PH142 Fall 2020 Final Exam SOLUTIONS

The exam is open book. This means you can use electronic or hard copies of notes from class. You may not use the internet to search for the answers or inform your answers. Using the internet is strictly prohibited and any evidence of this may result in a 0 on the exam.

While you take the exam, you are prohibited from discussing the test with anyone. Evidence of cheating may result in a 0 on the exam.

No copies of the exam are to be saved.

I affirm that I have read and agree to the above statements.

Berkeley's code of conduct is here: https://sa.berkeley.edu/code-of-conduct. See Section V and Appendix II for information about how UC Berkeley defines academic misconduct. In particular the sections on cheating and plagiarism.

Question 1: Honesty Statement

Question 2: 1 point

Question 3: 1 point

Question 4: 1 point

Question 5: 1 point

Question 6: 3 points

Question 7: 2 points

Question 8: 2 points

Question 9: 7 points

Question 10: 6 points

Question 11: 6 points

Question 12: 3 points

Question 13: 7 points

Question 14: 7 points

Question 15: 5 points

Question 16: 6 points

Question 17: 1 point

Bonus!: 1 point

Total: 59 points + 1 bonus point

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(grid)
library(broom)
library(here)
```

```
## here() starts at /Users/kelseymaccuish/ph142-dev
```

Question 1

1. Write your initials to affirm that you have read and agree to the above statements:

Question 2 [1 point total]

2.1 [1 point] Which of the following is FALSE about the Plus 4 method? Pick one statement only.

a) The Plus 4 method adds exactly two imaginary trials to the sample size.
b) The Plus 4 method is a good choice when n is at least 10 and the confidence level is at least 90%.
c) The Plus 4 method is a simplification of the Wilson Score Interval method.
d) The Plus 4 method can be used for the comparison of two proportions when the dataset is small.

```
# SOLUTION: a) The Plus 4 Method adds four imaginary trials to the sample size.
```

Question 3 [1 point total]

3.1 [1 point] True or False. The chi-square test statistic will always be positive.

a) True
b) False

```
# SOLUTION: a) True. The squared term in the numerator always makes
# the chi-square test statistic positive.
```

Question 4 [1 point total]

4.1 [1 point] Although bootstrap confidence intervals are powerful, there are limitations when:

a) making bootstrap CIs around only the standard deviation.
b) the sample is not SRS from the underlying population.
c) the underlying distribution is not Normally distributed.
d) the sample size is small.

```
# SOLUTION: b) the sample is not SRS from the underlying population.
# Bootstrap relies on only the assumption that the sample is a SRS from the underlying population.
```

Question 5 [1 total]

5.1 [1 point] The side effects (nausea or no nausea) of an anti-viral drug, Remdesivir, as a potential treatment for COVID-19 are being tested against a placebo. 100 patients enrolled in a study were randomly assigned to placebo or Remdesivir. At a significance level of 0.05, we performed a _____ to see if there is enough evidence to conclude that the treatment is independent of the side effect of nausea. We got a p-value of 0.0012, indicating there _____ a relationship between the treatment and nausea.

a) two-sample t-test, does NOT exist
b) two-sample t-test, exists
c) Chi-square test, does NOT exist
d) Chi-square test, exists

```
# SOLUTION: d)
# we are using Chi-square test to test for the independence between two variables.
# Since the p-value is smaller than 0.0012, we are rejecting the
# null hypothesis that they are independent.
```

Question 6 [3 points total]

6.1 [3 points] For each statistical test, fill in the blanks to indicate the types of variables that are needed to conduct it. A key word bank has been provided to help you. Note, the words can be used more than once or not at all.

Key word bank: Continuous, Categorical, Response, Explanatory, Groups, Two, K $>$ 2, K $\geq$ 2, Dependent, Independent

- Independent two sample t test:
  _____ _____ samples with a _____ _____ variable
- ANOVA:
  _____ _____ of a _____ variable and a _____ variable
- Linear regression introduced in part I of the course:
  _____ _____ variable and _____ _____ variable
- Chi-sq test for independence:
  _____ _____ of a _____ _____ variable and a _____ variable
- Wilcoxon Sign Rank: _____ _____ samples with a _____ _____ variable

```
# SOLUTION:
#* Two sample t test: Two independent samples with continuous response variable
#* ANOVA: K > 2 groups of a categorical variable with continuous variable. K>2 is also okay.
#* Linear regression: Continuous explanatory variable and continuous response variable
#* Chi-sq test for goodness of fit: K>=2 groups of a categorical variable and categorical outcome.
#* Wilcoxon Sign Rank: Two dependent samples with continuous response variable.
```

Question 7 [2 points total]

7.1 [2 points] Below is ANOVA output for an analysis of 3 distinct groups. Fill in the blanks labeled A-D to 4 decimal places where applicable. Blanks labeled "NA" do not need to be filled in.

term	df	sumSq	meanSq	statistic	p.value
treatment	A	3.766	C	D	0.0159
Residuals	27	B	0.3886	NA	NA

A:

B:

C:

D:

```
# A: <p> 2 (3 groups minus 1)
# B: <p> 10.4922 = 27*0.3886
# C: <p> 1.883 = 3.766/B = 3.766/10.4922
# D: <p> 4.8456 = C/0.3886 = 1.883/0.3886
```

Question 8 [2 points total]

8.1 [2 points] State and explain one reason why Tukey HSD would be used when conducting an ANOVA test.

```
# SOLUTION: Tukey HSD maintains an overall error rate of some alpha level,
# typically 5%.
# It is needed to correct for multiple testing.
```

Question 9 [7 points total]

Having natural red hair is rare (occurring in approximately 2% of individuals) and is the phenotype for a mutation in the melanocortin 1 receptor. There is research that suggests people with red hair have a lower pain tolerance and thus often require higher doses of pain medication and anesthesia. We are interested in examining if there is an association between hair color and pain tolerance.

1125 people volunteered to be part of a pain study, of which 1046 had brown or black hair, 56 had blonde hair, and 23 had red hair. Each person was administered subcutaneous lidocaine

(a numbing agent injected into the skin) and was administered electric shocks on a scale of 0 to 30 milliampere (mA). Each person was told to hold a button down and release the button when the shocks became too painful. The mA at which the button was released was recorded for each person.

Based on these data, we conducted a hypothesis test in R. The results are as follows:

term	estimate	std.error	statistic	p.value
(Intercept)	19.87	1.03	19.29	0.0000000625
blonde	-4.62	3.98	-1.16	0.322
red	-9.85	4.64	-2.12	0.0987

9.1 [1 point] Not applicable for Spring 2022 exam

9.2 [1 point] For the research question of interest, state the null hypothesis.

H_0 :

```
# SOLUTION: H_0: There is no difference in pain tolerance between
# those with brown hair and those with red hair. beta_2 = 0
```

9.3 [2 points] What conclusion can you make about having red hair and pain tolerance based on the model output at an alpha level of 0.05?

```
# SOLUTION: There is no significant difference between
# pain tolerance in those with red hair and those with brown hair.
```

9.4 [3 points] What is the average milliampere (mA) at which persons of each hair color released the button to indicate they reached their upper bound of pain tolerance?

a) mA for people with blonde hair:
b) mA for people with brown hair:
c) mA for people with red hair:

```
# SOLUTION:
# A: 19.87
# B: 15.25
# C: 10.02
```

Question 10 [6 points total]

The LA County Department of Public Health has asked you to analyze data to test whether there are disparities in the proportion of COVID-19 deaths by race/ethnicity. These data show the number of cumulative COVID-19 deaths by race/ethnicity in Los Angeles County as of November 2020:

Race/Ethnicity	Latinx	White	Asian	Black	Unknown	Total
Number of Deaths	9213	5757	2309	1385	167	18831

The % of the LA county population each race/ethnic group is as follows:

Race/Ethnicity	Latinx	White	Asian	Black	Other	Total
Percent of population	48.6%	26.1%	15.4%	9.0%	0.9%	100%

Is there evidence that the deaths are unevenly distributed by race/ethnicity?

10.1 [2 points] What are the null and alternative hypotheses for this statistical test?

H_0 :

H_A :

```
# SOLUTION:
# H0: proportion of deaths in each race/ethnic group is the same as the proportion
# of each group in the county; p_latinx = 48.6%; p_white = 26.1%;
# p_asian = 15.4%; p_black = 9.0%; p_other = 0.9%

# HA: At least one of these p_k differ from what is stated in H0,
# where k is latinx, white, asian, black, or other.
```

10.2 [1 point] What kind of statistical test would you use?

a) Two Sample t Test
b) Chi-Square Test of Independence
c) Paired T test
d) Chi-Square Goodness of Fit Test
e) None of the Above

```
# SOLUTION: d) Chi-Square Goodness of Fit
```

10.3 [2 points] Calculate the appropriate test statistic by hand. Show your work.

```
# SOLUTION:
# chi^2 = [(9213-9151.866)^2 / 9151.866] + [(5757-4914.891)^2 / 4914.891] +
# [(2309 -2899.974)^2 / 2899.974]+
# [(1385-1694.79)^2 / 1694.79] + [(167-169.479)^2 / 169.479] =
# 321.7887
```

10.4 [1 point] Write code to calculate the p value based on this test statistic.

```
# SOLUTION: pchisq(q = 321.7887 , df = 5 - 1, lower.tail = FALSE)
```

Question 11 [6 points total]

You conduct a study to test whether there is a significant difference in effectiveness of a brand name antidepressant, Zoloft, and its generic form, Sertraline Hydrochloride. One way to test drug effectiveness is to test the extent that it is absorbed in the blood. You collect blood absorption data from 20 individuals at two time points: 10 individuals are randomly assigned to the generic drug first and 10 are assigned to the brand name drug first. After a washout period to completely eliminate the first drug from the blood, each individual is given the other drug. You would like to test whether the drugs differ significantly in blood absorption.

Here is a summary of the data you collected:

```
summary
```

```
## # A tibble: 1 x 6
##   mean_zoloft mean_sert mean_diff sd_zoloft sd_sert sd_diff
##         <dbl>     <dbl>     <dbl>     <dbl>   <dbl>   <dbl>
## 1       2048.     2086.       -37      860.    643.   1071.
```

11.1 [1 point] What are the null and alternative hypotheses for this test?

H_0 :

H_A :

```
# SOLUTION:
# H0: There is no difference in the blood absorption of Zoloft compared
# to Sertraline Hydrochloride; diff = 0.
# HA: There is a difference in the blood absorption of Zoloft compared
# to Sertraline Hydrochloride; diff does not = 0.
```

11.2 [1 point] What are the degrees of freedom for the most appropriate statistical test?

a) 35.183
b) 19
c) 20
d) 18

```
# SOLUTION: b) 19. The data are paired so df = 20-1
```

11.3 [2 points] Calculate the 95% confidence interval for the difference in blood absorption between Zoloft and Sertraline Hydrochloride by hand (show your work for full credit). Use 2.09 as your critical value.

```
# SOLUTION:

# lowerbound: -37 - 2.09*(1070.622/sqrt(20)) = -537.3426
# upperbound: -37 + 2.09*(1070.622/sqrt(20)) = 463.3426
```

11.4 [2 points] Is the associated p-value greater or less than 0.05 based on this 95% confidence interval? How do you know? What can you conclude about the difference in blood absorption between Zoloft and Sertraline Hydrochloride?

```
# SOLUTION: The p-value is greater than 0.05.
# The confidence interval crosses the value of 0, which means there is no evidence
# against the null that the true difference in the mean blood absorption
# between Zoloft and Sertraline Hydrochlroide is 0.
# There is evidence that the difference is 0.
```

Question 12 [3 points total]

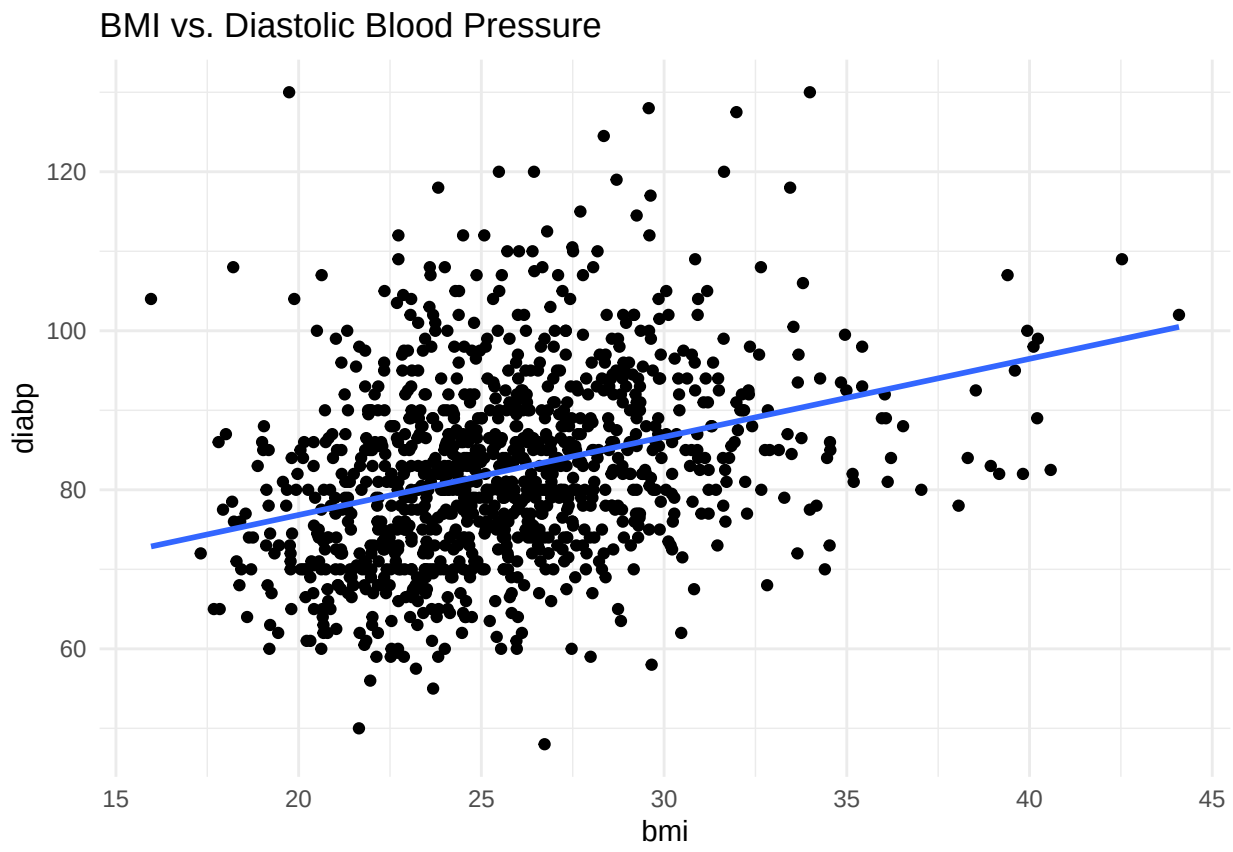
In your own words, what are the steps for constructing a 95% bootstrap confidence interval for a standard deviation? You may assume you have a simple random sample with n observations and your sample has a standard deviation of s. Be as detailed, but concise, as you can in 5 sentences or less.

```
# SOLUTION:
# [0.5 mark] Calculate your sample standard deviation.
# [1] Need to say many/or specific large number and need to say with replacement: Take many/1000/10k re
# [0.5] For each bootstrap sample, calculate the standard deviation.
# [0.5] Construct a histogram of all of the calculated sample standard deviations.
# [0.5]Do not necessarily need to say using the quantile function, but do need to say with percentiles
# Using the quantile function in R, find where the 2.5 percentile and 97.5% are.
# These are your upper and lowerbounds for the 95% confidence interval.
```

Question 13 [7 points total]

The Framingham Study is a population-based, observational cohort study initiated by the United States Public Health Service in 1948 to prospectively investigate the epidemiology and risk factors for cardiovascular disease. Suppose you are interested in the relationship between BMI and Diastolic Blood Pressure. Below is a scatter plot depicting the relationship between body mass index (BMI) and Diastolic Blood Pressure from the study along with its line of best fit. Following the scatter plot is the output of a linear regression model, named `model`, with the explanatory variable `bmi` and the response variable `diabp` = Diastolic Blood Pressure. The data set has 4415 observations and is called `framingham`.

```
suppressWarnings(print(ggplot(data=framingham, aes(x=bmi, y=diabp)) + geom_point() +
                       geom_smooth(method='lm', se = F) + labs(title='BMI vs. Diastolic Blood Pressur
```




```
tidy(model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   57.2      2.23        25.6 9.42e-114
## 2 bmi            0.982    0.0860      11.4 1.33e- 28
```

13.1 [1 point] Write code that would generate the above plot. Remember: the variables are `bmi`, `diabp`, `model`, `framingham`. You may assume all necessary packages are loaded.

```
# SOLUTION:
# ggplot(data=framingham, aes(x=bmi, y=diabp)) +
#    geom_point() + geom_smooth(method='lm') +
#    labs(title='BMI vs. Diastolic Blood Pressure')

# alt. solution for line of best fit: geom_abline(aes(yintercept = 57.2, slope = 0.982))
```

13.2 [1 point] Write the null and alternative hypotheses in words or using notation.

H_0 :

H_A :

```
# SOLUTION: Null hypothesis: $b=0$. Alternative hypothesis:$b\neq0$
# There is no relationship between bmi and diastolic blood pressure.
# There is a relationship between bmi and diastolic blood pressure.
# Alt solution: BP is independent of BMI. BP is dependent on BMI.
```

13.3 [2 points] Compute a 95% confidence interval by hand for the statistic of interest. Use a critical value of 1.96. Show your work and round your answers to two decimal places.

```
# SOLUTION: 0.982 +/- 1.96*0.0860 = (0.81, 1.15)
```

13.4 [3 points] Interpret both the BMI slope coefficient and the 95% confidence interval. Use this information to make a conclusion about your null hypothesis in the context of this study.

```
# SOLUTION: The estimate for the slope coefficient of BMI is 0.982 (95% CI: 0.81 to 1.15).

# This indicates for every one unit increase in BMI,
# diastolic blood pressure increases by 0.98 units.
```

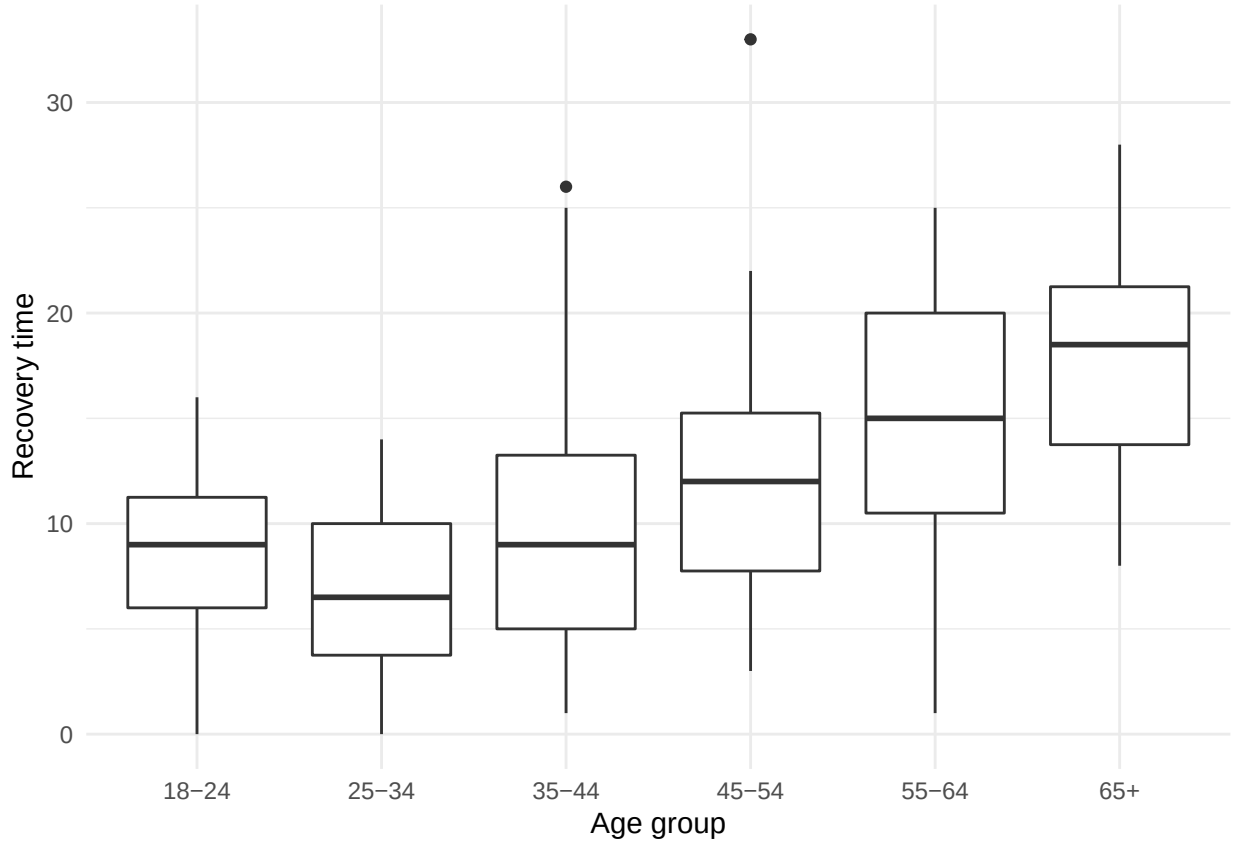
```
# The confidence interval is (0.81 to 1.15). If we were to take 100 random samples
# from the population and constructed this statistic 100 times,
# we would expect 95 of them to contain the true value
# of the slope between BMI and Diasp.

# Since 0 is contained in this interval, we fail to reject the
# null hypothesis that there is no relationship bmi and diastolic blood pressure.
```

Question 14 [7 points total]

You are interested in seeing if the mean recovery time for different age groups who have tested positive for COVID-19, presented with symptoms, and recovered is different. You are given the following data where `recovery` is in days for a simple random sample of 20 people in each of six age groups in the Bay Area. *Note: This data is contrived and not from any source.* Here is the `head()` of the data for reference, some summary statistics, and boxplots. Note that in the data set, there is one row of data per person.

```
##   age_group recovery
## 1     18-24       13
## 2     18-24       16
## 3     18-24       11
## 4     18-24        4
## 5     18-24        6
## 6     18-24       12
```

```
## # A tibble: 6 x 3
##   age_group mean_recovery_time sd_recovery_time
##   <chr>                  <dbl>            <dbl>
## 1 18-24                   8.45             4.36
## 2 25-34                   6.65             4.12
## 3 35-44                  10.2              7.01
## 4 45-54                  12.7              7.07
## 5 55-64                  14.4              7.12
## 6 65+                    17.6              5.47
```



14.1 [1 point] What parametric test would be appropriate for answering this question? Write the null and alternative hypotheses in words or notation. Hint: parametric tests are the ones that have corresponding chapters in the textbook.

```
# SOLUTION: ANOVA
# There is no difference in the average recovery time for the age groups.
# At least one of the groups has a different average recovery time.
```

14.2 [1 point] What non-parametric test would answer the same question?

```
# SOLUTION: Kruskal Wallis
```

14.3 [2 points] Knowing the names of the variables are `age_group` and `recovery`, and the dataset is called `covid`, write R code for conducting both tests you wrote in questions 14.1 and 14.2.

```
# SOLUTION:
# aov(recovery~age_group, data = covid)
# kruskal_wallis(recovery~age_group, data = covid)
```

14.4 [3 points] Which of the two tests is more appropriate for this data? Justify your choice by discussing the assumptions for each test. Note if you need more information determine if an assumption is met.

```
# SOLUTION:
# ANOVA is a better fit for this data. All of the distributions are roughly symmetric and
# normally distributed according to the boxplots.
# None of the groups has a standard deviation more than
# 2 times another one, and 20 samples in each category meets
# the sample size requirement.
```

Question 15 [5 points total]

For this question, suppose you have data on two groups of people – half exposed to treatment A and half exposed to treatment B. For each treatment group, you measure a continuous outcome variable of interest. You want to know if the assigned treatments affects the outcome variable.

15.1 [1 point] State the null and alternative hypotheses for a permutation test used to examine this question. You may use words or notation. If using words, be careful with your word choice.

H_0 :

H_A :

```
# SOLUTION:
# Null: mu_1 = mu_2
# (no difference in the population means of the two groups)
# Alternative: mu_1 != mu_2
# (there is a difference in the population means of the two groups)
```

15.2 [2 points] In a permutation test, we reshuffle the labels in our data many times. Keeping in mind your hypotheses above, why do we reshuffle the labels? Explain in 1-3 sentences.

```
# SOLUTION:
# If we reshuffle the labels, and the null hypothesis is true,
# we should see no difference in the distribution of the data than in our original sample

# Alt description: reshuffling the labels breaks any association between
# the the treatment group and the outcome variable.
# Thus we reshuffling to see what the association would look like in a
# case where we *know* the null to be true --
# so we can compare it to what we saw in the unshuffled dataset.
# If what we saw in unshuffled data is very different,
# then this provided evidence against null in favour of alternative.
```

15.3 [2 points] Say you reshuffle your labels and compute your statistic 10,000 times. State two ways you can calculate the p-value for your test statistic. Hint: one of the ways has R do more of the work for you and requires a specific package.

```
# SOLUTION:
# You can use the infer package and the `get_p_value()` function.
# You can count up how many reshufflings resulted in a test statistic
# more extreme than yours and divide that by 10000.
```

Question 16 [6 points total]

You are hired by a womens' health clinic to assist them in conducting a study. One of the primary questions of this study is to identify a difference in probability of infertility in women with and without endometriosis. You randomly sample 10,000 patients of this clinic and they agree to answer questions about their health. Below are the results for two questions regarding endometriosis and infertility.

	Endometriosis	No Endometriosis	Total
Infertile	605	1022	1627
Fertile	873	7500	8373
Total	1478	8522	10000

We learned about a few ways to make inference about whether there is a difference between the probability of infertility between two groups of people. For questions 16.1, 16.2, and 16.3, think about the first way we covered – this method allows us to conduct a hypothesis test and compute confidence intervals.

16.1 [1 point] State the null and alternative hypotheses for this test.

H_0 :

H_A :

```
# SOLUTION: H_0: p1 - p2 = 0
# H_A: p1 - p2 != 0
```

16.2 [2 point] Calculate the test statistic by hand. Show your work to receive full credit. Round to 4 decimal places.

```
# p1_hat - p2_hat/se = 0.4093-0.1199/sqrt(0.00016 + 0.00001238)
# = 22.0422

# se = sqrt(p_hat(1-p_hat)*(1/n1 + 1/n2))
# where p_hat = 1627/10000
```

16.3 [2 points] Construct a 95% confidence interval. Show your work to receive full credit.

```
# SOLUTION: p1_hat - p2_hat +- 1.96*se =
# 0.4093-0.1199 += 1.96*sqrt(0.00016 + 0.00001238) =
# (0.1636, 0.3150)

# se = sqrt(p1_hat(1-p1_hat)/n1 + p2_hat(1-p2_hat)/n2)

# could also do plus 4
```

16.4 [1 point] List one other method you could have used to conduct this hypothesis test.

```
# SOLUTION:
# could have done a permutation test
# chi-square test for independence
```

END