Week 4 Practice Exam

Name: Matthew L. Pergolski

Instructions: This is a “low stakes” (i.e., not graded) learning assessment of your comprehension of the first four weeks of this course*.* Compose brief answers to each of the following six questions, typing your response in *italics* below each question. Try to complete the whole exam in an hour. When you are done, check your answers against the key that the instructor will distribute. To make best use of the test, it is important to complete the whole test before consulting the key. If you have unanswered questions about your responses after examining the key, feel free to submit your exam to the LMS for feedback.

1. Why do we collect samples of data rather than collect data from whole populations?  
    *It is often impractical to access data from the entire population.*
2. Describe the conceptual connection between (“mu,” the population mean) and (“x-bar,” a sample mean). Are they always the same? Or are they always different? Or something else?  
    *While keeping concepts like the law of large numbers and central tendency in mind, the sample mean will begin to converge towards the population mean as the number of samples and trials increase.*
3. A large retail chain conducted a study of new cashier productivity by examining item UPC code scanning data from samples of new cashiers at each of 853 different locations. The mean time between item scans across all collected data was 2.3 seconds. What would a histogram of the raw data look like? If you calculated a mean for each of the locations and plotted a histogram of those means what would that look like? Would the two distributions look the same?  
    *Roughly, I would predict that the raw data would show somewhat of a normal distribution for all data points associated with the 853 locations. The peak point of this distribution would be near the 2.3 value.*

*The distribution of means would also look similar to a normal distribution. Although, for example, the tails of one graph could be thicker/thinner than the comparing distribution, they would, in essence, take on the same shape.*

1. Your boss at the social media marketing company asks you to conduct an A/B test on two different banner ad configurations. Each of the two banners is deployed on 98 highly popular web pages during a one-hour test period:   
     
   A banner: mean of 13.23 clicks (per 1000 impressions) across n=98 pages.   
   B banner: mean of 13.94 clicks (per 1000 impressions) across n=98 pages.   
     
   The 95% confidence interval for the mean difference is as follows:   
    *-0.83 < (mean difference, A - B) < -0.58*.   
     
   Answer the following questions about that confidence interval:   
   1. What is value that is at the **center** of the confidence interval – in other words what is the *point estimate* of the mean difference in clicks (per 1000 impressions) between A and B?  
       *13.23-13.94 = -.71*
   2. Does this confidence interval contain the population mean difference somewhere within its span?  
       *We cannot prove that the true population mean difference lies within the confidence interval; however, if this test was run 100 times, we’d be confident that the true mean difference would appear within the interval range 95 times.*
   3. Which banner ad do you prefer (A or B) and why?  
       *Assuming the results are statistically significant (i.e., the p-value is less than alpha), the test indicates that the first mean has a negative advantage against the second, thus meaning that B banner would be slightly preferred in this situation.*
   4. Write a brief paragraph that provides an interpretation of the point estimate and the confidence interval for your boss. Your boss is an expert at marketing, but knows very little about statistics. Make sure that you clarify the connections between (“mu,” the population mean), (“x-bar,” a sample mean), and the confidence interval. *Through analyzing a subset of our population data, we can calculate a mean and use further statistical techniques to generalize insight about the entire population mean. When comparing two groups, we can use a t-test to construct an analysis of the difference of sample means. Our t-test will show a range of numbers that may include this population mean difference. However, we can’t ever prove that this difference will be listed within the interval; we are only confident it will show in long-run with repeated testing – say, in 95 out of 100 tests.*
   5. Your boss tells you to run the same experiment 99 more times next week, calculating a new confidence interval each time. After completing this project, you now have a collection of 100 different confidence intervals, each of which was constructed in the same fashion and with the same sample sizes, but each from new data samples: What, if anything, can you say about this collection of confidence intervals?

*Since this test was run through the same fashion, we can conclude that we’d be confident the true population mean difference would lie within the confidence interval in 95 out of 100 of the tests that were run.*