Week 4 Practice Exam Key

**Please keep this key confidential! The ability of future students to study effectively from the practice exam depends upon the key not being available until students have completed the exam.**

Instructions: Check your answers when you are done against this answer key. If you have unanswered questions about your answers 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 usually impractical to collect complete populations of data, either because of inaccessibility of the whole population, or the expense of doing so.*
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?  
    *In the frequentist perspective, population parameters are fixed values that are unknown. In this view, mu represents the true, but generally unknown value of the population mean. Xbar is the result of using an estimator on a sample of data to try to approximate the value of mu. A good estimator, when tested over many replications, will produce results that have the smallest possible amount of variance around mu and that are, as a group, neither systematically lower nor higher than mu.*
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?  
    *The histogram of raw data might could probably have a strong mode somewhere below 2.3 seconds with a long right-hand tail. The histogram of the distribution of sampling means would probably be quite normal and centered on 2.3 seconds. There’s not enough information in the problem statement to know the variability of the sampling distribution, but its variance would be smaller than the variance of the raw data distribution.*
4. 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 the mean difference 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?  
       *About -0.705 clicks.*
   2. Does this confidence interval contain the population mean difference somewhere within its span?  
       *Impossible to say.*
   3. Which banner ad do you prefer (A or B) and why?  
       *B produces more clicks per 1000 impressions.*
   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 little about statistics.  *We analyzed the difference in clicks per 1000 impressions between n=98 pages with banner A and n=98 pages with banner B. Results showed a mean difference of -0.705 clicks, indicating that on average banner A generated about 7 fewer clicks for 10,000 impressions than B. We constructed a 95% confidence interval around this mean difference, which ranged from -0.83 clicks to -0.58 clicks.* ***Note that this confidence interval may or may not contain the true population value.*** *The width of the confidence band, about plus or minus 0.125 clicks, gives some indication of amount of uncertainty around the point estimate of -0.705 clicks. To reduce this uncertainty, we would have to increase sample sizes, reduce variability in click rates within groups, or both.*
   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 have a collection of 100 confidence intervals, each of which was constructed in the same way and with the same sample sizes, but each from new data samples: What, if anything, can you say about this collection of confidence intervals?   
       *All else being equal, and assuming no biases in measurement or sampling, about 95 out of the 100 constructed intervals would contain the true population mean value. It is impossible, however, for us to know which confidence intervals contained the population mean and which did not!*