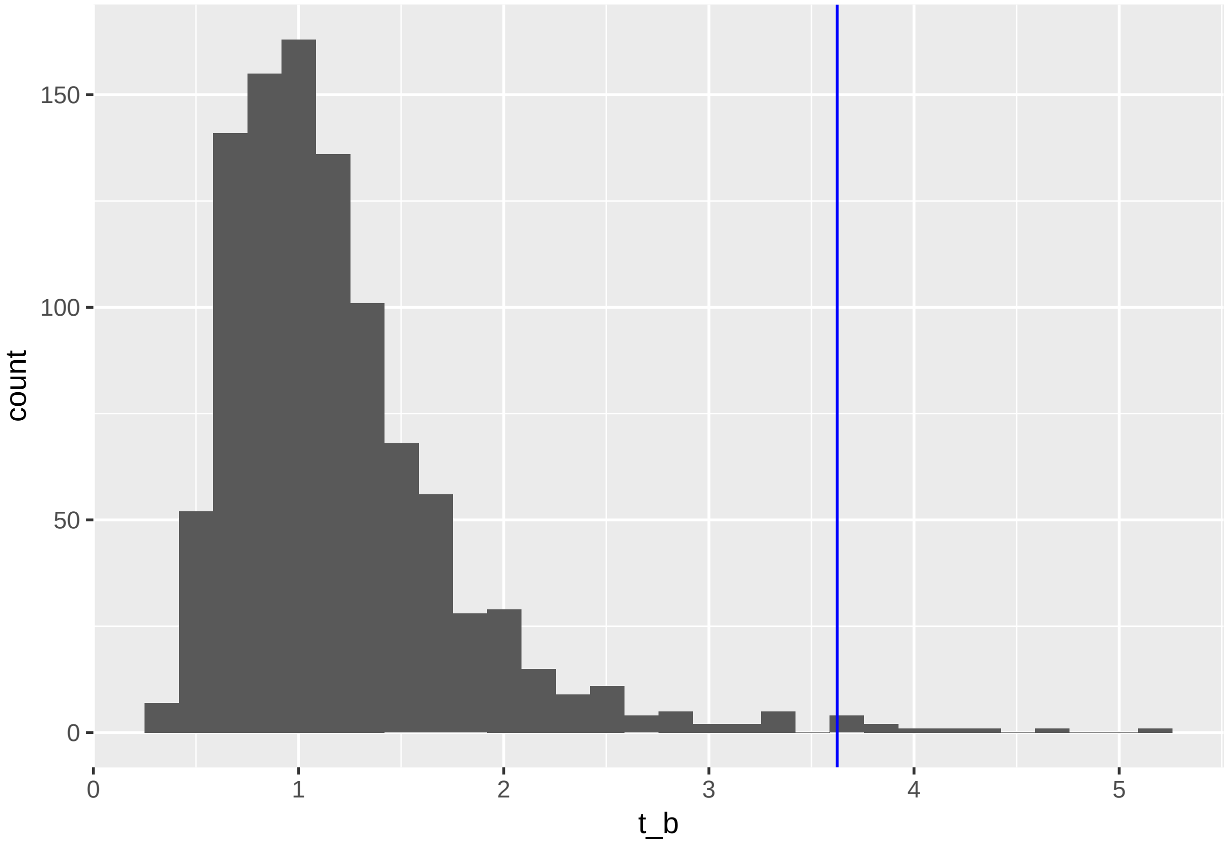
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The model seems to work be reasonalbe since the observed value is close to what we expected it to be with out test statistic.

```
t_b <- c()
for(i in s){
  theta.b.posterior <- rgamma(1, (12*n) + sum.yb, n + n_b)
  y_b.pred.datasets <- rpois(n_b, theta.b.posterior)
  t_b <- c(t_b, mean(y_b.pred.datasets)/var(y_b.pred.datasets))
}

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

ggplot(t_b.df, aes(t_b)) + geom_histogram() + geom_vline(xintercept = mean(y_b)/var(y_b), color = 'blue
```

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



For strain B, we do not have as favorable results by this metric. Our observed value is far from what our test statistic is expected to be.