ST 551 Homework 6

Luis Garcia Lamas

Fall 2022

Instructions

This assignment is due by November 28th 11:59 PM, November 14th on Canvas via Gradescope. **You should submit your assignment as a typed PDF which you can compile using the provide .Rmd (R Markdown) template.** Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, grammatically correct sentences for your solutions.

Question 1 (28 points)

Explore the effects of transforming your data: suppose you want to test a hypothesis about the equality of two different population means, but the underlying populations are quite skewed. As we discussed in class, some sources would then recommend that you consider transforming your data to make it look *more normal* (or at least more symmetric)

Part A. (8 points: For each of the following scenarios, perform 10,000 simulations to explore the rejection rates on the original data (let $\alpha = 0.05$), and when you transform your data using a log-transformation. In both cases (untransformed and transformed) perform a t-test to test for equality of means. Write a brief summary of your findings.

```
N <- 10000
m <- 200
n <- 500
df <- 4
shape <- .5
scale <- 21.75
alpha <- .05

p_values <- rep(0,N)
p_values_transformed <- rep(0,N)
count1 <- 0
count2 <- 0

for (i in 1:N) {

  sampleA <- rchisq(m,df)
  sampleB <- rgamma(n,shape = shape, scale = scale)
  logsampleA <- log(sampleA)
  logsampleB <- log(sampleB)

  p_values[i] <- t.test(sampleA,sampleB)$p.value
  p_values_transformed[i] <- t.test(logsampleA,logsampleB)$p.value

  if(p_values[i] <= alpha) {count1 <- count1 + 1}
  if(p_values_transformed[i] <= alpha) {count2 <- count2 + 1}

}

count1/N
```

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

```
count2/N
```

```
## [1] 0.0474
```

Sample 1 Dist	Sample 2 Dist	Untransformed	Transformed
Chi-square(4), $m = 20$	Gamma(shape=0.5, scale=8), $n = 50$	.0495	.8067
Chi-square(4), $m = 200$	Gamma(shape=0.5, scale=8), $n = 500$	.0562	1.00
Chi-square(4), $m = 20$	Gamma(shape=0.5, scale=21.75), $n = 50$	.9379	.0529
Chi-square(4), $m = 200$	Gamma(shape=0.5, scale=21.75), $n = 500$	1.00	.0507

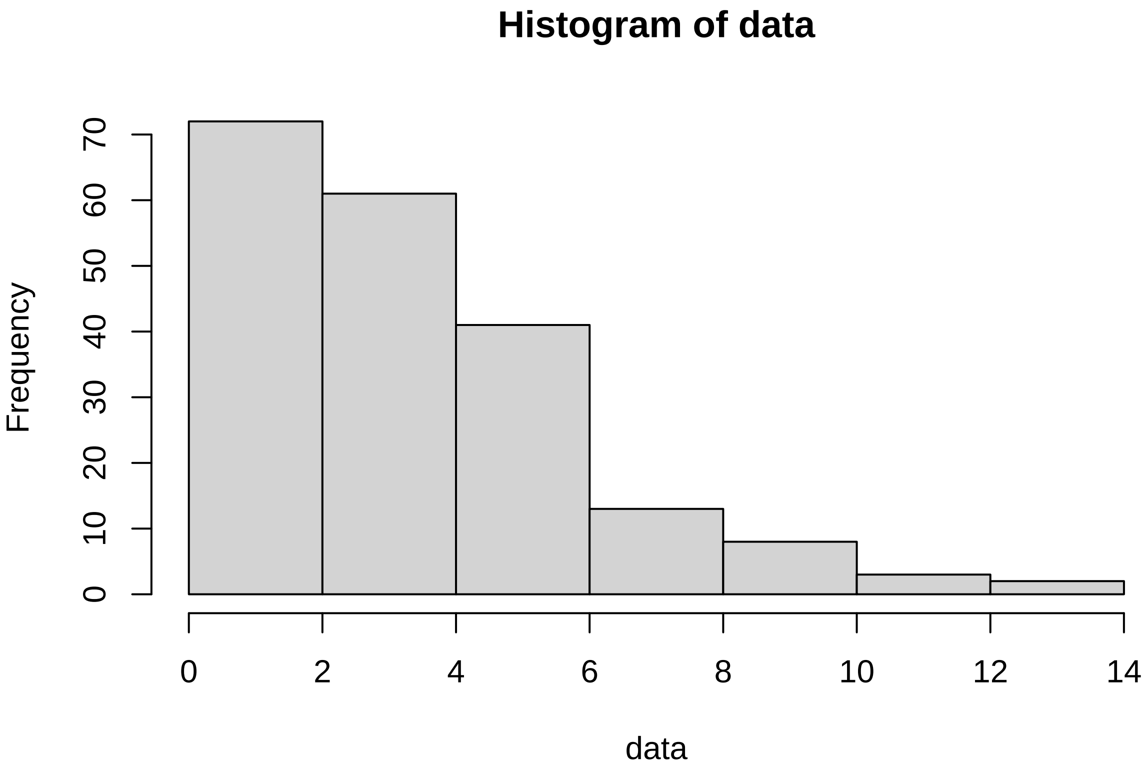
We know the means of a chi squared distribution should be its df so in this case it will always be 4. While for gamma it will be shape multiplied which scale. In this case it will also be 4. So for the first 2 rows of this table we should expect in a t test that we get a rejection rate of about .05 since that is our specified alpha. This is the opposite for the last 2 rows which the gamma distribution will have a mean of 10.875. Therefore, we would expect reject most fo the time in a large number of simulations, and this is what we've observed in the table as well.

Once we have the transformed the data, it tells a different story. We actually have high rates of rejection with the transformed data when the means are in fact equal (first 2 rows). And the opposite is true for the last 2 rows when the population mean is different. This kind of shows why transforming data and looking at the p-values isn't always reliable in providing the result we were looking for.

Part B. (9 points): Make histograms of the following samples (listed below) then answer the questions: (1) Do the untransformed samples look skewed? (2) Do you think these might be settings where transformations would be recommended? (3) Does the log transformation "fix" the skewness problem?

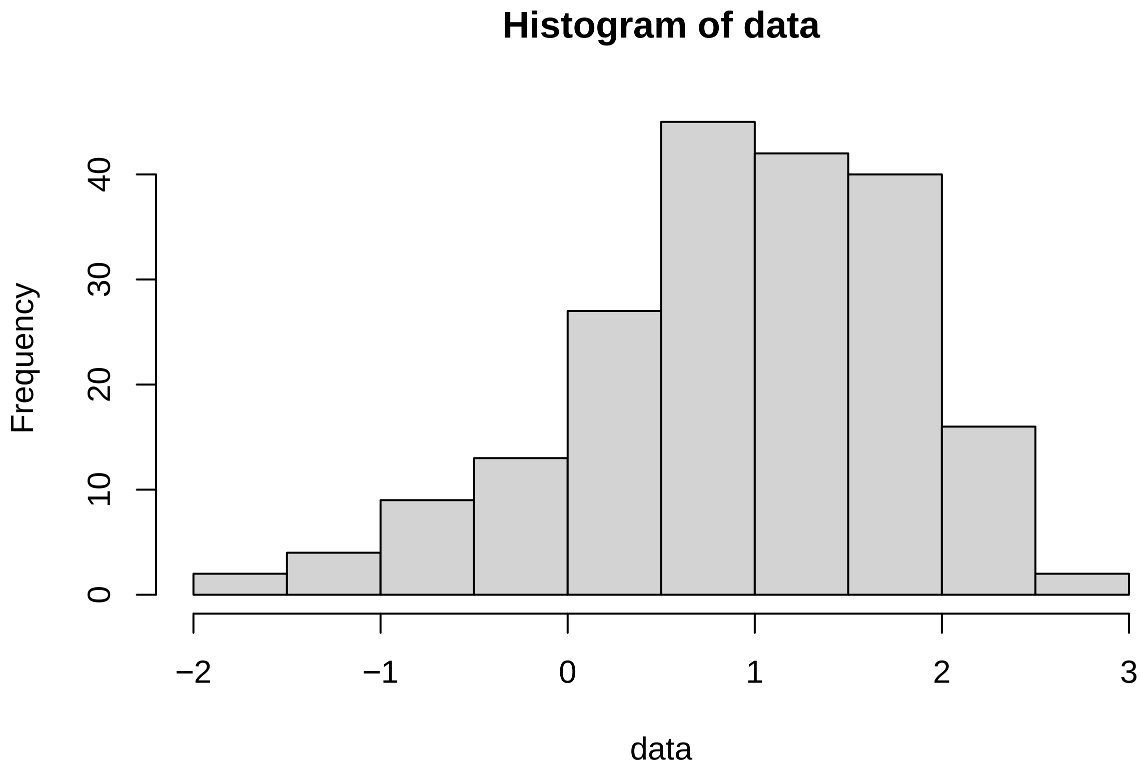
1. A single sample of size 200 from the Chi-squared(df=4) distribution.

```
data <- rchisq(200,4)
hist(data)
```



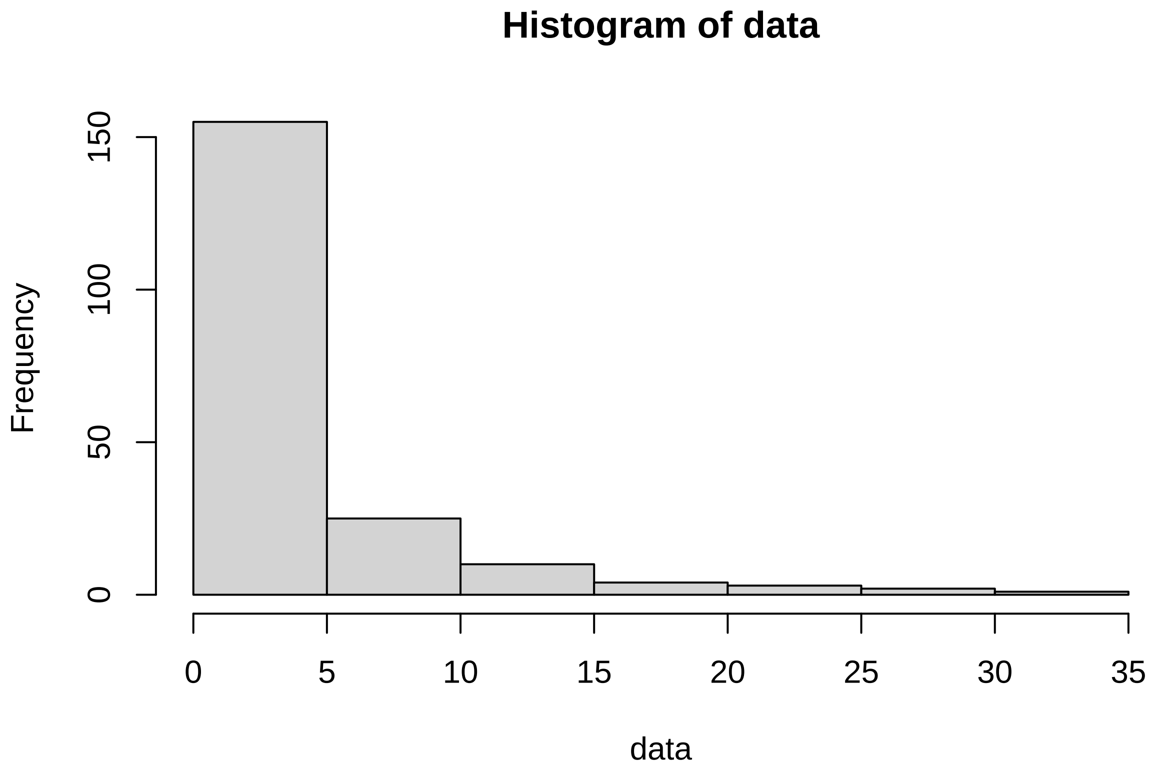
2. A single sample of size 200 from the Chi-squared(df=4) distribution, after a log-transformation.

```
data <- log(data)
hist(data)
```



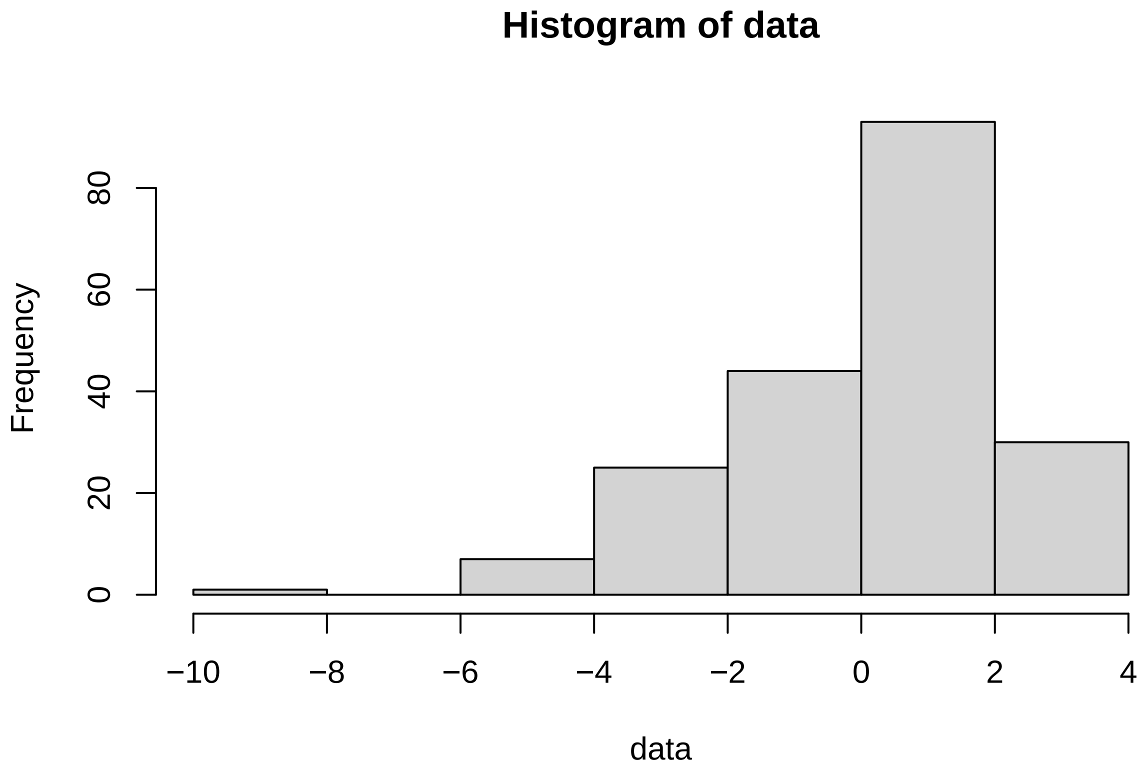
3. A single sample of size 200 from the Gamma(shape=0.5, scale=8) distribution.

```
data <- rgamma(200, shape = .5, scale = 8)
hist(data)
```



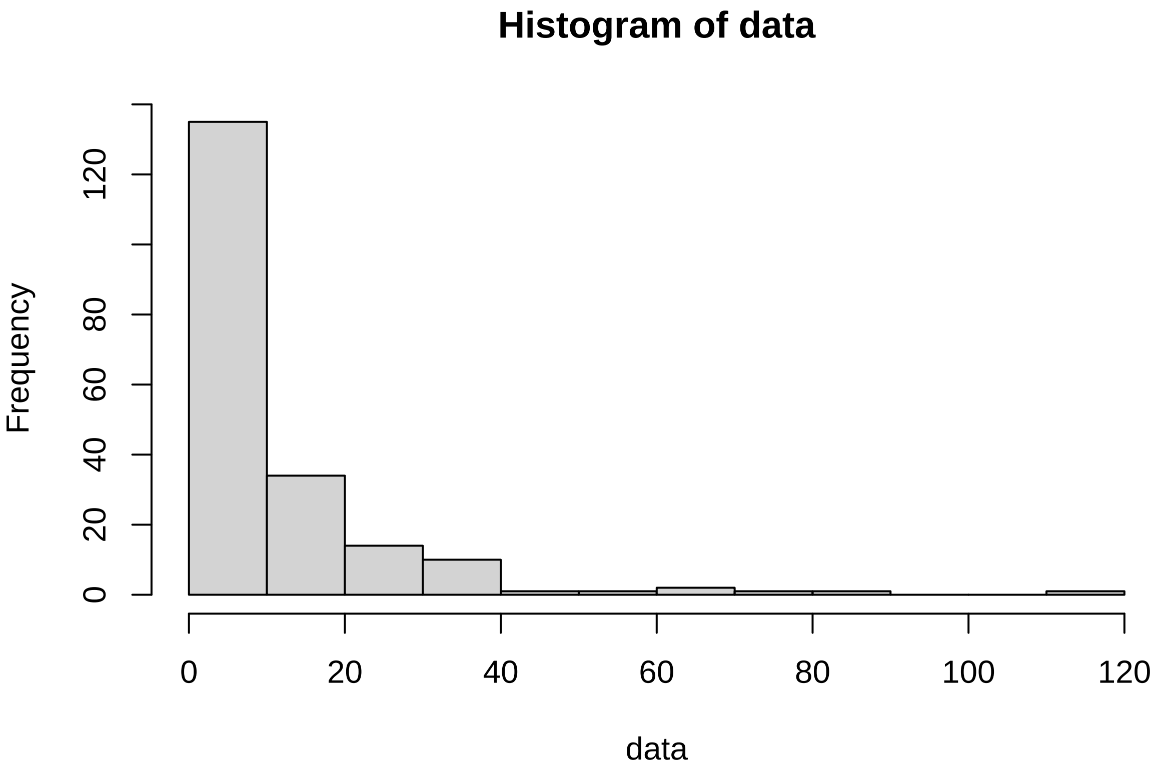
4. A single sample of size 200 from the Gamma(shape=0.5, scale=8) distribution, after a log-transformation.

```
data <- log(data)
hist(data)
```



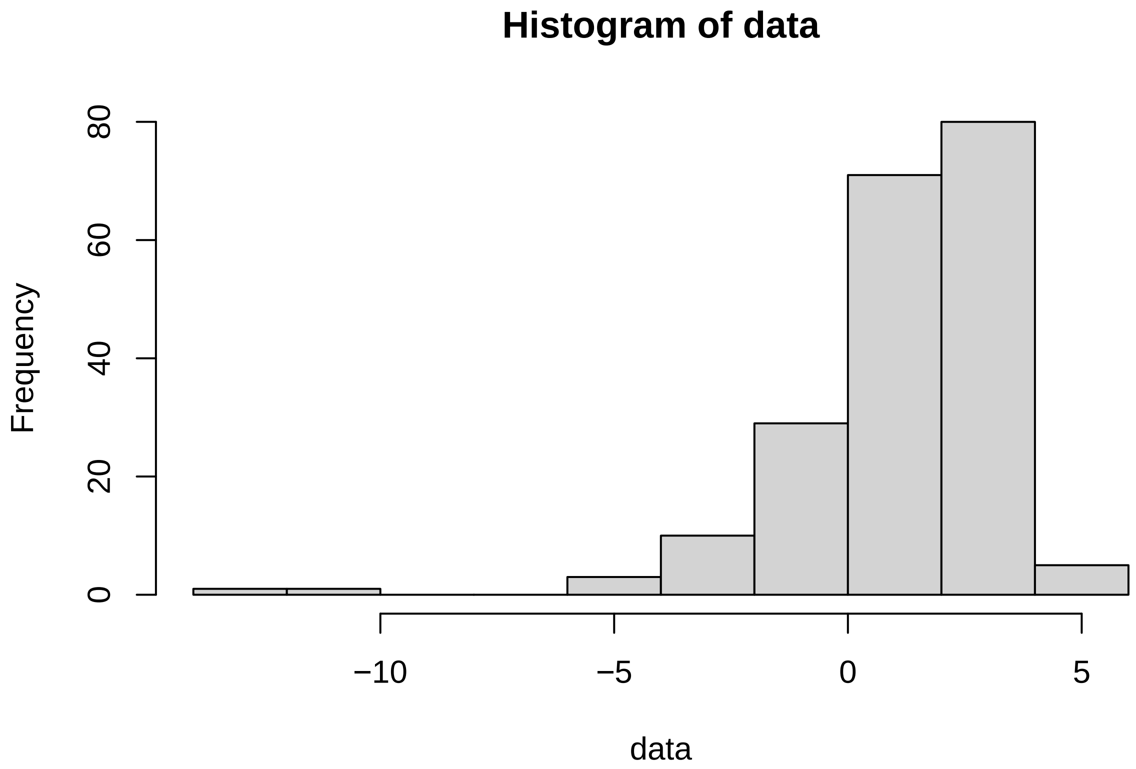
5. A single sample of size 200 from the Gamma(shape=0.5, scale=21.75) distribution.

```
data <- rgamma(200, shape = .5, scale = 21.75)
hist(data)
```



6. A single sample of size 200 from the Gamma(shape=0.5, scale=21.75) distribution, after a log-transformation.

```
data <- log(data)
hist(data)
```



(1) Do the untransformed samples look skewed?

Yes, they all look pretty rightly skewed.

(2) Do you think these might be settings where transformations would be recommended?

Yes, if we didn't know the underlying distribution between the two data samples some scientist/statisticians might recommend transformations. This of course comes with the risk of transforming data and misinterpreting results.

(3) Does the log transformation "fix" the skewness problem?

It does seem to help but it definitely does not solve the problem. For the chi squared distribution it did a fairly good job of centering it and removing skewness from the original data but it is still fairly skewed. This is even more obvious with the gamma distributions as even after transformation they look highly skewed.

Part C. (6 points): Fill in the true population means for each of the following distributions (You can look up how to compute these values using wikipedia, your theory textbook, or by simulating a huge (HUGE!) number of samples).

Population Distribution	True Mean (Untransformed)	True Mean (Log-transformed)
Chi-square(4)	4	1.1158
Gamma(shape=0.5, scale=8)	4	.1159
Gamma(shape=0.5, scale = 21.75)	10.875	1.116

Part D. (5 points): Comment on the rejection rates in the table from part A., in light of the means you calculated in part C. If the original (untransformed) population distributions have the same mean, what happens when we log-transform the data? If we fail to reject the hypothesis that the log-transformed data come from populations with the same mean, does that mean that the means are equal on the original scale?

The results in Part C seem to support our results in Part A. The transformed means of the gamma distribution with scale = 21.75 and the chi squared distribution have a very close mean to each other. Therefore, the .0507 rejection rate being close to alpha makes sense for the last row of the table for the transformed data. It also shows that distributions with the same mean before transformation have different means after being transformed, supporting the results in the first 2 rows. Overall, it seems that transforming data isn't the greatest thing to do without strong assumptions.

Question 2 (16 points)

In each of the following scenarios, indicate (1) which probabilities you can estimate and (2) give the appropriate estimate when applicable. Be sure to consider if the sampling scheme is *multinomial*, *prospective binomial*, or *retrospective binomial* sampling when deciding if it's possible to estimate the desired probability.

Part A. (3 points): A researcher collects a sample of 100 Corvallis residents and asks them whether they are a member of the public library (Yes/No), and whether they are vegetarian (Yes/No)

	Library Member		
Vegetarian	Yes	No	TOTAL
Yes	18	2	20
No	37	53	90
TOTAL	55	55	110

1. P(Library Member)?

   a. Can this probability be estimated? Yes.
   b. If so, what is the estimate? $P(\text{Library Member}) = \frac{55}{110} = .5$

2. P(Vegetarian | Non-Library Member)?

   a. Can this probability be estimated? Yes.
   b. If so, what is the estimate? $P(\text{Vegetarian} \mid \text{Non-Library Member}) = \frac{2}{53}$

3. P(Vegetarian)?

   a. Can this probability be estimated? Yes.
   b. If so, what is the estimate? $P(\text{Vegetarian}) = \frac{20}{110} = .1818$

Part B. (4 points): A researcher samples 100 smokers and 100 non-smokers, and asks them whether they have ever run a marathon.

	Smoker		
Marathon	Yes	No	TOTAL
Yes	5	15	20
No	95	85	180
TOTAL	100	100	200

1. P(Marathon Runner)?

   a. Can this probability be estimated? No.
   b. If so, what is the estimate?

2. P(Smoker)?

   a. Can this probability be estimated? No.
   b. If so, what is the estimate?

3. P(Smoker | Marathon Runner)?

a. Can this probability be estimated? No.
b. If so, what is the estimate?

4. P(Marathon Runner | Smoker)?

   a. Can this probability be estimated? Yes!
   b. If so, what is the estimate? $P(\text{Marathon Runner} \mid \text{Smoker}) = \frac{5}{100}$

Part C. (4 points): A researcher samples 100 marathon runners and 100 people who have never run a marathon, and asks them whether they smoke

	Smoker		
Marathon	Yes	No	TOTAL
Yes	2	98	100
No	8	92	100
TOTAL	10	190	200

1. P(Marathon Runner)?

   a. Can this probability be estimated? No.
   b. If so, what is the estimate?

2. P(Smoker)?

   a. Can this probability be estimated? Yes.
   b. If so, what is the estimate? $P(\text{Smoker}) = \frac{8}{100}$

3. P(Smoker | Marathon Runner)?

   a. Can this probability be estimated? Yes.
   b. If so, what is the estimate$P(\text{Smoker} \mid \text{Marathon Runner}) = \frac{2}{100}$

4. P(Marathon Runner | Smoker)?

   a. Can this probability be estimated? No.
   b. If so, what is the estimate?

Part D. (5 points): A researcher samples 20 undergrads and 80 grad students at OSU, and asks them whether they prefer skiing or snowboarding. Assuming probability that they prefer skiing or snowboarding instead of probabiliy of being a skier or snowboarder or both.

Preference	Undergrad	Grad	TOTAL
Ski	5	25	30
Snowboard	15	55	70
TOTAL	20	80	100

1. P(Undergrad)?

   a. Can this probability be estimated? No.
   b. If so, what is the estimate?

2. P(Skier)?

a. Can this probability be estimated? No.
b. If so, what is the estimate?

3. $P(\text{Skier} \mid \text{Grad})$?

    a. Can this probability be estimated? Yes.
    b. If so, what is the estimate? $P(\text{Skier} \mid \text{Grad}) = \frac{25}{80} = .3125$

4. $P(\text{Grad} \mid \text{Snowboarder})$?

    a. Can this probability be estimated? No.
    b. If so, what is the estimate?

5. If we know that the true proportion of the OSU population that is undergrad is 80%, can you estimate the population proportion of students at OSU who prefer skiing?

    a. Can this probability be estimated? Yes
    b. If so, what is the estimate? If we multiply the resulting probabilities to the proportion of the undergraduate students to graduate students we can get this probability: $P(\text{Skier}) = P(\text{Undergraduate})P(\text{Skier} \mid \text{Undergraduate}) + P(\text{Graduate})P(\text{Skier} \mid \text{Graduate})$ $P(\text{Skier}) = (.8)(\frac{5}{20}) + (.2)(\frac{25}{80}) = .2625$ $P(\text{Skier}) = .2625$

Question 3 (10 points)

I perform a study to test whether a friend can tell the difference between instant coffee and freshly ground drip coffee. The results are summarized in the table below.

> Note: I actually did this experiment in grad school with some friends! But we compared pour-over coffee vs. coffee from a Mr. Coffee machine or something.

	True Coffee Type		
Guess	Instant	Fresh Ground Drip	TOTAL
Instant	7	5	12
Fresh Ground Drip	3	5	8
TOTAL	10	10	20

Part A. (5 points): Perform Fisher's exact test (by hand) to test the null hypothesis that my friend's responses are random guesses vs. the alternative hypothesis that my friend is actually able to tell the difference (i.e. gets the correct answer more often than would be expected by chance). Enumerate all of the tables *more extreme* than the observed table, and calculate the probability of each under the null hypothesis to obtain the p-value.

```r
data <- rbind(c(7, 5),
       c(3, 5))
obs_prob <- dhyper(data[1][1], m = 10, n = 10, k = 12)
#hypergeometric distribution with correct parameters

all_table_probs <- dhyper(0:12, m = 10, n = 10, k = 12)
#observed probabilities for all k less than 12

p_val <- sum(all_table_probs[all_table_probs <= obs_prob])
p_val
```

```
## [1] 0.6499166
```

```r
#sum of the probabilities that are less than or equal to the kth probability

fisher.test(data)$p.value
```

```
## [1] 0.6499166
```

```r
#checking
```

Part B. (5 points): Perform Pearson's chi-squared test (by hand) to test the null hypothesis that my friend's responses are random guesses. What is the alternative hypothesis tested by Pearson's chi-squared test?

```
row1 <- c(7, 5)
row2 <- c(3, 5)
data_observed <- rbind(row1,row2)

a <- sum(row1)*(row1[1]+row2[1])/(sum(row1,row2))
b <- sum(row1)*(row1[2]+row2[2])/(sum(row1,row2))
c <- (row1[1]+row2[1]) - a
d <- (row1[2]+row2[2]) - b
#calculating expected values using method

erow1 <- c(a,b)
erow2 <- c(c,d)
data_expected <- rbind(erow1,erow2)
#data saved
X <- 0

for (i in 1:4) {

  X <- X + (data_observed[i] - data_expected[i])^2/data_expected[i]
}
#calculating our test statistic for the chi squared distribution

p_val <- 1 - pchisq(X, df = 1)
#calculating p-value of our test statistic.

X
```

```
## [1] 0.8333333
```

```
p_val
```

```
## [1] 0.3613104
```

```
chisq.test(data_observed, correct = FALSE)$statistic
```

```
## Warning in chisq.test(data_observed, correct = FALSE): Chi-squared approximation
## may be incorrect
```

```
## X-squared
## 0.8333333
```

```
chisq.test(data_observed, correct = FALSE)$p.value
```

```
## Warning in chisq.test(data_observed, correct = FALSE): Chi-squared approximation
## may be incorrect
## [1] 0.3613104
```

```
#verifying values
```

The alternative hypothesis is that the homogeneity of proportions is different across these groups.