Week 4 Practice Exam

Name: Maya Mileva

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?

*Sampling is a process of drawing a subset of elements from a “master” set – often referred to as population. We draw samples because it is impossible to look at the whole population (totality of units from which a sample can be drawn). If we try to collect data from the whole population, some of the information can be bias, nonreachable.*

1. 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?

*Sample Mean implies the mean of the sample derived from the whole population randomly. Population Mean is nothing but the average of the entire group.*

*The mean value of the population is usually unknown. We can use the sample eman as an estimate of the population mean.*

1. 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?  
    *The histogram of the raw data would have a normal distribution. If we calculate the mean for each location and plot them and do it large number of times thanks to the Central Limit Theorem the mean of the sampling distribution will match the mean of the underlying population. We have to consider The Law of the Large numbers too(if we run a stat. process a large number of times it will generally converge on a particular results). So that means that the distribution of those mean will be normal(bell-curved)*
2. 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?  
       *around - 0.70*
   2. Does this confidence interval contain the population mean difference somewhere within its span?  
      *We can`t say, there is 95% chance that it does, but we don`t know is we going to get those 5% out pf the range*
   3. Which banner ad do you prefer (A or B) and why?  
       *Banner B generated more clicks*
   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.

*We analyzed the difference in the 2 banners A and B across 98 pages during one-hour test period. Sample mean for banner A is 13.23 and for banner B is 13.94 so results showed mean difference .71 clicks more for banner B.*

*We constructed 95% confidence interval around this mean difference. Which ranges from -0.83 to -0.58. This confidence interval may or may not contain the true population value. The width of the confidence interval is about plus or minus 0.13, give some indication of amount of uncertainty around the point of estimate -0.7. To reduce this uncertainty we would have to increase sample size, reduce variability.*

* 1. 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?   
      *Out of 100 confidence intervals 95 % would contain the population mean difference.*