Homework 3

PSTAT 115, Fall 2021

Due on November 14, 2020 at 11:59 pm

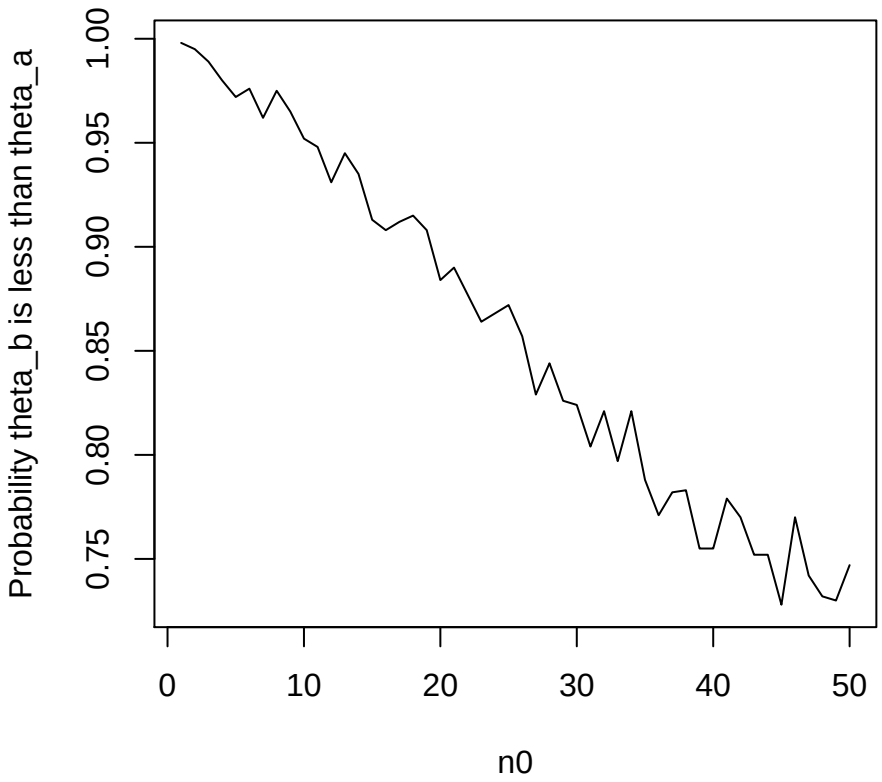
Note: If you are working with a partner, please submit only one homework per group with both names. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

Problem 1. 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 $\theta_A \sim$ gamma$(120, 10)$ and $\theta_B \sim$ 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 $\theta_A \sim$ gamma$(120, 10)$ and $\theta_B \sim$ 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.

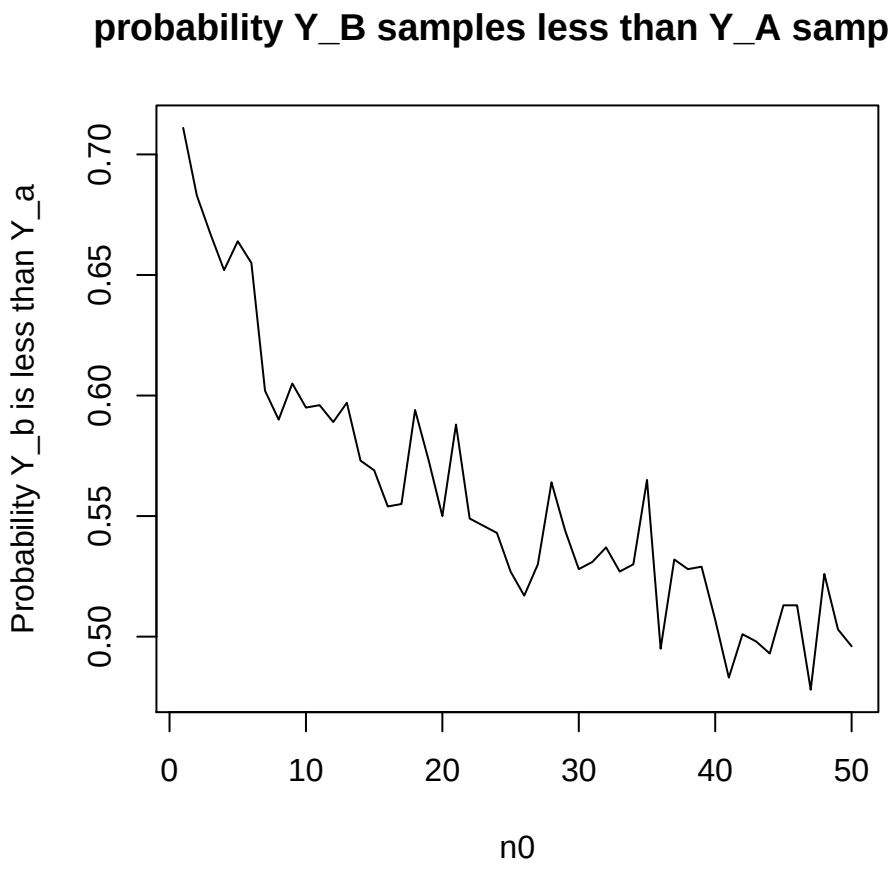
```
set.seed(50)
y_A <- c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
y_B <- c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)
n0 <- 50
t_s <-numeric(n0)
for(i in 1:n0){
  theta_a <- rgamma(n=1000,120+sum(y_A),10+length(y_A))
  theta_b <- rgamma(n=1000,12*i+sum(y_B),i+length(y_B))
  t_s[i] <- mean(theta_b < theta_a)
}
plot(1:n0,t_s,"l",lty=1,xlab="n0",ylab="Probability theta_b is less than theta_a")
```



As the number of n_0 increases, the probability that θ_b will be less than θ_a decreases. It can be seen that the conclusions are not sensitive to the prior distribution of θ_b as n_0 increases each time θ_b has a smaller chance of being greater than θ_a.

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)
t_s <-numeric(n0)
set.seed(1)
for(i in 1:n0){
  theta_a <- rgamma(n=1000,120+sum(y_A),10+length(y_A))
  theta_b <- rgamma(n=1000,12*i+sum(y_B),i+length(y_B))
  y_a <- rpois(1000,theta_a)
  y_b <- rpois(1000,theta_b)
  t_s[i] <- mean(y_b < y_a)
}
plot(1:n0,t_s,"l",lty=1,xlab="n0",ylab="Probability Y_b is less than Y_a",main="probability Y_B samples
```



Looking at this graph we can see that the probability results shrinks towards 0.5. This leads us to believe this is more sensitive to prior distribution θ_b compared to that of the event in part a).

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?

$\{\tilde{Y}_B < \tilde{Y}_A\}$ is the number of mice that have tumors forming. We check this to see if while the prior distribution increases so does the amount of mice found to have tumors in group B. $\{\theta_B < \theta_A\}$ is in terms of the chances that a mice will have a tumor. As that distribution increases there was not a great change in these events compared to the events found in part b).

2. 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?

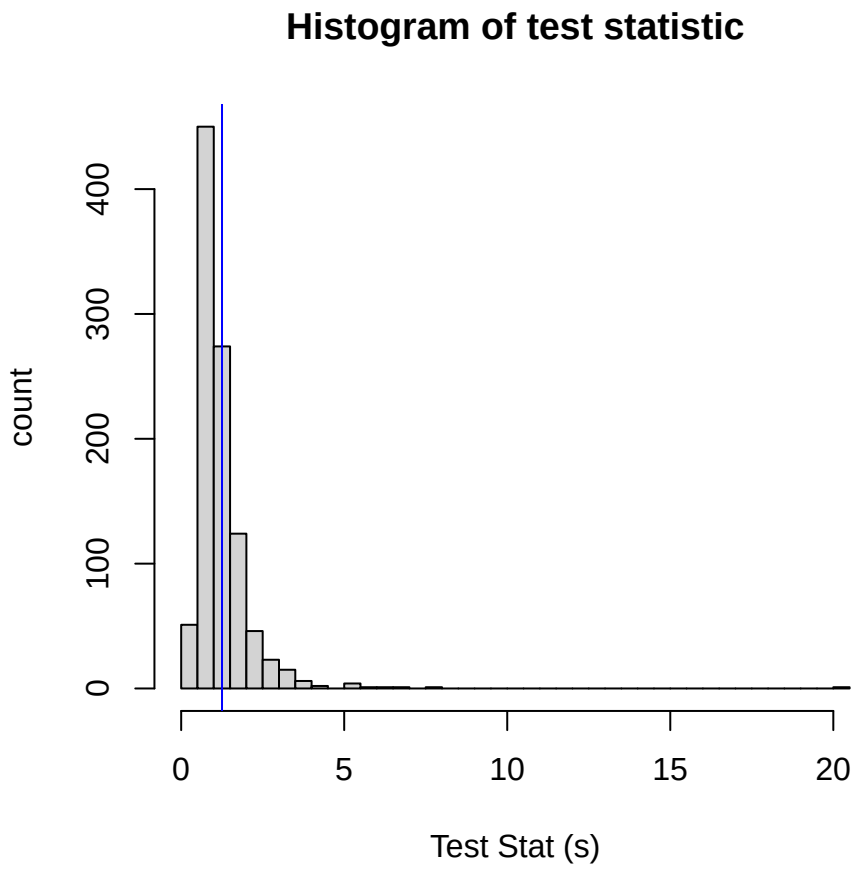
```
y_A = c(12, 9, 12, 14, 13, 13, 15, 8, 15, 6)
n_A <- 10
S <- 1000
t_s <-numeric(n0)
```

```r
set.seed(1)
for(s in 1:S){
  theta_a <- rgamma(n=n_A,120+sum(y_A),10+length(y_A))
  y_a <- rpois(n_A,theta_a)
  t_s[s] <- mean(y_a)/var(y_a)
}
```

A typical value of $t^{(s)}$ should be 1 since the mean and variance of a poisson distribution is lambda. So when we divide the mean by the varaince it should result in 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).

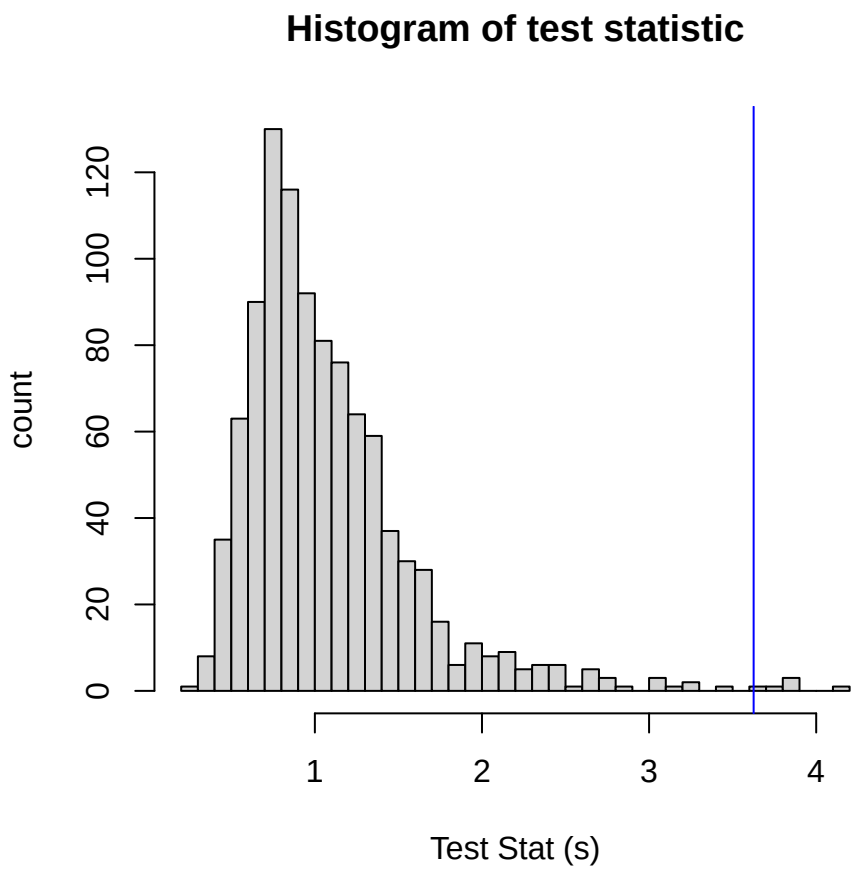
```r
hist(t_s,main="Histogram of test statistic",xlab="Test Stat (s)", ylab="count",breaks=50)
abline(v=mean(y_A)/var(y_A),col="blue")
```



Looking at the histogram, it would seem that our test statistic is relatively close to that of the observed test statistic. I believe that the poisson model is appropriate for this data.

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.

```r
y_B = c(11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)
n_B <- 13
S <- 1000
t_s <-numeric(n0)
set.seed(1)
for(s in 1:S){
  theta_b <- rgamma(n=n_B,12+sum(y_B),1+length(y_A))
  y_b <- rpois(n_B,theta_b)
  t_s[s] <- mean(y_b)/var(y_b)
}
hist(t_s,main="Histogram of test statistic",xlab="Test Stat (s)", ylab="count",breaks=50)
abline(v=mean(y_B)/var(y_B),col="blue")
```



Looking at this histogram, it would appear that the poisson model is not appropriate for this data. The average for our test stat was around 1 where as the actual observed test stat was more like 3.75.

3. Interval estimation with rejection sampling.

a. Use rejection sampling to sample from the following density:

$$p(x) = \frac{1}{4}|\sin(x)| \times I\{x \in [0, 2\pi]\}$$

Use a proposal density which is uniform from 0 to 2π and generate at least 1000 true samples from p(x).

Compute and report the Monte Carlo estimate of the upper and lower bound for the 50% quantile interval using the `quantile` function on your samples. Compare this to the 50% HPD region calculated on the samples. What are the bounds on the HPD region? Report the length of the quantile interval and the total length of the HPD region. What explains the difference? Hint: to compute the HPD use the `hdi` function from the `HDInterval` package. As the first argument pass in `density(samples)`, where `samples` is the name of your vector of true samples from the density. Set the `allowSplit` argument to true and use the `credMass` argument to set the total probability mass in the HPD region to 50%.

b. Plot $p(x)$ using the `curve` function (base plotting) or `stat_function` (ggplot). Add lines corresponding to the intervals / probability regions computed in the previous part to your plot using them `segments` function. To ensure that the lines don't overlap visually, for the HPD region set `y0` and `y1` to 0 and for the quantile interval set `y0` and `y1` to 0.01. Make the segments for HPD region and the segment for quantile interval different colors. Report the length of the quantile interval and the total length of the HPD region, verifying that indeed the HPD region is smaller.

Answer to part a) and b)

```
library("HDInterval")
n <- 1000
p_den <- function(x) {
0.25 * abs(sin(x)) * dunif(x, 0, 2 * pi)
}
q_den <- function(x) { dunif(x, 0, 2 * pi)}
d_ratio <- function(x) { p_den(x) / q_den(x) }
M <- optimize(d_ratio, lower = 0, upper = 2*pi, maximum = TRUE)$objective
sample <- runif(n, 0, 2*pi)
accept <- runif(n) < (d_ratio(sample) / M)
samples <- sample[accept]
  # hd_region is the result of calling hdi function
  hd_region <- hdi(density(samples), allowSplit = TRUE, credMass = 0.5)
  print(hd_region)
```

```
##         begin      end
## [1,] 1.066547 2.039832
## [2,] 4.208379 5.471943
## attr(,"credMass")
## [1] 0.5
## attr(,"height")
## [1] 0.193253
```

```
  print(sprintf("Total region length: %.02f", sum(hd_region[, "end"] - hd_region[,"begin"])))
```

```
## [1] "Total region length: 2.24"
```

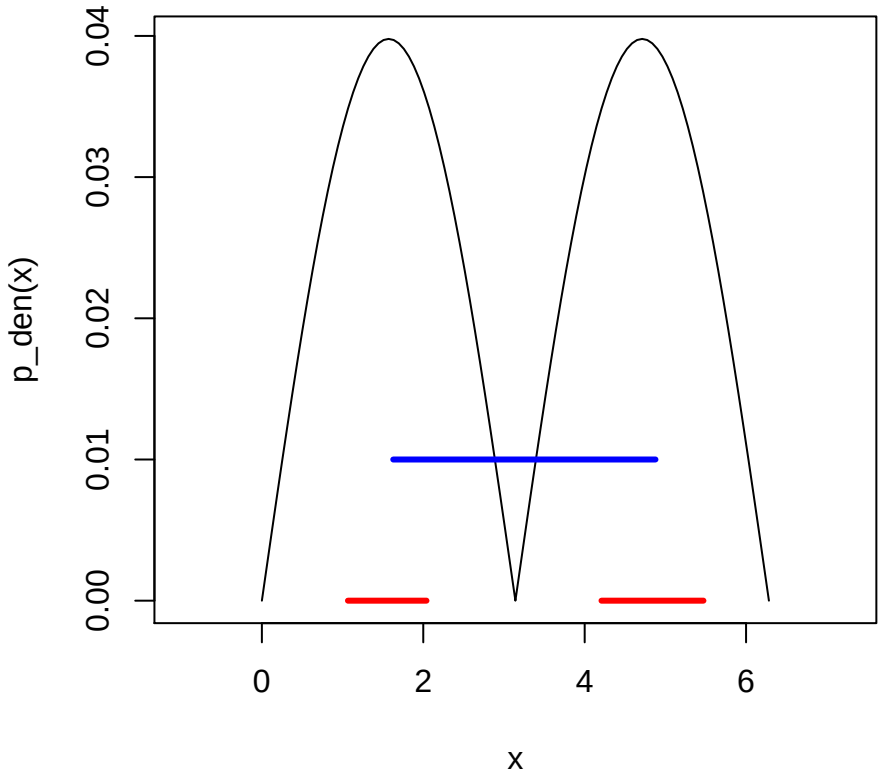
```
  quantile_interval <- quantile(samples, c(0.25, 0.75))
  print(quantile_interval)
```

```
##      25%      75%
## 1.627722 4.877534
```

```
  print(sprintf("Total region length: %.02f", quantile_interval[2] - quantile_interval[1]))
```

```
## [1] "Total region length: 3.25"
```

```
curve(p_den(x), from=0, to=2*pi, xlim=c(-1, 2*pi + 1))
segments(x0=hd_region[1,1], y0=0, x1=hd_region[1,2], y1 = 0, col="red", lwd=3)
segments(x0=hd_region[2,1], y0=0, x1=hd_region[2,2], y1 = 0, col="red", lwd=3)
segments(x0=quantile_interval[1], y0=0.01, x1=quantile_interval[2], y1=0.01, col="blue", lwd=3)
```




```
quantile_interval[2] - quantile_interval[1]
```

```
##      75%
## 3.249812
```

```
sum(hd_region[,"end"] - hd_region[,"begin"])
```

```
## [1] 2.236848
```

The HPD can be seen as smaller than the quantile interval. This is due to the HPD containing the highest point of density. So when looking for 50% of the distribution, the HPD requires a smaller interval compared to that of the quantile interval. The area underneath HPD is much more compared to the quantile interval since the quantile does not always contain the highest point of density.