**Problem Set #1: Basic Statistics and Introduction to R**

(1) (4 pts) The following data are numbers of protozoa sampled from a microcosm and counted on a hemacytometer: 6, 11, 4, 5, 7, 3, 5, 1, 5, 6. Calculate the statistics listed below for this sample of protozoan densities. Do these calculations “by hand” using a calculator.

mean =

variance =

SD =

SE =

CV =

median =

mode =

# (2a) (4 pts) Being the insightful biologist that you are, you notice that protozoa seem more dense at the bottom of the microcosm, perhaps because there is more food available there. You want to know if there is statistical support for this casual observation. You sample 20 replicate microcosms and measure the densities of protozoa. In 10 of the microcosms, you take the sample from the top and in the other 10 microcosms, you take the sample from the bottom. The data are as follows:

Top (# per uL): 3, 1, 0, 5, 4, 3, 6, 3, 4, 7

Bottom (# per uL): 3, 12, 3, 4, 7, 8, 7, 5, 15, 9

# Using R, calculate the following statistics for each of the two sites:

Top Bottom

Mean \_\_\_\_\_\_ \_\_\_\_\_\_

Standard deviation \_\_\_\_\_\_ \_\_\_\_\_\_

Variance \_\_\_\_\_\_ \_\_\_\_\_\_

95% Confidence Interval \_\_\_\_\_\_ \_\_\_\_\_\_

(2b) (5 pts) Make a publication-quality bar graph in R that presents means and standard errors for each group (top vs bottom). Provide a figure legend that describes the graph and includes a statement about whether you think protozoa densities differ between the top and bottom of the microcosm.

(3a) (3 pts) The csv file named "kelp bass gonad mass" contains the weights of gonads from several hundred kelp bass collected by Dr. Mark Steele’s lab. Estimate the mean, median, *σ2*, *σ*, CV, skewness, and kurtosis.

(3b) (3 pts) What effect would adding 5.0 to each observation of gonad mass have on the values of the mean, median, *s2*, *s*, CV, skewness, and kurtosis? (You don’t need to show the new values, but just describe how the statistics have changed.)

(3c) (3 pts) What would be the effect of adding 5.0 and then multiplying by 10.0?

(3d) (3 pts) Make a histogram of all raw observations (untransformed values) in the kelp bass gonad mass data set. Do these data look relatively normal or not? Add the histogram below.

(3e) (3 pts) Convert all raw observations in the kelp bass data set into *Z-*scores. Make a histogram of this new data set. How does this histogram differ from the one for the raw observations? Add the new histogram below.

(3f) (3 pts) Use the original kelp bass gonad data to create a Normal Probability Plot. Do the data appear to be normally distributed? Add the plot below.

(4) (5 pts) Round the following numbers to three significant figures and state their implied limits before and after rounding.

|  |  |  |  |
| --- | --- | --- | --- |
| number | implied limits | rounded to 3 SF | implied limits |
| 106.5 |
| 0.068191 |
| 3.049 |
| 2.03456 x 106 |
| 2.914 |
| 20.15000 |

(5) (5 pts) For each of the following questions, define the *sampling unit* and the *statistical population*.

(a) What proportion of blue whales in the Pacific Ocean are reproductively mature?

statistical population: sampling unit:

(b) How many mitochondria per cell?

statistical population: sampling unit:

(c) How many seeds per white flowered plant?

statistical population: sampling unit:

(d) How many bacteria per 1mL in a sewage treatment plant?

statistical population: sampling unit:

(e) How much time do bees spend each time they visit a flower?

statistical population: sampling unit:

(f) How many bees visit flowers in a 5-minute observation period?

statistical population: sampling unit:

(6a) (5 pts) Carla (former MS student in Peter Edmunds’ lab) sampled the weights (in grams) of 30 individuals of the coral, *Agaricia agaricites*. The data are available in the file “Agaricia.csv”. Are the data normally distributed? Does log-transformation improve the normality or not? Support your answer with whatever graph(s) you think are appropriate.

(6b) (4 pts) Use the Agaricia data set to estimate the mean ± 95% CI of the untransformed data sample by resampling the data with bootstrapping (just use 1000 resamplings). Plot the frequency distribution of estimates for the mean and indicate the 95% confidence intervals on the plot.