Fall 2023 Final Exam

The exam is closed book and closed notes. You are allotted **two** double sided "cheat sheet" which may contain typed or handwritten notes. You may also use a calculator. Your phone is not allowed as a calculator. Using any resources outside of the aforementioned items is strictly prohibited.

While you take the exam, you are prohibited from discussing the test with anyone. If you are taking the test after your classmates, you are also prohibited from talking to them about the test before you take it. Evidence of cheating may result in a 0 on the exam and be reported to the Student Conduct Board.

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 note the sections on cheating and plagiarism.

UC Berkeley Honor Code
"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." Please carefully read the statements below, and indicate your understanding and intent to adhere to the UC Berkeley Honor code by typing your name in the space below. I agree not to engage in any of the following behaviors:

- Copying or attempting to copy from others during an exam or on an assignment.
- Communicating answers with another person during an exam.
- Pre-programming a calculator or other personal electronic device to contain answers, or using other unauthorized information for exams.
- Using unauthorized materials, i.e. prepared answers.
- Allowing others to do an assignment or a portion of an assignment for you, including the use of a commercial term-paper service.
- Submitting the same assignment for more than one course without prior approval of all the instructors involved.
- Collaborating on an exam or assignment with any other person without prior approval from an instructor.
- Taking an exam for another person or having someone take an exam for you.
- Altering a previously graded exam or assignment for the purpose of a grade appeal or of gaining points in a re-grading process.
- Submitting an electronic file the student knows to be unreadable or corrupted instead of a completed assignment.

Write your name and SID below.

Enter your name:

Enter your SID:

INSTRUCTIONS:

Hand write your responses using a pencil or pen in the space provided. Give your responses ONLY in the space provided. Do not write your responses outside of the provided text boxes. The additional space provided, including space on the back side of the exam can be used as scratch paper but will not be graded.

Phones should be turned OFF prior to the start of the exam and secured in your backpack or another secure location. Do not leave your phone or other electronic devices out. If you need to leave the room for any reason during the exam please flag a GSI and let them know prior to exiting the room. Time will still accrue when you leave the room.

The length of the final exam is 180 minutes. If you finish early and are satisfied with your work you may leave early. Hand in your exam to a GSI, who will verify that they received it.

- Unless otherwise specified in the question, format your answers according to the following guidelines:
  - present your answers rounded to two decimal places
  - present proportions as % values (40.50% rather than .405)
- All logs are natural log base e

Exam Format:

Short Format Questions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 [12 points]
Long Format Question: 11 [6 points]
Long Format Question: 12 [5 points]
Long Format Question: 13 [7 points]
Long Format Question: 14 [7 points]
Extra Credit Questions: EC1, EC2, EC3 [4 points]

Short Format Questions:

Fill in the box clearly next to your responses. If your answer choice is unclearly marked, we will not count it towards your answer choice.

There is only one answer choice unless marked with [Select All That Apply]

1. [1 point] In a study comparing the lengths of hospital stays between two groups (treatment and control), the data is non-normally distributed. What is the most suitable statistical test for determining if there is a significant difference in hospital stays between the two groups?

- ☐ A. Wilcoxon Sign Rank Test
- ☐ B. Wilcoxon Rank Sum Test
- ☐ C. Chi-square test of independence
- ☐ D. Friedman Test

```
# Solution: Correct. B) Wilcoxon Rank Sum Test. Since the treatment and control group are two
#independent samples, we would choose the Wilcoxon Rank Sum Test.
```

2. [1 point] Which of these four plots is designed investigate the normality of a random variable? Select ALL that apply.

- ☐ A. Histogram
- ☐ B. Bar graph
- ☐ C. Boxplot
- ☐ D. Scatterplot
- ☐ E. QQ-plot

```
# Solution: Correct- A. Histogram and E. Q-Q Plot + C. Boxplot
```

3. [1 point] Which of the following elements are definitely related to the power to detect the difference between mean values in two different groups? Select ALL that apply.

- ☐ A. The true difference between two means
- ☐ B. The desired Type I error
- ☐ C. Sample size
- ☐ D. The shape of the distribution of the response variable
- ☐ E. The standard deviation of the response variable

```
#Answer: [A,B,C,E].
```

4. [1 point] In a simple random sample, which of the following will be affected by the sample size when testing whether the mean of a study population against a hypothesized null value(that is, $H_0 : \mu = \mu_0$? Select ALL that apply.

- ☐ A. The true difference true mean (μ) and the null value (μ_0).
- ☐ B. Type I error if null is true.
- ☐ C. Power if null is false.
- ☐ D. The distribution of the sample averages in repeated samples (sampling distribution of $\bar{x}$).

```
# Solution:
# c and d.  a is a fixed unknown fixed difference, and b should not be affected under a
# consistent test. c. is obvious.  and d. because of CLT and SD of sample average being a
# function of n.
```

5. [1 point] For which of the following situations is the use of a one-sample t-test appropriate.

- ☐ A. To test whether a nominal (categorical) variable has the same distribution as a known population.
- ☐ B. To test whether two independent samples have the same mean.

☐ C. To test whether the mean difference is different from 0 of two measurements under different conditions (say treated and untreated) *on the same person*.

☐ D. To test the means across several independent groups (e.g., three age groups) are difference.

```
#Answer: Correct. C in a paired design, one reduces observation to a difference and then
#testing the mean of a single random variable to a natural null value.
```

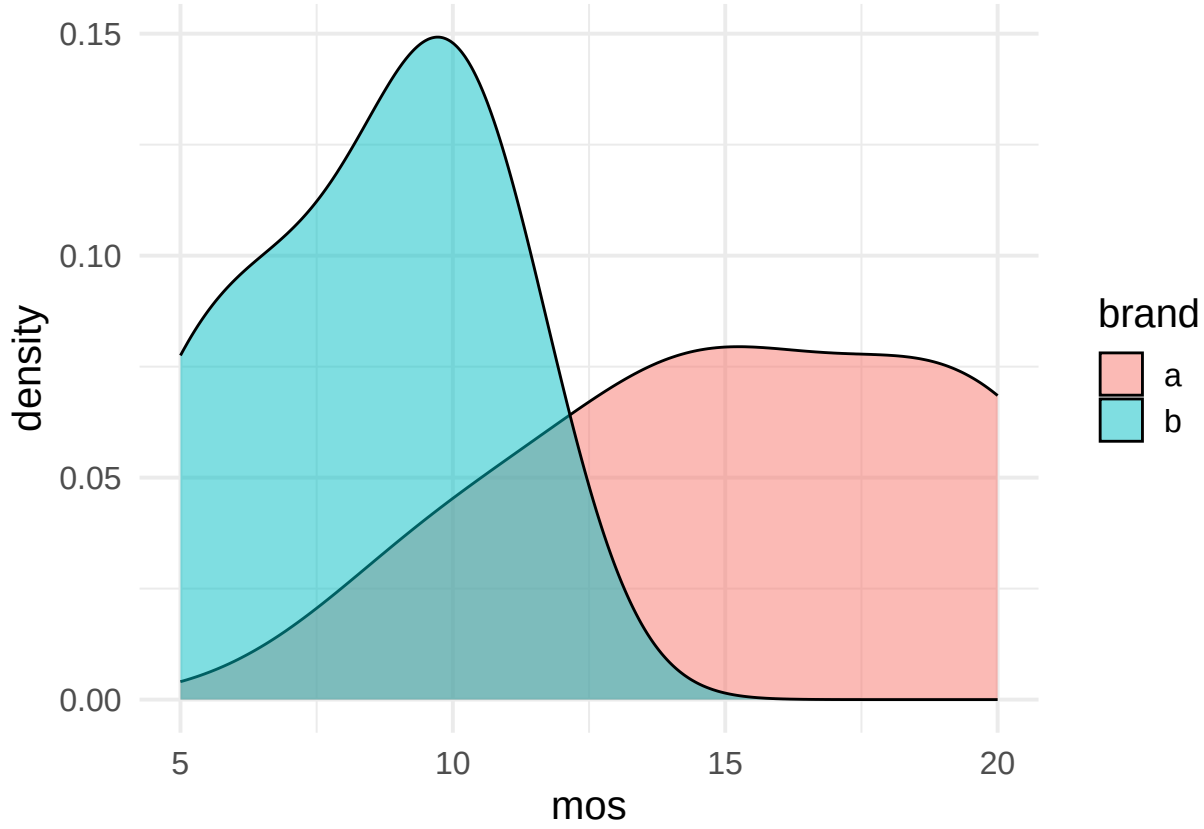
6. [1 point] True or False: When using Tukey's HSD in R, we get adjusted p-values that account for multiple testing for all pairwise comparison. The unadjusted (not accounting for multiple testing) p-value would be larger than the adjusted p-value.

☐ A. True
☐ B. False

```
#Answer: B (False)  The unadjusted p-value would be SMALLER than the adjusted p-value.
#So the statement is false.
```

7. [2 point]

Imagine you are hired to test the efficacy of different mosquito repellents. You recruit 10 students and randomly assign them to brand a or brand b and then bring them on a camping trip. At the end of the trip, you count the number of moquito bites the students have. Here are the data you have collected:



7.1 [1 point] What test should you use to evaluate your data?

☐ A. Simple two-sample t-test
☐ B. paired t-test
☐ C. using the permutation distribution to get p-value for the test statistic which is the difference of the averages in the two groups
☐ D. two sample test of proportions
☐ E. chi-square test for independence

```
# solution: (c) wilcox rank sum; can't rely on normality
# 0.5 credit for (a) two sample t test
```

7.2 [1 point] What is precise null hypothesis?

- ☐ A. Difference in means = 0
- ☐ B. The distribution of the mosquito bites is the same in the two treatment groups
- ☐ C. Mean in group a = mean in group b
- ☐ D. Mean difference = 0

```
# solution: (b) the observations all come from the same underlying distribution
# 0.5 credit for (c) mean in group a = mean in group b
```

8 [1 point] Weight Watchers has hired Andy to evaluate their program effectiveness. They recently implemented a new program and they have sampled the data for 100 patients on the new program, 24 of them had an equal or reduced body weight at 6 months follow up. In the past, the proportion of people who experienced weight loss at 6 months in a Weight Watchers program is known to be 20%. The alternative hypothesis of interest is that the new program increase the proportion with weight loss? What test should Andy use to determine if the new program increases the proportion with weight loss?

- ☐ A. 1-sided two-sample t-test
- ☐ B. 2-sided two-sample t-test
- ☐ C. chi-squared test of independence
- ☐ D. 1-sided, one sample test of proportions
- ☐ E. 2-sided, one sample test of proportions

```
# solution: (d)- 1-sided, one sample test of proportions
# 1-sided, one sample test of proportions. This is a one-sample proportion test since
# Andy is evaluating the **proportion** of people who experienced weight loss in 6 months.
#It is a 1-sided test since our alternative is testing if the new program **increase**
#the proportion with weight loss.

# 0.5pt for (e)
```

9 [2 point]

Julia is working for a company that is developing a new COVID booster vaccine. They want to ensure that the individuals that participate in the trial are a good representation of the population in the Bay Area. Julia pulls the census numbers for the distribution of race/ethnicity in the Bay Area and compares it to the 500 currently recruited participants. Self identified race/ethnicity has been captured in categories based on the US Census.

9.1 [1 point] What test should Julia use to test whether the study participants are representative of the Bay Area in terms of race/ethnicity?

- ☐ A. two-sample t-test
- ☐ B. paired t-test
- ☐ C. chi-squared test of independence
- ☐ D. chi-squared goodness of fit
- ☐ E. one sample test of proportions
- ☐ F. two sample test of proportions

```
# solution: (d)- chi squared goodness of fit; because we're comparing the observed and
#expected proportions for a categorical variable with more than 2 levels
```

9.2 [1 point] What is the null hypothesis of this test in words?

```
# solution: The distribution of the proportions in the sample is the same as
#that in the population
```

10. [1 point] I am fitting a linear regression of an outcome variable Y against a covariate X for a group project. My friend provided scatterplots of the residuals of the linear model ($r = Y - (\hat{a} + \hat{b}X)$)) against X. What assumptions is that plot meant to address? Select ALL that apply

- ☐ A. Whether Y is normally distributed.
- ☐ B. Whether there is a linear relationship between X and Y.
- ☐ C. Is the residual variance constant in X.
- ☐ D. Correlation between X and Y.
- ☐ E. Predictive power of the model.

```
#Solution: (b) and (c). We are referring to a residual vs. fitted plot. A can be checked via a
#histogram or a QQ plot. B and C can both be seen with a residual vs. fitted plot. D is through
#a regression model, or correlation coefficient.
```

Long Format Questions:

Question 11 [4 points]

A town has recently discovered that some of its residents have developed Disease A. The following 2 way table has been created:

Age Group	Disease Present (Yes)	Disease Absent (No)	Total
Under 18 years	120	A	200
18-64 years	B	450	750
65 years and older	80	C	150
Total	D	E	1100

11.1 [2 points] Calculate and put the numbers next to their corresponding letter below:

- A_____
- B_____
- C_____
- D_____
- E_____

```
#Solution: A: 80, B: 300, C:70, D:500, E: 600
```

11.2 [2 point] Calculate the marginal distribution of Age and Disease in this sample population. Round to the nearest 2nd decimal place.

- P(Age less than 18)_____
- P(Age is from 18 to 64)_____
- P(Age is greater than 65)_____
- P(Disease)_____
- P(noDisease)_____

```
#Solution:
#   Under 18:200/1100 = 0.18
#   18-64: 750/1100 = 0.68
#   65+: 150/1100 = 0.14
#   Disease:500/1100 = 0.45
#   No Disease:600/1100 = 0.55
```

11.3 [2 points] Calculate the conditional distribution of having the disease within each age group. Round to the nearest percentage place.

- P(Disease given Age less than 18)_____
- P(Disease given Age is between 18 and 64)_____
- P(Disease given age is 65 or greater)_____

```
#Solution:
#   18- = 120/200 = 60%
#   18-64 = 300/750 = 40%
#   64+ = 80/150 = 53%
```

Question 12 [5 points]:

Researchers have discovered a condition called Tomorrow Syndrome that affects individuals' ability to complete tasks promptly, and people with this syndrome exhibit an irresistible urge to delay tasks until "tomorrow." A public health researcher takes a random sample of 1,000 UC Berkeley students and found that 275 students were diagnosed with Tomorrow Syndrome (TS). Existing data suggested a 30% of people in California have TS. Her interest in testing whether or not the data she collected are statistically consistent with the California population

12.1 [1 point] Using standard notation, rephrase the statement above specifying the null and alternative hypotheses.

H_0: _____

H_A: _____

```
#H_0: mu = 0.30
#H_A: mu is not equal to 0.30
```

12.2 [2 points] If she had repeated the same experiment (random sampling UC Berkeley students) what do you think would be the exact distribution of the number of subjects in these repeated samples (assuming that UC Berkeley students have the same underlying probability of TS)? Be as precise as possible.

```
# If X is the count, then X ~ binomial(1000,0.3)
```

12.3 [1 point] What would be a good approximate distribution based on the central limit theorem (CLT)

```
# X ~ N(300,1000*0.3*0.7)
# or Normal distribution
```

12.4 [1 point] Assuming the sample is large enough so that one can rely on the CLT, provide the formula for a 95% confidence interval (again, assume that the data had 275 students with TS). Note, you do not need to the calculations beyond providing the formula. Note, that the 97.5% quantile of the N(0,1) (standard normal) distribution is 1.96.

```
#Answer: 275 +- 1.96*sqrt(1000*0.275*0.725)
```

Question 13 [7 points]:

A study is conducted to estimate the potential benefit (as defined by mean blood pressure of the population) of a new blood pressure drug (new drug) relative to an existing treatment (old drug) . A simple random sample of 100 patients with existing diagnosed high blood pressure were first randomized one of the two treatments. After being on the treatment for 30 days, their blood pressure was measured. After a period of time given to have the randomized drug, all subjects were given the drug (new or old) that they did not receive originally. Thus, if a patient was randomized the new drug at the first time, they were then given the old drug. After 30 days on the second drug, their blood pressure was measured again. let X_1 be the blood pressure (BP) on the old drug and X_2 on the new drug. The first few rows of the resulting data is shown below.

```
head(dat)
```

```
##   id       X1       X2
## 1  1 94.69581 84.51703
## 2  2 94.72237 84.08115
## 3  3 64.38372 49.32096
## 4  4 76.75756 69.74395
## 5  5 74.40345 54.97552
## 6  6 97.16679 91.95313
```

13.1 [1 point] Write out the null hypothesis and the alternative hypothesis if the alternative of interest is the new drug results in lower mean blood pressure relative to the old drug.

```
#H0: mu2=mu1, or mu2-mu1 = 0
#H1: mu2-mu1 < 0
```

13.2 [2 point] Write out the code to do the appropriate test consistent with the above.

```r
# t.test(dat$X2,dat$X1,alternative = "less", paired = T)
# can ignore alternative = "two.sided" and mu = 0

# -0.5pt for syntax errors
```

Consider the following summary statistics and quantile of the t-distribution:

```
library(dplyr)
sum.stat <- dat %>%
  summarise(mndiff = mean(X2-X1),sddiff=sd(X2-X1),n=n())
round(sum.stat,2)
```

```
##   mndiff sddiff   n
## 1  -2.18    9.9 100
```

```
# 97.5% quantile of the t-dist with df=n-1=99
round(qt(0.95,df=99),2)
```

```
## [1] 1.66
```

13.3 [2 points] Calculate the t-statistic for the desired test to 2 decimal places. Show your work.

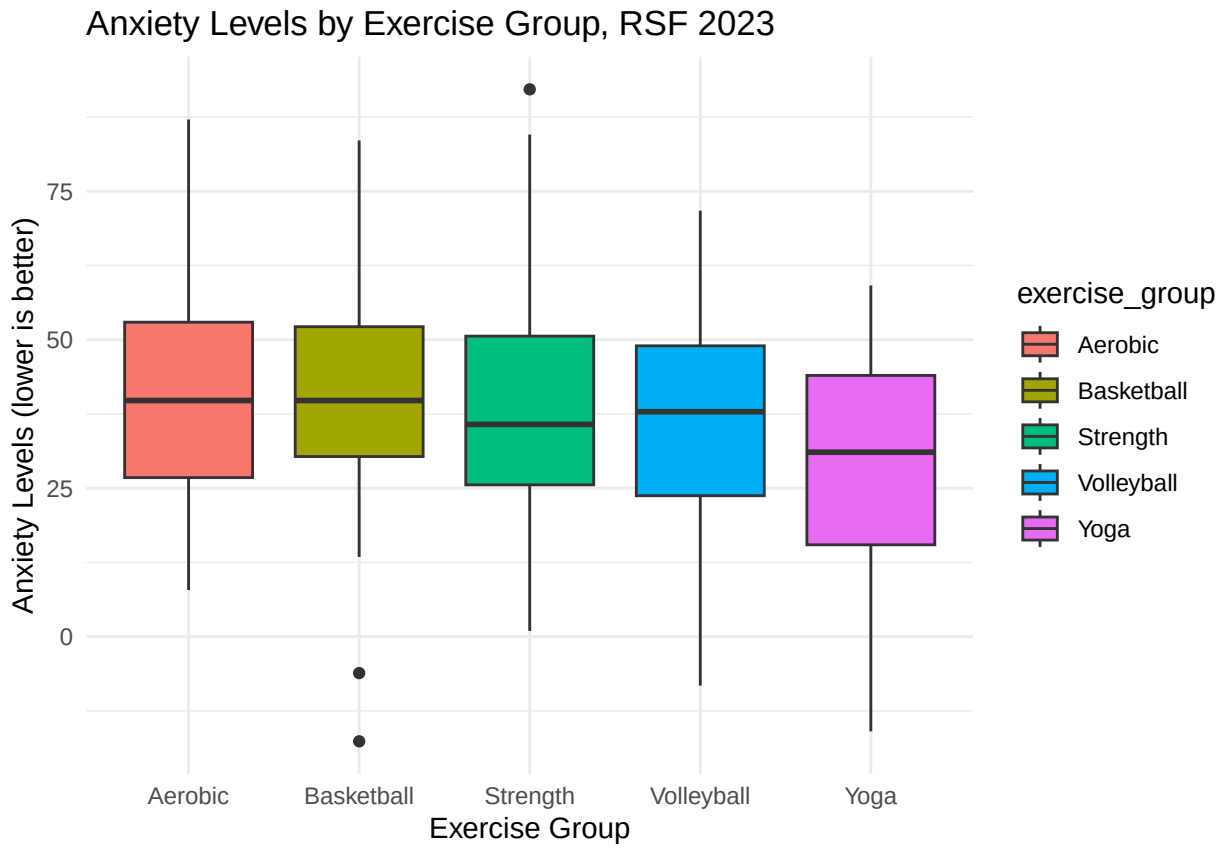
```
 # Solution: -2.18/(9.9/sqrt(100)) = -2.20
```

13.4 [2 points] Based on above stated null and alternative hypothesis, along with the quantile of the t-distribution provided, provide a reasonable guess of the resulting p-value of the desired test.

```
# Anything with p < 0.05 full credit. Describes that we would expect the p-value to be <.05
```

Question 14 [7 points]:

A study randomly selects to participate in one of five physical activities for a period of one month, when an instrument is used to measure anxiety (continuous response variable). Fifty (50) people are assigned to each exercise for a total of $n = 250$ subjects enrolled. After data collection, the data can be displayed in the boxplot below.



The first few rows are displayed below (the subjects are in random order):

```
##    exercise_group anxiety_levels
## 1        Strength       12.40222
## 2      Volleyball       41.11230
## 3      Volleyball       45.03370
## 4        Strength       38.91733
## 5         Aerobic       28.79335
## 6         Aerobic       67.86853
## 7      Basketball       38.67477
## 8      Volleyball       52.75135
## 9      Basketball       52.73001
## 10     Basketball       22.92211
## 11           Yoga       44.06208
## 12        Aerobic       36.79469
## 13        Aerobic       72.33919
## 14     Basketball       33.14654
## 15        Aerobic       28.00891
```

14.1 [1 point] If I know the anxiety response variable is normally distributed within each exercise group, and that the variance is constant across groups, what type of test is optimal here?

```
#Solution: ANOVA
```

14.2 [2 pointS] Given the variables `exercise_group` and `anxiety_level` and the dataset `exercise_data`,fill in the blanks to complete the code with output shown below.

```
## # A tibble: 2 x 6
##   term              df  sumsq meansq statistic p.value
##   <chr>          <dbl>  <dbl>  <dbl>     <dbl>   <dbl>
## 1 exercise_group     4  4495.  1124.      2.94  0.0211
## 2 Residuals        245 93572.   382.     NA    NA
```

anova_results <- aov(A ~ B , data = exercise_data)

C (anova_results)

A:_____
B:_____
C:_____

```
# solution: A: anxiety_levels
#B: exercise_group
#C: tidy
```

14.3 [2 point] Based on the output below, what would you conclude about the relationship of exercise to anxiety. Write no more than 2 sentences.

```
## # A tibble: 2 x 6
##   term              df  sumsq meansq statistic p.value
##   <chr>          <dbl>  <dbl>  <dbl>     <dbl>   <dbl>
## 1 exercise_group     4  4495.  1124.      2.94  0.0211
## 2 Residuals        245 93572.   382.     NA    NA
```

```
# solution: Correct-  p-value is < 0.05, reject the null hypothesis, in favor of the alternative
#hypothesis that there is a difference.

#Example: Our p-value is less than .05, so we reject the null hypothesis that there is no difference
#in mean anxiety levels, in favor of our alternative hypothesis that there is a difference.
```

14.4 [1 point] Below, we use a post-test procedure to see which pairwise comparisons of exercise groups are significant. Which pair of exercises have the most significant different in their mean anxieties?

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = anxiety_levels ~ exercise_group, data = exercise_data)
##
## $exercise_group
##                              diff       lwr         upr     p adj
## Basketball-Aerobic     -1.244740 -11.98635  9.49686678 0.9977605
## Strength-Aerobic       -3.875779 -14.61739  6.86582791 0.8589614
## Volleyball-Aerobic     -6.009544 -16.75115  4.73206299 0.5391050
## Yoga-Aerobic          -12.015157 -22.75676 -1.27354941 0.0197333
## Strength-Basketball    -2.631039 -13.37265  8.11056833 0.9620272
## Volleyball-Basketball  -4.764804 -15.50641  5.97680341 0.7403090
## Yoga-Basketball       -10.770416 -21.51202 -0.02880899 0.0490209
## Volleyball-Strength    -2.133765 -12.87537  8.60784229 0.9823578
## Yoga-Strength          -8.139377 -18.88098  2.60222989 0.2310047
## Yoga-Volleyball        -6.005612 -16.74722  4.73599481 0.5397572
```

```
#Solution: Yoga vs. Aerobic

#Also accepted: Yoga-Aerobic and Yoga-Basketball
```

14.5 [1 point] If your goal was to minimize anxiety, based only on this data, what exercise should you choose?

```
#Solution: Correct. Yoga - Aerobic has the smallest p-value, and from the test, we know
#that compare to Aerobic, Yoga could reduce stress level. So should choose Yoga
```

Extra Credit

EC1 [2 points] Assume one collects a random sample of subjects and records, during a single day: 1) how many times they thought about statistics, and 2) whether or not they had a headache. One wishes to estimate and derive inference for the ratio of the mean number of statistical thoughts in a day in those reporting headaches versus those without. How one could one derive a 95% confidence interval for this ratio without having a formula for the standard error for this estimator (ratio of averages)?

```
# 1pt - Mentions Bootstrap or permutation
# 1pt - Reasonable interpretation of the bootstrap process and how to construct
#the 95% CI using bootstrap quantile information.

# Example answer:
#I will use the bootstrap method:
#1. Resample from the collected subjects with replacement, getting a new group of
  #subjects with each subject having two answers for the two questions.
#2. In the new group, calculate the ratio.
#3. repeat step 1&2 for n times and we have n ratios' values.
#4. choose the 2.5% & 97.5% quantiles of the n rations and it'll be the 95% CI.

#Only one partial credits for permutation test. The permutation test can be used to test
#the null value of the ratio but can't be used to construct the 95% CI.
```

EC2 [1 point] What is a natural null value for the population value of this quantity (ratio of means)?

```
# Correct- Null value=1.  "Naturally" thinking of statistics is independent from having a headache.
#That is always the form of null hypothesis.

# Partial credit. Any positive vaule smaller than 1. Thinking of statistics has less headaches!
#Partial credits will be giving because this answer shows your love for statistics.
```

EC3 [1 point] Do you think the null is true? (no wrong answer)

```
# Correct. No wrong answers unless blank
```