Math 227 Second Exam Due: November 8 at 2:00 pm

This is a take-home exam. ***Please insert your answers into this document,*** *You may use a calculator, your notes, my notes, and R. You may not discuss any aspect of this exam with anybody but me until after 2:00 pm on November 8. Neatness and clarity will be appreciated. Do show the major steps in your work. Correct answers may not receive full credit without supporting work. Please present your answers in the form of complete sentences. Insert only as much R code and output as is necessary to support your answers. Please write and sign the Honor Pledge [I have neither given nor received help on this exam] on your exam.*
Honor Pledge

I have neither given nor received help on this exam.

Name: **Ellen Stanton**

Please do not write on this page.

1. The information in the data set *acs.csv* comes from the American Community Survey, a survey of individuals conducted annually by the Census Bureau. The variables in the data set are (i) race, (ii) whether or not the respondent is a US citizen, (iii) whether or not the respondent has health insurance, and (iv) the number of hours worked in the prior week. (This survey is for 2011, prior to passage of the Affordable Care Act). I believe that the individuals in this data set are from the New England states.

(a) What proportion of the responses were from non-US citizens?

```
tally(~USCitizen,data=acs,margin=T)
USCitizen
   no   yes Total
   35   470   505
35/505
[1] 0.06930693
```

The proportion of responses from non-US citizens is .069.

(b) Obtain a 96% confidence interval for the proportion of all adults living in New England who are non-US citizens.

First, I will check the two conditions to ensure that I can compute a confidence interval.

```
.069*505
[1] 34.845
(1-.069)*505
[1] 470.155
```

There are at least ten "successes" and ten "failures", since 470 respondents are US citizens and 35 are not US citizens. Because both 470 and 35 are greater than or equal to ten, the first condition is met. We will assume the sample was randomly selected, so the second condition is met. I will now proceed in finding the confidence interval.

```
phat <- 35/505
z <- qnorm(.98)
n <- 505
se <- sqrt(((phat)*(1-phat))/(n))
me <- z*se
lb <- phat-me
ub <- phat+me
lb;ub
[1] 0.04609596
[1] 0.0925179
```

The 96% confidence interval for the proportion of all adults living in New England who are non-US citizens is .046 to .093.

(c) Provide the correct interpretation of your interval.

If we were to repeat this survey, taking a random sample of 505 adults from New England, and compute a 96% confidence interval for the proportion of all adults living in New England who are non-US citizens (p), 96% of the resulting confidence intervals would contain p. We hope that our confidence interval, .046 to .093, is one of the 96% and not one of the 4% of confidence intervals that don't contain p.

(d) A 2011 estimate by the Department of Immigration and Naturalization, suggested that between 8% and 8.5% of adult residents in New England are non-US citizens. Is this estimate consistent with your interval in part (b)? Explain. What level of significance is implied in your answer?

The estimate suggests that the population proportion of adult residents in New England who are non-US citizens is somewhere between .080 and .085. This is consistent with my confidence interval, because my interval .046 to .093 includes the full range of the estimate, .080 to .085. My answer implies a level of significance of .04 or 4%, since my confidence interval was computed as a 96% confidence interval.

(e) According to the Bureau of the Census, there were 12,500,000 adult residents in New England (in 2011). Obtain an estimate for T, the total number of non-US citizens resident in England in 2011.

```
phat*t
[1] 866336.6
```

An estimate for the total number of non-US citizens residing in New England in 2011 is 866,367.

(f) Obtain a 96% confidence interval for T.

```
lb*t
[1] 576199.5
ub*t
[1] 1156474
```

A 96% confidence interval for the total number of non-US citizens residing in New England in 2011 is 576,200 to 1,156,474.

2. The data set *simmons.csv* contains information about 1649 Simmons freshmen. One variable contains the student's graduating class (2005, 2010, 2015, and 2020). The other variable contains an indication of whether (Yes) or not (No) the student represents the first generation in her family to attend college. The Dean of Students is interested in how, if at all, the percentage of students who are first generation has changed over time.

(a) Which is the response variable in this case?

The response variable is the percentage of hypothetical Simmons freshmen who are first generation.

When I say "hypothetical Simmons freshmen", I'm referencing the population of eligible students who, had they applied to Simmons, would have been accepted and would have attended. For the purposes of this problem, I will consider the data we have here from actual Simmons freshmen as a sample from the population of hypothetical Simmons freshmen.

(b) Obtain output that shows the percentage (or the proportion) that are first generation by class.

```
round(tally(firstgen~class,data=simmons,format="percent",margin=T),2)
        class
firstgen   2005   2010   2015   2020
   no     57.21  58.21  70.98  75.87
   yes    42.79  41.79  29.02  24.13
   Total 100.00 100.00 100.00 100.00
```

(c) Write a brief paragraph that respond to the Dean's interest and which summarize the nature of the relationship between these two variables.

```
42.79-24.13
[1] 18.66
18.66/42.79
[1] 0.4360832
```

There is a negative relationship between the percentage of hypothetical Simmons freshmen who are first generation and their graduating class, which shows that over time, the percentage of hypothetical freshmen at Simmons who are first generation has decreased. In the class of 2005, 42.79% of freshmen in our sample were first generation, but this percentage decreased by 1 point to the class of 2010 (41.79%), then by 12.77 points to the class of 2015 (29.02%), and finally by 4.89 points to the class of 2020 (24.13%). This makes for a total decrease in 18.66 percentage points from the sample for class of 2005 to the sample for class of 2020. In 2005, over ⅔ of the sample freshmen were first generation, but now only ¼ of the class of 2020 are. This decrease in percentage of first generation freshmen in our sample represents 43.6% of the percentage of first generation freshmen in 2005, since the 18.66 percentage points lost are 43.6% of the initial (2005) value of 42.79 percent.

(d) Are the differences in the sample percentages (or proportions) statistically significant? Perform the appropriate Chi-square test (with the correction, please). Explain your hypotheses and your conclusion.

Null Hypothesis H_0: In the population of hypothetical Simmons freshmen, the percentage distribution of first generation status is the same for all class years.

Alternative Hypothesis H_a: In the population of hypothetical Simmons freshmen, the percentage distribution of first generation status differs significantly by class year.

First, I will check the conditions. We will assume that every individual appears in only one cell, since a student can't be both first generation and not first generation, and although technically someone could have been a freshman more than once if they dropped out and then came back later, we'll assume that hasn't happened. We will also assume that the samples were taken randomly, so the second condition is met. Finally, we can see in the following table of expected counts that all expected counts are greater than 5.

```
chisq.test(simmons$firstgen,simmons$class)$expected
                simmons$class
simmons$firstgen     2005     2010     2015    2020
             no  268.8636 256.3736 294.5009 264.262
             yes 140.1364 133.6264 153.4991 137.738
```

All three conditions for a chi-square test are met, so we can proceed with the test.

```
chisq.test(simmons$firstgen,simmons$class)

      Pearson's Chi-squared test

data:  simmons$firstgen and simmons$class
X-squared = 46.818, df = 3, p-value = 3.8e-10
```

The p-value from our Chi-squared test is 3.8x10^-4, so we reject the null hypothesis at the 1% level of significance. The data strongly suggest that the percentage distribution of first generation status among hypothetical Simmons freshmen varies by class year.

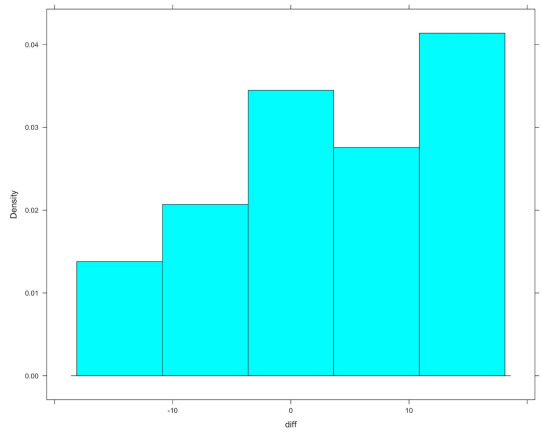
3. Does your your pulse rate tend to be higher when you are taking a quiz than when you are sitting in a lecture? The data in the file *pulserates.csv* are pulse rates collected on 20 students (at UC Davis) in a Psychology class lecture and then from the same students during an in-class quiz in the same course. You may regard these students as a random sample of UC Davis undergraduates.

(a) Summarize these data numerically and graphically. Write a brief paragraph summarizing your results

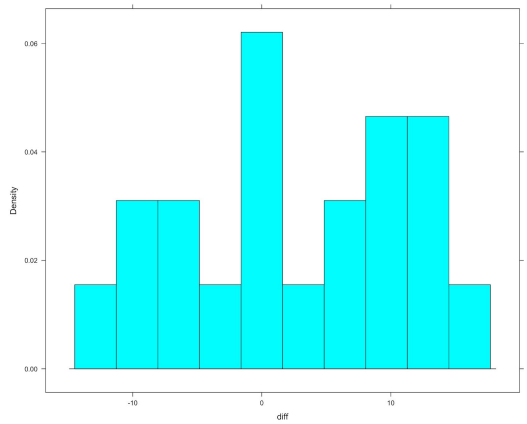
```
pulserates$diff <- pulserates$quiz-pulserates$lecture
```

A note about the variable diff: here, diff was created to be each student's pulse during the quiz minus their pulse during the lecture. So, positive values of diff show that their pulse was higher during the quiz, whereas negative values of diff show that their pulse was higher during the lecture.

```
favstats(~diff,data=pulserates)
 min Q1 median Q3 max mean       sd  n missing
 -12 -3    2.5 11  17  2.7 8.838314 20       0
histogram(~diff,data=pulserates)
```



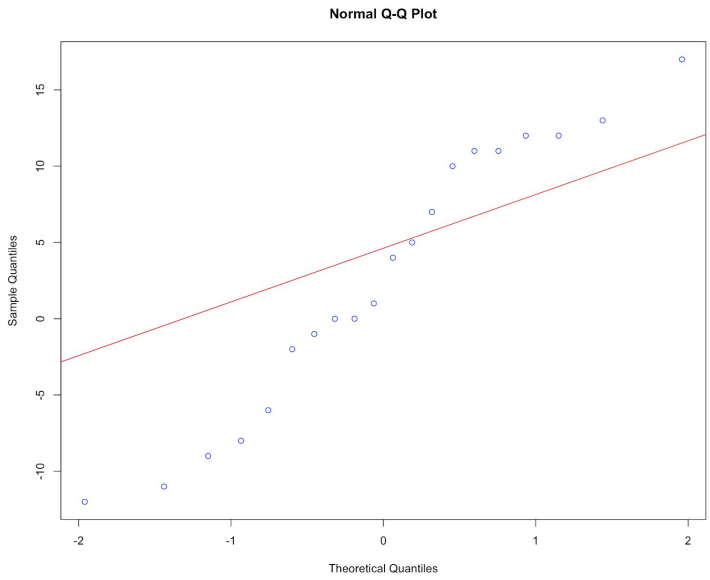

```
histogram(~diff,data=pulserates,nint=10)
```



The variable diff has a mean value of 2.7 units (presumably bpm) and a median value of 2.5 units. The mean and median are close to each other, as would be the case for a symmetrical distribution, but the mean being slightly larger could indicate that the distribution of diff is slightly skewed right. However, the first histogram generated instead shows a distribution that is skewed left. A second histogram divided along more intervals shows instead that the distribution is fairly symmetrical with a slight skew to the left. Keeping in mind that the sample size is small and that the number of intervals for a histogram can change how the shape of the distribution appears, it does seem as if the distribution of diff is fairly symmetrical with a slight left skew.

(b) Perform one graphical check and the S-W test for normality. Report your conclusion.

```
qqnorm(pulserates$diff,col="blue")
qqline(oc$diff,col="red")
```




```
shapiro.test(pulserates$diff)
```

```
      Shapiro-Wilk normality test

data:  pulserates$diff
W = 0.9453, p-value = 0.3013
```

In the Q-Q plot, the points are not very close to the line, and the distances of the points from the line do seem to vary systematically in the way that a distribution slightly skewed left would. This fits with the graphs presented in part (a) which show that the distribution of diff is slightly skewed to the left. The Shapiro-Wilk test for normality has a p-value of .30, so we can't reject the null hypothesis that the distribution is Normal at the .05 level of significance, though the p-value does reflect as well that the distribution is nowhere near perfectly Normal. All of this leads to the conclusion that we can perform tests as if this data is Normal, while understanding that a larger sample size may or may not produce a normal distribution.

(c) Perform a one-sided t test on your data. Be sure to carefully state your hypotheses and your conclusion.

First, we check the conditions. As stated in part (b), the distribution of the differences is Normal enough that we can proceed with testing, so the first condition is met. In this case, μ_d is the mean of the differences in pulse rates. We will consider this psychology class as a random sample of UC Davis undergraduates, so the second condition is met.

Null Hypothesis H_0: $\mu_d = 0$. There is no difference between pulse rates during a quiz and pulse rates during a lecture.
Alternative Hypothesis H_a: $\mu_d > 0$. Pulse rates are higher during a quiz than they are during a lecture.

```
t.test(pulserates$quiz,pulserates$lecture,paired=T,alt="greater")
```

```
      Paired t-test

data:  pulserates$quiz and pulserates$lecture
t = 1.3662, df = 19, p-value = 0.09392
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.7172973        Inf
sample estimates:
mean of the differences
                    2.7
```

The p-value from the one-sided t-test is .094, so we fail to reject the null hypothesis that pulse rates are the same for quizzes and lectures at the 5% level of significance. The average difference in pulse rates during a quiz and during a lecture is not statistically significant at the 5% level of significance. The data suggest that pulse rates are not significantly different during quizzes and lectures.

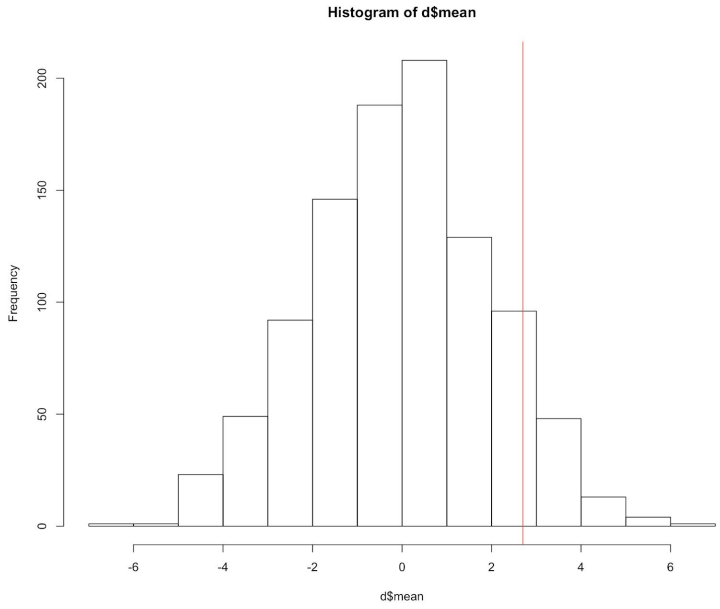
(d) In reaching your conclusion in part (c), what type of hypothesis testing error might you have made? Make sure your answer is about increases in pulse rates.

I might have made a type II hypothesis testing error in part (c) because I failed to reject the null hypothesis, and a type II error takes places if I failed to reject the null hypothesis but I should have because it does not accurately represent the population. This is like failing to convict someone who is guilty. In this case, it would mean that there is indeed a higher pulse rate during quizzes for UC Davis undergraduates, but my test failed to draw this conclusion.

(e) Perform a one-sided permutation test on these data using 999 simulations. Be sure to carefully state your hypotheses and your conclusion.

Null Hypothesis H_0: $\mu_d = 0$. There is no difference between pulse rates during a quiz and pulse rates during a lecture.
Alternative Hypothesis H_a: $\mu_d > 0$. Pulse rates are higher during a quiz than they are during a lecture.

```
d <- do(999)*mean(pulserates$diff*resample(c(-1,1),20))
hist(d$mean)
abline(v=2.7,col="red")
```




```
k <- sum(d$mean>=2.7)
(k+1)/1000
[1] 0.1
```

The p-value from the one-sided permutation test is .1, so we fail to reject the null hypothesis that pulse rates are the same during quizzes and lectures at the 5% level of significance. The average difference in pulse rates during a quiz and during a lecture is not statistically significant at the 5% level of significance. The data suggest that pulse rates are not significantly different during quizzes and lectures.

(f) Perform a one-sided Sign test on these data. Be sure to carefully state your hypotheses and your conclusion.

Null Hypothesis H_0: The median difference in pulse rates during quizzes and lectures is 0. M=0. Alternative Hypothesis H_a: The median difference in pulse rates during quizzes and lectures is greater than 0. M>0.

```
sum(pulserates$diff>0)
[1] 11
sum(pulserates$diff<0)
[1] 7
sum(pulserates$diff==0)
[1] 2
1-pbinom(10,18,.5)
[1] 0.2403412
```

We fail to reject the null hypothesis at the 5% level of significance because the p-value of .240 is greater than .05. The data suggest that the median difference in pulse rates during quizzes and lectures (quizzes-lectures) for UC Davis undergraduates is not significantly greater than 0.

(g) Why, in general, would you prefer the t test and the permutation test to the Sign test when analyzing paired data?

The sign test is wasteful because it ignores the magnitude of the differences for each pair and instead only looks at the signs of the differences. The t test and the permutation test both consider the magnitude of the differences, which sometimes has an effect on whether or not we reject the null hypothesis.