Math 227 Final Exam Due: December 19 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 December 19. 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 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**

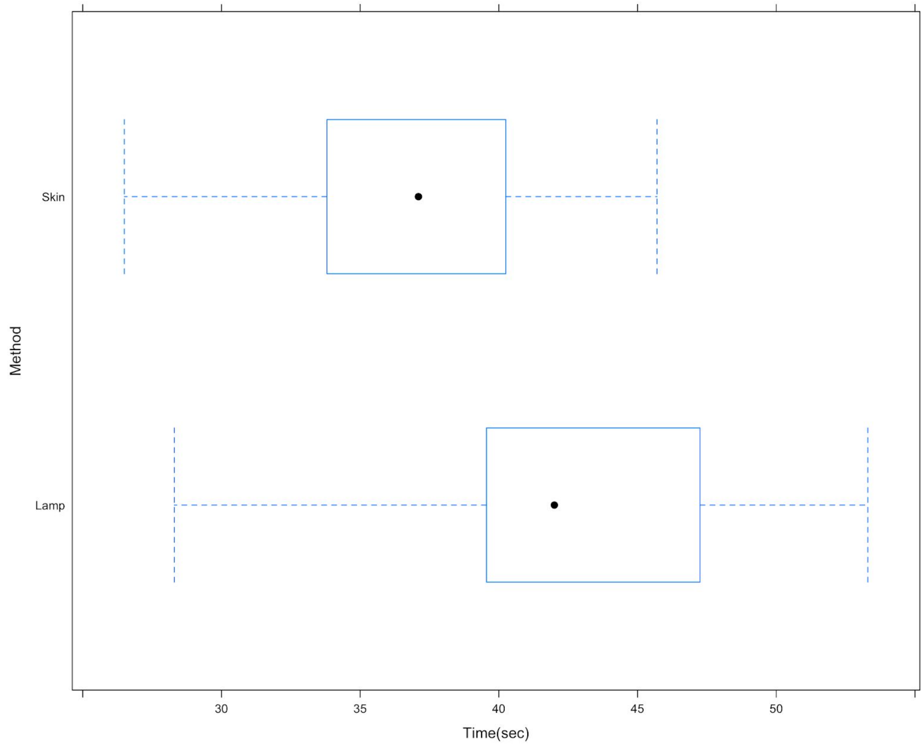
Please do not write on this page.

1. In a recent experiment at Beth Israel Hospital in Boston, 30 mothers were randomly assigned to one of two groups ("lamp' or 'skin'). The control group had their new-born infant warmed to 98ºF under a heating lamp; infants in the experimental group were warmed skin-to-skin (on the mother's chest). In this context, less time is better than more time. The warming times and the group identifications are in the data set *warming.csv*.

(a) Obtain both numerical and graphical summaries of the data. Write a brief account of the results.

```
favstats(Time~Method,data=warming)
  Method  min    Q1 median    Q3  max  mean    sd  n missing
1   Lamp 28.3 39.55   42.0 47.25 53.3 42.32 6.199 15       0
2   Skin 26.5 33.80   37.1 40.25 45.7 37.09 4.838 15       0
```

```
bwplot(Method~Time,data=warming,xlab="Time(sec)",ylab="Method")
```



The mean warming time for the Lamp method is 42.32 seconds with a standard deviation of 6.199 seconds, which is slower than the mean warming time for the Skin group, 37.09 seconds with a standard deviation of 4.838 seconds. Similarly, the median warming time for the Lamp group (42.0 seconds) is slower than the median warming time for the Skin group

(37.1 seconds). A comparison of boxplots of warming times for the two treatment groups confirms the general trend that newborns in the Lamp group warmed slower overall than those in the Skin group. Although the Lamp group has a greater spread and range than the Skin group, both extreme values and all quartiles for the Lamp group are larger times, and thus slower warming, than those for the Skin group.

(b) State the null and the alternative hypotheses for the appropriate test; define any parameters that you refer to.

The parameter µL is the population mean warming time in seconds of newborns warmed using the Lamp method and the parameter µS is the population mean warming time in seconds of newborns warmed using the Skin method.

H_0: µL - µS = 0. There is no difference between mean warming times for newborns warmed using the Lamp and Skin methods.
H_A: µL - µS ≠ 0. There is a difference between mean warming times for newborns warmed using the Lamp and Skin methods.

We will perform a two-sided Welch two-sample t-test.

Note- We could have chosen a one-sided test, as the data show that newborns warmed with the Lamp method have slower (larger) warming times than those in the Skin method. But, since that result was found after performing the experiment and not necessarily something we expected beforehand, we will perform the more conservative two-sided test.

(c) Paste in the output necessary to test the null hypothesis.

```
t.test(Time~Method,data=warming)

      Welch Two Sample t-test

data:  Time by Method
t = 2.6, df = 26, p-value = 0.02
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.063 9.403
sample estimates:
mean in group Lamp mean in group Skin
             42.32              37.09
```

(d) Carefully state your conclusion in context.

At the 5% level of significance, we reject the null hypothesis. The data suggest that the mean warming time for newborns warmed using the Lamp method is significantly different from the mean warming time for newborns warmed using the Skin method. Looking at our sample specifically, it appears as if warming times are slower (greater) for newborns in the Lamp method as compared to the Skin method.

Note- Had we chosen a one-sided test, our p-value would have been half of what we found for the two-sided test, so we also would have rejected the null hypothesis. Having performed a more conservative test and *still* found statistical significance at the 5% level was still worthwhile, though, given that results favoring either of the two "sides" of the two-sided test (Lamp>Skin and Skin>Lamp) seemed plausible to our limited knowledge *before* performing the experiment.

(e) Explain (in terms of warming times) which hypothesis-testing error you may have made in reaching your conclusion in part (c).

We may have made a Type I hypothesis-testing error, which would mean we rejected the null hypothesis but in actuality it is true so we should not have. In the context of warming times, this would mean we rejected the null hypothesis (that mean warming times are the same for newborns using the Skin and Lamp methods) but in actuality the mean warning times for the Skin and Lamp groups are indeed the same.

(f) Verify the calculation of the lower bound of the 95% confidence interval provided as part of the output. (Do use the values from the output in part (c).

```
lamp <- filter(warming,Method=="Lamp")
skin <- filter(warming,Method=="Skin")
sampdiff <- mean(lamp$Time)-mean(skin$Time)
sL <- 6.199
sS <- 4.838
nL <- 15
nS <- 15
SE <- sqrt((sL^2/nL)+(sS^2/nS))
t <- qt(.975,26)
me <- t*SE
lb <- sampdiff-me
lb
[1] 1.06
```

The value obtained here, 1.06, is approximately the same as the value for the lower bound of a 95% confidence interval provided in the output of the Welch test, 1.063.

(g) Explain (in context) what conditions are necessary for your test to be valid.

The first condition is that both samples are randomly selected from their respective populations. In context, this means that the newborns in the experiment warmed using the Lamp method represent a random sample of all newborns warmed via the Lamp method, and the newborns in the experiment warmed using the Skin method represent a random sample of all newborns warmed via the Skin method. Because the 30 newborns were randomly assigned to one of the two warming methods, we can view this condition as having been met.

The second condition is that the response variable is approximately normal in both populations. In context, this means that the distribution of warming times is approximately normal in both the population of all newborns warmed with the Lamp method and the population of all newborns warmed with the Skin method. We must test for Normality to see if this condition is met.

(h) Produce a formal test for Normality and state your conclusion.

We will perform the Shapiro-Wilks test for Normality twice, once for each population.

First, for the population of newborns warmed with the Lamp method:

H0: The distribution of warming times in the population of newborns warmed with the Lamp method is Normal.
Ha: The distribution of warming times in the population is not Normal.

```
shapiro.test(lamp$Time)
```

```
      Shapiro-Wilk normality test

data:  lamp$Time
W = 0.96, p-value = 0.7
```

We cannot reject the null hypothesis at any reasonable level of significance. The sample data is consistent with the distribution of warming times for newborns warmed with the Lamp method being Normal.

Second, for the population of newborns warmed with the Skin method:

H0: The distribution of warming times in the population of newborns warmed with the Skin method is Normal.
Ha: The distribution of warming times in the population is not Normal.

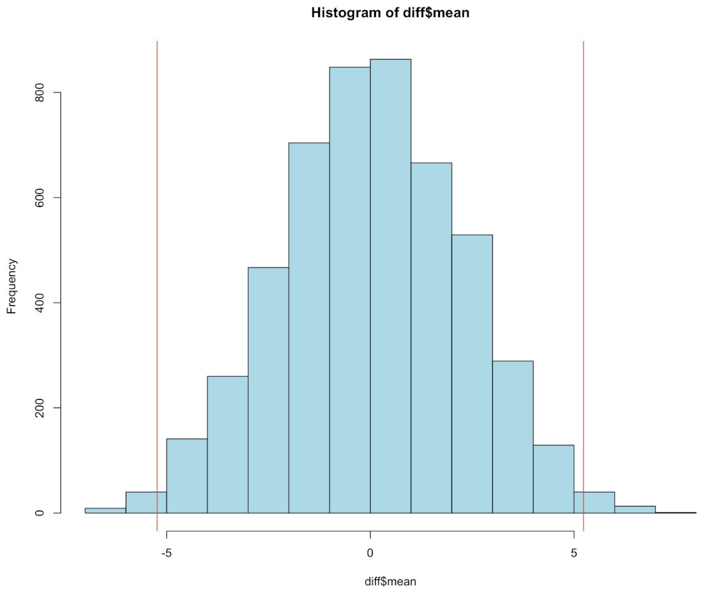
```
shapiro.test(skin$Time)

	Shapiro-Wilk normality test

data:  skin$Time
W = 0.98, p-value = 0.9
```

We cannot reject the null hypothesis at any reasonable level of significance. The sample data is consistent with the distribution of warming times for newborns warmed with the Skin method being Normal.

(i) Use R to perform a randomization test based on the difference in means. Use 4999 random assignments. Paste in output showing the result of your test. What is the p-value? What conclusion does this suggest?

We will perform a two sample, two-sided permutation test of the hypotheses stated in part (b).
H_0: μL - μS = 0. There is no difference between mean warming times for newborns warmed using the Lamp and Skin methods.
H_A: μL - μS ≠ 0. There is a difference between mean warming times for newborns warmed using the Lamp and Skin methods.

```
d <- do(4999)*mean(sample(warming$Time,15))
#Y = (15*YL+15*YS)/30
#YS=(30Y-15YL)/15
sum(warming$Time)
[1] 1191
#YS=(1191-15YL)/15
e <- (1191-15*d)/15
diff <- d-e
hist(diff$mean,col="lightblue")
abline(v=c(-5.233,5.233),col="red")
```



```
k <- sum(abs(diff$mean >= 5.233))
pvalue <- (k+1)/5000
pvalue
[1] 0.0084
```

With a p-value of .0084 we can reject the null hypothesis at the 1% level of significance. The data suggest that the mean warming time for newborns warmed with the Lamp method is significantly different than the corresponding mean warming time for newborns warmed with the Skin method. Looking at our sample, it appears as if warming times are slower (greater) for newborns in the Lamp method as opposed to the Skin method. A result such as the one in our experiment's sample is unlikely to occur in a population where the null hypothesis is true and mean warming times are the same for Skin and Lamp groups.

(j) Obtain a 95% bootstrap percentile confidence interval for the difference in population means (μ_L - μ_S). Use 10000 replications.

```
mean(lamp$Time)-mean(skin$Time)
[1] 5.233
d_lamp <- do(10000)*mean(resample(lamp$Time,15))
d_skin <- do(10000)*mean(resample(skin$Time,15))
di <- d_lamp$mean - d_skin$mean
q <- c(.025,.975)
qdata(~di,p=q)
      quantile     p
2.5%     1.407 0.025
97.5%    9.013 0.975
lb <- 2*5.233 - 9.013
```

```
ub <- 2*5.233 - 1.407
lb;ub
[1] 1.453
[1] 9.059
```

A 95% bootstrap adjusted percentile confidence interval for the difference in population means (μL-μS) is 1.453 to 9.059. In 95% of random samples of this size, the newborns in the Lamp method group will experience a slower average warming time than those in the Skin group by a difference within this range. This confidence interval does not contain 0, so we can be 95% confident that the difference between the two true means of warming time by each method (Lamp-Skin) is greater than 0.

(k) This is a bonus question worth 5 points

Obtain an 80% bootstrap percentile confidence interval for the ratio in population means (μ_L / μ_S). Use 10000 replications.

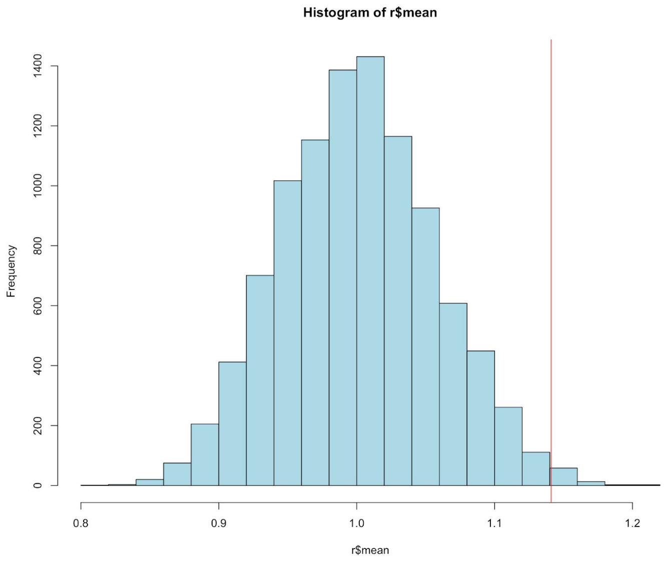
```
mean(lamp$Time)/mean(skin$Time)
[1] 1.141
dr <- d_lamp$mean/d_skin$mean
q <- c(.1,.9)
qdata(~dr,p=q)
    quantile   p
10%    1.071 0.1
90%    1.215 0.9
lb <- 2*1.141-1.215
ub <- 2*1.141-1.071
lb;ub
[1] 1.067
[1] 1.21
```

An 80% bootstrap adjusted percentile confidence interval for the ratio in population means (μL/μS) is 1.067 to 1.21. In 80% of random samples of this size, the ratio is within this range. This ratio does not include 1, so we are 80% confident that the ratio between the true population means (Lamp/Skin) is greater than 1. This would mean that μL>μS.

Perform a one-sided permutation test of the null hypothesis that the ratio $\mu_L / \mu_S = 1$ against the alternative H_A: $\mu_L / \mu_S > 1$. Use 9999 replications.

```
d <- do(9999)*mean(sample(warming$Time,15))
#YS=(1191-15YL)/15
e <- (1191-15*d)/15
r <- d/e
```

```
hist(r$mean,col="lightblue")
abline(v=c(1.141),col="red")
```



```
k <- sum(r$mean>=1.141)
pvalue <- (k+1)/10000
pvalue
[1] 0.0072
```

With a p-value of .0072, we reject the null hypothesis at the 1% level of significance. The data suggest that the ratio of population means (Lamp/Skin) is significantly greater than 1. A result such as the one in our experiment's sample is unlikely to have occurred if the null hypothesis were true and mean warming times were the same for Lamp and Skin warming methods.

2. The Simmons nurses who organized the study described in Q. 1 were pleased with the results but regarded it as a pilot study. They are writing a proposal for a larger scale randomized experiment but need help with the sample size calculations. They would like the number of new-born babies in each warming group to be large enough so that there is an 85% chance of detecting a difference of μ_L - μ_S = 2 minutes.

(a) Please advise! (You will need to enter a guesstimate for the common standard deviation (σ) in your code. Combine the two sample values (for S) from the pilot.)

```
sd <- (6.199+4.838)/2
```

```
sd
[1] 5.518
> power.t.test(n=NULL,delta=2,sd=5.518,power=.85,alternative="one.sided")

     Two-sample t test power calculation 

              n = 110.1
          delta = 2
             sd = 5.518
      sig.level = 0.05
          power = 0.85
    alternative = one.sided

NOTE: n is number in *each* group
```

The nurses should have 111 newborns in each warming group. This way, there'll be an 85% chance of detecting a difference (Lamp-Skin) of 2 minutes, if they're testing with the standard significance level of .05.

(b) How would your advice change if the nurses anticipate that 10% of those mothers enrolled in the experiment will be unable to participate for one reason or another?

```
111/.9
[1] 123.3
```

I'd advise that they take a sample large enough so that if only 90% of those enrolled in each group participate, there will still be 111 newborns in each warming group. Thus, I advise that they choose 124 newborns for each warming group.

3. Election officials in Kings County, CA were alarmed at the low voter turn-out in their county at the last gubernatorial election in the state and sought ways of improving it. They decided to conduct a randomized experiment to compare methods for encouraging registered voters to actually vote in the up-coming election for governor. The experiment involved 600 registered voters. The 600 were randomly divided into three groups of (roughly) 200. All 600 received the standard notice of the election with details of the candidates and voting places. Group A received no other information from the County Board of Elections. Those in Group B received a personalized letter from the Board pointing out the low turnout in the past and urging that the recipient vote. Those in group C received a personalized letter urging the recipient to vote but also contained a promise to publish and circulate a list of people in the neighborhood who did not vote. A note was made of the names of the 600 registered voters that did vote. The results are in the data set ~~voters.csv~~. **(voting.csv)** You are to compare the three groups.

(a) Perform a descriptive analysis—that focuses on both percentages and odds ratios.

```
voting$voted <- factor(voting$voted,levels=c("Yes","No"))
tally(voted~group,data=voting,margin=T)
       group
voted     A   B   C
  Yes    58  67  89
  No    143 129 114
  Total 201 196 203
58/201 #A
[1] 0.2886
67/196 #B
[1] 0.3418
89/203 #C
[1] 0.4384
```

The percentage of those in group A who voted is 28.86%. Of those in group B, 34.18% voted, which is higher than the percentage for group A. Finally, for group C, 43.84% of individuals voted, which is higher than the percentages for both groups A and B. Since C received more information than B which received more information than A, it appears as if the more experimental "treatment" a group received, the higher the percentage of its members who voted.

```
58/143 #OA
[1] 0.4056
67/129 #OB
[1] 0.5194
89/114 #OC
[1] 0.7807
.7807/.5194 #OC/OB
[1] 1.503
.7807/.4056 #OC/OA
[1] 1.925
.5194/.4056 #OB/OA
[1] 1.281
```

The odds that a member of each group voted follows the same trend as the percentages of each group who voted, where members of group B had a higher odds of voting than members of group A, and members of group C had a higher odds of voting than both those of A and B. Since the meaning of odds is most interesting when they're compared relationally, we will look at the odds ratios. The odds of voting for a member of group B was 1.281 times the odds of voting for a member of group A. The odds of voting for a

member of group C was 1.925 the odds of voting for a member of group A and 1.503 the odds of voting for a member of group B. Again, this fits with the trend we've identified.

(b) Perform an inferential analysis—that focuses on a Chi-Square analysis.

We will perform a Chi-square Test for Independence. First, we must check the conditions. The independent count condition is met, as no individual was a part of multiple groups or voting statuses. The random sample condition is met because registered voters were randomly assigned to treatment groups.

```
chisq.test(voting$group,voting$voted)$expected
            voting$voted
voting$group   Yes    No
           A 71.69 129.3
           B 69.91 126.1
           C 72.40 130.6
```

Finally, the expected count condition is met because all expected counts were greater than 5.

H0: The probability of voting is the same for members of all three groups. Voting status (yes or no) is independent from group status (A, B, or C).
Ha: The probability of voting is different for members of each experimental treatment group. Voting status is not independent from group status.

```
chisq.test(voting$group,voting$voted)

      Pearson's Chi-squared test

data:  voting$group and voting$voted
X-squared = 10, df = 2, p-value = 0.006
```

We can reject the null hypothesis at the 1% level of significance. The data suggest that the probability of a group member voting is not the same for each treatment group. Looking at our sample data, it's likely that more information provided in an experimental treatment, the higher the probability of a member voting.

(c) Write a brief summary of your results in parts (a) and (b).

The result of the Chi-square Test for Independence are consistent with our initial analysis of percentages and odds ratios. There does seem to be a relationship between the treatment

groups and whether or not group members voted. The county's interventions probably did make a statistical difference in whether or not registered voters bothered to vote, with more interventions in the form of more information and "treatment" leading to higher probabilities and odds of members voting.

4. A headline in a recent edition of *USA Today* stated "Attending religious services lowers blood pressure more than tuning in to religious TV or radio, a new study shows."

The study referred to was conducted by the U.S. National Institute of Health. The study followed 2391 people aged 65 or older for six years. The article described one of the study's principal finding: "People who attended a religious service once a week and prayed or studied the Bible once a day were 40% less likely to have high blood pressure than those who don't go to church every week and prayed and studied the Bible less". The study reported a p-value of < 0.001 for the Chi-Square test.

(a) Explain how you know this study is observational rather than a randomized experiment.

If this study were a randomized experiment, the people would have been randomly assigned to a different treatment group involving some combination of attending church, praying, and studying the Bible, and the treatment would have been imposed on the people. That is not the case in this study, as no treatments were imposed. Individuals' status as having regular religious habits or not and status as with or without high blood pressure were simply observed.

(b) A colleague is impressed with the tiny p-value and concludes that she should start attending her church more regularly because of her high blood pressure. Explain carefully (and preferably with examples) why she should not jump to this conclusion.

She should not jump to this conclusion because of the possibility of confounding variables. A confounding variable is something that is initially distributed differently between the two groups and distorts the estimate of causal effect of the explanatory variable on the response variable. In this case, it would be a variable that has different distributions over the two groups, which I will simplify by referencing as religious and nonreligious people. The difference has to be something that has an effect on someone's blood pressure. A possible example is stress levels. Is it plausible that religious people and nonreligious people have generally different stress levels? Yes, it is, since religious people are more likely to believe that everything happens according to god's plan and thus would be less stressed out about things in life because they have a faith that everything will work out well as long as they

remain devoted to religion. Nonreligious people could quite plausibly have higher stress levels in general than religious people, since they're likely lacking this faith. Stress levels definitely have an impact on blood pressure, as people with high stress levels are more likely to have high blood pressure. So, despite the small p-value of the Chi-Square test, we can't be sure that differences in high blood pressure status are directly caused by differences in status of religious habits, since such a clear confounding variable exists. There are other possible confounding variables that are similarly likely to be responsible for some of the differences in blood pressure between religious and nonreligious people. It's plausible that religious people are less likely to be alcoholics, since spirituality is an important component to common treatments of alcoholism such as AA. Alcoholism and general high levels of alcohol consumption are another cause of high blood pressure, so this meets the requirements to be a confounding variable as well, since it's plausibly distributed differently for religious people than it is for nonreligious people, and it has an impact on increasing one's blood pressure. Thus, the colleague should not jump to the conclusion of attending church in the hopes of lowering her blood pressure.