THE ROLLINS SCHOOL OF PUBLIC HEALTH

OF EMORY UNIVERSITY

MSCR 500 & 533, FALL 2019

Final Exam & Data Analysis (Take-home)

NAME: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**READ THESE INSTRUCTIONS CAREFULLY BEFORE BEGINNING:**

1) **YOU MUST WORK ON YOUR OWN. You must not give aid to nor receive aid from anyone except the course instructor. You must not discuss any aspect of this quiz, in any manner, direct or indirect, with anyone except the course instructor.**

2) If the wording of a question seems ambiguous or incorrect, contact Azhar Nizam for clarification, as soon as possible. He will attempt to correct any wording or typographical errors, but otherwise will answer only if it is possible to do so without revealing solutions to problems.

3) This is an open-notes and open-book quiz.

4) Unless a problem prohibits it, you are allowed to use SAS.

5) Show all details of any manual work in order to maximize your chances of getting partial credit.

6) Write/type answers in the spaces provided.

7) **This exam is due on Dec. 17, 2019, at 1pm**. *Please print out your exam*, and place it in Azhar’s mailbox (2nd floor, GCR bldg.)

8) Read the additional instructions in the boxes below.

**For Q. 5-6: provide all requested information, including:**

- hypotheses

- definitions of population parameters used

- name of hypothesis test, & justification for its use

- distribution of the test statistic

- test statistic formula & calculation (except for non-parametric tests –just show the value)

- p-value calculation details. (except for non-parametric tests –just show the value).

- decision (reject or fail to reject H0)

- conclusion (stated in the context of the problem)

You may use SAS or do the problems manually –your choice. If using SAS attach your well-formatted and documented SAS program; if manually, show all details of any calculations / p-value determinations!

**For Q.1-4: choose the most appropriate procedure(s) from the list below.**

-One-sample z-test for a population mean

-One-sample t-test for a population mean

-Paired t-test

-Wilcoxon Signed Rank Test

-Unpooled t-test comparing two population means

-Pooled t-test comparing two population means

-Wilcoxon Rank Sum Test (used in place of two-sample t-tests if sample size is not large enough for those tests)

-Chi-square test of homogeneity

-Chi-square test of independence

-Fisher’s Exact Test

-None of the above or cannot be determined with

the information given

**Do not use any statistical software other than SAS. (No, you can’t use**

**Excel either! Basically, for any kind of computational/statistical work, if you’re going to use software, use SAS --don’t use any other app/software.)**

**This exam consists of 8 pages (including this cover page)**

1. Blood-pressure measurements taken on the left and right arms of a person are assumed to be comparable. To test this assumption, 40 people are randomly sampled, and systolic blood-pressure (SBP) readings are taken simultaneously on both arms by two observers. Assume that the two observers are comparable in skill and experience, and assume that left arm SBP, right arm SBP, and the difference between left and right arm SBP are all symmetric but distinctly non-normal. The most appropriate procedure for testing whether or not the two arms give comparable results is:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Weight loss was recorded for a random sample of 10 people who had been taking a weight-loss drug. The population of weight losses is likely to be strongly skewed. The most appropriate procedure for testing whether the true average weight loss is greater than 10 lbs is:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Investigators measured the before-breakfast plasma citrate concentrations for 110 randomly sampled people. The average was 119 μmol/l. Suppose that plasma citrate concentrations are known to follow a normal distribution. The most appropriate procedure for testing whether the true average before-breakfast plasma citrate concentration for the population is greater than 115μmol/l is:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. A study was conducted to analyze the relationship between vasectomies and prostate cancer. It was found that 7 out of 2130 randomly sampled men who had not had a vasectomy had prostate cancer, while, in an independent random sample, 11 out of 2200 men with vasectomies had prostate cancer. The most appropriate hypothesis test method for analyzing the relationship of interest is:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q.5** A microbiologist measured the growth of two strains of a bacterium – a mutant strain and a non-mutant strain – using mouse cells in petri dishes. Nine randomly selected mice were used. Nine pairs of petri dishes were used. Within each pair of dishes, cells from the same mouse were used; in one of the dishes, the mouse’s cells were exposed to the non-mutant strain, in the other dish the same mouse’s cells were exposed to the mutant strain. Hence, the data were paired by mouse. The sample results are shown below. Each number in the second and third columns represents the total growth in 24 hours of the bacteria in a single dish. A priori, the researcher suspected that the *mutant* strain would grow faster. Test, at α=0.05, whether the researcher’s suspicion is true. All populations involved can be assumed to be normally distributed. (Continued on next page)

Mouse Non-mutant Strain (A) Mutant Strain (B)

===== ================ =============

1 180 100

2 160 97

3 140 80

4 62 6

5 82 31

6 73 110

7 43 7

8 36 55

9 110 100

**Hypotheses: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Define Parameter(s): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Name of the test: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Justification for using this particular test: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Test Statistic (if reading from SAS output, simply indicate that you are doing so, and report the value; if calculating manually, show all calculation details):**

**P-value (if reading from SAS output, simply indicate that you are doing so and report the value; if determining manually, show all details, including the probability expression being evaluated):**

**Decision and Conclusion: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Q.6** A major court case on the health effects of drinking contaminated water took place in the town of Woburn, MA. A town well in Woburn was contaminated by industrial chemicals. During the period that residents drank water from this well, there were 16 birth defects among 414 births. In years when the contaminated well was shut off and water was supplied from other wells, there were 5 birth defects among 228 births. The plaintiffs suing the firm responsible for the contamination claimed that these data showed that the proportion of birth defects was higher when the contaminated well was in use. Was the true proportion of birth defects higher when the contaminated well was in use? Perform the appropriate hypothesis test at the α = 0.05 level.

**Name of the test: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Hypotheses: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Define Parameter(s): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**­­\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**α = 0.05**

**Justification for using this particular test: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Test Statistic (if reading from SAS output, simply indicate that you are doing so and report the value; if calculating manually, show all calculation details):**

**P-value (if reading from SAS output, simply indicate that you are doing so and report the value; if calculating manually, show all details, including the probability expression being evaluated):**

**Conclusion: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Q.7** Researchers studied the effect of dietary supplementation of calcium on blood zinc levels. Blood zinc concentrations (mg/ml) were measured in pairs of rats. In each pair: (i) the rats came from the same litter, and (ii) one rat received a dietary supplement of calcium, and the other rat did not. The data (blood zinc levels) for the pairs of rats is stored in the Excel file ‘zinc.xls’. Your tasks: use the SAS import wizard and SAS programming statements to:

1. Import the data from Excel into a temporary SAS data set named ‘zinc’.
2. Add the variable ‘DIFF’ to the ‘zinc’ data set, where, for each pair of rats, DIFF is equal to the difference in zinc concentrations for the calcium-treated rat and the non-treated rat.
3. Add the variable ‘PAIR’ to the ‘zinc’ data set where PAIR = 1 for the first pair of rats, PAIR=2 for the second pair, etc. (Hint: google help on the *\_*N\_ ‘internal’ variable in SAS…that might help)
4. Add a variable called ‘SEX’ to the ‘zinc’ data set, which is equal to 0 for the first 10 pairs, 1 for the remaining pairs. (The first 10 pairs were female rats, the remaining pairs were male).
5. Save the resulting data set as a permanent SAS data set named ‘zinc\_xxxxx’ (substitute your first name for xxxxx).
6. Perform the necessary descriptive and inferential statistical analyses to complete the information on the following page.
7. Write (or copy ‘n paste, or otherwise attach) the (well-formatted and documented) program that you wrote for tasks 2-6 in the box below

For the variable DIFF, complete this page:

1. Do there appear to be any extreme outliers, implausible data values or missing values?

(circle one) YES NO

If YES, then fill in the following information:

**Current Data Value Description of Problem, and Suggested**

**Pair # For DIFF Corrective Action, if any**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(You may or may not need all –or even any-- of the rows above)

1. Provide the following information for the variable ‘DIFF’:

Mean: \_\_\_\_\_\_\_\_\_\_\_\_ Standard Deviation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Median: \_\_\_\_\_\_\_\_\_\_\_\_ Interquartile Range: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Skewness: \_\_\_\_\_\_\_\_\_\_\_\_ Kurtosis: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Comment on the appearance of any graphs you created to assess the normality of DIFF:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

90% confidence interval for the true average DIFF: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

90% confidence interval for the true average DIFF for female rats: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

90% confidence interval for the true average DIFF for male rats: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Very briefly, describe the distribution of DIFF, both overall and by sex:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Q.9** In an exercise physiology study, the relationship between oxygen consumption (the dependent variable) and several potential predictors of oxygen consumption was to be studied. In a random sample of 31 subjects, the following variables were measured:

**Oxygen consumption** (% of available oxygen consumed, as measured during 1.5 mile treadmill run. The more fit a person is, the greater their oxygen consumption will be during physical activity)

**Sex** (M, F)

**Age** (years)

**Weight** (kg)

**Run time** (time, in minutes, taken to run 1.5 miles on treadmill)

**Maximum pulse** (highest pulse rate during the run)

**Performance** (a subjective measure of the subject’s fitness, assigned by an interviewer; larger values mean the subject was assessed as having a higher fitness level)

The data for this study are contained in the Excel file named ‘fitness2019’ (posted on Canvas). Use SAS to perform a complete analysis of oxygen consumption regressed on *one or more of the remaining (independent) variables*\*. Use a 5% significance level. Turn in:

1. A report of your analysis (3 pages max). In the report you should:
   1. Include a table of descriptive statistics for each variable.
   2. State the regression model
   3. Briefly describe the method(s) used to arrive at your model.
   4. State the estimate of the model.
   5. Report the p-value for the overall test for the model in d.
   6. Report a measure of fit of the model.
   7. Report the p-values for the partial tests for the model in d.
   8. Report the key results of any diagnostics work performed for the model.
   9. Include a very short but clear and well-written summary describing the significant relationships.
   10. Cut and paste into your answers any graphs from SAS that are useful and relevant, but do not cut and paste any other SAS output. Anything you cut-n-paste must be referred to somewhere in your report! Any tables must be beautifully typed up by you.
2. Attach your well-formatted, well-documented, concise and error-free SAS code to this exam too. Just the program –don’t attach the log or any output pages.

**\***Deciding which variable(s) to include in your model is up to you! Here are some possible strategies:

1) Put all of the independent variables into the model, and report which relationships are significant and which are not. Often not a great strategy, especially when the # of independent variables is large!

2) Put all of the independent variables into the model, and examine the partial t-test p-values for each one. Identify the variable with the largest p-value that is >= 0.05, and remove it from the model –i.e., re-run the regression without that variable. Repeat this process until all variables in the model have a partial t-test p-value that is < 0.05. (Note: this model selection strategy is known as ‘backward elimination’).

3) Use subject matter knowledge –i.e., your own scientific expertise, and/or a literature search – to identify the independent variables that are most interesting/most likely to be related to oxygen consumption, and include only those in your model. (Note: this can be a difficult strategy to implement if you don’t have scientific expertise in the area).

**Q. 10 You’re not going to like me for this! However, do this I must.**

**It looks like we’re not going to be able to cover as much applied logistic regression as I would like to cover before the end of the semester.**

**I know you have covered this topic in Epi; but it would be useful for you to see it from the Biostats perspective as possible, and for you to see a simple implementation of it in SAS. So, for Question 10 in this final exam, please:**

**1) read the posted logistic regression notes**

**2) watch the posted videos**

**3) get your hands on the keyboard and practice typing and running the SAS demos for PROC LOGISTIC that are shown in the notes.**

**Please do all of this by the start of the Spring semester.**

**There is nothing to turn in for this question. You’re on the honor system! Simply insert your name below, in order to affirm that you will follow the instructions above, and the points for this question are yours.**

**“I < *insert your name here* > will complete all of the work for question 10 before the Spring 2020 semester begins.”**