Due: midnight, Sep 20 (Thursday labs); midnight, Sep 21 (Friday labs)

Lab 4: Estimating parameters from a random sample

Objectives

1) Estimate the mean and variance from a random sample.
2) Plot histograms of random samples.
3) Plot estimated probability distributions.
4) Interpret sample size effects on estimates.
5) Estimate the standard error of the mean.
6) Estimate the confidence interval of the mean.
7) Interpret sample size effects on the standard error of the mean.
8) Use the R functions `hist`, `rnorm`, `mean`, `var`, and `sd`.

Exercises

1. Imagine that you are investigating the monitor lizards that inhabit Kangaroo island off the coast of Australia. Ultimately, you are interested in knowing whether these lizards are different in length from the mainland population. For now, however, you are simply trying to describe the probability distribution of the lengths of the island lizards. The file entitled "lizard.csv" contains data on body length (in cm) from a random sample of Kangaroo Island monitor lizards. Write a script that does the following:
   - Import the data.
   - Estimate the mean of length by direct calculation.
   - Estimate the mean with R's built-in function.
   - Estimate the variance of length by direct calculation.
   - Estimate the variance with R's built-in function.
   - Estimate the standard deviation of length by direct calculation.
   - Estimate the standard deviation with R's built-in function.
   - Estimate the standard error of the mean length, and calculate the approximate 95% confidence interval.

   Answer:
   - Is the estimated mean likely to equal the population parameter? Why or why not?
   - Which is a better descriptor of the variation in length of lizards, the standard deviation or the standard error? Why?
   - Which is a better descriptor of the uncertainty in the estimated mean length of lizards? Why?

2. Graphical displays of data can be very useful for understanding a problem. In this exercise you will make a graph showing two things: the data you analyzed in exercise 1, and your best estimate of the probability density function (PDF) of the lizard lengths. Write a script that does the following:

- Make a histogram of the lizard lengths. Make it show probability density, not frequency. Give it informative axis titles and an overall title.
- Add a curve to the histogram showing your estimate of the probability density function of lizard length. To do this, you will make use of the parameter values you estimated in exercise 1.

3. The plot that you made in exercise 2 shows the estimated probability distribution of lizard lengths. However, you also need to know the distribution of a different random variable: the *average* of a sample of *n* lizard lengths.
   - Add another curve to the histogram showing the probability density function of the average of 15 lizard lengths.
   - Add one more curve showing the probability density function of the average of 25 lizard lengths.
   - Add a legend to the plot.

   Answer: Compare the location and dispersion of the distributions of the three variables: length, average length for a sample of 15, average length for a sample of 25. Interpret the pattern you see to explain why the average is an accurate estimator of the mean, and how sample size effects the quality of the estimate.

4. Males of the European rhinoceros beetle have a single horn on the head that they use in territorial competition with other males. Imagine that you know that horn length is normally distributed with a mean of 12 millimeters and a standard deviation of 2 millimeters. (This perfect knowledge is unlikely in reality, but we will assume it here for the sake of making predictions.)
   - What is the probability that a randomly sampled male will have a horn less than 11 mm long?
   - What is the probability that the *average* length of a random sample of 10 horns will be less than 11 mm?
   - Explain the difference in the two probabilities you just calculated.
   - What is the probability that the average length of a random sample of 10 horns will be greater than 12.5 mm?
   - What is the probability that the average length of a random sample of 50 horns will be greater than 12.5 mm?
   - Explain the difference in the two probabilities you just calculated.
   - What are the 0.025 and 0.975 quantiles of average horn length, for a sample of 10?

5. Assume that you know that the weight of lodgepole pine cones follows a normal distribution with a mean of 6.2 g and a variance of 0.25 g^2. (Again, this perfect knowledge is unlikely in reality, but we assume it here to explore effects of sample size.) Write a script that does the following:
   - Generate a sample of 10 randomly sampled cone weights.
   - Use your sample to estimate the mean, variance, and standard deviation using R's built-in functions.

- Use your sample to estimate the standard error of the mean, and calculate the approximate 95% confidence interval.

Answer:

- Run the script again. Are the boundaries of the confidence interval the same as before? Why or why not?
- Did your confidence interval include the true value of the mean? If you repeated this exercise 100 times, roughly how many times would you expect the interval to include the true mean?
- Now run your script once again, but first make one change: use a sample size of 90 instead of 10. Compare the new confidence interval to the first one you generated. What happened to the interval when you changed sample size, and why?
- Assuming that you did not already know the true value of the mean, which of your two estimates would you trust more, and why?

6. For each of the following, state the *expected* effect of increasing sample size. Did you see the expected effect in your own calculations in exercise 5? *Note: This question does not require any additional R script or output.*
   - Estimate of the mean weight.
   - Estimate of the variance of weight.
   - The estimated standard error of the mean weight.
   - The width of the 95% confidence interval.

Assignment: Turn in a single Microsoft Word document with a heading that includes your name, the lab number, and your section (day and time). Create a subheading for each exercise. For each exercise, include (1) your annotated R script, (2) your annotated output, and (3) your answer to the final "Answer" item for each exercise (where applicable). In addition to the word document, you must also submit a separate .R file with all of your R scripts.