Final Exam Take-Home Component

Fall 2016

**Before you begin:**

* This is a take-home exam. You may use any reference materials you wish (class notes, lectures, homeworks, sample problems, etc.), including the computer/internet. However, you must work on this exam **by yourself** and you **may not discuss this exam with anyone**. Emails to the professor for clarification of questions are the only exception to the “no discussing” rule.

**General Instructions:**

* Complete each problem to the best of your ability. Partial credit will be given where appropriate.
* You **must use this Word document as the template** into which you **type** your answers. There are no exceptions to this rule.
* For this exam you are **required to use Stata**. Some problems will require you to include Stata output by cutting-and-pasting into this Word document – this instruction will be explicitly indicated.
* When you are asked to provide a “summary of your conclusion” this **must be in the context of the problem**; in other words, writing “reject the null” is **not** appropriate.

**Submission Instructions:**

* The exam is due by **Monday, Dec 12, 2016, 11:59pm.**
* **No late assignments will be accepted.**
* To submit, upload your exam to the “Final Exam - Take-Home Component” Dropbox on Carmen. Please submit your exam as **one file**.

This exam has a total of 22 questions, and a total of 100 points possible.

Make sure you complete all questions (and all parts of all questions) on the exam.

Good luck!

**Honor Code Signature:**

Joshua Wu

I pledge on my honor that I have **neither given nor received unauthorized aid on this examination**. Typing my name on the line above serves as a signature of this pledge.

As part of a homework assignment, you (and your classmates) completed a short online survey. The resulting data are available in the Stata dataset **classdata.dta**. The survey questions with corresponding response choices and variable names are in the file **PUBHBIO6210\_Data\_Collection.pdf**. Using these two files, answer the following questions. Note that you will need to reference the PDF file in order to be able to understand the data!

***Note:*** *These are the same data you used on the midterm take-home component.*

1. (5 pts) For each of the variables below, state the name of the variable type (choices are: continuous, discrete, ordinal, nominal, binary).

|  |  |
| --- | --- |
| **Variable** | **Type of Variable** |
| (a) hand | Binary |
| (b) courses | Discrete |
| (c) age | Discrete |
| (d) miles | Continuous |
| (e) pet | Binary |

**2**. (5 pts) The target population in which we are interested is all graduate students taking a CPH biostatistics class (online or in-person); call this population “CPH biostatistics students.” For the purposes of the analyses required for this exam, we will treat this sample of 103 students as a **random sample** from this population. Select **one** variable on the survey and argue whether or not you think this sample is representative of the population (all “CPH biostatistics students”) with respect to this one variable.



Graduate students usually take between 12-18 credit hours. However, there are occasions where graduate students also have full time jobs, in which case their credit hours would be significantly lower (6 credit hours). Seeing the distribution of credit hours, this sample is representative of the population, since it shows many people taking less than full time credit hours, indicating work on the side and many within the graduate student credit hour range.

3. (12 pts) We are interested in the relationship between age and number of text messages among this population. Start by summarizing these two variables by completing the table below. You do not have to provide Stata output.





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **SD** | **Median** |
| Age | 103 | 26.80583 | 5.703556 | 25 |
| Text messages per average day | 103 | 42.50485 | 53.93421 | 25 |

4. (4 pts) Describe what you know about the shape of the distribution of age based on the summary statistics in the table above.

We can see that the distribution is not normal, and that there is a slight right skew, since the mean is larger than the median. That would also indicate that it is not symmetrical.

5. (4 pts) Create a scatterplot that has age on the X-axis and text messages on the Y-axis, and paste this scatterplot below. Describe what (if any) pattern you see.





There does not seem to be a pattern for association between age and text based off this scatter plot.

6. (6 pts) Perform the appropriate hypothesis test to test for an association between age and text messages and **include Stata output for the test**. Write a one sentence summary of your conclusion and include the p-value. You do not have to write out all the steps of the test.



Hypothesis: H0: β1 = 0

Ha: β1 ≠ 0

Test Statistic: t = = = -2.97

P-value: From Stata -> 0.004 -> less than 0.05 ->Reject null hypothesis.

**There is evidence to suggest that age is significantly associated with number of texts a day (p = 0.004).**

7. (4 pts) Use your results from the previous question to calculate the predicted number of average text messages sent in a day for someone who is 36 years old.

Texts = 114.44 -2.683722(age)

114.44-2.683722(36) = 114.44-96.613992 = **17.83 texts a day** for someone who is 36 yrs old.

8. (4 pts) Based on this model, what is the estimated difference in number of average text messages per day between someone who is 36 years old and someone who is 21 years old? Be sure to indicate which age has the higher number of messages.

36-21 = 15. 15\*2.683722 = **40.25583 texts a day difference.** The **21 yr old** will have the higher number of messages.

9. (2 pts) What proportion of the variability in number of text messages is explained by age?

R2 = **0.0805**. **8.05% of the variability in number of text messages can be explained by age.**

10. (3 pts) **Based on the scatterplot in question #5**, are there any assumptions necessary for performing the analysis you did for question #6 that are questionable for these data? Cite specific evidence to support your answer.

There are 4 assumptions. Linearity, Independence, Normality, and Equal Variance. However, if we just look at the scatter plot, **Linearity** is the most important one. As we can see, the scatter plot in question 5 is definitely not linear; it looks like a graph of 1/x. Since that is the case, we probably should have transformed the data first and see if the data is then linear.

11. (3 pts) Another possible characteristic that is related to number of text messages is whether or not an individual has ever lied about their age. To begin investigating this possible relationship, create side-by-side boxplots showing the distribution of text messages separately for those who have lied about their age and those who have not. (Note: you have seen the necessary Stata code in a lecture.) Paste the boxplot below. Briefly describe what you see in the plots, specifically as it relates to comparing those who have lied about their age and those who have not.



Lied about age on the Right

In the group that lied, there is a higher median but less spread (Q3-Q1) than the group that did not lie. The group that did not lie had more outliers, but did not have the highest number of text, which belongs to the “did Lie” group with 300 texts.

12. (4 pts) Calculate and report the mean of the number of text messages in an average day separately for those who have lied about their age and those who have not and report them here. (Note: again, you have seen this code in a lecture.) **Include Stata output to show where you got your answers.**



Got answer here

Mean number of text messages for those that did not lie: **37.36**.

Mean number of text messages for those that did lie: **52.08**.

13. (6 pts) Is there evidence that the mean number of text messages in an average day is significantly different for CPH biostatistics graduate students who have and have not lied about their age based on this sample? Perform the appropriate test and **include Stata output for the test**. Write a one sentence summary of your conclusion and include the p-value.You do not have to write out all the steps of the test.



With a p-value of 0.1878, we conclude that the mean number of text messages in an average day is **not significantly different** for CPH biostatistics graduate student who have and have not lied about their age.

14. (4 pts) Report the 95% confidence interval for the difference in text message means (be sure to be clear about the order of your subtraction), and explain how it relates to the result of your hypothesis test in the previous problem.

Based on the test in Q13, the 95% CI for the different is (-36.75278, 7.302531). This is for those who did not lie minus those who did lie. As we can see, **0 is included in this CI**. With 0 in the CI, that means that there is a good chance that there would not be differences in text message means, which would let us conclude that the mean difference was not significantly different.

15. (2 pts) The OSU Veterinary clinic group would like to know what proportion of students in CPH biostatistics classes prefers dogs (as opposed to cats). What is the estimate of the proportion of students who prefer dogs? **Include Stata output to show where you got your answer.**

Got answer here



The estimate proportion of students who prefer dogs is **76.70%**.

16. (6 pts) One source claims that 85% of people prefer dogs to cats. Based on the sample you have, is there evidence that the proportion of CPH biostatistics students who prefer dogs is significantly different from 85%? Perform the appropriate hypothesis test and **include Stata output for the test.** Write a one sentence summary of your conclusion and include the p-value.You do not have to write out all the steps of the test.



With a p-value of 0.0183, we conclude that there is evidence to support the claim that the proportion of CPH biostatistics students who prefer dogs **is significantly different from 85%.**

17. (4 pts) Are men more likely to prefer dogs than women? As a first step towards answering this question, calculate what percent of men prefer dogs and what percent of women prefer dogs. **Include Stata output to show where you got your answers.**

Got answer from here

**85.71% of men** prefer dogs and **72.06% of women** prefer dogs. Using this sample, men are more likely to prefer dogs than women.

18. (4 pts) Compute a measure of association between sex and preferring dogs and interpret this quantity. There may be more than one appropriate measure; pick just one that is appropriate.



With a chi2 value of 2.4, which corresponds to a p-value of 0.12, we can **conclude that pet preference between dogs and cats are not associated with sex.**

19. (6 pts) Do we have evidence of a difference in how likely men and women are to prefer dogs? Perform the appropriate test and i**nclude Stata output for the test**. Write a one sentence summary of your conclusion and include the p-value.You do not have to write out all the steps of the test.



With a p-value of .1205, we can conclude that there is evidence to suggest that there is **no difference** in how likely men and women are to prefer dogs.

20. (3 pts) As a final analysis, we will look at how easy CPH biostatistics students believe it is to manipulate statistics to show a particular point of view. Calculate the mean ease of manipulating statistics and also create a corresponding 95% confidence interval. **Include Stata output to show where you got your answers.**



Got answer from here

The mean is **65.79612** and the CI is **(61.74253,69.8497)**.

21. (3 pts) Next, we will consider if this variable is associated with self-reported math knowledge. We want to consider math knowledge as a binary variable – (1) “strong math knowledge” vs (0) “comparatively weaker math knowledge.” Define “strong math knowledge” as self-reported math knowledge either “high” or “very high” (and thus that “comparatively weaker math knowledge” includes all other responses). Calculate the percentage of students in this sample who report “strong math knowledge.” **Include Stata output to show where you got your answers.**



Got answer from here

The percentage of students in this sample who report “strong match knowledge” is **34.95%**.

22. (6 pts) Is there evidence that the mean ease of manipulating statistics is significantly different between those with “strong math knowledge” compared to all others? Use Stata to generate any variables you need (if you did not already) and perform the appropriate test. I**nclude Stata CODE AND OUTPUT you used.** Write a one sentence summary of your conclusion and include the p-value.You do not have to write out all the steps of the test.

Note: If you cannot figure out how to create a “strong math knowledge” variable, you can instead compare ease of statistics between left-handed and right-handed individuals for a 3 point penalty.





With a p-value of .7466, we conclude that the mean ease of manipulating statistics is **not significantly different** between those with “strong math knowledge” and those who do not.