Module 4: Data Exploration
Data Exploration

Now that you have your data downloaded from the Streams Project database, the detective work can begin! Before computing any advanced statistics, we will first use descriptive statistics to examine the distribution of your data.

The following topics are covered in this module:

1) **Overview of descriptive statistics**
2) **Central tendency**
3) **Dispersion**
4) **Visualizing your data**
Calculating the descriptive statistics outlined in this module may be the extent of your analysis or the first step towards a more in-depth analysis as outlined in Module 5.

Descriptive statistics help you describe your data in terms of its distribution. To examine your data’s distribution you will need a measure of **central tendency** and **dispersion**. These measures are illustrated in the frequency histogram below which shows the count of data for each interval of 100 MPN (Most Probably Number) of E.coli helping you visualize the frequency of data occurring over the range of values in the dataset:

*A histogram helps you visualize the frequency of values in your dataset occurring over a series of defined intervals. We will show you how to create a histogram later in the module.*
Data Exploration

1) Overview of Data Exploration

Before we calculate measure of central tendency and dispersion, let’s look at what we mean by distribution. The ideal distribution of data is called the “normal distribution.” For a normal distribution, all measures of central tendency (you will see there are several!) are the same, and there is an equal number of observed data points on either side of these measures of central tendency.

A histogram is a great way of initially visualizing your data’s distribution because you can get a sense of the central tendency and dispersion of data around that center before calculating any statistics. The following histograms illustrate normal distributions as well as non-normal, or “skewed,” distributions for comparison:

For advances statistical tests, it is important to determine if your distribution is normal or otherwise, as this will affect the type of statistical test you chose to use. The statistical tests in this tutorial assume your data is normally distributed.
Ambrose et al. (2002:22) describe “central tendency” as “what usually happens…”

*If we measure E.coli at all our stream sites, what amount of E.coli do we usually measure?*

Having a measurement that reflects what usually happens allows us to compare individual sample data points to the “usual” value of data observed represented by a measure of central tendency. The following are three measurements of central tendency:

**Mean** = *the sum of observed data points divided by the number of data records*

**Median** = *the middle data point (or average of the two middle data points if there is an even number of observations) when all data points are lined up in either ascending or descending order*

**Mode** = *The most frequently occurring data point value in a dataset*

On the next page you will see examples of how these three measurements would be calculated for a sample E.coli dataset.

Continued...
Data Exploration

2) Central tendency

Given our sample dataset, we would calculate the mean, median, and mode as follows:

**Mean** = \( \frac{(410+263+310+476+388+417+345+402+379+379)}{10} = 377 \)

**Median** = \( 263, 310, 345, 379, 379, 388, 402, 410, 417, 476 \) > \( \frac{381+388}{2} = 385 \)

**Mode** = 379 which occurs twice while other values occur only once

379

These values all suggest that the center of our data, or the most frequent values of E.coli measured in our streams, falls around 370 – 390 MPN.
We have a measure of central tendency for our sample dataset (let’s use the mean), but how do we describe how the rest of the data falls around our mean? If the following bell curves (think of them as smoothed out histograms) represent data with the centerline as your measure of central tendency, how do we describe the different ways that the data falls around their centerline?

Now that you have a visual on what we mean by “dispersion,” the following page helps you calculate statistics used to quantify the nature of the dispersion around the mean.
Data Exploration

3) Dispersion

The following are measurements of dispersion used to quantify the spread of data about the data’s center:

**Range** = *The highest measurement – the lowest measurement in the dataset*

**Variance** = *A cumulative measure of individual data points’ distance from the mean. The following equation is used to calculate variance:*

\[ s^2 = \frac{\sum x^2 - \left(\frac{\sum x}{n}\right)^2}{n - 1} \]

*Where:*

- \( s^2 \) = variance
- \( x \) = individual values in of the dataset
- \( n \) = the number of data points in the dataset
**Standard Deviation** = *Somewhat of an average deviation of the data from the mean. Standard deviation is calculated as the square root of the variance:*

\[ s = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n - 1}} \]

**Where:**
- \( s \) = standard deviation
- \( x \) = individual values in the dataset
- \( n \) = the number of data points in the dataset

On the next page you will see examples of how these three measurements would be calculated for our sample E.coli dataset.
Given our sample dataset, we would calculate the following values for range, variance, and standard deviation:

**Range** = 476 (highest value) – 263 (lowest value) = 213

**Variance** =

\[
\sigma^2 = \frac{\sum 410^2 + 263^2 + 310^2 \ldots - \left( \frac{\sum 410 + 263 + 310 \ldots}{10} \right)^2}{10 - 1}
\]

\[
\sigma^2 = \frac{410^2 + 263^2 + 310^2 \ldots - (410 + 263 + 310 \ldots)^2}{10 - 1}
\]

**Standard deviation** =

\[
s = \sqrt{\frac{\sum 410^2 + 263^2 + 310^2 \ldots - \left( \frac{\sum 410 + 263 + 310 \ldots}{10} \right)^2}{10 - 1}}
\]

\[
s = \sqrt{\frac{410^2 + 263^2 + 310^2 \ldots - (410 + 263 + 310 \ldots)^2}{10 - 1}}
\]

The standard deviation, variance and mean are all metrics commonly used in more advanced statistical tests, some of which will be described in Module 5.
In the first part of this module you learned how to calculate two types of parameters to help you describe the distribution of your data:

- **Measure of central tendency**
- **Dispersion**

When you are presenting these parameters, it is helpful to provide a visual of your data’s distribution in addition the numbers. The remainder of this module will help you create a histogram and/or a boxplot depicting your data’s distribution. The module will conclude by helping you describe your combined visual and statistical results.
Data Exploration

4) Visualizing your data

First, we will revisit our **frequency histogram**, which shows the frequency of values in your dataset over a series of intervals that covers the range of your dataset. See the link below to see how to create a histogram for your data in excel.

Click on the video icon to watch a video on how to create a histogram using Microsoft Excel.
Data Exploration

4) Visualizing your data

A box plot also illustrates the distribution of your data. A box plot is made up of the following values derived from your dataset: median, minimum value, maximum value, quartile 1 value, and quartile 2 values. The following graph illustrates these components with descriptions following:

Definitions:

Maximum Value = the largest value in the dataset
Minimum Value = the smallest value in the dataset
Median = the middle value of the dataset when all values are lined up either ascending or descending
Quartile 1 = the median value of all values less than and excluding the median of the entire dataset
Quartile 3 = the median value of all values greater than and excluding the median of the entire dataset
Inter-Quartile Range (IQR) = Quartile 3 – Quartile 1

Click on the video icon to watch a video on how to create a box-plot using Microsoft Excel
Data Exploration

4) Visualizing your Data

Remember to discuss whether or not your data is normally distributed, or perhaps is skewed to the left or right. Your graphs, in addition to your descriptive statistics, help you communicate your findings, so be sure to include both!

**Histogram**

- Data is equally dispersed on either side of the data's center (peak of the histogram).

**Box Plot**

- Data is equally dispersed on either side of the median (center line of the box).

**Normal Distribution**

- There is more data on the lower end of the data range with values tapering to the higher end.

**Left-Skewed**

- The median is visibly closer two one the lower end of values.

**Right-Skewed**

- The median is visibly closer two one the higher end of values.
It is often the case that after exploring your data with descriptive statistics you want to modify or refine your central question – that’s fine!
SUMMARY

- Descriptive statistics help describe your data’s distribution

- A measure of central tendency and dispersion are needed to describe your data’s distribution statistically

- Ideally your data fits the descriptions of a normal distribution with data distributed evenly on either side of the measure of central tendency.

- The following are measures of central tendency: mean, median and mode

- The following are measure of dispersion: range, variance, and standard deviation

- Histograms and box plots can help you illustrate your data’s distribution

- Your descriptive statistics, histograms and/or box plots together help you describe the nature of your data

- After exploring your data using descriptive statistics it’s good to reflect on your question and modify or refine it as needed.