WEEK 3 READING QUESTIONS

Sonja Glasser

1. *PLOTS: Histogram, Scatterplot, Cleveland dotplot, Boxplot, QQ plot, coplot. Which of the plot types show every data point?*

   Scatterplot, Cleveland dotplot, QQ plot, and coplot are plots that show every data point. You can also specify in the boxplot argument in R to show your data points on top of the box plot figure.

2. *Which of the plot types show aggregated or summarized data?*

   Plots that show aggregated or summarized data include box plots and histograms.

3. *Conditional plot, conditioning variable, and related terms occurred throughout the Zuur and McGarigal readings. Explain what a conditional variable means in the context of graphical data exploration*.

   A conditional plot shows the relationship between two variables conditioned on another variable – the conditioning variable. This sort of plot is useful when an independent and dependent variable have a strong relationship that may hide the effects of other independent variables on the dependent.

4. *List* at least three *of the common measures of spread or dispersion that were mentioned in the readings.*

   Three common measures of spread or dispersion are variance, standard deviation, and coefficient of variation.

5. *Choose* two of the measures *in your list and explain how they capture different aspects of the concept of spread.*

   The standard deviation is the root mean squared deviation from the mean or expected value. When your data is normally distributed, the number of standard deviations (+ & –) from the mean gives us an idea of how the data is spread. For example, + and – one standard deviation from the mean encompasses 68% of the data value points.

   The coefficient of variation is the standard deviation divided by the mean. This makes it so we can compare the spread of variable measured on different scales. The coefficient of variation can be expressed as percentages. This shows the variability in relation to the mean of the population.

6. *Consider a dataset that you have collected or worked with.*

   If you haven't worked much with existing datasets hypothesize a dataset that you might collect for your research.

   List two of the important reasons to perform data exploration (numerical and/or graphical). For each of the two reasons you identify, describe the quantities, or plots you would use and the insight you would gain.

In my hypothesized dataset I am looking at bee pathogen load depending on flower species visited. There are a few factors to consider, such as time of season of observed data point, site of the data point and species of bee.

Data exploration is important to see if the relationships you are testing show a pattern or if there are other variables measured that may have an even stronger or correlative effect on the dependent variable of interest. I would do a pairs plot to see how all the variables are interacting. Data exploration is also useful for seeing if there are outlier values in your dataset. These outliers can affect your model fit and could potentially be outliers due to human error, which can then be revised. You could potentially use a box and whisker plot to check for extreme outlier points.

Potentially the pathogen load might be more correlated with site as less correlated with flower species, this would be interested to visualize seeing as it is not included in the project's question and should be considered. Additionally, looking for the outliers would be a good way to check if I accidently entered in a data point incorrectly, or if the outlier data point should be considered in the model fit.