Final Exam

Brody Vogel

12/12/2017

Part I

1

a)

If $f(x) = c(A^2 - x^2)$ on the interval $[-A, A]$ is a probability distribution, then $\int_{-A}^{A} c(A^2 - x^2)dx = 1$, by an extension of the Law of Total Probability.

Going step by step: $\int_{-A}^{A} c(A^2 - x^2)dx = 1$

$c\int_{-A}^{A}(A^2 - x^2)dx = 1$

The integral: $\int(A^2 - x^2)dx = A^2x - \frac{x^3}{3} + Constant.$

The definite integral, then, is: $\frac{3A^3}{3} - \frac{A^3}{3} - (\frac{-3A^3}{3} - \frac{-A^3}{3}) = \frac{2A^3}{3} + \frac{2A^3}{3} = \frac{4A^3}{3}$.

Factoring c back in, $1 = \frac{4A^3c}{3} <==> \frac{3}{4A^3} = c$.

So, for this to be a true probability distribution, $c = \frac{3}{4A^3}$

b)

Because expectation is linear, and by the Central Limit Theorem, $E[\overline{X}] = E[\frac{X_1}{n} + ... + \frac{X_n}{n}] = \frac{nE[X]}{n} = E[X]$. In this case, $E[X] = \int_{-A}^{A} x\frac{3}{4A^3}(A^2 - x^2)dx$.

After calculus, $E[X] = \frac{-3(A^2-A^2)^2}{16A^3} - (\frac{-3(A^2-(-A^2))^2}{16A^3}) = 0 - 0 = 0$.

Furthermore, $Var(X) = E[X^2] - (E[X])^2$, as a simplification of the second central moment of a random variable, and $Var(\overline{X}) = \frac{Var(X)}{n}$, because $Var(\overline{X}) = Var(\frac{X_1}{n} + ... + \frac{X_n}{n}) = \frac{Var(X)}{n}$.

(Or, put simply, $Var(\overline{X}) = (SE_{\overline{X}})^2 = (\frac{\sigma_X}{\sqrt{n}})^2 = \frac{Var(X)}{n}$).

Moving on, $E[X^2] = \int_{-A}^{A} x^2\frac{3}{4A^3}(A^2 - x^2)dx$

After more calculus, $E[X^2] = (\frac{A^3}{4A} - \frac{3A^5}{20A^3}) - (\frac{-A^3}{4A} - \frac{-3A^5}{20A^3}) = (\frac{A^2}{4} - \frac{3A^2}{20}) - (\frac{-A^2}{4} - \frac{-3A^2}{20}) = \frac{2A^2}{4} - \frac{6A^2}{20} = \frac{A^2}{5}$.

So $Var(X) = \frac{A^2}{5} - 0^2 = \frac{A^2}{5}$, and so, using what we know from above, $Var(\overline{X}) = \frac{A^2}{5n}$.

Finally, for n large, the distribution of $\overline{X}$ is approximately Normal with $\mu = E[X] = 0$ and $\sigma^2 = \frac{A^2}{5n}$, by the Central Limit Theorem.

2

a)

$$P(Y_i = y) = \begin{cases} p + (1-p)e^{-\lambda}, & y = 0 \\ (1-p)e^{-\lambda}\frac{\lambda^y}{y!}, & y > 0 \end{cases}$$

b)

The likelihood function, accordingly, is:

$L(\lambda|Y_i s = 0,0,1,1,2,2,2,4,4,5) = [p + (1-p)e^{-\lambda}]^2 \times (1-p)^8 \times e^{-8\lambda} \times \frac{\lambda^{\sum_{i=3}^{10} Y_i}}{\prod_{i=3}^{10} Y_i!}$, which, from the given, is:

$L(\lambda|Y_i s = 0,0,1,1,2,2,2,4,4,5) = [\frac{1}{3} + \frac{2e^{-\lambda}}{3}]^2 \times (\frac{2}{3})^8 \times e^{-8\lambda} \times \frac{\lambda^{21}}{552960}$.

The log likelihood:

$ln(L(\lambda|Y_i s = 0,0,1,1,2,2,2,4,4,5)) = 2ln(\frac{1}{3} + \frac{2e^{-\lambda}}{3}) + 8ln(\frac{2}{3}) - 8\lambda + 21ln(\lambda) - ln(552960)$.

Derivative of the log likelihood:

$\frac{d}{d\lambda} ln(L(\lambda|Y_i s = 0,0,1,1,2,2,2,4,4,5)) = \frac{-4}{e^{\lambda}+2} + 0 - 8 + \frac{21}{\lambda} - 0 = \frac{-4}{e^{\lambda}+2} - 8 + \frac{21}{\lambda}$.

Finding maximum:

$0 = \frac{-4}{e^{\lambda}+2} - 8 + \frac{21}{\lambda}$

<==>

$\lambda = \frac{-4\lambda}{8e^{\lambda}+16} + \frac{21}{8}$.

Using the iterative method:

Guess: $\lambda = 2.5$; Output: 2.53836

Guess: $\lambda = 2.53836$; Output: 2.53838

So, $\hat{\lambda}_{MLE} \approx 2.54$.

3

The simulations:

```
set.seed(runif(1, 1000, 2000))
## Generate random numbers for the sample
helper1 <- function() {
  samp <- sample(1:101, 15, replace = F)

  ## Here's m
  m <- median(samp)

  ## Find how often the median of the bootstrapped sample is equal to m
  outputs <- c()
  for (x in 1:20000) {
    boot <- sample(samp, 7, replace = T)
    outputs <- c(outputs, median(boot))
  }


max(table(outputs)[toString(m)]) == max(table(outputs))
}

## This tells the percentage of simulations in which the most
  ## frequent bootstrapped median was equal to m
outcome <- replicate(100, helper1())
length(outcome[outcome == T])/length(outcome)
```

```
## [1] 0.98
```

So, from the samples, it looks like it is the case that E[X] always equals m, as the median was most likely to be equal to m in over 90 percent of the simulations (this number rises for higher numbers of simulations).

(Also, I used the specifics in the simulation - a sample of size 15 and bootstrapped sample of 7 - to reduce the number of simulations needed to get the data to converge.)

From theory, we know the expected value of a random variable is the average, long-run (over many trials) value of said random variable. So, in this case, we know the expected value of the median of a bootstrapped sample is the long-run median of that bootstrapped sample. In other words, the expected value is the median of an infinitely large bootstrapped sample. This sample would be: $[\infty$ *repetitions* x_1, ∞ *repetitions* $x_2, ..\infty$ *repetitions* $x_{\frac{n}{2}+1}, ...\infty$ *repetitions* $x_n]$. The median of this sample is, accordingly, $x_{\frac{n}{2}+1}$, which is equal to m.

4

The simulations:

```
set.seed(runif(1, 1000, 2000))
ys <- c()
while (length(ys) < 500) {
  x <- rnorm(1, mean = 0, sd = 1)
  if (x > 1) {
    ys <- c(ys, x^-1)
  }
}

mean(ys)
```

```
## [1] 0.703
```

```
var(ys)
```

```
## [1] 0.02942911
```

The rule of thumb for getting a good estimation of a population's distribution is to use $n = 30$ simulations. According to Chihara and Hesterberg, though, it's safer to use many more than that; they say it can take more than 2000 simulations for a highly-skewed distribution. I looked at the histogram of 1000 samples of Y, and it didn't show much skew, so I went with half as many simulations - 500.

This produced a simulated distribution with $E[Y] = \mu_{sim} \approx .70$ and $Var(Y) = \sigma^2_{sim} = .03$.

My reasoning for believing these to be correct is as follows:

—E[Y]—

- 1 is exactly 1 standard deviation from the mean of X~N(0,1), in the positive direction. We know from the standard normal table that ≈ 84.2 percent of the standard normal distribution falls to the left of 1 standard deviation in the positive direction.
- Using qnorm(.842), we find the lower bound of our forthcoming integral to be 1.002712.
- We know that $f(X|X > 1) = \frac{f(X)}{Pr(X>1)} = f(x) \times Pr(X > 1)^{-1} = f(x) \times .158^{-1} = f(x) \times 6.32911$.
- We also know that $E[g(f(x))] = \int g(x)f(x)$.
- Putting the last three points together, $f(y) = \int_{1.002712}^{\infty} Pr(X > 1)^{-1} g(f(x))$, where $g(x) = x^{-1}$.

- So, because we know $E[Y] = \int_{1.002712}^{\infty} g(y) \times f(y)$, $E[Y] = \int_{1.002712}^{\infty} 6.32911 \times y^{-1} \times \frac{1}{\sqrt{2\pi}} \times e^{\frac{-y^2}{2}} \approx .7023$, which is the same as the result of my simulations.

—Var(Y)—

- We know $Var(Y) = E[Y^2] - (E[Y])^2$.
- We just found $E[Y]$, and we know $E[Y^2] = \int y^2 f(y)$, which in this case will be $E[Y^2] = \int g(y)^2 f(y) = E[Y^2] = \int (y^{-1})^2 f(y)$
- In this case, then, $Var(Y) = \int_{1.002712}^{\infty} 6.32911 \times (y^{-1})^2 \times \frac{1}{\sqrt{2\pi}} \times e^{\frac{-y^2}{2}} - (\int_{1.002712}^{\infty} 6.32911 \times y^{-1} \times \frac{1}{\sqrt{2\pi}} \times e^{\frac{-y^2}{2}})^2 = .52317 - .7023^2 = .029945$, which is also the same as the result of my simulations.

Part II

5

H_0 : The age of the mother and the gestation period have no effect on each other; that is, they are independent.

H_a : The age of the mother and the gestation period are *not* independent; that is, there's reason to believe they have an effect on each other.

a)

From the permutation test, it looks like the mother's age and the gestation period *are* independent. To set up the test, I found the average gestation period for mothers aged: 19 and under, 20-24, 25-29, 30-34, 35-39, and over 40. I then summed the absolute values of the differences between those averages and the overall average gestation period. This was my test statistic: The summation of the differences between the overall gestation period and the average gestation period specific to an age group. My test statistic, then, had a value of $\approx .389$. I then ran a permutation test that randomly assigned gestation periods to 6 appropriately-sized groups and used the just-mentioned method to generate a distribution of 10,000 test statistics. The p-value corresponding to the test statistic from the true data, then, was .72, which is very large. This means that, in the simulated distribution, over 72% of the test statistics were at least as extreme as that from the data. This - and the histogram - provide evidence that the two variables are indeed independent.

```
set.seed(runif(1, 1000, 2000))

## Load the data
births <- read.csv('/Users/brodyvogel/Desktop/Data/NCBirths2004.csv')

## Group average
gestAVG <- mean(births$Gestation)

## Subgroups
a <- births[births$MothersAge %in% c('under 15', '15-19'), 7]
b <- births[births$MothersAge == '20-24', 7]
c <- births[births$MothersAge == '25-29', 7]
d <- births[births$MothersAge == '30-34', 7]
e <- births[births$MothersAge == '35-39', 7]
f <- births[births$MothersAge %in% c('40-44', '45-49'), 7]

## Test Statistic
testStat <- abs(mean(a) - gestAVG) + abs(mean(b) - gestAVG) + abs(mean(c) - gestAVG) +
            abs(mean(d) - gestAVG) + abs(mean(e) - gestAVG) + abs(mean(f) - gestAVG)
```

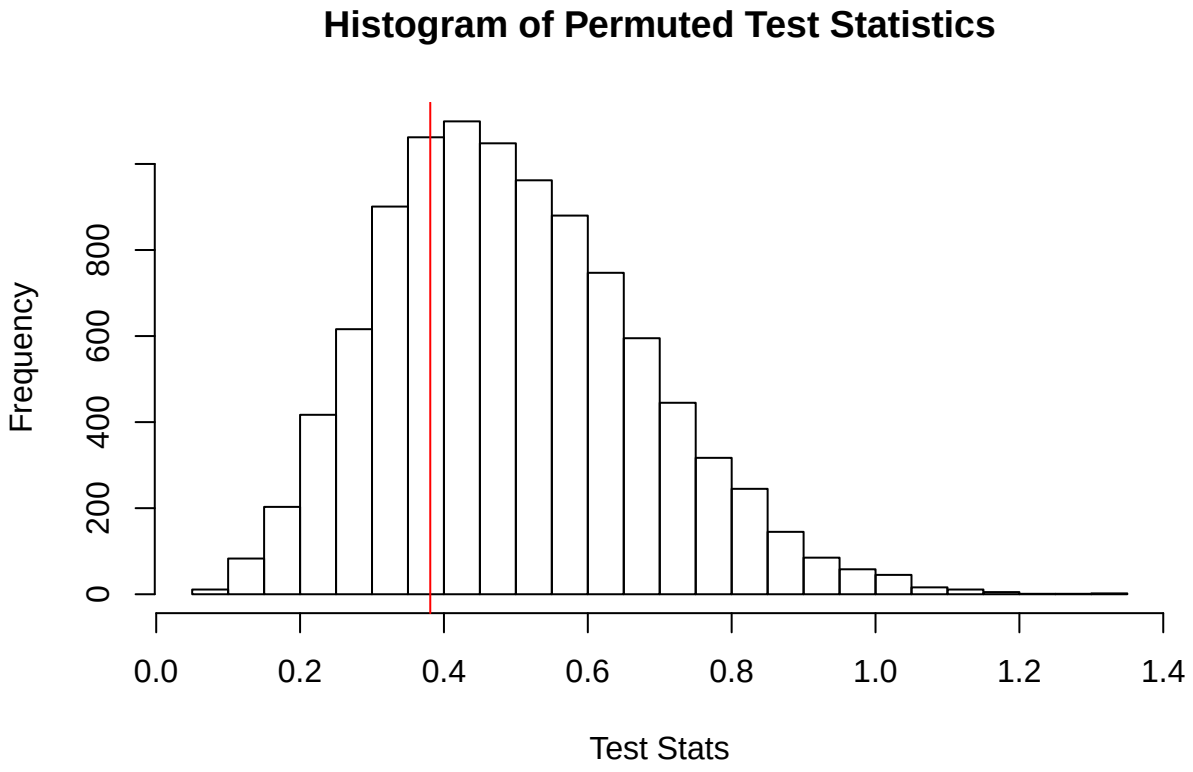
```
testStat
```

```
## [1] 0.3809769
```

```
## Permutations
testStats <- c()
for (x in 1:10000) {
  samp <- sample(births$Gestation, length(births$Gestation), replace = F)

  g <- samp[1:length(a)]
  h <- samp[length(a):(length(a) + length(b))]
  i <- samp[(length(a) + length(b)):(length(a) + length(b) + length(c))]
  j <- samp[(length(a) + length(b) + length(c)):(length(a) + length(b) + length(c) + length(d))]
  k <- samp[(length(a) + length(b) + length(c) + length(d)):(length(a) + length(b) + length(c) + length
  l <- samp[(length(a) + length(b) + length(c) + length(d) + length(e)):(length(a) + length(b) + length

  holder <- abs(mean(g) - gestAVG) + abs(mean(h) - gestAVG) + abs(mean(i) - gestAVG) +
            abs(mean(j) - gestAVG) + abs(mean(k) - gestAVG) + abs(mean(l) - gestAVG)
  testStats <- c(testStats, holder)
}

## p-value from the permutation tests
p1 <- length(testStats[testStats >= testStat])/length(testStats)

p1
```

```
## [1] 0.7127
```

```
## Histogram with line representing the test statistic
hist(testStats, main = 'Histogram of Permuted Test Statistics', xlab = 'Test Stats', breaks = 20)
abline(v = testStat, col = 'red')
```



b)

I set up the X^2 test a bit differently from the permutation test: To get expected cell counts, I used the actual variables instead of percentages. This means that my two-way table had 6 columns (37, 38, 39, 40, 41, and 42-week gestation periods) and 6 rows (mothers aged 19 and under, 20-24, 25-29, 30-34, 35-39, and over 40). Each of the week buckets have more than 5 values, so none were merged. I merged the 'under 15' age group with the '15-19', and the '40-44' with the '45-49', because the 'under 15' and '45-49' groups had only 2 values. I then found the expected counts of each (row, column) pair.

Performing the X^2 test (summing the $X^2 = \sum \frac{(Expected-Observed)^2}{Expected}$ statistics) produced a X^2 value of 21.49 with (6-1)(6-1) = 25 degrees of freedom. Using the X^2 Approximation, this corresponds to a p-value of .33, which is large. Thus, as in part (a), there is evidence that the age of mothers and gestation period *are* independent.

```
table(births$Gestation)
```

```
##
##  37  38  39  40  41  42
##  84 219 327 265 104  10
```

```
table(births$MothersAge)
```

```
##
##    15-19    20-24    25-29    30-34    35-39    40-44    45-49 under 15
##      110      279      278      222       95       21        2        2
```

```
g <- births[births$MothersAge %in% c('under 15', '15-19'), ]
h <- births[births$MothersAge == '20-24', ]
i <- births[births$MothersAge == '25-29', ]
j <- births[births$MothersAge == '30-34', ]
k <- births[births$MothersAge == '35-39', ]
l <- births[births$MothersAge %in% c('40-44', '45-49'), ]
```

```
## Function to build the test statistic
Generator <- function (df) {
  testStat1 <- 0
  for (num in c(37,38,39,40,41,42)) {
    exp <- length(births[births$Gestation == num, 7]) * length(df$Gestation)/length(births$Gestation)
    trueVal <- length(df[df$Gestation == num, 7])
    testStat1 <- testStat1 + ((trueVal - exp)^2/exp)
  }
  return(testStat1)
}

realTestStat <- Generator(g) + Generator(h) + Generator(i) + Generator(j) + Generator(k) + Generator(l)

realTestStat
```

```
## [1] 21.48776
```

```
## p-value
pchisq(realTestStat, df = 25)
```

```
## [1] 0.3348785
```

Thus, I cannot reject H_0 : The age of the mother and the gestation period have no effect on each other; that is, the tests suggest they are independent.

6

H_0 : There is no difference in the median birth weights of babies born to smoking versus non-smoking mothers.

H_A : There is a noticeable difference in the median birth weights of babies born to smoking versus non-smoking mothers.

Graphically, we can see the distributions of birth weights of babies born to smoking versus non-smoking mothers are quite different. It looks like babies born to non-smoking mothers weigh significantly more (~250) than those born to smokers.

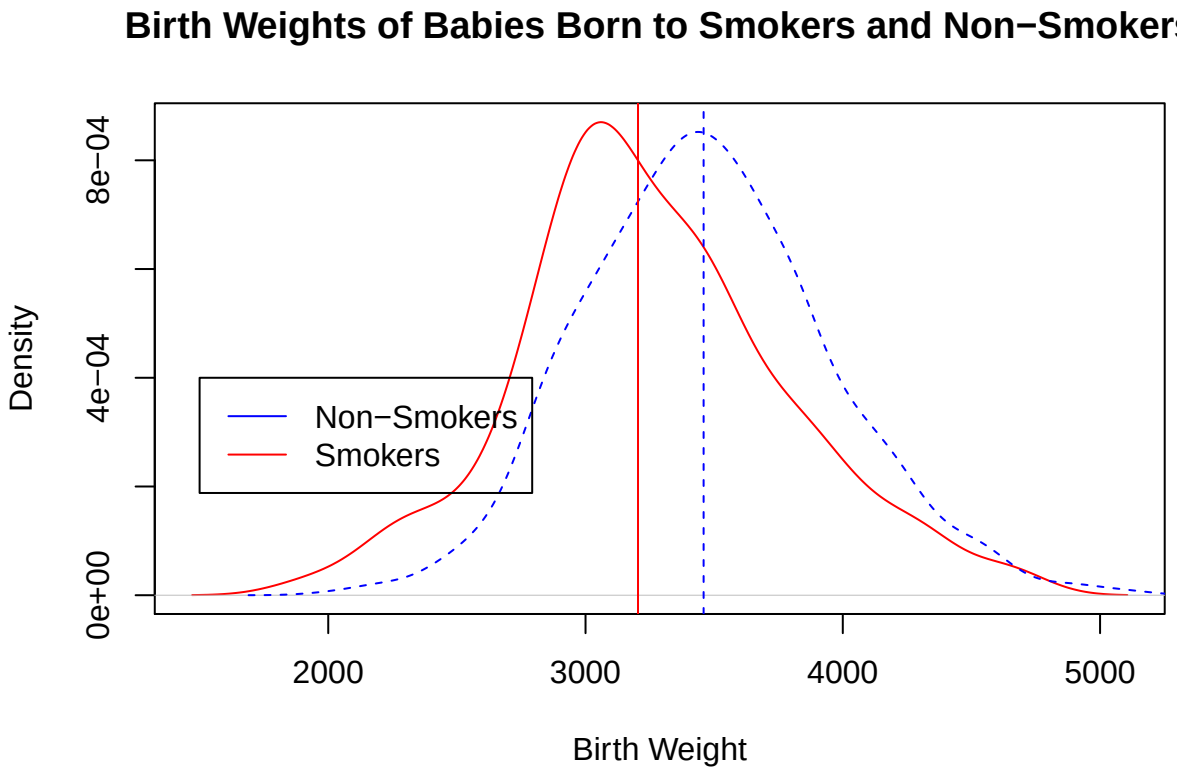
As for statistical analyses, I first performed a permutation test. As it turns out, the absolute difference in the medians of smokers and non-smokers is 255. This is the test statistic - the absolute difference in medians of the weights of babies born to smoking versus non-smoking mothers. I used the absolute value of the difference as my test statistic because we are only testing whether $Median_{smoke} = Median_{non-smoke}$. If the weights of babies were not affected by their mothers' smoking habits, we'd expect the difference in the medians of the two groups to be close to 0. Whether 255 is, in this case, sufficiently 'close' to 0 can be answered by the permutation test. As can be seen in the histogram, from the test, the p-Value of a test statistic of 255 was effectively 0, which tells us that it is very unlikely that the difference in the medians of the two groups can be accredited to chance.

I then built a bootstrapped confidence interval for: the absolute difference in medians, the median baby weight of babies born to smokers, and the median baby weight of babies born to non-smokers. I decided to use boostrapped intervals because I could not be sure that the distributions were not skewed; and, because there are a lot of data, it is better to let the data 'speak for themselves'. Furthermore, I used a simple bootstrapped percentile method - as opposed to more robust methods for boostrapping the medians - because the large amount of data in both categories nullified the added value of using something like the t method. Subsequently, when I ran the simulations the 95% bootstrapped confidence interval for the absolute difference in medians was [113, 368]. Because this interval is far from including 0, that means we can be very confident that the medians are not the same. This is echoed by the other intervals I found. The 95% bootstrapped

confidence interval for the median weight of babies born to smoking mothers was [3090, 3345], while that of their counterparts born to non-smoking mothers was [3416, 3487]. These do not overlap. Although this is not as good of a test as the confidence interval for the difference in medians, the non-intersection of the two intervals does add more evidence in support of our believing the medians are, in fact, different.

Thus, both graphical and statistical analyses lead me to feel confident in rejecting H_0 : There is no difference in the median birth weights of babies born to smoking versus non-smoking mothers; that is, it looks like there is a statistically significant difference in the weights of babies born to smoking versus non-smoking mothers. More specifically, it would appear that babies born to non-smoking mothers outweigh those born to smokers by somewhere between 150 and 300.

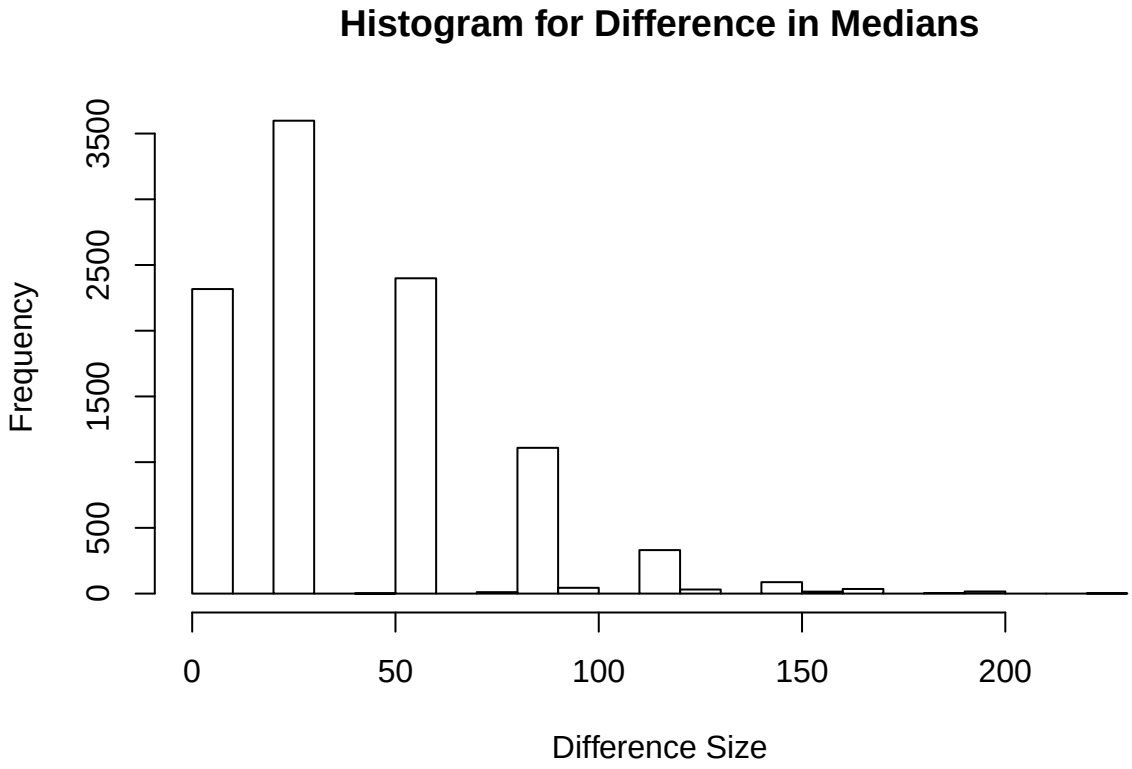
```
set.seed(runif(1, 1000, 2000))

## Subgroups of babies born to smoking and non-smoking mothers
smoke <- births[births$Smoker == 'Yes', 6]
noSmoke <- births[births$Smoker == 'No', 6]

## Visual investigation
plot(density(smoke), col = 'red', main = 'Birth Weights of Babies Born to Smokers and Non-Smokers', xlab
lines(density(noSmoke), col = 'blue', lty = 'dashed')

abline(v = median(smoke), col = 'red')
abline(v = median(noSmoke), col = 'blue', lty = 'dashed')

legend(x = 1500, y = .0004, legend = c('Non-Smokers', 'Smokers'), col = c('blue', 'red'), lty = 1)
```




```
## The test statistic
testStat2 <- abs(median(smoke) - median(noSmoke))

testStat2
```

```
## [1] 255
```

```
## Permutation test statistics for difference in medians
testStats2 <- c()
for (x in 1:10000) {
  samp1 <- sample(nrow(births), length(noSmoke), replace = F)
  helper <- births[samp1, ]
  helper1 <- births[-samp1, ]

  testStats2 <- c(testStats2, abs(median(helper$Weight) - median(helper1$Weight)))
}

hist(testStats2, breaks = 20, main = 'Histogram for Difference in Medians', xlab = 'Difference Size')
abline(v = testStat2, col = 'red')
```




```
## p-value
length(testStats2[testStats2 >= testStat2])/length(testStats2)
```

```
## [1] 0
```

```
## Bootstrapped confidence intervals for individual medians and difference in medians
testStats3 <- c()
testStats4 <- c()
testStats5 <- c()
for (x in 1:10000){
  helper2 <- sample(smoke, length(smoke), replace = TRUE)
  helper3 <- sample(noSmoke, length(noSmoke), replace = TRUE)
  testStats3 <- c(testStats3, abs(median(helper2) - median(helper3)))
  testStats4 <- c(testStats4, median(helper2))
  testStats5 <- c(testStats5, median(helper3))
}
```

```r
quantile(testStats3, c(.025, .975))
```

```
##  2.5% 97.5%
##   113   369
```

```r
quantile(testStats4, c(.025, .975))
```

```
##  2.5% 97.5%
##  3090  3345
```

```r
quantile(testStats5, c(.025, .975))
```

```
##  2.5% 97.5%
##  3416  3487
```

7

For the first four intervals - (37,38), (38,39), (39,40), (40,41) - I used the bootstrapped t method for building a confidence interval for $\mu_1 - \mu_2$. I chose to build boostrapped confidence intervals instead of classical intervals because, as in question 6, there were enough data in each bucket to let the data 'speak for themselves'. Moreover, I chose the bootstrapped t method because, as was just mentioned, there were a decent amount of data in each gestation period bucket, but not enough to get the desired accuracy using a simple percentile bootstrapped method. Accordingly, the confidence intervals from my simulations were:

95% Confidence Intervals for Weight Gain per Week of Gestation:

(Week 37 - 38): [175.37, 360.58]

(Week 38 - 39): [72.64, 253.29]

(Week 39 - 40): [30.79, 209.14]

(Week 40 - 41): [17.27, 198.38]

For the last interval - (41, 42) - there were not enough data in the week 42 bucket (n = 10) to justify a bootstrapped approach. Instead, I used a classical approach for building a confidence interval for a difference in means: $(\overline{x}_1 - \overline{x}_2) \pm t * \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$. Any results, accordingly, should be taken with a grain of salt. That said, the final confidence interval was:

95% Confidence Interval for Weight Gain per Week of Gestation:

(Week 41 - 42): [-329.6, 330.36]

As was predicted, this doesn't make a lot of sense. In fact, if we look at the data, the average weight of babies born after a 42-week gestation period was actually 7 less than that of babies born after 41 weeks. So, really, there isn't enough data to make any sort of conclusions about the difference in baby weights after 42 versus 41 weeks of gestation.

```r
set.seed(runif(1, 1000, 2000))

## Find the size of the subgroups
for (x in c(37,38,39,40,41,42)) {
  print(length(births[births$Gestation == x, 7]))
}
```

```
## [1] 84
## [1] 219
## [1] 327
```

```
## [1] 265
## [1] 104
## [1] 10
```

```
## Build a function to make the process of finding bootstrapped t confidence intervals easier
bstrap <- function(x,y){
  testStats6 <- c()
  for (z in 1:10000){
    group1 <- births[births$Gestation == x, 6]
    group2 <- births[births$Gestation == y, 6]

    helper4 <- sample(group1, 200, replace = TRUE)
    helper5 <- sample(group2, 200, replace = TRUE)

    new <- ((mean(helper5) - mean(helper4)) - (mean(group2) - mean(group1)))/sqrt(var(helper4)/length(g

    testStats6 <- c(testStats6, new)
  }
  lower <- unname(quantile(testStats6, .025))
  upper <- unname(quantile(testStats6, .975))
  lower1 <- (mean(group2) - mean(group1)) - upper*sqrt(var(group1)/length(group1) + var(group2)/length(g
  upper1 <- (mean(group2) - mean(group1)) - lower*sqrt(var(group1)/length(group1) + var(group2)/length(g
  interval <- c(lower1, upper1)
  print(interval)
}

bstrap(37,38)
```

```
## [1] 173.1024 360.8636
```

```
bstrap(38,39)
```

```
## [1]  72.26568 249.96713
```

```
bstrap(39,40)
```

```
## [1]  32.49374 207.65229
```

```
bstrap(40,41)
```

```
## [1]  19.74252 201.14689
```

```
## Classical test for the last k
testStat3 <- mean(births[births$Gestation == 42, 6]) - mean(births[births$Gestation == 41, 6])
tstar <- qt(.975, df = 9)
SE <- sqrt(var(births[births$Gestation == 42, 6])/10 + var(births[births$Gestation == 42, 6])/104)
testStat + (tstar*SE)
```

```
## [1] 330.3592
```

```
testStat - (tstar*SE)
```

```
## [1] -329.5972
```

Bonus

H_0 : Tobacco use by mothers has no effect on the gestation period.

H_a : Tobacco use by mothers shortens the gestation length period.

I first explored the data visually. The boxplot, though, didn't provide any insight.

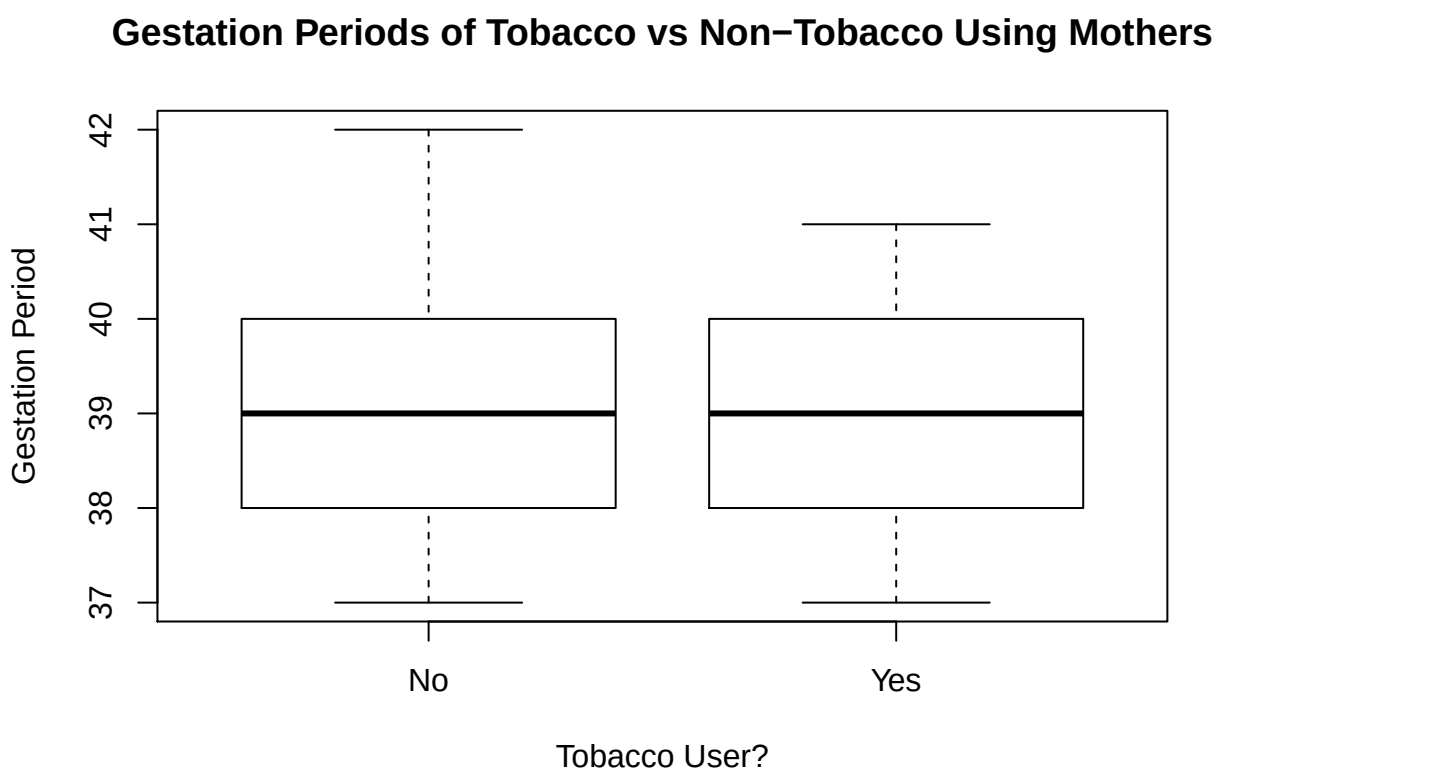
To test this hypothesis, I'll use a permutation test and run a one-tailed t-test. The permutation is suitable because there is a relatively large sample of data. And, for the t-test, that'll be used as extra evidence for or against rejecting H_0.

My test statistic will be $\overline{x}_{non-smoker} - \overline{x}_{smoker}$. From the permutation test, the p-value of getting a test statistic as extreme as that produced by the data is .046. This is right on the edge of statistical significance at the $\alpha = .95$ level.Furthermore, the one-tailed t-test produced a p-value of .041, which is, again right on the edge of statistical significance. When I examined the 95% confidence interval from the permutation test, though, I found it to be [-.225, .220].

So the results of the tests were somewhat mixed. On the one hand, the permutation test and one-tailed t-test produced test statistics that were significant at the $p = .05$ level. But, on the other hand, the confidence iterval included 0. Also, the means from the data are $\overline{x}_{non-smoker} = 39.14$ and $\overline{x}_{smoker} = 38.94$, which - based on the eye test - are too close to spur any conclusions. I think these two red flags - the confidence interval containing 0, and the small difference between the means - is enough to keep me from rejecting H_0.

In conclusion, I cannot reject H_0 : Tobacco use by mothers has no effect on the gestation period. There is not enough evidence in the data to confidently say tobacco use shortens the gestation period.

```
set.seed(runif(1, 1000, 2000))

## Get the relevant subgroups
smoke1 <- births[births$Tobacco == 'Yes', 7]
noSmoke1 <- births[births$Tobacco == 'No', 7]

## Explore data visually with a boxplot
boxplot(births$Gestation~births$Tobacco, main = 'Gestation Periods of Tobacco vs Non-Tobacco Using Moth
```



```
## Test statistic
testStat6 <- mean(noSmoke1) - mean(smoke1)

testStats6 <- c()

## Permutation Test
for (x in 1:10000){
  samp3 <- sample(nrow(births), length(smoke1), replace = FALSE)
  helper4 <- births[samp3, ]
  helper5 <- births[-samp3, ]
  testStats6 <- c(testStats6, mean(helper5$Gestation) - mean(helper4$Gestation))
}

## Various analyses using the results of the permutation test
testStat6
```

```
## [1] 0.2000341
```

```
length(testStats6[testStats6 >= testStat6])/length(testStats6)
```

```
## [1] 0.046
```

```
quantile(testStats6, c(.025, .975))
```

```
##       2.5%      97.5%
## -0.2251149  0.2304019
```

```
## Test for True Difference in means

## Welch's Approximation of df
v = (var(smoke1)/length(smoke1)+var(noSmoke1)/length(noSmoke1))^2/(var(smoke1)^2/(length(smoke1)^2*(leng

## t-test
pt(((mean(smoke1)-mean(noSmoke1))/sqrt(var(smoke1)/length(smoke1) + var(noSmoke1)/length(noSmoke1))), df
```

```
## [1] 0.04055942
```