Biostatistics Problem Set 1

1. Complete Assignment Problem 22, Chapter 2. (11 pts)

2. Refer to Example 2.3C. For all three parts below, make sure to follow the principles of good graph construction from Chapter 2. (9 pts)
    A. Create the boxplot shown in Figure 2.3-4 (just get the shape right, don't worry about the colors, text size, etc.). You can refer to the Whitlock and Schluter R pages for the code to do this.
    B. Ignoring the hemoglobin variable, make the appropriate chart to display just the male population data by country.
    C. Ignoring the country variable, make the appropriate chart to display just the hemoglobin concentration data.

3. Complete Assignment Problem 19, Chapter 3. Hint: See p. 71. (4 pts)

4. Calculate the mean, variance, standard deviation, and median of the abundance data presented in Example 2.2B. See the R Pages on the Whitlock and Schluter website to download this data, or you can type the data into R manually (see Basic Intro to R page on the Whitlock and Schluter website for an example of how to do this). (6 pts)

5. Complete Assignment Problem 19, Chapter 5 (8 pts)

6. Presume that in our enormous tomato field, 80% of tomato plants were alive (as opposed to dead) and 30% of tomato plants had a type of parasite in their roots (as opposed to no parasite) at the end of the season. (6 pts)
    A. Create a contingency table describing the expected percentage of plants with all four combinations of the "living" and "parasite" variable, *assuming* that "living" and "parasite" are independent variables.
    B. Create two additional contingency tables showing some form of dependence between these two variables. Be sure that your table still has the correct marginal alive/dead and parasite/no parasite probabilities that are described above.

7. Complete Assignment Problem 18, Chapter 4, skipping part c. (5 pts)

8. As we've discussed, larger sample sizes give us a more precise estimator of a population parameter, which is reflected in the width of the sampling distribution (and its standard error) getting smaller. Imagine we were to make our sample size so large that it was equal to the size of the entire population. What would be the relationship between Y_bar and mu, and what would the standard error of the mean be? (4 pts)

Species abundance distributions are the name for the probability distributions giving the probability that a randomly selected species from a community will have some number of individuals. Species abundance distributions can take many shapes, one of which is thought to be an exponential distribution. Presume that for a particular community (for example, all of the bird species of a small island), this probability distribution is an exponential distribution with a parameter rate (sometimes called lambda) equal to 0.05. (26 pts)

A. Use the function ceiling(rexp(n=100, rate=0.05)) to draw a random sample of 100 species abundances from this population distribution. The rexp function here draws random numbers from an exponential distribution, while the ceiling function around it rounds the random number up to the next integer, as appropriate for counts of abundance. Make an appropriate graph to display this sample. What is being plotted on the x-axis and y-axis?
B. What is the mean and standard deviation of this frequency distribution?
C. Repeat (b) but instead draw a random sample of 1,000 species. What is the mean and standard deviation of this frequency distribution, and how does it compare to your answer in (b)?
D. Simulate a sampling distribution of the mean species abundance, taking 10,000 random samples from this probability distribution, each of which contains 100 species abundances. Make an appropriate graph to display the sampling distribution. What is being plotted on the x-axis and y-axis?
E. What is the mean of this sampling distribution? How does it compare to the mean of the population-level probability distribution, which is 20?
F. What is the standard deviation of this sampling distribution, and what is the other name for this quantity?
G. Repeat (b) by taking 10,000 random samples again, but now each containing 500 species. Make an appropriate graph to display this new sampling distribution.

H. What is the standard deviation of the sampling distribution in (e), and how does it compare to your answer in (d)? Explain why these are different.

10. Complete a hypothesis test of the question posed in Practice Problem 4, Chapter 7. Perform this hypothesis test three ways. Make sure to write out all six steps in hypothesis testing, as described in the class, for the first test a. below. For b. and c., you can refer back to your answer in a. for steps that are identical. (21 pts)
    A. Using a binomial test with a simulated null distribution (like our first "toad handedness" example).
    B. Using a Chi-squared goodness of fit test with a simulated null distribution (like our second "toad handedness" example).
    C. Using R to perform the Chi-squared test
    D. How do the P-values from your three hypothesis tests compare to each other? Which one do you think is the most accurate and why?