Exploring Marine Primary Productivity with Descriptive Statistics and Graphing in Excel

Statistics Background Info (Definitions)

Descriptive Statistics – Descriptive statistics is the branch of statistics focused on various ways to organize, summarize, and display data. Descriptive statistics can also refer to the actual calculations and graphs created.

Outlier – An outlier is a value in your data set that obviously does not fit with the data set in some way. Outliers have different causes – they may be due to a typo, a sensor malfunction, or a highly unusual (but valid) event occurring as part of your data set. Because of the differing causes, it's impossible to have a single rule with which to identify them all. You have to use your judgement and assessment of the situation to identify them. Although outliers are often high/low values in a data set, keep in mind that not all extreme values are outliers; not all outliers are extreme values.

Dependent & Independent Variables – When looking at paired data, one variable is called the *independent variable* and the other is the *dependent variable*. The *dependent variable* (often called y) is the variable whose values are suspected to change depending on the value of the *independent variable* (often called x). Another way to think of the choice of variables is the *independent variable* is the input or explanatory variable. The *dependent variable* is the output and can be thought of as the response variable. If you suspect the value of y depends on the value of x, x is chosen as the in*dependent variable* and y is chosen as the *dependent variable*.
Independent Variable: *x, "explanatory variable"*
Dependent Variable: *y, "response variable"*

Mean – The mean is a common way to measure the center of a data set. To find the mean, you add up all the numbers in the data set, then divide by the number of values. That is: $\text{mean} = \frac{\Sigma x}{n}$. The mean is commonly called an "average"; however, statisticians prefer to use the more precise term, mean. The mean of an entire population is denoted by μ (the Greek letter mu), and the mean of a sample is denoted by $\bar{x}$ (read as "x-bar").

Median – The median is another commonly used measurement of the center of a data set. In any data set, 50% of the values lie above the median, 50% below. Although you'll be using Excel to find the median, it's instructive to see the manual definition. Another name for the median is the second quartile, denoted Q_2.

To find the median:
1) Write the data in order from smallest to largest. Let n=size of the data set.
2) If n is odd, the median is the number exactly in the middle of the sorted data set.
3) If n is even, the median is the mean of the two numbers in the middle.

Standard Deviation – The standard deviation is a measure of how "spread out" the data are. The larger the standard deviation of a data set, the more variation there is in that data set. *Informally,* you can think of the standard deviation as the average distance the data points are from the mean. [Note: this is not formally true, but is close enough to provide a good way to conceptualize it.] Formally, we have two different formulas:

- Standard deviation of a population $= \sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ (N is the population size)
- Standard deviation of a sample $= s = \sqrt{\frac{(x-\bar{x})^2}{n-1}}$ (n is the sample size)

Note that the standard deviation formula is different whether your data set is from an entire population or a sample, so be sure to use the appropriate command in Excel.

Quartiles – Every data set has three numbers (known as quartiles) — Q_1, Q_2, Q_3 — that divide the data set into 4 parts.

- Q_1 (the first quartile) divides the bottom 25% of the data from the upper 75% of data
- Q_2 (the second quartile, or median) divides the bottom 50% of the data from the upper 50% of data
- Q_3 (the third quartile) divides the bottom 75% of the data from the upper 25% of the data

Five-Number Summary – A five-number summary is a quick way to present a summary of a data set.

The five-number summary includes: the lowest value in the data set (the minimum), Q_1, Q_2, Q_3, and the highest value in the data set (the maximum). Five-number summaries are often shown in graphical form as a box-and-whisker plot.

Iterate – To iterate means to perform repeatedly. In the context of Excel, *iterating* means to use your mouse and/or keyboard shortcuts to repeat the calculations of a previously-entered formula across multiple rows and/or columns so that you don't need to retype formulas multiple times.

Exploring Marine Primary Productivity with Descriptive Statistics and Graphing in Excel
Marina McLeod, Jennifer Olson, Wendy Houston

This material is based upon work supported by the National Science Foundation under Grant No.1919613. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.