**SPEA V506**

**Midterm Exam – Summer 2018**

Turn in your answers as a word document (you may include pictures or screenshots of hand-written work if that is easier. Show your work for all of questions below and make your final answer clear (circle, bold, etc..). For Part II, please include both your R code and the relevant output.

You are allowed to use any of the Canvas materials for this exam but are not allowed to consult other web sites or materials, or to discuss the exam with any other students. If you have questions about the exam, please email me or the TAs.

Please complete the following agreement form and turn this in as part of answer document.

**Take Home Midterm Exam Agreement Form**

**Date:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Directions:**  **Read and initial next to each statement, type your name below as indicated**. Your initial located next to each item below signifies your understanding and compliance with the instructions for this take home examination.

**\_\_\_\_\_\_ I understand that I may not copy or distribute this exam to anyone.**

**\_\_\_\_\_\_ I understand that I must return the examination by the stated deadline July 1, 2018 11:59 pm EST and that late exams will not be accepted.**

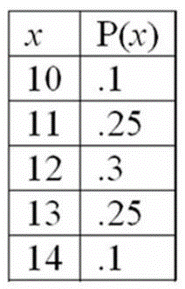
**\_\_\_\_\_\_ I understand this is a test and I cannot copy answers from other students. I understand that I may not discuss this exam with other students (current or previous). Furthermore, I cannot seek assistance, of any kind, from anybody other than the TAs or Instructor.**

**\_\_\_\_\_\_\_ I understand that I CAN consult resources including the textbook for the class, my own hand-written notes, the course presentations, or any of the resources on the Canvas course site. I understand that I CANNOT consult the web beyond Canvas.**

**\_\_\_\_\_\_ I understand that violation of this agreement will result in an F on this exam and it cannot be replaced, it will be averaged in (as a 0%) with my other class scores. Violations of this agreement will be handled according to IU’s policies on academic misconduct.**

**(Type your name here-on the line): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Part I: Short Answer Questions**

1. A company was studying the demographics of their customers. As part of the study, they collected the following variables: political party, marital status, credit rating (low, medium, high), annual income, and age. Label each variable as qualitative or quantitative, discrete or continuous, and nominal, ordinal, interval, or ratio.
2. When describing a variable, we often want to know its central tendency (i.e. mean, median, mode). What are the advantages and disadvantages of each of these measures? Under what circumstances would it be best to use each measure?
3. A study by the National Park Service revealed that 50% of the vacationers going to the Rocky Mountain region visit Yellowstone Park, 40% visit the Grand Tetons, and 35% visit both. What is the probability that a vacationer will visit at least one of these magnificent attractions?
4. Why is randomization/random assignment such an important tool to establish causality?
5. Briefly describe (in your own words – that means no copying and pasting directly from lecture slides) how the following types of bias might influence data collected from a survey: a) selection bias, b) nonresponse bias, c) social desirability bias.
6. What is the mean (expected value) for the random variable x in the following probability distribution)?  
     
    
7. For a normal distribution, what is the likelihood (expressed as a percentage) that a random variable is within plus or minus two standard deviations of the mean? What about three standard deviations?
8. In a management trainee program, 80% of the trainees are female, while 20% are male. Ninety percent of the females attended college; 78% of the males attended college. A management trainee is selected at random. What is the probability that the person selected is a female who attended college?
9. In a large metropolitan area, past records revealed that 30% of all the high school graduates go to college. From 20 graduates selected at random, what is the probability that exactly 8 will go to college?
10. What is the area under the normal curve between z = -1.0 and z = -2.0? What about between z=1.0 and z=3.0? What about z=-1.0 and z=3.0?
11. The mean score of a college entrance test is 500; the population standard deviation is 75. The scores are normally distributed. What percent of the students scored below 320?
12. The Office of Student Services at a large state university maintains information on the study habits of its full-time students. Suppose we select a random sample of 35 current students and find a sample mean of 21.5 hours spent studying per week, with a sample standard deviation of 5.6 hours. Based on this survey data, test the claim that the average number of hours students spend studying per week is greater than 18 (use the 0.05 level of significance).
13. A total of 500 voters are randomly selected and asked whether they plan to vote for the Democratic incumbent or the Republican challenger in an upcoming statewide election. Of the 500 surveyed, 350 said they would vote for the Democratic incumbent. Using the 0.95 confidence level, what are the confidence limits for the proportion that plan to vote for the Democratic incumbent?
14. Suppose that the Monroe County school district has proposed a small sales tax increase to help cover costs associated with the construction of a new middle school. The tax will only go into effect if a majority (more than 50%) of voters approve the increase during an upcoming vote. The local newspaper conducted a poll drawn from a random sample of Monroe County voters, and found that 320 out of 600 respondents indicated that they supported the new tax. The school district superintendent wants to know whether the new sales tax will pass, and has tasked you with analyzing the survey results.

a) Derive and test the appropriate hypotheses which will answer the superintendent’s question. Use the p-value method, with α=0.05. Briefly (2-3 sentences) interpret the results.

b) When you present your results to the superintendent, she asks how much confidence you place in the estimate for the proportion of voters supporting the initiative. To answer this question, construct a confidence interval for the proportion of voters in favor of the new tax, using α = .05. Briefly interpret the results.

1. Suppose that you are a TA for wonderful, amazing, incredibly smart statistics professor. This professor has asked you to determine if former SPEA V506 have the same mean income after graduation compared to former students who took stats in a different department at IU. To answer this question, you randomly sample 20 former V506 students and ask them to report their annual income. Using this sample data, you determine that the average income for former SPEA V506 students is $68,456, with a sample standard deviation of $9,568. Next, you compile a random sample of 25 former students who took statistics in a different department at IU. The mean annual income within this sample is $60,350, with a sample standard deviation of $9,378. You may assume that the population standard deviation for former V506 students is equal to that for former students who took stats in other departments.

Assuming a normal distribution for both of these populations, derive and test the appropriate hypotheses that will allow you to answer your statistics professor’s question. Use α = 0.01. Briefly (2-3 sentences) interpret your results.

**Part II – R Programming Questions**

The following questions require that you download the “university\_data.csv” file from canvas and import this file into R. This dataset contains information on the amount charged for in-state tuition and fees (intuitfees) in the 2014-2015 school year, along with variables indicating the name of the institution (instnm), state (stateabbr) and whether the institution is a public or private university (sector). Note that not every institution in the country is included in this dataset (some institutions had missing data or did not respond to the survey). Please be sure to include both your code and the relevant output in your answers.

1. Calculate the following descriptive statistics for the in-state tuition and fees variable for the entire dataset: a) mean, b) standard deviation, c) number of observations, d) standard error of the mean.
2. Now calculate those same descriptive statistics, but disaggregated by sector (i.e. compute the mean, standard deviation, number of observations, and standard error of the mean separately for public institutions and for private institutions).
3. Construct histograms for in-state tuition and fees for both public and private universities (i.e. make two histograms, one for each sector). In constructing these histograms, change the color of the bars to be grey, and be sure to label both the X and Y axis. Add a title to each histogram to identify which group of universities it is describing.
4. Because public universities receive state subsidies to help offset costs, they are generally thought to charge less in tuition and fees than their private counterparts.
   1. Conduct a hypothesis test, at the 0.05 significance level, to evaluate this hypothesis (that public universities have **lower** in-state tuition and fees than private institutions). **\*\*\*Note this is NOT a single sample test. You should treat public and private institutions as separate samples, and thus should do a two-sample test). Briefly interpret the results.**
   2. Next, construct a 95% confidence interval around the estimated **difference** in tuition and fees between the two groups (hint: remember that confidence intervals are never one-tailed). Briefly interpret the results.