STAT511 Take Home Final Exam (Section 001, 2019)

Due by Wednesday 12/18/19 by 11:59pm (midnight).

Instructions:

1. **Students are expected to work independently on the exam. Do NOT discuss the exam with anyone else (including other students). Do NOT post questions or comments about the exam to Canvas.**
2. **You may use any software, reference or on-line resource that you find helpful.**
3. **For some questions, there may be more than one possible analysis or graph that could be used for full credit. Choose one approach (making a reasonable choice and justifying if needed).**
4. Please use the markdown or Word template provided from Canvas. Note that both templates have R code appearing at the end a section. This will help me when grading.
5. Both templates include the honor pledge. I will consider your printed name as your signature.
6. If you have questions about the exam, you can see me during office hours or email me directly (hess@stat.colostate.edu). I will not respond to email after 4pm on Wed 12/18.
7. I expect the answers (graphs, tables, discussion, etc) to be cleanly formatted and labeled. Keep responses brief, but use complete sentences where appropriate. I reserve the right to deduct points for poor formatting, grammar, etc. It is not my responsibility to "sift through" R output to find your answers.
8. Use $\alpha = 0.05$ and/or 95% confidence where needed.
9. As a general rule, give answers to at least 2 decimal places.
10. Several questions ask about the "direction" of association. For a correlation or regression analysis, you can simply say "positive" or "negative". When comparing two means/proportions/etc, you can indicate which group has the higher mean/proportion/etc.
11. All questions are worth 4 points except where noted. Maximum score is 100.

Questions 1 through 9 (Exercise, 38 pts): The many benefits of exercise are well known. But in this study, a CSU researcher investigated the association between CACS (coronary artery calcium score) and high-volume endurance training in middle-aged athletes. CACS is a measure of hardened arterial plaque, with increased CACS associated with increased heart attack risk. Exercise is quantified using METh (metabolic equivalent hours) based on a seven day physical activity diary. This is an observational study, including a total of n = 43 subjects. The data is available from Canvas as **Exercise.csv**.

Sex = sex (M, F)
Age = age (in years)
METh = metabolic equivalent hours per week
Group = CON (METh < 60) or ATH (METh $\geq$ 60)
CACS_score = coronary artery calcium score (numerical)
CACS_01 = 0 (CACS_score = 0) or 1 (CACS_score > 0)

Notes:

- Some of the subjects in this study are considered "ultra-endurance" athletes. For example, some of the athletes competed in IronMan races.
- You should read the questions carefully because the roles of response (Y) and predictor (X) change.
- Some of the conclusions for this data may be surprising.

1. Create a summary table including the following information by Group: number of subjects, proportion positive for CAC (CACS_01 = 1), mean age, min METh, median METh, max METh, (**6 pts**)
2. Create a summary graph for CACS_score. Also provide the 6 number summary (min, Q1, median, mean, Q3, max).
3. Considering the information from the previous questions, why do you think an analyst might consider treating CACS as binary (CACS_01) instead of continuous (CACS_score)? (**2 pts**)
4. Of females, what proportion (or percentage) are positive for CAC (CACS_01 = 1)? Of males, what proportion are positive for CAC (CACS_01 = 1)?
5. Treating CACS_01 (Y) as the response and Sex (X) as the predictor, run an appropriate analysis. Summarize the results in a few sentences. Be sure to (1) state the name of the method used (be specific), (2) make a conclusion in the context of the problem (including a p-value or CI), (3) direction of association (even if not "statistically significant"). (**6 pts**)
6. Create a summary graph of CACS_score (Y) versus Group (X). (**2 pts**)
7. Treating CACS_score (Y) as the response and Group (X) as the predictor, run an appropriate analysis. Summarize the results in a few sentences. Be sure to (1) state the name of the method used (be specific), (2) make a conclusion in the context of the problem (including a p-value or CI), (3) direction of association (even if not "statistically significant"). (**6 pts**)
8. Treating CACS_01 (Y) as the response and METh (X) as the predictor, run a logistic regression analysis. Provide a (model based) estimate of odds ratio and corresponding 95% CI. Interpret the estimated odds ratio in the context of this problem. (**6 pts**)
9. Explain why a formal analysis of METh (Y) vs Group (X) is NOT of interest. (**2 pts**)

Questions continue on the next page…

Questions 10 and 11 (Chocolate, 10 pts): This group of questions uses information from the article "Eating dark and milk chocolate: a randomized crossover study of effects on appetite and energy intake" (Nutrition and Diabetes (2011) 1, e21; doi:10.1038/nutd.2011.17). The participants in the study were n = 16 healthy, normal-weight men. On two different test days, a "snack" either milk chocolate (100g) or dark chocolate (100g) was served. 135 minutes later lunch was served and total calories consumed (including both chocolate and lunch measured in MJ) was recorded for each subject. Since this is a crossover design, each subject received both treatments (in a randomized order). The goal of the study is to compare mean calories consumed for the two treatments.

	Milk Choc	Dark Choc	Diff = Milk-Dark
n	16	16	16
mean	7.69	7.06	0.63
SE	0.39	0.36	?

Notes:

- The full article is NOT required to answer these questions, but is available from Canvas if you are curious.
- I pulled the information for this question from Figure 3 using WebPlotDigitizer. Please use the information provided above as there may be slight discrepancies compared to what is given in the paper.
- Be sure to show your code or "work" for these two questions.
- Question 11 is challenging. Do the best you can. I will consider giving partial credit.

10. Suppose an investigator is interested in running a similar study but using a two-sample design (unlike the original paired study design). Use the summary statistics provided to choose an appropriate sample size (per group) to achieve 80% power. No need to write this up, just provide the required sample size per group. (**6 pts**)

11. If we had the SE for the differences (missing from the table), it would be possible to calculate power corresponding to a paired design. The only additional "clue" given in the paper is that the p-value for the (two-sided) paired t-test is p = 0.01. Use this information to calculate the SE of the differences. Note: You do NOT need to calculate power.

Questions continue on the next page…

Questions 12 through 18 (Guns, 52 pts): This group of questions uses information from the 2015 Washington Post article "Zero correlation between state homicide rate and state gun laws". This is an observational study, including a total of n = 50 states. The original data has been supplemented with some state level demographic variables. The data is available from Canvas as **GunData.csv**.

State = State name
Region = Midwest, Northeast, South, West
BradyScore = numerical score based on number of state gun laws, with low scores meaning a low level of gun restrictions and high scores meaning a high level
HomicideRate = number of homicides per 100,000 people based on Justice Department data
PerUrban = percentage of population that lives in urban areas
Poverty = 3-year average for the estimated poverty rate
MedAge = median age
PerDgr = percentage of population with a bachelor's degree

Notes:

- I suggest using code something like this to import the data using the state name as the row.name. Obviously you will need to specify the full file path location.
  `GunData <- read.csv("GunData.csv", row.names = 1)`
- The full article is NOT required to answer these questions, but link is provided from Canvas if you are curious.
- I have omitted Washington DC from the data. So your results will not exactly match the article.
- Some of the conclusions for this data may be surprising.

12. Create a summary graph for BradyScore. Which state has the lowest value? Which state has the highest value?

13. Create a summary graph for HomicideRate. Which state has the lowest value? Which state has the highest value?

14. Fit a regression model using HomicideRate (Y) as the response and BradyScore (X) as the predictor.

A. Provide a summary graph of Homicide Rate (Y) versus BradyScore (X). Overlay the fitted regression line. (**4 pts**)
B. Provide a (model based) estimate of the slope and corresponding p-value or CI. (**2 pts**)
C. Based on your answer to the previous question, what can we conclude about an association between HomicideRate and BradyScore? Briefly discuss. (**2 pts**)
D. What proportion of variation in HomicideRate is explained by the model? (**2 pts**)

15. Consider the plot of residuals versus fitted values for model from the previous question. You do NOT have to include this plot in your exam.

A. What state has the largest magnitude residual? Are there more or fewer homicides in this state then you would expect based on the model?

B. Run a formal outlier test for the state you identified above. Provide the Bonferroni adjusted p-value and make a conclusion in the context of the problem. Explain why the Bonferroni adjustment is appropriate here.
C. Just for this question, omit the state you identified from part A. Re-run the model from question 14. Identify the slope and provide a corresponding p-value or CI. Compared to question 14B, do your conclusions change (in terms of "statistical significance or direction)? See hint below.

16. Now consider Pearson correlations between HomicideRate versus each of the 4 "demographic" variables (PerUrban, Poverty, MedAge, PerDgr).
A. Provide a pairwise correlation matrix of all 5 variables. See hint below.
B. Which demographic variable has the strongest magnitude correlation with HomicideRate? (**2 pts**)

17. Now fit a multiple regression model HomicideRate (Y) as the response and BradyScore plus the demographic variable you identified in the question 16B as predictors.
A. Using this model, provide a (model based) estimate of the slope corresponding to BradyScore and corresponding p-value or CI. (**2 pts**)
B. What proportion of variation in HomicideRate is explained by this model? (**2 pts**)

18. Fit a one-way ANOVA model using HomicideRate (Y) as the response and Region (X) as the predictor.
A. Provide a boxplot of Homicide Rate (Y) versus Region (X). (**2 pts**)
B. Provide the ANOVA table.
C. Run Tukey adjusted pairwise comparisons and summarize your findings. Be sure discuss the direction of any differences you find. (**6 pts**)

Hint for 15C: There are (at least) two ways to do this:

- We can use the update() function to update an existing model. In this example, I fit an initial model (Fit1), then update the model (Fit2) dropping the 10th observation.

```
Fit1 <- lm(Y ~ X, data = ExData)
Fit2 <- update(Fit1, subset = -10)
```

- We can use the subset() function to subset rows/observations from the dataset. In this example, I retain observations with HomicideRate > 2.

```
GunSS <- subset(GunData, HomicideRate > 2)
```

- Note find a way to check that you have "subsetted" correctly.

Hint for 16A: It may help to create a dataset with just the 5 variables of interest for this question. Here are several options.

- Subset columns by number. In this example, I retain columns 3 through 7.

```
SmallData <- GunData[ , 3:7]
```

- Subset columns by name using base R. For example:

```
SmallData <- subset(GunData, select=HomicideRate:PerDgr)
```

- Subset columns by name using tidyverse. For example:

```
library(tidyverse)
SmallData <- select(GunData, HomicideRate:PerDgr)
```